

Electronic Theses and Dissertations, 2004-2019

2007

Modeling And Partitioning The Nucleotide Evolutionary Process For Phylogenetic And Comparative Genomic Inference

Todd Castoe
University of Central Florida

 Part of the [Biology Commons](#)

Find similar works at: <https://stars.library.ucf.edu/etd>

University of Central Florida Libraries <http://library.ucf.edu>

This Doctoral Dissertation (Open Access) is brought to you for free and open access by STARS. It has been accepted for inclusion in Electronic Theses and Dissertations, 2004-2019 by an authorized administrator of STARS. For more information, please contact STARS@ucf.edu.

STARS Citation

Castoe, Todd, "Modeling And Partitioning The Nucleotide Evolutionary Process For Phylogenetic And Comparative Genomic Inference" (2007). *Electronic Theses and Dissertations, 2004-2019*. 3111.
<https://stars.library.ucf.edu/etd/3111>

MODELING AND PARTITIONING THE NUCLEOTIDE EVOLUTIONARY PROCESS FOR
PHYLOGENETIC AND COMPARATIVE GENOMIC INFERENCE

by

TODD A. CASTOE

B.S. SUNY – College of Environmental Science and Forestry, 1999

M.S. The University of Texas at Arlington, 2001

A dissertation submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Biomolecular Sciences
in the Burnett College of Biomedical Sciences
at the University of Central Florida
Orlando, Florida

Spring Term
2007

Major Professor: Christopher L. Parkinson

© 2007 Todd A. Castoe

ABSTRACT

The transformation of genomic data into functionally relevant information about the composition of biological systems hinges critically on the field of computational genome biology, at the core of which lies comparative genomics. The aim of comparative genomics is to extract meaningful functional information from the differences and similarities observed across genomes of different organisms. We develop and test a novel framework for applying complex models of nucleotide evolution to solve phylogenetic and comparative genomic problems, and demonstrate that these techniques are crucial for accurate comparative evolutionary inferences. Additionally, we conduct an exploratory study using vertebrate mitochondrial genomes as a model to identify the reciprocal influences that genome structure, nucleotide evolution, and multi-level molecular function may have on one another. Collectively this work represents a significant and novel contribution to accurately modeling and characterizing patterns of nucleotide evolution, a contribution that enables the enhanced detection of patterns of genealogical relationships, selection, and function in comparative genomic datasets. Our work with entire mitochondrial genomes highlights a coordinated evolutionary shift that simultaneously altered genome architecture, replication, nucleotide evolution and molecular function (of proteins, RNAs, and the genome itself). Current research in computational biology, including the advances included in this dissertation, continue to close the gap that impedes the transformation of genomic data into powerful tools for the analysis and understanding of biological systems function.

This dissertation is dedicated to my wife, my family, my close friends, and to the memory of Kenneth and Mary Van Norstrand. Their love, support, respect, encouragement, and genuine depth of character are the only reasons that this particularly extraordinary process has been possible. My appreciation for their roles in my life cannot be overstated, and this dissertation stands as one humble anecdotal testament of the strength of their conviction and virtue. Despite the copyright on the preceding page, the credit for this work belongs to them.

ACKNOWLEDGMENTS

My advisor, Christopher Parkinson, has somehow found the patience to endure many years of dealing with all of my personality quirks, the significance of which is obvious and enormous for anyone who knows me well. He has always supported my independence and facilitated my growth as a biologist eager to choose his own course. He has done his best to encourage and enable my success and scientific development, and we have shared many fun experiences together, and I sincerely thank him for his guidance, patience, and friendship. The UCF Department of Biology has supported my work in many ways, including both financial and moral support. I was supported by a fellowship from the UCF Graduate School, and indirectly through support received by Chris Parkinson in the form of his startup funds from the UCF Biology Department, and from his National Science Foundation grant (DEB-0416000).

My committee, including Laurie Von Kalm, Henry Daniel, and Jack Ballantyne has consistently provided a very important platform of encouragement and support throughout the entire process and I sincerely thank them for their positive role throughout my dissertation work. Even beyond academic relationships, the personal relationships that I have had the pleasure of sharing with both Laurie and Henry have been particularly important to my keeping focused and enthusiastic, and for this I thank them both.

I would like to thank David Pollock and members of his lab (especially Zhijie Jiang and Wanjun Gu) for their collaborative support on mitochondrial genome projects. Rather than competing with one another, collectively we have enjoyed a rewarding collaborative experience that has resulted in many positive things being accomplished. Their help and encouragement,

both on specific projects and in general, have played key roles in the success of many of the mitochondrial genome-based projects included here and to come. Zhijie Jiang also has delivered a number of important favors and I sincerely appreciate his patience, understanding, and much-needed help.

Throughout my time at UCF I have been particularly fortunate to have overlapped with a large number of truly exceptional people, both in the lab, and in the program. These close friends have tremendously enriched my experience here and their constant support and friendship have played key roles in my enjoyment and success. A number of great people in the lab have had the misfortune of co inhabiting 425 (and earlier labs) with me, all of which played major roles in my development as a scientist, and also acted as important cheerleaders providing constant encouragement over the years, one cup of coffee at a time. Josh Reece always talked me into a camping trip when I needed a break. Jack Degner always talked me into a beer when he knew I needed one, and our mutual love for Tom Waits songs effectively erased thoughts of scientific failures of the preceding days, or at least erased previous thoughts altogether. Matt (dirty red) Herron was around back in the good old days when we thought 1 million MCMC generations deserved a Nature paper; wouldn't that be nice if things stayed that way eh Matt? Håkon (makemecupacoffee) Kalkvik provided many much needed coffee breaks, sober safe rides home, and lots of great chocolate to fuel my scientific fires. Jenna, the little sister of the lab, always provided some much needed sarcasm and a good laugh to reduce the ambient tension in the air. Allyson Modra – finally a coffee lover in the lab – actually helped me take less coffee breaks and I probably owe her for getting out of here on time. Juan (el gordo) Daza has been a great friend,

drinking buddy, fellow naïve biogeographer, and snake lover, and together we have shared many good times and avoided some paramilitary.

I owe a huge debt of gratitude to Robert Ruggiero, my long-time roommate and old friend. A majority of my academic career has been shared with Rob and it is often difficult to disassociate my path from him and his impact on it. Above all, I need to thank him for his relentless support, solid advice, and for making this otherwise tedious and frustrating process an exciting and enriching experience. I am a better scientist because of my years of interaction and friendship with Rob. Sam and Erin Fox, our close friends and confidants, positively distracted us in many memorable and enjoyable ways, including introducing us to many new forms of adventure: surfing, clothing optional rooftop bars, and a previously underrated brand of cheap whiskey - Dickel. We definitely miss all these characters and look forward to meeting them soon somewhere above 10,000 ft. in the Rockies, one day soon with Ph.Ds of their own.

My cohorts in the Biomolecular Sciences Ph.D. program happened to be an amazing group of people who made these 5 years particularly special. Even if we don't see each other as much as we used to, my relationship with them is as close as it ever was, and this way it will remain. My closest confidant and buddy, George Kryazis, has remained a steadfast unflinching bedrock of support, advice, and encouragement throughout this process, and I really hope that he gets rich someday (and shares) so we both don't have to waste beautiful expensive bottles of scotch complaining about science any longer. Claudia, whom I essentially look to as a big sister, really set the bar pretty high around these parts (gee thanks) and proved to us all upfront that finishing a PhD was possible. Max Chen, Mangala Soundarapandian, and Paul Cohill have comprised an excellent group of study partners, commiserators, entertainers, and loyal friends

throughout the last five years and I will surely miss their faces and personalities, hopefully not for too long.

In addition to those above, many others shared great times in Orlando, predominantly spent in the dingy booths of a lesser-known hot-spot named Underground Bluz or at our place – the home of the Cosmopolitan. With Walter Sotero, Pierre Decayette, Amit Dhingra, and Tanya Vargas, we spent many fun and memorable evenings together, as these folks rounded out a makeshift home-away-from home and novel extended family.

Make new friends but keep the old, the first are silver, the latter gold. There is surely some truth to this, and many old friends from the beginning of my graduate career in Texas have remained close friends, collaborators, and invaluable scientific resources. Eric and Karin Smith have remained very close friends for many years, and over many miles. Eric always appropriately reminded me about the reasons for doing all this in the first place, including the joys of drinking coffee in the cloud forest, and the excitement that never fades when flipped rocks reveal hidden snakes. I thank them both for the physiological support and guidance that I needed most throughout this process – a means of staying grounded. Jon Campbell and Paul Chippindale have remained supportive of my work over the years and their support has meant a lot. Mahmood Sasa, Ron Bonett, Jesse Meik and Walter Schargel have strongly positively contributed to my understanding and enjoyment of science. Brice Noonan requires special thanks for his recent assistance with supercomputing and other analytical help that proved too complex for me to do alone, and I look forward to our collaborative work being born and set free. Carol Spencer has remained a close friend, collaborator, and champion of my success; her constant enthusiasm and support have been very important. Tiffany Doan has collaborated on numerous

projects and made important suggestions, edits, etc. to many more. Her advice, support, help, and encouragement has provided much needed inertia for completing my Ph.D.

A number of other people have provided important help over the years with advice, analytical help, reviewing manuscripts, etc. These include: Andrew Crawford, Andrew Rambaut, Chris Simon, Jack Sites, Jack Sullivan, John Fauth, Pedro Quintana-Ascencio, Eric Hoffman, Cristina Calestani, Ulrich Kuch, Bill Lamar, Manuel Varela, Antonio Ramírez, Marco Antonio López-Luna, Roberto Mora, Luís Canseco, Manuel Acevedo, Matt Brandley, Tom Devitt, Allan Larson, John Wiens, Ron Gutberlet, Mike Harvey.

My exceptionally close relationships with my early and closest friends Rick Constantino, David Greenblatt, Glenn Darwell, and Marty Willms have been critical to my motivation and enjoyment of life. Never has there been more wholesome enriching companions for nearly sinking canoes in 40 degree lakes, climbing mountains, late night walks, innumerable happy-hours, nearly drowning, jumping off cliffs, tropical parties, military checks at gunpoint, or amoebic dysentery. These guys have made even the worst of times the absolute best!

Finally, it is difficult to summarize the debt of gratitude that I owe my wife, Jill Castoe, and my family. I have dedicated this dissertation to them because, above all, none of this would have been possible without their love and support in so many ways. My family has supported me 110% since the beginning, no matter what choices I made, and I look to them as role models of humility, strength, caring, and support. From the beginning, my mother has not only made every sacrifice necessary, but also every sacrifice possible to provide me with all that she could, while providing incredible strength, support, and love. My wife, Jill, is an incredible person that combines strong and deep personal characters with the humor and enthusiasm it has taken to

keep me afloat. Jill has directly sacrificed more than anyone (even myself) to facilitate my career in science. Her patience, understanding, help, encouragement and love have made this process very easy for me, and I cannot imagine how I would have done this without her, and I cannot thank her enough.

Collectively, my wife, family, and friends listed here are the reason I stand here today. Their struggle to keep me alive and moving through this process, both emotionally and physically, has been the most dominating force dictating both my presence and my achievements. Thank you.

TABLE OF CONTENTS

LIST OF TABLES	XVI
LIST OF FIGURES	XIX
CHAPTER 1 – INTRODUCTION	1
A Foundation for Comparative Genomics	1
Complex Modeling of the Nucleotide Evolutionary Process	2
Vertebrate Mitochondrial Genomes as a Model System for Comparative Genomics....	6
References.....	11
CHAPTER 2 – DATA PARTITIONS AND COMPLEX MODELS IN BAYESIAN	
ANALYSIS: THE PHYLOGENY OF GYMNOPHTHALMID LIZARDS.....	20
Introduction.....	20
Methods.....	26
DNA Sequences Used.....	26
Sequence Homology and Alignment	28
Phylogeny Estimation Using Maximum Parsimony.....	31
Bayesian Phylogeny Estimation	32
Results.....	37
Parsimony Phylogenetic Reconstruction	37
Mitochondrial Gene MCMC Analyses	40
C-mos (Nuclear Gene) MCMC Analyses.....	42
Combined MCMC Analyses.....	44
Comparison Among Phylogenetic Reconstructions	58

Discussion.....	59
Model Selection and Evaluation.....	59
Effects of Partitioning Gamma and Using Auto-correlated Rate Variation.....	62
Taxonomic Considerations and Alterations.....	67
References.....	71
CHAPTER 3 – MODELING NUCLEOTIDE EVOLUTION AT THE MESOSCALE: THE PHYLOGENY OF THE NEOTROPICAL PITVIPERS OF THE PORTHIDIUM GROUP (VIPERIDAE: CROTALINAE).....	80
Introduction.....	80
Modeling nucleotide evolution at the mesoscale.....	80
Systematics of the Neotropical pitvipers of the Porthidium group.....	83
Theoretical and empirical scope of this study.....	85
Materials and methods.....	86
Taxon sampling.....	86
DNA sequencing and sequence alignment.....	90
Phylogenetic reconstruction.....	91
Results.....	98
Dataset characteristics and individual gene phylogenies.....	98
Maximum parsimony phylogenetic analysis.....	98
Bayesian MCMC model selection and evaluation.....	99
Effects of model choice on Bayesian phylogenetic hypotheses.....	106
Bayesian MCMC phylogenetic results under the best-fit model.....	108

Discussion.....	109
Model partitioning in Bayesian MCMC analyses.....	109
Suggestions and prospects for complex Bayesian MCMC modeling and model testing.....	111
Relationships and taxonomy of the Porthidium group	115
References.....	122
 CHAPTER 4 – BAYESIAN MIXED MODELS AND THE PHYLOGENY OF PITVIPERS (VIPERIDAE: SERPENTES).....	
Introduction.....	132
Pitvipers and their contemporary systematics.....	132
Challenges and strategies for resolving pitviper phylogeny	135
Materials and methods	138
Taxon sampling.....	138
DNA sequencing and sequence alignment	145
Phylogenetic reconstruction.....	146
Results.....	152
Properties of the dataset.....	152
Maximum Parsimony phylogenetic analyses.....	153
Selection, evaluation, and comparison of Bayesian MCMC models.....	156
Bayesian phylogenetic hypotheses based on 10x partitioned model	158
Differences in MCMC phylogenetic estimates between 1x and 10x partitioned analyses	163

Discussion.....	165
Strengths and limitations of complex partitioned models.....	165
Phylogeny and systematics of pitvipers.....	168
Future directions for pitviper systematics.....	175
References.....	176
 CHAPTER 5 – COMPARATIVE MITOCHONDRIAL GENOMICS OF SNAKES: EXTRAORDINARY SUBSTITUTION RATE DYNAMICS AND FUNCTIONALITY OF THE DUPLICATE CONTROL REGION	
Introduction.....	186
Material and Methods	191
Sampling, sequencing and annotation.....	191
Phylogenetic and sliding-window analyses	195
tRNA structure	199
Analysis of control region functionality	199
Results.....	201
Brief summary of the new complete snake mitochondrial genomes	201
Comparison of <i>A. piscovorus</i> genomes	205
Phylogenetics	208
Nucleotide frequencies and control region functionality.....	213
Gene length and stability of truncated tRNAs in snakes	215
Spatio-temporal substitution rate dynamics across mtDNA genes and regions	217
Discussion.....	226

Gene size reduction and control region functionality	227
Concerted evolution in and around the duplicate control regions	229
Potential impacts of genome architecture on genome replication and transcription	232
Comparative rates of molecular evolution.....	234
Conclusions.....	236
References.....	238
CHAPTER 6 – CONCLUSION.....	246
Advancing the Framework for Functional Comparative Genomics.....	246
Complex Modeling of the Nucleotide Evolutionary Process	247
Insight into the Evolutionary Process at the Genome Scale	254
References.....	257

LIST OF TABLES

Table 1. List of sequences used in this study. operational taxonomic units (OTUs) used in this study with GenBank accession numbers. Cells with an X indicate that gene sequence was not used in this study. (a) and (b) refer to individuals indicated in the figures. Museum accession numbers for specimens sequenced in this study are given. Acronyms for museums are: KU (University of Kansas), MHNSM (Museo de Historia Natural, Universidad Nacional Mayor de San Marcos, Lima, Peru), QCAZ (Museo de Zoología, Pontificia Universidad Católica del Ecuador, Quito, Ecuador) and UTA (University of Texas at Arlington).....	29
Table 2. Parametric composition of models tested in Bayesian MCMC analyses of the combined data.....	35
Table 3. Statistics for datasets used, including results from MP searches and suggested model from hierarchical ln likelihood ratio test (hLRT) criterion from ModelTest.....	38
Table 4. Parameter estimates for all mitochondrial gene and c-mos.....	41
Table 5. Parameter estimates for CNR-SSG model MCMC runs summarized as means with 95% credibility interval in parentheses.....	54
Table 6. Current phylogenetic classification of family Gymnophthalmidae.....	70
Table 7. Specimens used in this study including GenBank accession numbers.....	87
Table 8. Best-fit models selected by ModelTest for various partitions of the dataset based on both hLTR and AIC criteria. P1-6 refer to the six independent partitions of the dataset under the 6x-gene,codon model.....	95

Table 9. Description of complex partitioned models used in the analysis of the combined dataset.	101
Table 10. Mean and 95% credibility interval for each parameter sampled from the combined posterior distribution of three independent MCMC runs of the 6x-gene,codon model.....	105
Table 11. Taxon sampling with voucher information, locality data, and Genbank accession numbers for gene fragments. An asterisk is used to indicate novel sequences generated in this study.....	139
Table 12. Description of complex partitioned models used in the analysis of the combined dataset. Each partition identified below was allocated the model selected by AIC criteria estimated in MrModeltest.	147
Table 13. Results of AIC model selection conducted in MrModeltest for partitions of the dataset.	150
Table 14. Mean and 95% credibility interval (in parentheses) of model parameters from Bayesian phylogenetic analyses of the combined data set conducted under the 1x and 10x models. Parameter estimates for each model are based on a total of 9 x 10 ⁶ generations combined from three independent MCMC runs. Partitions of the 10x model (P1 – P10) are defined in Table 2.	159
Table 15. Table S1 - Primer sets used to amplify mitochondrial genome fragments in this study.	194
Table 16. Table S2 - Complete mitochondrial genomes used in this study, and associated Genbank accession numbers.....	196

Table 17. Estimated T_{AMS} values of genes for squamates. Two T_{AMS} values are given for each species of alethinophidian snakes; T_{AMS}^1 is estimated based on the assumption of exclusive CR1 usage, whereas T_{AMS}^2 is estimated based on exclusive CR2 usage. Genes that have alternative T_{AMS} estimates under different CR usage scenarios in alethinophidian mtDNAs are indicated in bold.....	200
Table 18. Detailed genome annotation of <i>Agkistrodon piscivorus</i>	203
Table 19. Detailed genome annotation of <i>Pantherophis slowinskii</i>	204
Table 20. Gene-specific polymorphisms observed between the two <i>Agkistrodon piscivorus</i> genomes (<i>Api1</i> and <i>Api2</i>).....	207
Table 21. Negative log likelihood values and Akaike weights (in parentheses) for individual origin of replication models and the mixed model, along with the most likely CR2 preference parameter in the mixed model, for alethinophidian snakes.	214

LIST OF FIGURES

- Figure 1. Strict consensus phylogram of two most parsimonious trees based on the equally-weighted maximum parsimony search including all four genes (c-mos, ND4, 12S, and 16S). Labels (a) and (b) indicate individuals of a species (see Appendix 1). For reference, labels on the right side represent the taxonomy presented by Pellegrino et al. (2001). Taxa that are not labeled have relationships that do not agree with that former taxonomy. 39
- Figure 2. Bayesian phylogenetic trees for the independent nuclear and mitochondrial data partitions. Labels (a) and (b) indicate individuals of a species (see Appendix 1). (A) Majority rule phylogram and posterior probabilities resulting from Bayesian analysis of all three mitochondrial genes combined (ND4, 12S, and 16S) based on a combined 3 million post burn-in generations under the GTR+I+G model of evolution. (B) Majority rule phylogram and posterior probabilities resulting from Bayesian analysis of nuclear c-mos gene data based on a combined 3 million post burn-in generations under the K80+G model of evolution. 43
- Figure 3. The ln likelihood scores of MCMC chains based on alternative models of evolution, sampled in 10,000 generation intervals for clarity of presentation. See text for descriptions of alternative models. 45
- Figure 4. The mean and 95% credibility interval for post burn-in ln likelihood scores of MCMC chains based on alternative models of evolution. 46
- Figure 5. The ln likelihood scores of MCMC chains based on alternative models of evolution, focusing on the period below 50,000 generations, sampled every 100 generations. 48

Figure 6. Bayesian phylogenetic tree and posterior probabilities for clades based on the combined, four-gene data set analyzed under the CNR-SSG model. Tree is based on the combination of all post burn-in generations resulting from three independent runs of the model, for a combined total of 3 million post-burn-in generations. 50

Figure 7. Plot of the posterior probabilities derived from the GTR+I+G MCMC analyses (all three runs combined) versus the posterior probabilities derived from the CNR-SSG model (all three runs combined). For comparison, a 1:1 line is plotted on the same axis. 52

Figure 8. Plots of selected parameters of the CNR-SSG model through generations. All three independent runs are plotted per graph to show common burn-in rates and similar parameter estimates. (a) Plot of parametric estimates of $r(A-G)$ from the GTR rate matrix. (b) Plot of the parametric estimates of $r(A-T)$ from the GTR rate matrix. (c) Plot of the gamma parameter estimates. (d) Plot of the site-specific rate multiplier for the ND4, rRNA (12S + 16S), and c-mos site specific partitions of the gamma parameter 53

Figure 9. Comparison of deviation of posterior probability estimates at intervals of generations compared to overall means from long MCMC runs (33 million generations) for the GTR+I+G and CNR-SSG models. Values represent the absolute deviation of posterior probability estimates (relative to overall mean for long MCMC run) averaged across all nodes receiving less than 100% posterior probability. 56

Figure 10. Plots of individual node posterior probability estimates over intervals of extended (33 million generation) MCMC runs of the GTR+I+G and CNR-SSG models. Numbers for nodes are given at the right; numbering of nodes is consistent between the two graphs for comparative purposes. 57

Figure 11. Majority-rule consensus of 144 equally-parsimonious trees (of 2587 steps) from heuristic maximum parsimony search based on 1405 bp. Bootstrap support for nodes is provided (values below 50% not shown). Bootstrap values of 100% are indicated with gray-filled circles..... 100

Figure 12. Flow chart illustrating the process of model selection among complex models tested for the analysis of the combined dataset. Statistics for models are given (A_w = Akaike weights, $2\ln B_{10}$ = $2\ln$ Bayes factor, RBF = Relative Bayes factor). For $2\ln B_{10}$ comparisons between models, M_i is represented by the model indicated by the arrowhead. See Table 9 for definitions of models..... 103

Figure 13. Majority rule consensus trees resulting from Bayesian MCMC phylogenetic reconstructions under two different models of nucleotide evolution (the favored partitioned model “6x-gene,codon” and the base unpartitioned 1x- GTR+ Γ +I). Nodal posterior probabilities are indicated; nodal posterior probabilities of 100% are indicated with a gray-filled circle. (A) Majority rule consensus phylogram based on a combined 9×10^7 post burn-in Bayesian MCMC generations of the favored “6x-gene,codon” partitioned model. (B) Majority rule consensus cladogram based on a combined 9×10^6 post burn-in Bayesian MCMC generations of the unpartitioned 1x-GTR+ Γ +I model (note: branch lengths are not informative in Fig. 13B). 107

Figure 14. Strict consensus cladogram of 12 equally-parsimonious trees obtained from maximum parsimony analysis of 2306 bp of mitochondrial DNA sequences (14816 steps, consistency index = 0.162, retention index = 0.568, homoplasy index = 0.838). Bootstrap support for

nodes above 50% is given adjacent to nodes; nodes receiving bootstrap support of 100% are indicated by gray-filled circles..... 154

Figure 15. Comparisons of means and 95% credibility intervals (CI) of selected nucleotide model parameters estimated from Bayesian MCMC analyses conducted under the 1x (unpartitioned) and the 10x (partitioned) models. Partitions of the 10x model are designated P1–P10 and correspond with Table 2. Gray-shaded bands indicate the 95% CI of parameters estimated under the 1x model. 160

Figure 16. Bayesian MCMC fifty-percent majority-rule consensus phylogram compiled from analyses of 2306 bp of mitochondrial DNA sequences analyzed under the best-fit “10x” partitioned model (see text for model definition and selection). Consensus phylogram and posterior probabilities (shown adjacent to nodes) were estimated from a total of 9×10^6 post-burn-in generations (from three independent MCMC runs). Nodes receiving posterior probability support of 100% are indicated by grey-filled circles; otherwise, posterior probability support for nodes based on the 10x model is shown in black print. Posterior probability estimates based on the unpartitioned 1x model that differed notably from those from the 10x model are shown in black rectangles with white print (black boxes with dashes indicate clades that were not present in the consensus topology of the 1x tree). 161

Figure 17. Annotated mitochondrial genome maps of *Agkistrodon piscivorus* and *Pantherophis slowinskii*. The two *Agkistrodon* samples (*Api1* and *Api2*) have identical annotations except for minor variations in gene length. Labels of genes outside the circle refer to genes transcribed from the light strand, and names within the circle represent genes transcribed from the heavy strand..... 202

Figure 18. Differences per site for homologous genes or groups of sites in the two *Agkistrodon* genomes and in the two viperid genomes. The differences per site are shown for a comparison of *Api1* and *Api2* (A), and for *Agkistrodon* (mean of *Api1* and *Api2*) and *Ovophis* (B). Differences are shown only for the longer protein-coding genes. For the control regions only (shaded black), differences are shown for each aligned site including indels (e.g., CR1+I), or excluding indels (e.g., CR1-I). For all other genes, indels are not included in the difference measure. The bars for 3rd codon positions (3rd Codon) and for all codon positions (All Codon) are summed over all protein-coding genes..... 206

Figure 19. Maximum likelihood phylogeny for vertebrate taxa included in this study. This phylogeny is based on all protein-coding and rRNA genes. Most branches have greater than 95% support for both NJ ML distance bootstrap and Bayesian posterior probability support (see Methods), and are not annotated with support values. Where support from either measure is less than 95%, the support values are indicated by ratios, with the ML bootstrap support on top and the Bayesian posterior probability support below in italics, except for two nodes with less than 50% support by either measure, which are indicated by a hollow circle. Other than for these two nodes, support values less than 50% are indicated with an asterisk (*)..... 210

Figure 20. Hypotheses for the relative timing of alterations in mitochondrial genome architecture and molecular evolution throughout snake phylogeny. The topological relationships among snakes and branch lengths shown are the same as in Figure 3. Major groups of snakes are indicated along with the approximate diversification time of the Alethinophidia..... 211

Figure 21. Comparison of gene lengths in snakes and other squamates. The total length is shown for all protein coding regions (A), tRNAs (B), and rRNAs (C). All snakes are in gray, while other squamates (lizards) are in black, and light gray and dark gray bars are drawn under snake species to indicate membership in the Colubroidea or Henophidia, respectively. ... 216

Figure 22. Phylograms based on the relative branch lengths for rRNA and protein-coding genes, topologically constrained based on the ML phylogeny (Figure 3). Branch lengths on this constrained topology were estimated using all rRNA genes (A) or all protein-coding genes (B). The substitution rate scale is the same in both trees..... 218

Figure 23. Comparison of branch lengths from different genes and gene clusters for mammals, snakes, and lizards. Branch lengths for each gene or gene cluster are shown based on the cumulative branch lengths within each clade (A), or based on the gene or gene cluster branch length estimated along the ancestral branch leading to each nominal clade (B). Mammals are shown in gray, snakes in black, and lizards in white fill. rRNA branch lengths have been multiplied by ten to make them visible in this figure compared to protein branch lengths..... 220

Figure 24. Plot of branch lengths obtained from rRNA versus various genes and gene clusters. Snake branches are indicated with filled circles, and non-snake tetrapod branches are indicated with an unfilled circle. The locations of selected snake branches are labeled (in bold) with arrows. Outlying non-snake branches are indicated and labeled in normal type. Genes and gene clusters shown are (A) COX1, (B) CytB, (C) COX2 + ATP6 + ATP8, (D) ND2, and (E) COX3 + ND3 + ND4L, (F) ND1, (G) ND4, (H) ND5, (I) ND6..... 222

Figure 25. Standardized substitution rates across the mitochondrial genome for selected branches or clusters. For each 1000 bp window applied to a set of branches, standardized substitution rates were obtained by first dividing by the median window value for that branch, and then subtracting this value from the average across all non-snake branches. This helps to visualize regions of the genome that are evolving at slower or faster rates, with the average tetrapod relative rate being zero. Branches or branch sets shown are (A) the ancestor of all snakes and the ancestor of the Alethinophidia; (B) the ancestor of the Colubroidea and the sum of all colubroid terminal branches; and (C) the ancestor of the Henophidia and the sum of all henophidian terminal branches..... 224

CHAPTER 1 – INTRODUCTION

A Foundation for Comparative Genomics

“He looked at the streak of rust on the stone and thought of iron ore under the ground. To be melted and to emerge as girders against the sky... waiting for the drill, the dynamite and my voice; waiting to be split, ripped, pounded, reborn...” (Rand, 1943).

The expanding yet fledgling field of comparative genomics exists at the interface of several historically unrelated fields, thus relying on advances in a number of otherwise disjunct areas of science – computer science, computational biology, biological modeling, biochemistry, structural biology, systems biology, molecular and cellular biology, classical genetics, statistical genetics, and population genetics. The technological aspects of comparative genomics involved in the collection of essentially infinite amounts of genome sequence data has excelled far past the fields of science that enable the interpretation of this limitless resource of biological information. The factors limiting the extraction of vast amounts of biological system information from available genomic data is, therefore, not the ability to collect the data. Instead, the current limitations are imposed by the infancy of the science involved in analyzing and distilling patterns of genomic diversity into relevant quanta of information that may be applied to: 1) our understanding of the makeup and function of biological systems and, 2) deciphering how biological genomic diversity directs biological functional diversity.

Comparative genomic data are also absolutely essential for bridging biological information from numerous model organisms to the organisms of ultimate interest, whether the ultimate interests are humans, important crops, disease vectors, or agricultural pests. In fact, exploiting these comparative data is so crucial to deciphering biological information from genomic sequence that the core of the current phase of the Human Genome Project is to obtain additional comparable genomes and to develop comparative genome analyses to distill functional information from the human genome (Collins et al., 2003). Collectively, the studies that comprise this dissertation directly target major limitations of genomic analysis by addressing and advancing the theoretical and practical issues required for distilling biologically meaningful information from comparative genomic data, while also identifying several significant examples of dogmatic patterns of system-wide functional genome evolution in the mitochondrial genomes of vertebrates.

Complex Modeling of the Nucleotide Evolutionary Process

The most critical component of comparative genomics is an understanding of the evolutionary relationships and temporal contexts for comparative data points (be they organisms, genomes, genes, proteins, or single nucleotides), or in other words, having robust information detailing the relationship among species and genes that are being compared. This is particularly critical because such an evolutionary comparative framework provides the context whereby the order and precise patterns of genomic change may be understood, and subsequent overall

patterns of genome change may be correlated to functional biological effects. The most fundamental analyses in the field of comparative genomics rely not only on a well-known organismal phylogeny, but also the ability to compare this organismal tree of life with the phylogenetic relationships of members of multi-gene families or homologous genome regions (e.g., regulatory regions). These types of studies are critical for our inference of how the changes in non-protein-coding regions (regulatory regions), and the expansion/contraction, diversification, and modification of gene families in different genomes results in similarities or modifications observed in organismal complexity, ontogeny, and overall biological system function – the fundamental goal of comparative genomics.

Incorporating genetic data from multiple genes, often from multiple genomes, is becoming standard in molecular phylogenetics, as is the use of complex model-based likelihood techniques to estimate phylogenetic relationships based on these data. Despite numerous authors advocating the superiority of using multiple loci (especially from multiple genomes) to reconstruct phylogenies (e.g., Pamilo and Nei, 1988; Wu, 1991), few have addressed theoretical and practical effects of modeling sequence evolution simultaneously for different genes (but see examples: Yang, 1996a; Caterino et al., 2001; Pupko et al., 2002; Nylander et al., 2004). Using a single model with a single set of parameters to account for nucleotide substitution over heterogeneous gene regions in a combined analysis may fail to accurately portray locus-specific or site-specific evolutionary patterns. For instance, protein-coding vs. rRNA genes may evolve under drastically different constraints because protein-coding genes commonly experience particularly elevated rates of substitution at the third positions of codons. Ribosomal RNA genes, on the other hand, may experience relatively slow rates of compensatory change over regions

corresponding to stem-forming secondary structures in the core of the molecule, yet generally more rapid rates in regions corresponding to functionally distinct loops and short-range stems (e.g., Dixon and Hillis, 1993; Simon et al., 1994; Muse, 1995; Hickson et al., 1996; Savill et al., 2001). Even among protein-coding genes or rRNA genes, different patterns of substitution rate heterogeneity may result from overall differential rates of evolution or differential functional constraints on particular regions within a gene (Hickson et al., 1996; Yang, 1996b; Moncalvo et al., 2000). Considering these potential variations in evolutionary rates and patterns across sites, genes and genomes, it seems logical that models of molecular evolution that account for evolutionary heterogeneity need be employed to reconstruct phylogenetic trees.

The recent shift in phylogenetic methodology towards Bayesian inference of phylogeny has heightened the importance of the use of more realistic evolutionary models. This is important for topological accuracy (e.g., Huelsenbeck, 1995; Huelsenbeck, 1997; Sullivan and Swofford, 2001) as well as accurate estimation of support via posterior probabilities (e.g., Buckley, 2002; Suzuki et al., 2002; Erixon et al., 2003). A major strength of Bayesian analyses is that posterior probability distributions of trees allow direct interpretation of the likelihood of a particular relationship recovered being true, given the data, the model, and the priors. However, because the accuracy of posterior probabilities in Bayesian phylogenetic methods relies inherently on the model, even models that do not affect the consensus topology may still have important effects on the posterior probability distribution of parameters, and thus on confidence regarding phylogenetic conclusions. Therefore, employing complex models that more realistically and precisely portray natural patterns of DNA evolution should produce less biased posterior probability estimates as long as the added parameters can be accurately estimated from the data.

The accuracy of posterior probability estimates in Bayesian phylogenetic reconstruction and the factors that may affect this accuracy remain unclear. Many studies suggest Bayesian posterior probabilities appear to be inflated compared to conventional bootstrap support (e.g., Leaché and Reeder, 2002; Cummings et al., 2003; Douady et al., 2003), although the accuracy of posterior probability support values in terms of both type I and type II error remains unresolved (e.g., Buckley, 2002; Suzuki et al., 2002; Wilcox et al., 2002; Alfaro et al., 2003). Despite this, evidence is accumulating that suggests a direct relationship between accuracy of posterior probabilities and model complexity whereby Bayesian analyses conducted with underparameterized models appear to experience higher error rates compared with parameter rich models (Suzuki et al., 2002; Wilcox et al., 2002; Erixon et al., 2003). Complicating the matter, the benefits of constructing and employing more realistic evolutionary models of DNA substitution are challenged by the potential for imprecise and inaccurate parameter estimation (including topology) resulting from overparameterization. Given ever-increasing computational power, in addition to the speed afforded by Bayesian Markov Chain Monte Carlo phylogenetic methods, the need for accurate models and model testing is apparent.

The first three main chapters of this dissertation (Chapters 2, 3, and 4) focus on the development of methods for employing complex models of nucleotide evolution, and the practical effects of using such complex (versus simple) models of evolution for the inference of phylogenetic trees. The results of these studies have tremendous ramifications for achieving accuracy, reliability, and precision in essentially all comparative genomics studies, as accurate phylogenetic inferences are absolutely critical for meaningful comparative genomics. These three studies analyzed datasets designed to estimate relationships among organisms (rather than

genes within gene families) because organismal phylogeny questions have the advantage of being able to incorporate larger genomic datasets (ie, multiple genes) per branch being placed in the tree. Thus, they represent preferred training datasets and ideal model systems for investigating these evolutionary modeling questions, and the results of these studies are exactly comparable to more restricted problems of estimating gene genealogies in which the available data used to estimate trees is limited to the gene of interest.

Vertebrate Mitochondrial Genomes as a Model System for Comparative Genomics

In addition to the gaps in our understanding of modeling nucleotide evolution of genomes discussed above, a significant unexplored aspect of genomic research remains in the linking of the biological significance of genome structure, genome nucleotide evolution, and genome function (at the molecular and system-wide scales). Genomic research has demonstrated the plasticity of genome structure at almost all levels yet how this diversity of structure relates to genome evolution at the level of nucleotide evolution is essentially uninvestigated. How these features affect nucleotide evolution and genome function require further attention, ideally with a model system that is both sufficiently complex yet still computationally tractable.

Organelar genomes (mitochondrial and plastid genomes) present a valuable opportunity to explore patterns of genome evolution on a computationally manageable scale, unlike the computationally unwieldy nuclear genomes of eukaryotes, and may demonstrate insightful patterns applicable to their nuclear counterparts. Accordingly, vertebrate mitochondrial genomes

(mtDNA) have been a prevalent model system for studying molecular evolution, phylogenetic reconstruction, and genome structure. Furthermore, mitochondrial genomes are the genetic repository for some of the most critical genes in eukaryotes that play primary roles in aerobic metabolism and are also directly linked to apoptosis. Collectively, the tremendous functional significance of vertebrate mitochondrial genomes, together with their small size and limited complexity make them ideal models for comparative genomics.

The versatility and prominence of vertebrate mitochondrial genomes stems from their compactness and manageable size for sequencing and analysis, well-characterized replication and transcription processes (e.g., (Clayton, 1982; Fernandez-Silva et al., 2003; Shadel and Clayton, 1997; Szczesny et al., 2003; see also Holt and Jacobs, 2003; Reyes et al., 2005; Yang et al., 2002), and the diversity of protein and structural RNA genes that they encode. Vertebrate mitochondrial genomes generally lack recombination and have a conserved genome structure, although instances of intramolecular recombination have been proposed (Piganeau et al., 2004; Tsaousis et al., 2005), and there are numerous examples of structural rearrangements (Cooper et al., 2001; Mindell et al., 1998; Sankoff et al., 1992). Despite extensive molecular studies, little is known regarding the ways in which genome architecture might affect the various aspects of genome function and evolution (including replication, transcription, and function of proteins and RNAs). Nevertheless, patterns linking mitochondrial genome structure, function, and nucleotide evolution have begun to emerge (Krishnan et al., 2004a; Krishnan et al., 2004b; Raina et al., 2005).

Across vertebrates, mitochondrial genome size and structure are generally conserved (relative to plant mitochondria for example; Adams et al., 2002; Albert et al., 1998; Cho et al.,

1998; Cosner et al., 2001; Stiller et al., 2003). Typically, vertebrate mitochondrial genomes are characterized by a size of ~17kb and a gene content including 13 protein-coding genes, 2 rRNA genes, 22 tRNA genes, and a control region (or D-loop) involved in initiation of DNA replication and transcription. A traditional view of mitochondrial genome stability highlights the dramatic structural plasticity of plant mtDNA contrasting a ‘conserved’ structure across animal mitochondrial genomes (Palmer et al., 2000). While plant mitochondrial structural rearrangements are dramatic, the literature over the last several years has demonstrated significant structural diversity among animal mitochondrial genomes (Arntdt and Smith, 1998; Beagley et al., 1996; Downton and Austin, 1999; Hickerson and Cunningham, 2000; Karabayashi et al., 2000; Ladoukakis and Zouros, 2001; Shao et al., 2001; including vertebrate: Cooper et al., 2001; Macey et al., 1998; Macey et al., 1999; Macey et al., 2000; Mindell et al., 1998), especially considering their smaller size (~17 kb) as compared to plant mitochondrial genomes (~100 – 1000 kb; Adams et al., 2002; Palmer et al., 2000).

Despite their small size and limited gene content, vertebrate mitochondrial genomes show a diverse array of heterogeneous patterns of evolution both within and among genes (Krakauer and Plotkin, 2002; McKenzie et al., 2003; Monclavo et al., 2000; Nielsen, 1997; Pesole et al., 1999; Pupko et al., 2002; Rand, 2001; Savolainen et al., 2002). This collection of genic regions evolving under a mosaic of patterns presents an ideal, underutilized system for testing hypotheses of the molecular evolution of diverse genomic regions, with broad applications to modeling larger genomes (including eukaryotic nuclear genomes). Elevated rates of molecular evolution characteristic of vertebrate mitochondrial genomes additionally provide a high degree

of variance, which adds tremendous detection and hypothesis testing power to comparative genomic analyses.

Insight into the evolutionary forces that govern the composition of biological systems often comes from the study of extreme examples, where otherwise subtle patterns become dramatic and obvious. The mitochondrial genomes of snakes contain a number of particularly unusual qualities and structural features compared to other vertebrates. Snake mitochondrial genomes have elevated evolutionary rates and contain truncated tRNAs (Dong and Kumazawa, 2005; Kumazawa et al., 1998). All snake species sampled to date, except the scolecophidian snake *Leptotyphlops dulcis*, have a duplicated control region (CR2) between NADH dehydrogenase subunit 1 (ND1) and subunit 2 (ND2), in addition to a control region (CR1) adjacent to 5'-end of the 12s rRNA, as it is in other vertebrates. These two control regions appear to undergo concerted evolution that acts to homogenize the nucleotide sequence of each duplicate copy within a given genome (Dong and Kumazawa, 2005; Kumazawa et al., 1996, 1998). This pattern has been likened to the situation in chloroplast genomes (Kumazawa et al., 1996, 1998), many of which contain inverted repeated regions that are also maintained via concerted evolution (Goulding et al., 1996). Interestingly, based on comparisons between chloroplast genomes (cpDNA) with and without the inverted repeat region, several studies have suggested that the presence of the inverted repeats in cpDNA has genome-wide effects on both structural (Palmer and Thompson, 1982; Strauss et al., 1988) and nucleotide (Perry and Wolfe, 2002) evolutionary patterns, suggesting a link in a different organellar genome (other than mtDNA) between genome structure and nucleotide evolution.

In snake mtDNA, the functionality of the two control regions in transcription and initiation of heavy strand replication is not clear, but since the nucleotide sequence of each is nearly identical, any functional features that are not dependent on surrounding sequences should be similar. In contrast, recent evidence suggest that initiation of heavy strand replication may be distributed across a broad zone, including cytochrome b (CytB) and NADH dehydrogenase subunit 6 (ND6; Reyes et al., 2005), indicating that CR2 may not function as effectively in this role.

Using vertebrate mitochondrial genomes as a model, a number of interesting questions arise that might be addressed through comparative genomic analysis, including: (1) does one or the other, or do both control regions function as origins of heavy strand DNA synthesis? (2) does the altered genome structure affect patterns of snake mtDNA molecular evolution? (3) when during snake evolution did various features arise, and do particular features appear to coincide? (4) do patterns of molecular evolution vary at different depths of phylogeny? and (5) is there any evidence or plausible rationale for selection as a causative agent in generating these differences in genomic structure and molecular evolutionary patterns? In chapter 4 we address these questions, that are broadly relevant to comparative genome biology, by conducting an extensive analysis of vertebrate mitochondrial genomes, focusing on the extreme examples observed in the mitochondria of snakes.

References

- Adams, K.L., Qiu, Y., Stoutemyer, M., Palmer, J.D., 2002. Punctuated evolution of mitochondrial gene content: High and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. *Proc. Nat. Acad. Sci. U. S. A.* 99, 9905–9912.
- Albert, B., Godelle, B., Gouyon, P.-H., 1998. Evolution of the plant mitochondrial genome: dynamics of duplication and deletion of sequences. *J. Mol. Evol.* 46, 155–158.
- Alfaro, M. E., Zoller, S., Lutzoni, F., 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and Bootstrapping in assessing phylogenetic confidence. *Mol. Biol. Evol.* 20, 255–266.
- Arntdt, A., Smith, M.J., 1998. Mitochondrial gene rearrangement in the sea cucumber genus *Cucumaria*. *Mol. Biol. Evol.* 15, 1009–1016.
- Beagley, C.T., Okada, N.A., Wolstenholme, D.R., 1996. Two mitochondrial group I introns in a metazoan, the sea anemone *Metridium senile*: One intron contains genes for subunits 1 and 3 of NADH dehydrogenase. *Proc. Nat. Acad. Sci.* 93, 5619–5623.
- Buckley, T. R., 2002. Model misspecification and probabilistic tests of topology: evidence from empirical data sets. *Syst. Biol.* 51, 509–523.
- Caterino, M. S., Reed, R.D., Kuo, M.M., Sperling, F.A.H., 2001. A partitioned likelihood analysis of swallowtail butterfly phylogeny (Lepidoptera: Papilionidae). *Syst. Biol.* 50, 106–127.
- Cho, Y., Qiu, Y., Kuhlman, P., Palmer, J.D., 1998. Explosive invasion of plant mitochondria by a group I intron. *Proc. Nat. Acad. Sci.* 95, 14244–14249.

- Clayton, D.A., 1982. Replication of animal mitochondrial DNA. *Cell* 28, 693–705.
- Cooper, A., Lalueza-Fox, C., Anderson, S., Rambaut, A., Austin, J., Ward, R., 2001. Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. *Nature* 409, 704–707.
- Collins, F.S., Green, E.D., Guttmacher, A.E., Guyer, M.S., 2003. A vision for the future of genomics research. *Nature* 422, 835–847.
- Cosner, M.E., Jansen, R.K., Palmer, J.D., Downie, S.R., 1997. The highly rearranged chloroplast genome of *Trachelium caeruleum* (Campanulaceae): multiple inversions, inverted repeat expansion and contraction, transposition, insertions/deletions, and several repeat families. *Cur. Genet.* 31, 419–429.
- Cummings, M.P., Handley, S.A., Meyers, 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.
- Dixon, M.T., Hillis, D.M., 1993. Ribosomal RNA secondary structure: compensatory mutations and implications for phylogenetic analysis. *Mol. Biol. Evol.* 10, 256–267.
- Dong, S., Kumazawa, Y., 2005. Complete mitochondrial DNA sequences of six snakes: Phylogenetic relationships and molecular evolution of genomic features. *J. Mol. Evol.* 61, 12–22.
- Douady, C.J., Delsuc, F., Boucher, Y., Doolittle, W.F., Douzery, E.J.P., 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Mol. Biol. Evol.* 20, 248–254.

- Dowton, M., Austin, A.D., 1999. Evolutionary dynamics of a mitochondrial rearrangement “hot spot” in the hymenoptera. *Mol. Biol. Evol.* 16, 298–309.
- Erixon, S.P., Britton, B., Oxelman, B., 2003. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Syst. Biol.* 52, 665–673.
- Fernandez-Silva, P., Enriquez, J.A., Montoya, J., 2003. Replication and transcription of mammalian mitochondrial DNA. *Exp. Physiol.* 88, 41–56.
- Goulding, S.E., Olmstead, R.G., Morden, C.W., Wolfe, K.H., 1996. Ebb and flow of the chloroplast inverted repeat. *Mol. Gen. Genet.* 252, 195–206.
- Hickerson, M.J., Cunningham, C.W., 2000. Dramatic mitochondrial gene rearrangements in the hermit crab *Pagurus longicarpus* (Crustacea, Anomura). *Mol. Biol. Evol.* 17, 639–644.
- Hickson, R., Simon, C., Cooper, A., Spicer, G., Sullivan, J., Penny, D., 1996. Conserved sequence motifs, alignment, and secondary structure for the third domain of animal 12s rRNA. *Mol. Biol. Evol.* 13, 150–169.
- Holt, I.J., Jacobs, H.T., 2003. Response: The mitochondrial DNA replication bubble has not burst. *Trends Biochem. Sci.* 28, 355–356.
- Huelsenbeck, J.P., 1995. Performance of phylogenetic methods in simulation. *Syst. Biol.* 44, 17–48.
- Huelsenbeck, J. P., 1997. Is the Felsenstein zone a fly trap? *Syst. Biol.* 46, 69–74.
- Krakauer, D.C., Plotkin, J.B., 2002. Redundancy, antiredundancy, and the robustness of genomes. *Proc. Nat. Acad. Sci. U. S. A.* 99, 1405–1409.
- Krishnan, N.M., Raina, S.Z., Pollock, D.D., 2004a. Analysis of among-site variation in substitution patterns. *Biological Procedures Online* 6, 180–188.

- Krishnan, N.M., Seligmann, H., Raina, S.Z., Pollock, D.D., 2004b. Detecting gradients of asymmetry in site-specific substitutions in mitochondrial genomes. *DNA Cell Biol.* 23, 707–714.
- Kumazawa, Y., Ota, H., Nishida, M., Ozawa, T., 1996. Gene rearrangements in snake mitochondrial genomes: Highly concerted evolution of control-region-like sequences duplicated and inserted into a tRNA gene cluster. *Mol. Biol. Evol.* 13, 1242–1254.
- Kumazawa, Y., Ota, H., Nishida, M., Ozawa, T., 1998. The complete nucleotide sequence of a snake (*Dinodon semicarinatus*) mitochondrial genome with two identical control regions. *Genetics* 150, 313–329.
- Kurabayashi, A., Ueshima, R., 2000. Complete sequence of the mitochondrial DNA of the primitive opisthobranch gastropod *Pupa strigosa*: systematic implication of the genome organization. *Mol. Biol. Evol.* 17, 266–277.
- Ladoukakis, E.D., Zouros, E., 2001. Direct evidence for homologous recombination in mussel (*Mytilus galloprovincialis*) mitochondrial DNA. *Mol. Biol. Evol.* 18, 1168–1175.
- Leaché, A.D., Reeder, T.W., 2002. Molecular systematics of the eastern fence lizard (*Sceloporus undulatus*): a comparison of parsimony, likelihood, and Bayesian approaches. *Syst. Biol.* 51, 44–68.
- Macey, J.R., Schulte, J.A. II., Larson, A., Papenfuss, T.J., 1998. Tandem Duplication Via Light-Strand Synthesis May Provide a Precursor for Mitochondrial Genomic Rearrangement. *Mol. Biol. Evol.* 15, 71–75.

- Macey, J.R., Schulte, J.A. II., Larson, A., Tuniyev, B.S., Orlov, N., Papenfuss, T. J., 1999. Molecular phylogenetics, tRNA evolution, and historical biogeography in anguid lizards and related taxonomic families. *Mol. Phylogenet. Evol.* 12, 250–272.
- Macey, J.R., Schulte, J.A. II., Larson, A., 2000. Evolution and phylogenetic information content of mitochondrial genomic structural features illustrated with acrodont lizards. *Syst. Biol.* 49, 257–277.
- McKenzie, M., Chiotis, M., Pinkert, C.A., Trounce, I.A., 2003. Functional respiratory chain analysis in murid xenomitochondrial cybrids expose coevolutionary constraints of cytochrome *b* and nuclear subunits of complex III. *Mol. Biol. Evol.*, 20, 1117–1124.
- Mindell, D.P., Sorenson, M.D., Dimcheff, D.E., 1998. Multiple independent origins of mitochondrial gene order in birds. *Proc. Natl. Acad. Sci. U. S. A.* 95, 10693–10697.
- Moncalvo, J.M., Drehmel, D., Vilgalys, R., 2000. Variation in modes and rates of evolution in nuclear and mitochondrial ribosomal DNA in the mushroom genus *Aminita* (Agaricales, Basidiomycota): phylogenetic implications. *Mol. Phylogenet. Evol.* 16, 48–63.
- Muse, S.V., 1995. Evolutionary analyses when nucleotides do not evolve independently. Pages 115–124 In *Current topics on molecular evolution* (M. Nei and N. Takahata, eds.). Institute on Molecular Evolution and Genetics, University Park, Pennsylvania.
- Nielsen, R., 1997. Site-by-site estimation of the rate of substitution and the correlation of rates in mitochondrial DNA. *Syst. Biol.* 46, 346–353.
- 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.

- Palmer, J.D., Adams, K.L., Cho, Y., Parkinson, C.L., Qiu, Y., Song, K., 2000. Dynamic evolution of plant mitochondrial genomes: Mobile genes and introns and highly variable mutation rates. *Proc. Nat. Acad. Sci. U. S. A.* 97, 6960–6966.
- Palmer, J.D., Thompson, W.F., 1982. Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. *Cell* 29, 537–550.
- Pamilo, P., Nei, M., 1988. Relationships between gene trees and species trees. *Mol. Biol. Evol.* 5, 568–583.
- Perry, A.S., Wolfe, K.H., 2002. Nucleotide substitution rates in legume chloroplast DNA depend on the presence of the inverted repeat. *J. Mol. Evol.* 55, 501–508.
- Pesole, G., Gissi, C., De Chirico, A., Saccone, C., 1999. Nucleotide Substitution Rate of Mammalian Mitochondrial Genomes. *J. Mol. Evol.* 48, 427–434.
- Piganeau, G., Gardner, M., Eyre-Walker, A., 2004. A broad survey of recombination in animal mitochondria. *Mol. Biol. Evol.* 21, 2319–2325.
- Pupko, T., Huchon, D., Cao, Y., Okada, N., Hasegawa, M., 2002. Combining multiple data sets in a likelihood analysis: which models are the best? *Mol. Biol. Evol.* 19, 2294–2307.
- Raina, S.Z., Faith, J.J., Disotell, T.R., Seligmann, H., Stewart, C.B., Pollock, D.D., 2005. Evolution of base-substitution gradients in primate mitochondrial genomes. *Genome Res.* 15, 665–673.
- Rand, A., 1943. *The Fountainhead*. Bobbs-Merril Company, Inc, NY, NY.
- Rand, D.M., 2001. The units of selection on mitochondrial DNA. *Ann. Rev. Ecol. Syst.* 32, 415–448.

- Reyes, A., Yang, M.Y., Bowmaker, M., Holt, I.J., 2005. Bidirectional replication initiates at sites throughout the mitochondrial genome of birds. *J. Biol. Chem.* 280, 3242–3250.
- Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B.F., Cedergren, R., 1992. Gene order comparisons for phylogenetic inference - evolution of the mitochondrial genome. *Proc. Natl. Acad. Sci. U. S. A.* 89, 6575–6579.
- Savill, N.J., Hoyle, D.C., Higgs, P.G., 2001. RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods. *Genetics* 157, 399–411.
- Savolainen, V., Chase, M.W., Salamin, N., Soltis, D.E., Soltis, P.S., López, A.J., Fèdrigo, O., Naylor, G.J.P., 2002. Phylogeny reconstruction and functional constraints in organellar genomes: plastid *atpB* and *rbcL* sequences versus animal mitochondrion. *Syst. Biol.* 51, 638–647.
- Shadel, G.S., Clayton, D.A., 1997. Mitochondrial DNA maintenance in vertebrates. *Annu. Rev. Biochem.* 66, 409–43
- Shao, R., Campbell, N.J.H., Barker, S.C., 2001. Numerous gene rearrangements in the mitochondrial genome of the wallaby louse, *Heterodoxus macropus* (Phthiraptera). *Mol. Biol. Evol.* 18, 858–865.
- Simon, C., Frati, F., Beckenbach, A., Brespi, B., Liu, H., Flook, P., 1994. Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved PCR primers. *Annals Entomol. Soc. Am.* 87, 651–701.
- Stiller, J.W., Reel, D.C., Johnson, J.C., 2003. A single origin of plastids revisited: convergent evolution in organellar genome content. *J. Phycol.* 39, 95–105.

- Strauss, S.H., Palmer, J.D., Howe, G.T., Doerksen, A.H., 1988. Chloroplast genomes of two conifers lack a large inverted repeat and are extensively rearranged. *Proc. Nat. Acad. Sci. U. S. A.* 85, 3898–3902.
- Sullivan, J., Swofford, D.L., 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? *Syst. Biol.* 50, 723–729.
- Suzuki, Y., Glazko, G.V., Nei, M., 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *Proc. Nat. Acad. Sci. U. S. A.* 99, 16138–16143.
- Szczesny, B., Hazra, T.K., Papaconstantinou, J., Mitra, S., Boldogh, I., 2003. Age-dependent deficiency in import of mitochondrial DNA glycosylases required for repair of oxidatively damaged bases. *Proc. Natl. Acad. Sci. U. S. A.* 100, 10670–10675.
- Tsaousis, A.D., Martin, D.P., Ladoukakis, E.D., Posada, D., Zouros, E., 2005. Widespread recombination in published animal mtDNA sequences. *Mol. Biol. Evol.* 22, 925–933.
- Wilcox, T.P., Zwickl, D.J., Heath, T.A., Hillis, D.M., 2002. Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. *Mol. Phylogenet. Evol.* 25, 361–371.
- Wu, C.I., 1991. Inferences of species phylogeny in relation to segregation of ancient polymorphisms. *Genetics* 127, 429–435.
- Yang, M.Y., Bowmaker, M., Reyes, A., Vergani, L., Angeli, P., Gringeri, E., Jacobs, H.T., Holt, I.J., 2002. Biased incorporation of ribonucleotides on the mitochondrial L-strand accounts for apparent strand-asymmetric DNA replication. *Cell* 111, 495–505.

Yang, Z., 1996a. Maximum likelihood models for combined analyses of multiple sequence data.

J. Mol. Evol. 42, 587–596.

Yang, Z., 1996b. Among-site rate variation and its impacts on phylogenetic analyses. Trends

Ecol. Evol. 11, 367–372.

CHAPTER 2 – DATA PARTITIONS AND COMPLEX MODELS IN BAYESIAN ANALYSIS: THE PHYLOGENY OF GYMNOPHTHALMID LIZARDS

Introduction

Incorporating genetic data from multiple genes, often from multiple genomes, is becoming standard in molecular phylogenetics, as is the use of complex model-based likelihood techniques to estimate phylogenetic relationships based on these data. Despite numerous authors advocating the superiority of using multiple loci (especially from multiple genomes) to reconstruct phylogenies (e.g., Pamilo and Nei, 1988; Wu, 1991), few have addressed theoretical and practical effects of modeling sequence evolution simultaneously for different genes (but see examples: Yang, 1996a; Caterino et al., 2001; Pupko et al., 2002; Nylander et al., in press). Using a single model with a single set of parameters to account for evolution over multiple loci in a combined analysis may fail to accurately portray locus-specific evolutionary patterns. For instance, protein-coding vs. rRNA genes may evolve under drastically different constraints because protein-coding genes commonly experience particularly elevated rates of substitution at the third positions of codons. Ribosomal RNA genes, on the other hand, may experience relatively slow rates of compensatory change over regions corresponding to stem-forming secondary structures in the core of the molecule, yet generally more rapid rates in regions corresponding to functionally unconstrained loops and short-range stems (e.g., Dixon and Hillis,

1993; Simon et al., 1994; Muse, 1995; Hickson et al., 1996; Savill et al., 2001). Even among protein-coding genes or rRNA genes, different patterns of substitution rate heterogeneity may result from overall differential rates of evolution or differential functional constraints on particular regions within a gene (Hickson et al., 1996; Yang, 1996b; Moncalvo et al., 2000). Considering these potential variations in evolutionary rates across sites, it seems logical that models of molecular evolution that account for heterogeneity with regard to among-site rate variation should be employed to reconstruct phylogenetic trees.

The recent shift in phylogenetic methodology towards Bayesian inference of phylogeny has heightened the importance of the use of more realistic evolutionary models. This is important for topological accuracy (e.g., Huelsenbeck, 1995; Huelsenbeck, 1997; Sullivan and Swofford, 2001) as well as accurate estimation of posterior probabilities (e.g., Buckley, 2002; Suzuki et al., 2002; Erixon et al., 2003). In general, it has been shown that likelihood methods are fairly robust to model choice in their estimation of topology (Yang et al., 1994; Posada and Crandall, 2001; Sullivan and Swofford, 2001). A major strength of Bayesian analyses is that posterior probability distributions of trees allow direct interpretation of the likelihood of a particular relationship recovered being true, given the data, the model, and the priors (although the robustness of posterior probabilities has not been thoroughly investigated). However, because the accuracy of posterior probabilities in Bayesian phylogenetic methods relies inherently on the model, models that do not affect the consensus topology may have notable effects on the posterior probability distribution of parameters, and thus on confidence regarding phylogenetic conclusions. Therefore, employing complex models that more accurately portray

DNA evolution should produce less biased posterior probability estimates as long as parameters can be accurately estimated from the data.

The accuracy of posterior probability estimates in Bayesian phylogenetic reconstruction and the factors that may affect this accuracy remain unclear. Many studies suggest Bayesian posterior probabilities appear to be inflated compared to conventional bootstrap support (e.g., Leaché and Reeder, 2002; Cummings et al., 2003; Douady et al., 2003). However, the accuracy of posterior probability support values in terms of both type I and type II error remains unresolved (e.g., Buckley, 2002; Suzuki et al., 2002; Wilcox et al., 2002; Alfaro et al., 2003). Despite this, evidence is accumulating that suggests a direct relationship between accuracy of posterior probabilities and model complexity whereby Bayesian analyses conducted with underparameterized models appear to experience higher error rates compared with parameter rich models (Suzuki et al., 2002; Wilcox et al., 2002; Erixon et al., 2003). However, benefits of constructing and employing more realistic evolutionary models of DNA substitution are challenged by the potential for imprecise and inaccurate parameter estimation (including topology) resulting from overparameterization. Given ever-increasing computational power, in addition to the speed afforded by Bayesian Markov Chain Monte Carlo phylogenetic methods, the need for accurate models and model testing is apparent.

Our primary goal in this paper is to concentrate on evaluation of alternative models which practically affect phylogenetic inference. Specifically, we designed our approaches to make final decisions about best-fitting models based on the effects they had on topology and posterior probability estimates. This is important because while some alternative, relatively parameter rich models may provide a better fit to the data, they may not result in alternative topologies or

significantly different posterior probability estimates. In such cases, our strategy would instead favor a model with fewer parameters that produced essentially the same topology and posterior probability support estimates.

In this study, the genes used to reconstruct phylogenies are diverse and include one protein-coding nuclear gene (*c-mos*), one protein-coding mitochondrial gene (ND4) and two rRNA mitochondrial gene fragments (12S and 16S). We focused on the construction and evaluation of models that utilize alternatively partitioned patterns of among-site rate variation to account for heterogeneous evolution of multiple loci in a combined phylogenetic analysis. Particularly, MrBayes v2.01 allows among-site rate variation (γ ; Yang, 1993) to be partitioned among defined sites (site-specific γ) as well as allowing the use of an auto-correlated γ parameter to account for local auto-correlation of among-site rates (Kimura, 1985; Schöniger and von Haeseler, 1994; Yang, 1995; Nielsen, 1997). These models allow γ parameter for among-site rate variation to be rescaled across partitions while using a single rate nucleotide substitution rate matrix for the entire data set. Along with conventional models of sequence evolution (e.g., GTR+I+G), we explore more complex models which partition the among-site rate variation in various ways among loci, in addition to those which employ an additional parameter for auto-correlation of site rate variation. We examine the phylogenetic hypotheses resulting from several alternative partitions of among-site rate variation and discuss their relevance to Bayesian support for clades and support for alternative topological placements of clades.

The taxonomic group examined in this study, lizards of the family Gymnophthalmidae, comprises a large radiation consisting of approximately 34 genera and 180 species occurring

throughout South America with relatively few species in Middle America (Pellegrino et al., 2001; Doan, 2003a). The family is composed of small to medium lizards that occur in a variety of habitats and occupy a wide range of niches. This lizard group has been poorly studied with many species unknown beyond their original descriptions.

Relationships of genera within the family Gymnophthalmidae are poorly understood. The most comprehensive and contemporary revision of the supergeneric classification of the family Gymnophthalmidae was made by Pellegrino et al. (2001). They reconstructed a phylogeny of 50 species in 24 genera (recently reduced from 26 by Doan, 2003a) using five genes (two nuclear and three mitochondrial). Based on their reconstruction they erected four subfamilies and four tribes. The subfamily Alopoglossinae, consists solely of *Alopoglossus*. The subfamily Gymnophthalminae contains 13 genera, divided into two tribes, the Heterodactylini (5 genera) and the Gymnophthalmini (8 genera). The subfamily Rhachisaurinae is monotypic, consisting of *Rhachisaurus brachylepis*, a new genus separated from *Anotosaura*. The final subfamily, the Cercosaurinae, consists of 20 genera divided into tribe Cercosaurini (14 genera) and tribe Eupleopini (6 genera).

Harris (2003) used c-mos sequences to reconstruct a phylogeny of the Squamata with concentrated taxon sampling in the Gymnophthalmidae. He primarily used Pellegrino et al.'s (2001) sequences but added a new sequence of *Proctoporus bolivianus*. Harris's (2003) reconstruction differed from Pellegrino et al.'s in the placement of *Ptychoglossus*, *Bachia*, *Arthrosaura*, and several smaller scale relationships. In addition, a teiid genus, *Tupinambis*, was nested within the Gymnophthalmidae as the sister to *Ptychoglossus* and *Alopoglossus* (although this relationship received bootstrap and posterior probability support below 50%).

Missing from Pellegrino et al.'s (2001) study were 10 genera. Whereas Pellegrino et al. sampled all genera of Alopoglossinae, Rhachisaurinae, and Gymnophthalmini, they lacked *Stenolepis* from the Heterodactylini, *Amapasaurus* from the Eupleopini, and eight genera from the Cercosaurini. The limited taxon sampling of the Cercosaurini renders conclusions about that tribe problematic (see Hillis, 1998) because only half of the genera and 18 of the approximately 121 species were sampled (14.9%).

In addition to limited taxon sampling, the separate gene partitions were often in conflict with regards to the positions and relationships of many key taxa (Pellegrino et al., 2001). Such conflicts put into doubt the subfamilial and/or tribal placement of genera such as *Ptychoglossus*, *Rhachisaurus*, *Bachia*, and *Neusticurus*. Our study addresses some of the problems suggested in the combined analysis of Pellegrino et al. (2001) and fills in some significant gaps in taxon sampling. Concentrating on the Cercosaurini, we add 19 new individuals, including 12 new species and one new genus, as well as an additional individual of *Ptychoglossus brevifrontalis* (a total of 73 additional sequences). With this greater taxon sampling we clarify the relationships and classification within family Gymnophthalmidae. In addition to adding more taxa, we utilize a Bayesian approach to the phylogeny to complement the standard parsimony and likelihood methods used by Pellegrino et al. (2001). We synthesize this information, emphasizing overall phylogenetic evidence, and propose an alternative hypothesis for the inter-generic relationships and taxonomy within the family.

The objectives of our study included: 1) evaluating the effects of partitioning among-site rate variation and among-site rate auto-correlation parameters on phylogenetic topology and posterior probabilities for relationships, 2) developing a robust strategy for choosing the best-fit

model for among-site rate variation considering practical effects on topology and posterior probability support, 3) identifying the most likely and robust phylogenetic hypothesis for relationships among gymnophthalmid lizards, and 4) re-evaluating the supergeneric classification of the family based on our best estimate of gymnophthalmid phylogeny.

Methods

DNA Sequences Used

A significant subset of the sequences used in this study is from Pellegrino et al. (2001) and Doan and Castoe (2003). Additional sequences of gymnophthalmid lizards were added to this dataset. Of the five genes used by Pellegrino et al. (2001), we chose to use and expand upon four: mitochondrial NADH dehydrogenase subunit 4 (ND4), mitochondrial small subunit rRNA gene (12S), mitochondrial large subunit rRNA gene (16S), and the nuclear oocyte maturation factor gene (*c-mos*). The nuclear small subunit rRNA gene (18S) used by Pellegrino et al. (2001) was omitted from our study for two reasons: 1) low phylogenetic signal apparent from the Pellegrino et al. (2001) study, and 2) the nuclear gene for 18S occurs in hundreds or thousands of copies per nuclear genome (e.g., humans, International Human Genome Sequencing Consortium, 2001; *Xenopus*, Pardue, 1974; Long and Dawid, 1980). Using sequences of this multi-copy gene to resolve relationships principally among species and genera within a family may increase the potential for recovery of misleading phylogenetic estimates based on incomplete gene

conversion among alleles at different loci or differential fixation of alleles among loci (Gasser et al, 1998; Gonzalez and Sylvester, 2001).

Laboratory methods for obtaining novel sequences used in this study are as follows. Where possible, two individuals of each taxon from distant sampling localities were added to the data set. Genomic DNA was isolated from tissue samples (liver or skin preserved in ethanol) using the Qiagen DNeasy extraction kit and protocol (Qiagen Inc., Hilden, Germany). The mitochondrial ND4 gene was amplified via PCR using the primers ND4 and LEU as described in Arévalo et al. (1994). Mitochondrial ribosomal small and large subunit genes (12S and 16S) were amplified as described in Parkinson et al. (1997) and Parkinson et al. (1999). The nuclear c-mos gene was amplified with primers G73 and G74 as described in Saint et al. (1998) and Pellegrino et al. (2001). Positive PCR products were excised from agarose electrophoretic gels and purified using the GeneCleanIII kit (BIO101). Purified PCR products were sequenced in both directions with the amplification primers (and for ND4, an additional internal primer HIS, Arévalo et al., 1994). Samples that could not be sufficiently sequenced directly were cloned using the Topo TA cloning kit (Invitrogen) according to the manufacturer's protocols. Plasmids were isolated from multiple clones per individual using the Qiaquick spin miniprep kit (Qiagen). Plasmids were sequenced using M13 primers (provided by Topo TA kit, Invitrogen) and, in some cases, the internal HIS primer for ND4. Purified PCR products and plasmids were sequenced using the CEQ D Dye Terminator Cycle Sequencing (DTCS) Quick Start Kit (Beckman-Coulter) and run on a Beckman CEQ2000 automated sequencer according to the manufacturers' protocols. Raw sequence chromatographs for sequences generated in this study were edited using Sequencher 3.1 (Gene Codes Corp.). In cases where gene fragments were

cloned, all sequences from a single individual were edited together. Novel sequences were deposited in GenBank. The GenBank accession numbers for each gene sequence used in this study (including novel sequences) are given in Table 1.

Sequence Homology and Alignment

Multiple sequence alignment was performed using ClustalW (Thompson et al., 1994). Initial alignments were conducted with a gap opening penalty of 10, a gap extension penalty of 1, and a transition weight of 0.5. For rRNA genes (12S and 16S), alternative multiple alignments were examined with gap opening and gap extension penalties ranging from 10 and 10 (respectively) to 1 and 1, including varying ratios in this range. Initial alignments for protein-coding genes (ND4 and c-mos) were rechecked based on the homology of their translated amino acid sequence using GeneDoc (Nicholas and Nicholas, 1997). The ND4 alignment was unambiguous and not edited manually and the c-mos alignment was slightly manually modified to maximize amino acid similarity over a short indel region within the alignment. Alternative automated alignments (from ClustalW) for rRNA genes (12S and 16S) were compared, along with estimates of secondary structures (Gutell, 1994; Gutell et al., 1994; Titus and Frost, 1996), to evaluate evidence for positional homology. Positions in rRNA genes where alignment was ambiguous were excluded. To minimize the effects of missing data resulting from incomplete sequences, all gene alignments were truncated at the 5' and 3' ends. The final alignment of all concatenated genes, including positions excluded from phylogenetic analyses, is available as supplemental data at <http://biology.ucf.edu/~clp/Lab/Lab.htm>.

Table 1. List of sequences used in this study. operational taxonomic units (OTUs) used in this study with GenBank accession numbers. Cells with an X indicate that gene sequence was not used in this study. (a) and (b) refer to individuals indicated in the figures. Museum accession numbers for specimens sequenced in this study are given. Acronyms for museums are: KU (University of Kansas), MHNSM (Museo de Historia Natural, Universidad Nacional Mayor de San Marcos, Lima, Peru), QCAZ (Museo de Zoología, Pontifica Universidad Católica del Ecuador, Quito, Ecuador) and UTA (University of Texas at Arlington).

OTU	Museum number	ND4	c-mos	12S	16S
<i>Alopoglossus atriventris</i>		AF420908	AF420821	AF420695	AF420746
<i>Alopoglossus carimicaudatus</i>		AF420909	AF420847	AF420693	AF420744
<i>Alopoglossus copii</i>		AF420865	AF420819	AF420692	AF420745
<i>Anotosaura spn</i>		AF420902	X	AF420682	AF420719
<i>Anotosaura vanzolinia</i>		AF420910	X	AF420670	AF420724
<i>Arthrosaura kockii</i>		AF420866	X	AF420680	AF420721
<i>Arthrosaura reticulata</i>		AF420894	X	AF420676	AF420722
<i>Bachia bresslaui</i>		AF420876	AF420860	X	AF420755
<i>Bachia dorbignyi</i>		AF420892	X	AF420688	AF420754
<i>Bachia flavescens</i>		AF420869	AF420859	AF420705	AF420753
<i>Calyptommatus leirolepis</i>		AF420874	AF420858	AF420683	AF420712
<i>Calyptommatus nicterus</i>		AF420903	AF420822	AF420684	AF420747
<i>Calyptommatus sinebrachiatus</i>		AF420873	AF420832	AF420685	AF420720
<i>Cercosaura argulus</i> (a)		AF420896	AF420838	AF420698	AF420751
<i>Cercosaura argulus</i> (b)		AF420893	AF420852	AF420696	AF420750
<i>Cercosaura eigenmanni</i>		AF420895	AF420828	AF420690	AF420728
<i>Cercosaura ocellata</i>		AF420883	AF420834	AF420677	AF420731
<i>Cercosaura quadrilineata</i>		AF420880	AF420830	AF420672	AF420717
<i>Cercosaura schreibersii albostrigata</i>		AF420882	AF420856	AF420658	AF420729
<i>Cercosaura schreibersii schreibersii</i>		AF420911	AF420817	AF420686	AF420749
<i>Colobodactylus dalcyanus</i>		AF420881	AF420844	AF420663	AF420736
<i>Colobodactylus taunayi</i>		X	AF420831	AF420662	AF420741
<i>Colobosaura mentalis</i>		AF420899	AF420842	AF420694	AF420726
<i>Colobosaura modesta</i>		AF420887	AF420845	AF420666	AF420733
<i>Colobosaura spn</i>		AF420868	AF420840	AF420667	AF420739
<i>Colobosauroides cearensis</i>		AF420886	AF420849	AF420659	AF420727
<i>Ecpleopus gaudichaudii</i>		AF420901	AF420855	AF420660	AF420738
<i>Gymnophthalmus leucomystax</i>		AF420906	AF420824	AF420675	AF420715
<i>Gymnophthalmus vanzoi</i>		AF420867	AF420827	AF420687	AF420743
<i>Heterodactylus imbricatus</i>		AF420885	AF420835	AF420661	AF420725
<i>Iphisa elegans</i>		AF420889	AF420843	AF420668	AF420714
<i>Leposoma oswaldoi</i>		AF420897	AF420854	AF420678	AF420723
<i>Leposoma percarinatum</i>		AF420898	X	AF420700	AF420735
<i>Micrablepharus aticolus</i>		AF420904	AF420826	AF420664	AF420718
<i>Micrablepharus maximiliani</i>		AF420875	AF420850	AF420657	AF420730
<i>Neusticurus bicarinatus</i>		X	AF420816	AF420671	AF420708
<i>Neusticurus ecpleopus</i>		AF420890	AF420829	AF420656	AF420748
<i>Neusticurus juruazensis</i>		AF420878	AF420857	AF420704	AF420758
<i>Neusticurus rudis</i>		AF420905	X	AF420689	AF420709
<i>Neusticurus strangulatus</i>	KU 21677		X		
<i>Nothobachia ablephara</i>		AF420900	AF420851	AF420669	AF420740
<i>Pholidobolus macbrydei</i>	KU 218406				
<i>Pholidobolus montium</i>		AF420884	AF420820	AF420701	AF420756
<i>Placosoma cordylinum</i>		AF420879	AF420823	AF420673	AF420734
<i>Placosoma glabellum</i>		AF420907	AF420833	AF420674	AF420742
<i>Procellosaurinus erythrocerus</i>		AF420870	AF420836	AF420679	AF420711
<i>Procellosaurinus tetradactylus</i>		AF420871	AF420818	AF420703	AF420713

OTU	Museum number	ND4	c-mos	12S	16S
<i>Proctoporus bolivianus</i> (a)	UTA R-51506	AY225175			
<i>Proctoporus bolivianus</i> (b)	UTA R-51487	AY225180			
<i>Proctoporus cashcaensis</i>	KU 217205		X		
<i>Proctoporus colomaromani</i>	KU 217209				
<i>Proctoporus guentheri</i> (a)	UTA R-51515	AY225185			
<i>Proctoporus guentheri</i> (b)	UTA R-51517	AY225169			
<i>Proctoporus orcesi</i>	KU 221772		X		
<i>Proctoporus simoterus</i>	KU 217207				
<i>Proctoporus succullucu</i> (a)	UTA R-51478	AY225171			
<i>Proctoporus succullucu</i> (b)	UTA R-51496	AY225177			
<i>Proctoporus unicolor</i>	KU 217211				
<i>Proctoporus unsaaca</i> (a)	UTA R-51477	AY225170			
<i>Proctoporus unsaaca</i> (b)	UTA R-51488	AY225186			
<i>Proctoporus ventrimaculatus</i>	KU 219838				
<i>Proctoporus cf. ventrimaculatus</i>	KU 212687				
<i>Proctoporus</i> sp. K19	QCAZ 879				
<i>Psilophthalmus paeminosus</i>		AF420872	AF420825	AF420702	AF420710
<i>Ptychoglossus brevifrontalis</i>		X	AF420848	AF420697	AF420757
<i>Ptychoglossus brevifrontalis</i>	MHNSM				
<i>Rhachisaurus brachylepis</i>		AF420877	AF420853	AF420665	AF420737
<i>Tretioscincus agilis</i>		AF420891	AF420837	AF420681	AF420732
<i>Tretioscincus oriximinensis</i>		AF420888	AF420846	AF420691	AF420752
<i>Vanzosaura rubricauda</i>		X	AF420839	AF420699	AF420716
<i>Cnemidophorus ocellifer</i>		AF420914	AF420862	AF420706	AF420759
<i>Kentropyx calcarata</i>		AF420913	AF420864	AF420707	AF420760
<i>Tupinambis quadrilineatus</i>		AF420912	AF420863	X	AF420761

automated alignments (from ClustalW) for rRNA genes (12S and 16S) were compared, along with estimates of secondary structures (Gutell, 1994; Gutell et al., 1994; Titus and Frost, 1996), to evaluate evidence for positional homology. Positions in rRNA genes where alignment was ambiguous were excluded. To minimize the effects of missing data resulting from incomplete sequences, all gene alignments were truncated at the 5' and 3' ends. The final alignment of all concatenated genes, including positions excluded from phylogenetic analyses, is available as supplemental data at <http://biology.ucf.edu/~clp/Lab/Lab.htm>.

Phylogeny Estimation Using Maximum Parsimony

We inferred phylogenies based on the maximum parsimony (MP) criterion in PAUP* v4.0b10 (Swofford, 2002) and Bayesian (Markov Chain Monte Carlo, MCMC) analysis in MrBayes v2.01 (Huelsenbeck and Ronquist, 2001). Phylogenetic inference was conducted, hierarchically, in three steps: 1) all genes individually, 2) intermediate partitions including both rRNA genes (12S+16S) and all mitochondrial genes (mtgenes), and 3) the combined concatenated dataset including all four genes. For MP analyses of independent genes and intermediate partitions, we conducted equally-weighted parsimony searches using the heuristic strategy with 200 random taxon addition sequence replicates. Settings for MP analyses were tree bisection-reconnection branch swapping, steepest descent off, and MULTREES option on (Swofford, 2002). For all individual genes and intermediate partitions (rRNA and mtgenes) we assessed support for clades using 200 nonparametric bootstrap pseudoreplicates (Felsenstein, 1985) with 20 random taxon addition sequence replicates implemented with PAUP*. For the combined MP analysis of all genes we searched for trees using equally-weighted parsimony

heuristic searches with 1000 random taxon addition sequence replicates and assessed clade support with 200 bootstrap pseudoreplicates with 200 random addition sequence replicates per bootstrap pseudoreplicate. We consider relationships that are supported by at least 70% bootstrap to be significantly resolved (Hillis and Bull, 1993).

Bayesian Phylogeny Estimation

ModelTest version 3.0 (Posada and Crandall, 1998) was used to infer the best-fit model of evolution for each gene data set (individual genes, intermediate partitions, and total combined data) based on hierarchical log-likelihood ratio tests comparing successively complex models (Huelsenbeck & Crandall, 1997; Posada & Crandall, 2001).

All MCMC phylogenetic reconstructions were conducted in MrBayes v2.01 (Huelsenbeck and Ronquist, 2001) with vague priors (as per the program's defaults) and model parameters estimated as part of the analyses. Three heated chains and a single cold chain were used in all MCMC analyses and runs were initiated with random trees, as per the program's defaults. Trees were sampled every 100 generations and majority rule consensus phylograms and posterior probabilities for nodes were assembled from all post burn-in sampled trees. Phylogenetic reconstructions for all data partitions were estimated using three independent runs to confirm that stationarity (or global optimality) was reached and that independent runs converged on similar stationary parameter estimates. Each of these data partition runs was conducted with a total of 1.4 million generations, 400,000 of which were discarded as burn-in, yielding 1 million post burn-in generations.

Each MCMC run for all individual gene and intermediate data partitions employed the model selected by ModelTest for that partition, or the nearest model to that model that could be implemented in MrBayes. The total combined data set was subjected to MCMC analyses under multiple alternative evolutionary models which differed in the way they parameterized among-site rate variation.

The most complex (parameter rich) model that ModelTest v3.0 can evaluate is a General Time Reversible (GTR; Tavaré, 1986) model with an estimated proportion of invariant sites (I) and gamma distributed among-site rate variation (G; Yang, 1993). MrBayes v2.10 is capable of employing more complex models than this GTR+I+G model. MrBayes allows among-site rate variation to be partitioned among user defined sites (site-specific gamma; SSG) as well as allowing the use of an auto-correlated gamma (A; e.g., Yang, 1995; Penny et al., 2001; Huelsenbeck, 2002) to account for auto-correlation of among-site rates. These two modifications of among-site rate variation may be used independently as well as simultaneously in a given model in MrBayes. In addition to the GTR+I+G model, we conducted combined data MCMC analyses with alternative models that partitioned gamma with (SAG) and without (SSG) accounting for auto-correlation (A) of rates. These two alternative ways to estimate among-site rate variation were invoked by the commands “rates = ssadgamma” and “rates = ssgamma” (respectively) in the MrBayes 2.0 command block. All models applied a single common GTR substitution rate matrix across all data and differed only in the way they modeled and partitioned among-site rates according to the following a priori partitions: GTR + auto-correlated gamma (GTR+AG); protein-coding genes vs. rRNA genes (PR-SSG and PR-SAG); nuclear vs. mitochondrial genes (NM-SSG and NM-SAG); c-mos vs. ND4 vs. rRNA genes (CNR-SSG and

CNR-SAG); all genes partitioned independently (4gene-SSG and 4gene-SAG). Table 2 provides a summary of the parametric content of each of these models.

Choosing among these models to identify the best model of evolution on which to base phylogenetic and taxonomic decisions was approached in several ways. Our goal was to find the model of evolution which best fit the data yet contained the fewest total parameters (the best-fit model). Specifically, our major criteria for identification of the simplest best-fit model included the demonstration of clear improvements of likelihood estimates under that model, along with a practical effect on topology and/or posterior probability support for clades. Therefore, we were not interested in more complex models which did not estimate a different topology or have significant effects on posterior probability estimates. Once a tentative model was chosen, this model was rigorously tested for overparameterization and unreliability (which would suggest it was not a candidate for the best-fit model).

We examined the burn-in plots of likelihoods for MCMC chains for each model to determine the rate of ascent to an apparent stationary plateau, in addition to the degree of overlap between models and superiority (based on chain likelihood values) of models, relative to the number of parameters they employed. To examine the relative improvement in likelihood scores with respect to model complexity, we compared the 95% credibility interval (CI) of MCMC chain likelihood scores between models. To calculate the 95% CI, we ranked all post burn-in tree estimates by ln likelihood and included the most likely 95% (Felsenstein, 1968; Huelsenbeck et al., 2002). Once a tentative best-fit was identified, we evaluated parameter burn-in plots of these models for evidence of identifiability of parameters by checking for commonality in parameter estimates among runs. We also examined the sensitivity of posterior probability

Table 2. Parametric composition of models tested in Bayesian MCMC analyses of the combined data.

Model Name	Nucleotide Substitution Matrix	Model Parameters in Addition to GTR Matrix	Gamma Parameter	Gamma Autocorrelation Parameter	Site (= Partition) Specific Gamma Parameters
GTR+I+G	GTR	2	+	-	-
GTR+AG	GTR	2	+	+	-
NM-SSG	GTR	3	+	-	2
PR-SSG	GTR	3	+	-	2
NM-SAG	GTR	4	+	+	2
PR-SAG	GTR	4	+	+	2
CNR-SSG	GTR	5	+	-	3
CNR-SAG	GTR	6	+	+	3
4gene-SSG	GTR	6	+	-	4
4gene-SAG	GTR	7	+	+	4

values to model complexity using Wilcoxon signed rank tests implemented with Statistica (StatSoft, 1993) to test for significant changes in posterior probability estimates between the chosen model and those which were proximal alternative best-fit models. For interpretation of phylogenetic inferences, we consider posterior probability values over 95% to be well-resolved.

In addition to the three independent MCMC runs (1.4 million generations each) conducted for each model, we conducted a single MCMC run for an extended number of generations (33 million generations) for the two main alternative best-fit models (GTR+I+G and CNR-SSG) for the combined data set. For each long MCMC run, the posterior probabilities for clades were monitored in intervals of two million generations to examine any trends and the overall precision associated with posterior probability estimates through generations of extended MCMC runs. Additionally, posterior probabilities of clades estimated from these long MCMC runs were compared to those estimated from the initial three MCMC runs per model (run for 1.4 million generations) to examine the effect that MCMC analysis strategy (multiple short runs vs. single long run) has on estimates of posterior probabilities. Posterior probabilities estimated from these long runs were also used to re-test for significant changes in posterior probability estimates derived from analyses under alternative models (as described above).

Results

A total of 1810 characters were included in the analysis (c-mos 408 bp; ND4 623 bp; 12S 331 bp; 16S 448 bp). Details of optimal trees selected by maximum parsimony and best-fit models of evolution selected by ModelTest (Posada and Crandall, 1998) are presented in Table 3. After preliminary phylogenetic reconstructions, we identified several apparent problems with the Pellegrino et al. (2001) dataset, including switching of taxon names and apparent contamination, which we rectified prior to final analyses.

Parsimony Phylogenetic Reconstruction

The total evidence (all four genes) equally-weighted parsimony reconstruction resulted in two most parsimonious trees of 6600 steps with 769 parsimony-informative characters and CI = 0.228, RI = 0.543, RC = 0.124, HI = 0.772 (Fig. 1). Six major clades were recovered, each with high bootstrap support (70–100%) and differing from the reconstruction of Pellegrino et al.

(2001). Whereas the earliest split within the Gymnophthalmidae in Pellegrino et al. (2001) was the divergence of a clade composed of the three *Alopoglossus* species from all others, we recovered a clade of *Alopoglossus* spp. and *Ptychoglossus brevifrontalis*. As explained in Appendix 2, an apparent taxon name error in the 12S and 16S data sets presumably resulted in the erroneous placement of *Ptychoglossus* in the Cercosaurinae.

Table 3. Statistics for datasets used, including results from MP searches and suggested model from hierarchical ln likelihood ratio test (hLRT) criterion from ModelTest.

	c-mos	ND4	12S	16S ^a	All rRNA	all mt	All protein	Total
Number of Characters	408	623	331	448	779	1402	1031	1810
Parsimony-informative	173	363	112	121	233	596	536	769
Number of Trees	666	30	5195	>120,000	5065	6	4	2
Optimal tree score	566	4365	749	739	1547	5990	4976	6600
CI	0.528	0.177	0.290	0.296	0.282	0.202	0.215	0.228
HI	0.472	0.823	0.710	0.704	0.718	0.798	0.785	0.772
hLRT selected model	K80+G	GTR+I+G	TrN+I+G	TrN+I+G	HKY+I+G	TVM+I+G	TVM+I+G	GTR+I+G
Proportion invariable sites	---	0.321	0.421	0.551	0.535	0.461	0.331	0.432
Gamma parameter	0.595	0.506	0.543	0.511	0.657	0.564	0.519	0.553
Ti:Tv ratio	2.67	---	---	---	2.651	---	---	---
Rate Matrix: r(A-C)	---	0.301	1	1	---	0.438	0.411	0.553
r(A-G)	---	6.907	12.086	3.023	---	3.531	4.370	4.101
r(A-T)	---	0.648	1	1	---	0.476	0.543	0.530
r(C-G)	---	0.429	1	1	---	0.179	0.630	0.384
r(C-T)	---	4.648	5.352	7.150	---	3.531	4.370	3.608
r(G-T)	---	1	1	1	---	1	1	1

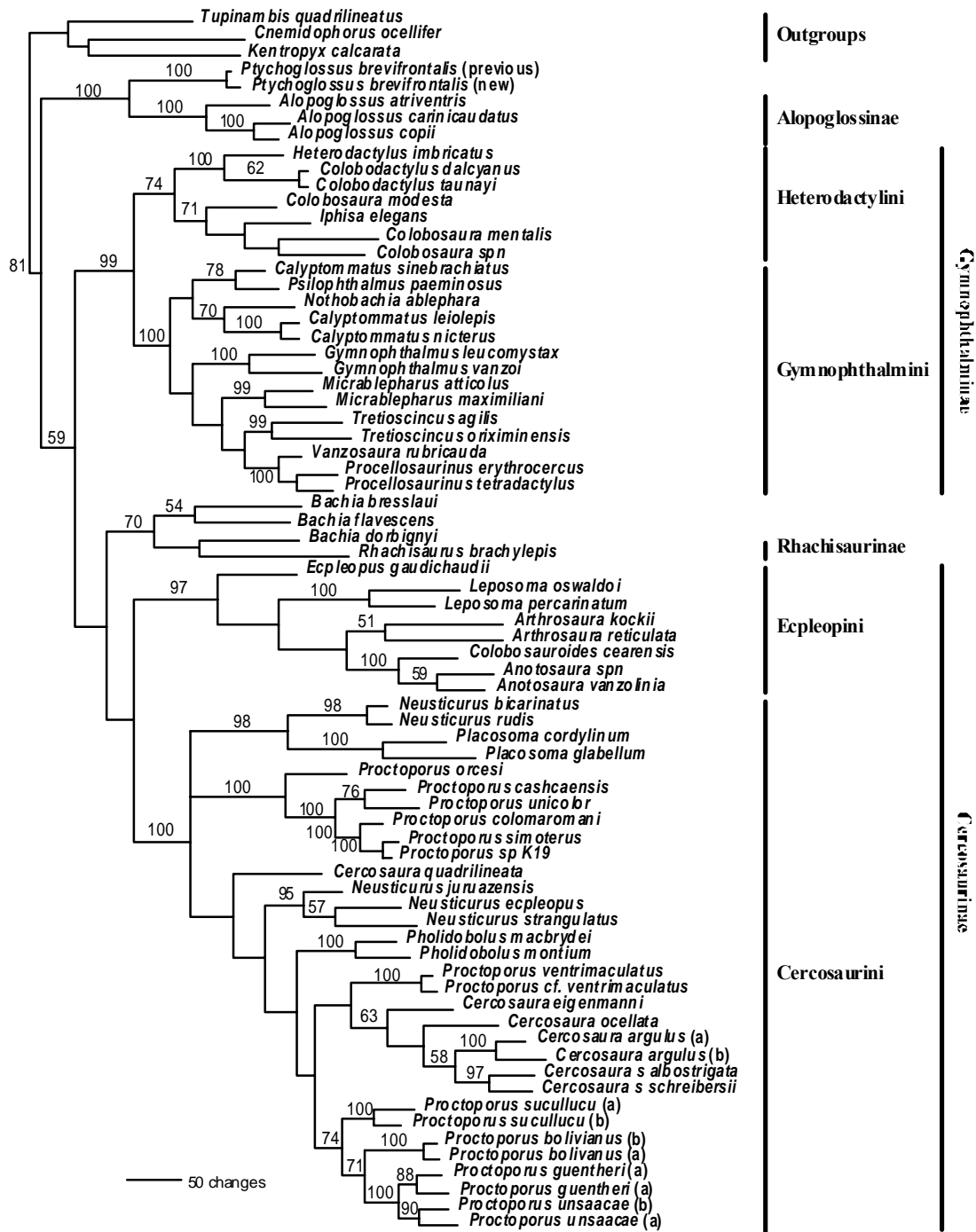


Figure 1. Strict consensus phylogram of two most parsimonious trees based on the equally-weighted maximum parsimony search including all four genes (c-mos, ND4, 12S, and 16S). Labels (a) and (b) indicate individuals of a species (see Appendix 1). For reference, labels on the right side represent the taxonomy presented by Pellegrino et al. (2001). Taxa that are not labeled have relationships that do not agree with that former taxonomy.

Mitochondrial Gene MCMC Analyses

Based on hierarchical log likelihood ratio tests (hLRT) of successively complex models of sequence evolution, ModelTest indicated the best-fit model for the combined mitochondrial dataset was the TVM+I+G (Table 3). This model of evolution, characterized by a five parameter nucleotide substitution rate matrix, is not currently available in MrBayes. Instead, the next best-fitting parameter rich model, which employs a general time reversible (GTR) six parameter nucleotide substitution rate matrix, was employed with proportion of invariant sites (I) and gamma distributed among-site rate variation (G). Parameter estimates derived from the combination of all post burn-in estimates from the three independent MCMC runs are summarized in Table 4. All three runs reached apparent stationarity (in estimates of substitution model parameters, as well as chain likelihood scores) prior to 50,000 generations, well before the conservative burn-in period of 400,000 generations.

The all mitochondrial data partition MCMC reconstruction (Fig. 2a) contrasts with the Pellegrino et al. (2001) reconstruction in the relative phylogenetic placement of major clades deep in the phylogeny, but posterior probability support for some the relationships was not high. As in the parsimony reconstruction described above, *Alopoglossus* and *Ptychoglossus* form a clade sister to the remaining gymnophthalmids. The next node splits the Ecleopini from the remainder of the taxa (but with low posterior probability support). Of the three MCMC runs, one differed slightly with regard to the structure of the remainder of the tree. Examination of this difference among runs revealed that the difference between the majority rule topologies from post burn-in MCMC trees resulted from an approximately 1% difference between runs in the posterior probability density supporting one relationship over another (both of which received

Table 4. Parameter estimates for all mitochondrial gene and c-mos.

	All mt genes – All runs	c-mos – All Runs
ln likelihood	-26274.1 (-26295.6 – -26254.2)	-3616.0 (-3634.3 – -3599.0)
pi(A)	0.397 (0.380 – 0.416)	0.259 (0.230 – 0.289)
pi(C)	0.283 (0.269 – 0.297)	0.265 (0.236 – 0.294)
pi(G)	0.077 (0.070 – 0.084)	0.243 (0.215 – 0.272)
pi(T)	0.243 (0.230 – 0.256)	0.233 (0.205 – 0.262)
r(A-C)	0.406 (0.310 – 0.524)	---
r(A-G)	4.236 (3.497 – 5.091)	---
r(A-T)	0.470 (0.351 – 0.604)	---
r(C-G)	0.201 (0.124 – 0.300)	---
r(C-T)	3.641 (2.910 – 4.459)	---
r(G-T)	1	---
Tv:Ti ratio	---	5.432 (4.490 – 6.534)
Gamma parameter	0.503 (0.462 – 0.547)	0.644 (0.517 – 0.803)
Proportion of invariable sites	0.419 (0.387 – 0.450)	---

posterior probabilities below 50%). Figure 2a depicts the consensus of those three runs that collapses nodes in conflict, creating a polytomy of *Rhachisaurus*, *Bachia*, the Gymnophthalminae, and the Cercosaurini minus *Bachia*. Even with the differences among the reconstructions, the lack of monophyly of the Cercosaurinae differs from Pellegrino et al. (2001; their Fig. 4) and our parsimony reconstruction (Fig. 1).

C-mos (Nuclear Gene) MCMC Analyses

Based on hLRTs of successively complex models of sequence evolution, ModelTest indicated the best-fit model for the combined mitochondrial dataset was the K80+G model (Table 3). Parameter estimates derived from the combination of all post burn-in estimates from the three independent MCMC runs, using a K80+G model, are summarized in Table 4. All three runs reached apparent stationarity (in estimates of substitution model parameters, as well as chain likelihood scores) prior to 50,000 generations.

The nuclear *c-mos* reconstruction (Fig. 2b) differs from that of Pellegrino et al. (2001), Harris (2003), our parsimony reconstruction (Fig. 1), and our mitochondrial reconstruction (Fig. 2a). As with Harris (2003), our parsimony reconstruction, and our mitochondrial DNA reconstruction, *Alopoglossus* and *Ptychoglossus* form a basal clade. Similar to our parsimony reconstruction, four additional major clades are formed, each with high posterior probability support for clade monophyly, but low support of the relationships among the clades. The Gymnophthalminae forms a monophyletic group with strong posterior probability support. The Cercosaurinae is not monophyletic because there is strong support for the Eupleopini being only distantly related to the Cercosaurini. Additionally, as in the parsimony reconstruction, *Bachia*

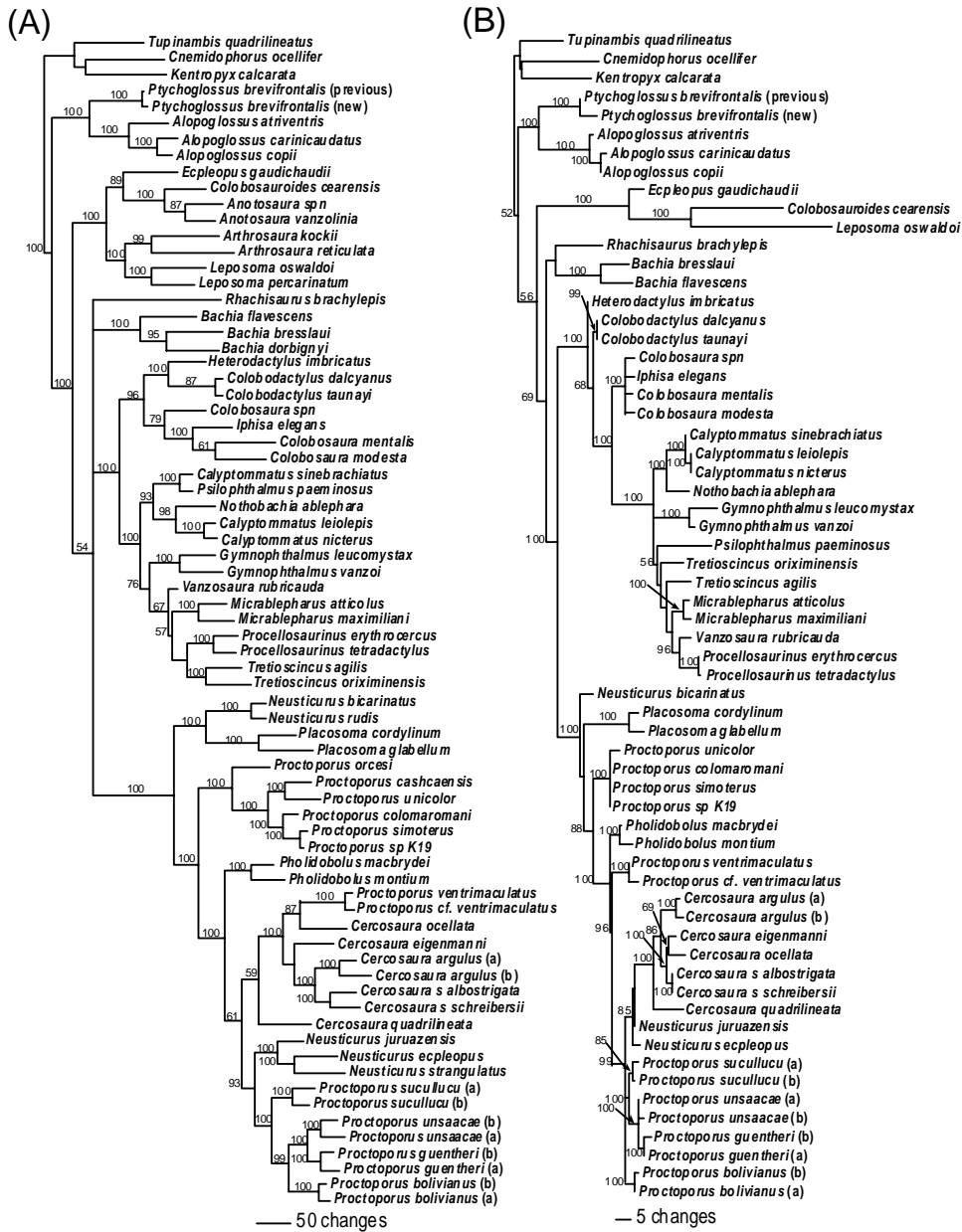


Figure 2. Bayesian phylogenetic trees for the independent nuclear and mitochondrial data partitions. Labels (a) and (b) indicate individuals of a species (see Appendix 1). (A) Majority rule phylogram and posterior probabilities resulting from Bayesian analysis of all three mitochondrial genes combined (ND4, 12S, and 16S) based on a combined 3 million post burn-in generations under the GTR+I+G model of evolution. (B) Majority rule phylogram and posterior probabilities resulting from Bayesian analysis of nuclear c-mos gene data based on a combined 3 million post burn-in generations under the K80+G model of evolution.

and *Rhachisaurus* form a clade. As in Pellegrino et al.'s (2001) maximum likelihood reconstruction, but differing from our parsimony reconstruction, tribe Heterodactylini is not monophyletic but is paraphyletic with respect to the Gymnophthalmini.

Combined MCMC Analyses

Based on the hLRT criterion for model selection, ModelTest chose the GTR+I+G model of nucleotide substitution for the combined data set (Table 3). Burn-in plots of likelihood scores of MCMC chains conducted with this model and alternative models are shown in Figure 3 and mean stationary values (with 95% credibility interval) across models are compared in Figure 4. A more detailed plot of the ascent of likelihood scores of chains toward stationarity for each model is shown in Figure 5. Although we only show burn-in plots for chains from one of three individual MCMC analyses for each model, no single run (under a particular model) was noticeably different with regard to burn-in time, parameter estimate mean, or credible interval at stationarity.

Based on the post-burn-in plateau of chain likelihood values observed in Figures 3 and 4, the GTR+I+G model appears to out-perform models which partition among-site rate variation between either nuclear vs. mitochondrial genes (NM-SSG and NM-SAG) or between protein-coding vs. ribosomal RNA genes (PR-SSG and PR-SAG). The GTR+AG model resulted in chain likelihood scores which were markedly lower than those estimated under the GTR+I+G model. Two classes of models which partition among-site rate variation into either three or four classes appeared to result in clear improvements in the likelihood scores of stationary chains when compared to the GTR+I+G model: models which partitioned among-site rate variation

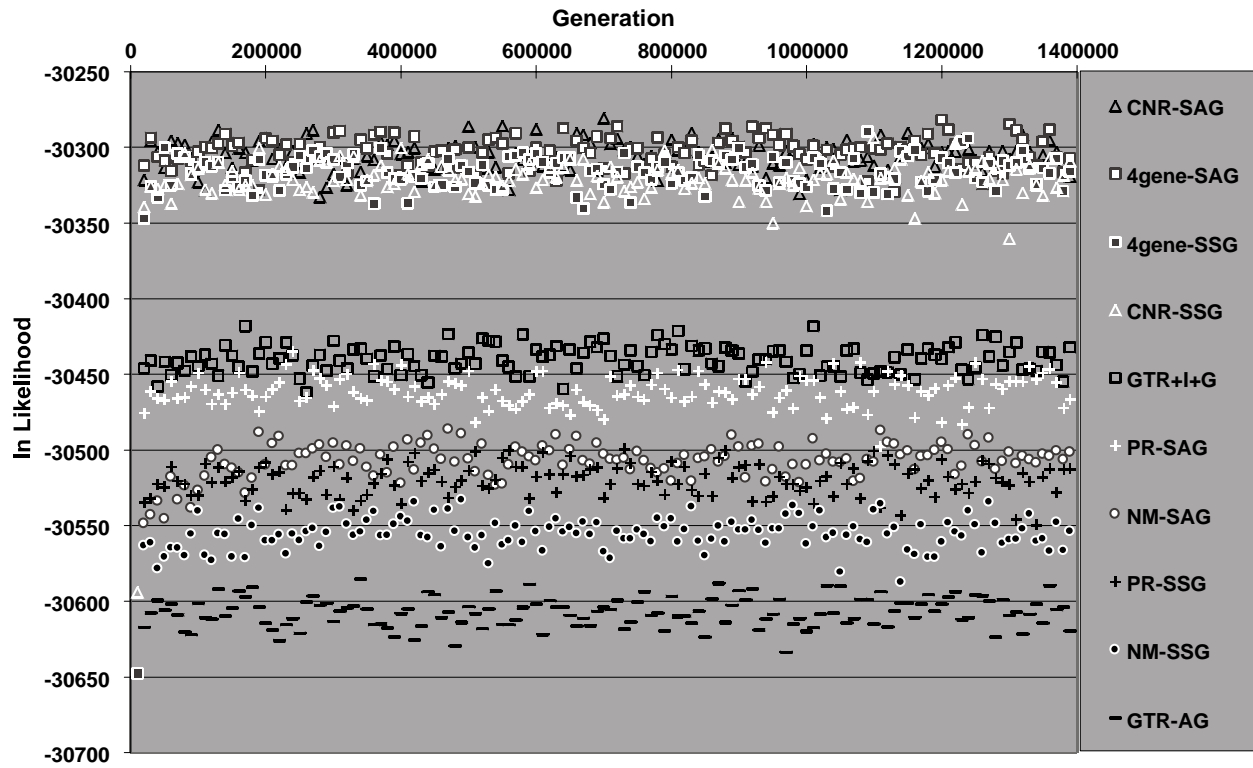


Figure 3. The In likelihood scores of MCMC chains based on alternative models of evolution, sampled in 10,000 generation intervals for clarity of presentation. See text for descriptions of alternative models.

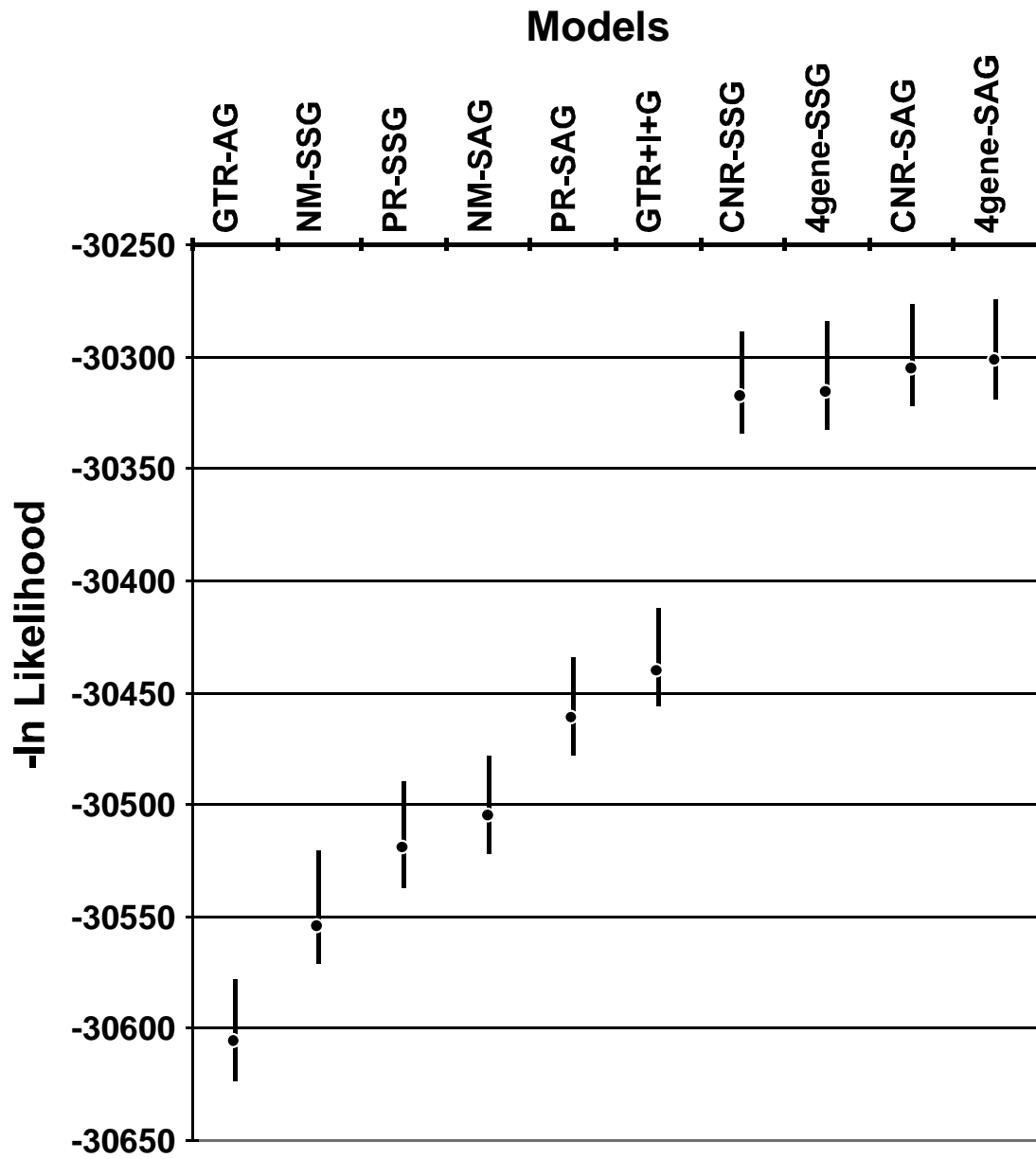


Figure 4. The mean and 95% credibility interval for post burn-in \ln likelihood scores of MCMC chains based on alternative models of evolution.

among c-mos (nuclear protein-coding) vs. ND4 (mitochondrial protein-coding) vs. ribosomal RNA genes (CNR-SSG and CNR-SAG; three site partitions) and those which partitioned rate variation among all individual genes (4gene-SSG and 4gene-SAG; four site partitions). Within this group of models with either three or four partitions of among-site rate variation (with and without auto-correlated gamma), no single model clearly outperformed any other based on estimates of stationary chain likelihood scores (Fig. 3). From Figure 5 we observe that all models, including those with three or four partitions for among-site rate variation, achieve stationarity rapidly by approximately 30,000 generations (although we conservatively discarded trees prior to 400,000 generations as burn-in).

Consensus topologies estimated from post-burn-in generations were identical among multiple independent runs under a particular model (Fig. 6). We found a general correlation between topology and model fit (inferred based on relative values of stationary chain likelihood scores), whereby the two models with the lowest range of ln likelihood scores (mean ln likelihood < -30,550) for chains produced slightly different topologies compared with all models resulting in chains with higher ln likelihoods (mean ln likelihood > -30,550). Analyses of the combined data employing all models except NM-SSG and GTR+AG recovered the identical topology. The analyses under the NM-SSG and GTR+AG models recovered a topology identical to the others except for a swap in the relative branching order with respect to two clades (*Rhachisaurus* + Gymnophthalminae and Eupleopini; neither rearrangement received high posterior probability support), in addition to a modification affecting the phylogenetic position of *Proctoporus ventrimaculatus* + *P. cf. ventrimaculatus*.

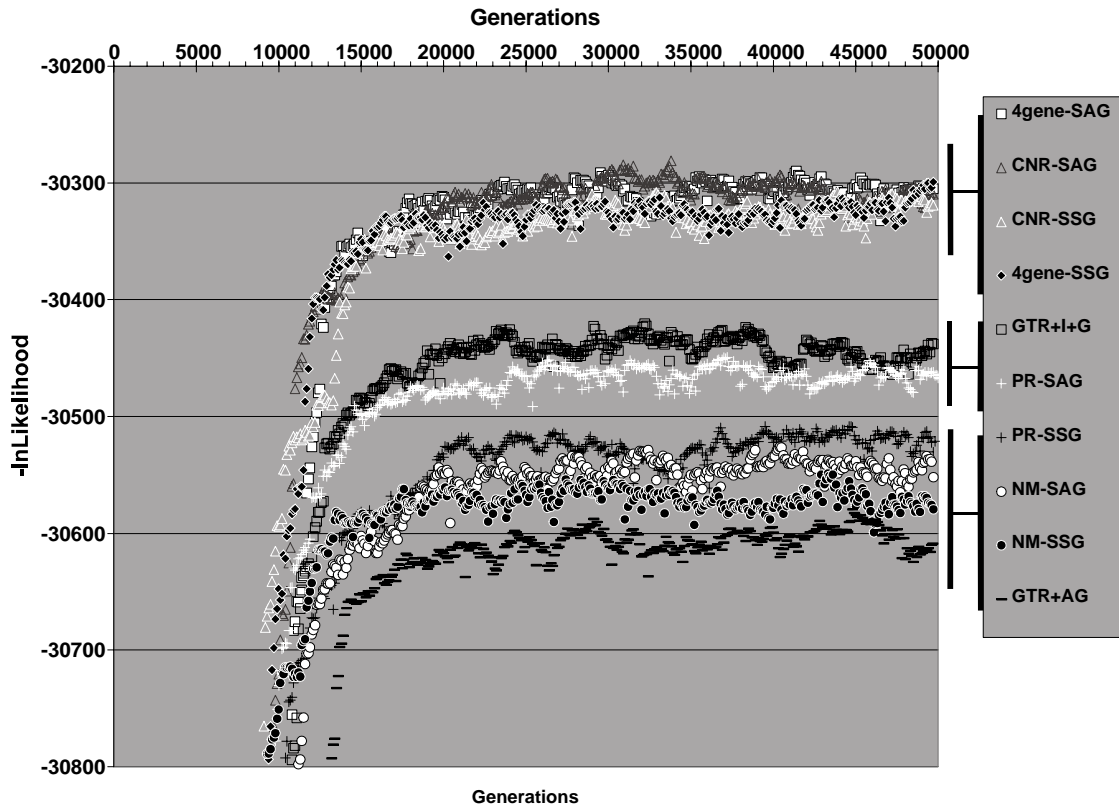


Figure 5. The \ln likelihood scores of MCMC chains based on alternative models of evolution, focusing on the period below 50,000 generations, sampled every 100 generations.

Based on our a priori criteria for initial identification of the preferred evolutionary model as that which contained the fewest number of parameters while demonstrating a clear optimization of overall chain likelihood, we chose the CNR-SSG model. To examine the effect of model choice on the cumulative posterior probabilities for clades, we tested for significant changes in posterior probabilities between the CNR-SSG model and all other models that were found to be as good or better than the GTR+I+G model (including GTR+I+G, CNR-SAG, 4gene-SSG, and 4gene-SAG) with Wilcoxon signed rank tests. For these, posterior probabilities for matched nodes were pairs for comparison. Tests comparing the overall change in posterior probabilities between the CNR-SSG model and other models with three or four gamma partitions (with and without auto-correlated gamma) were not significant. However, the GTR+I+G model was found to produce, overall, significantly lower estimates for posterior probabilities of clades than the preferred model (CNR-SSG; $z = 2.173$, $p = 0.029$). This trend is demonstrated by the relationship between posterior probabilities from the GTR+I+G model and the CNR-SSG model, plotted against one another in Figure 7. Overall, a majority of nodes plotted in this figure fall above the 1:1 line, indicating higher nodal support resulting from the CNR-SSG model. It is also important to note, however, that several nodes did decrease in posterior probability support under the CNR-SSG model.

As described above, burn-in plots of \ln likelihood of MCMC chain scores from all independent MCMC runs under the CNR-SSG model are essentially identical with a rapid and direct approach to a common stationary plateau (not shown). To investigate burn-in and common estimates at stationarity for the parameters of the independent (1.4 million generation) CNR-SSG model runs, burn-in plots of the reversible rate of A-G and A-T substitutions as well

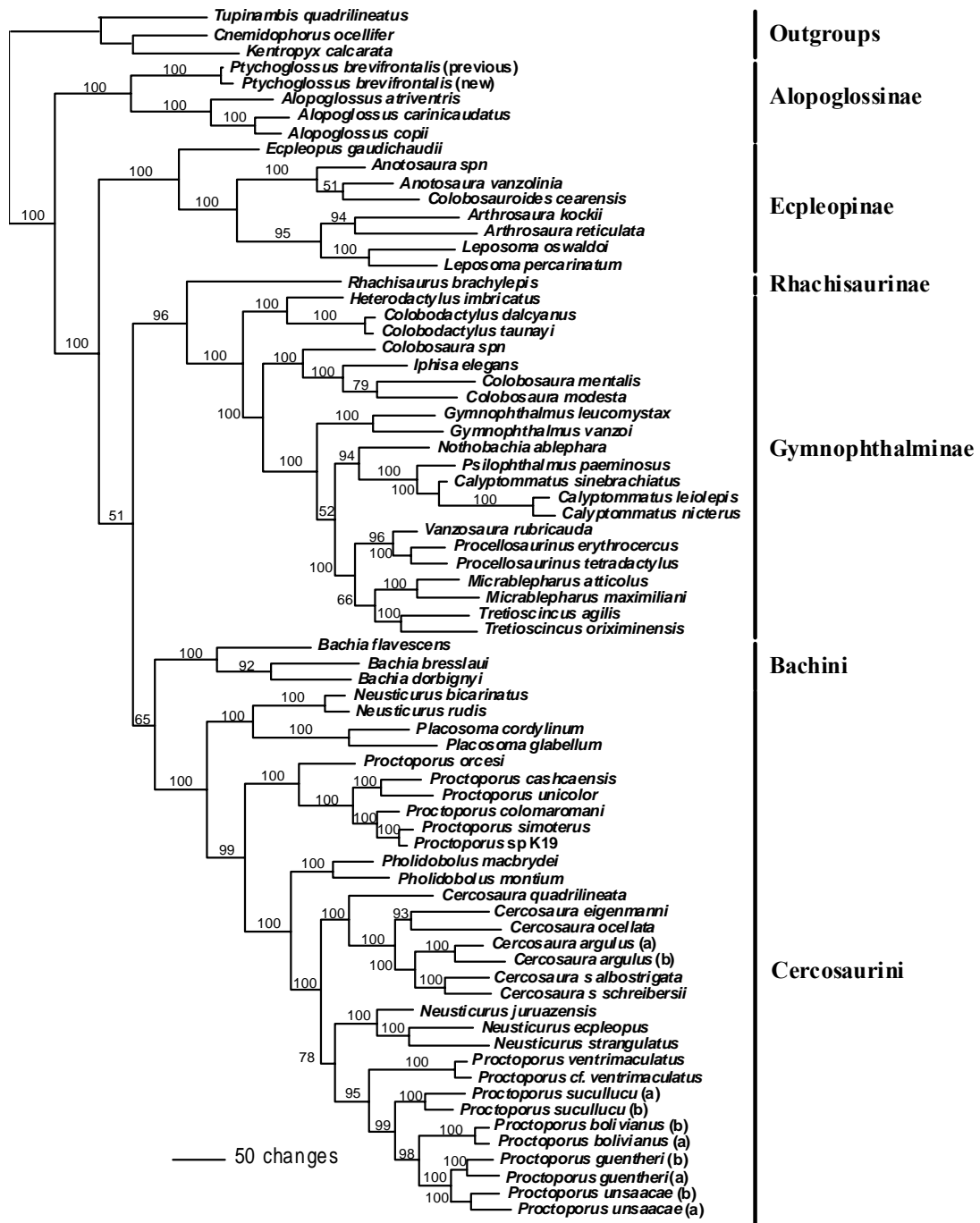


Figure 6. Bayesian phylogenetic tree and posterior probabilities for clades based on the combined, four-gene data set analyzed under the CNR-SSG model. Tree is based on the combination of all post burn-in generations resulting from three independent runs of the model, for a combined total of 3 million post-burn-in generations.

as the gamma parameter and site-specific partition rate parameters are shown in Figure 8. Similar to burn-in plots of likelihood tree scores (Fig. 5), all parameters appear to approach stationarity rapidly (in less than 50,000 generations) and oscillate around a common stationary value (across independent runs). Because all three (1.4 million generation) independent runs of our preferred model (CNR-SSG) appear to reach common stationary estimates of parameters, produce identical topologies, and nearly identical posterior probability estimates, hereafter we report only results based on the combination of all 3 million post burn-in generations pooled from the three independent runs of MCMC analyses using the CNR-SSG model. Parameter values, with 95% credibility intervals, resulting from MCMC CNR-SSG model analyses are given in Table 5.

Posterior probability estimates derived from post burn-in generations from the single long MCMC run (33 million generations) of the GTR+I+G and CNR-SSG models were very similar to estimates based on the combination of the three shorter (1.4 million generation) runs. Considering only clades supported with less than 100% posterior probability support, the long MCMC run of the GTR+I+G model produced estimates that were, on average, 1.05% different from the short run estimates, compared with 0.40% for the CNR-SSG model. Given the almost identical posterior probability estimates (within 1%) derived from the single long MCMC run under the CNR-SSG model as compared with those previously estimated from the combination of the three short MCMC runs of this model, we retain the use of posterior probabilities derived from the three short runs for further discussions of the phylogeny. Estimates of model parameters derived from this long CNR-SSG MCMC run are given in Table 5.

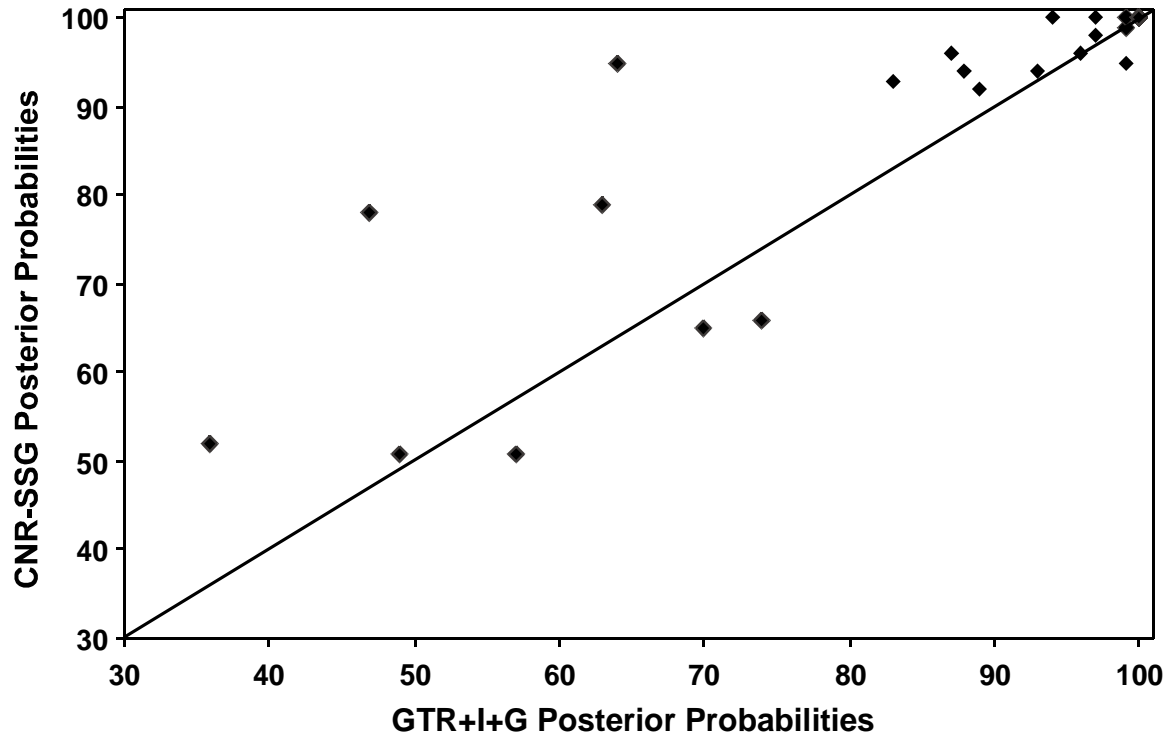


Figure 7. Plot of the posterior probabilities derived from the GTR+I+G MCMC analyses (all three runs combined) versus the posterior probabilities derived from the CNR-SSG model (all three runs combined). For comparison, a 1:1 line is plotted on the same axis.

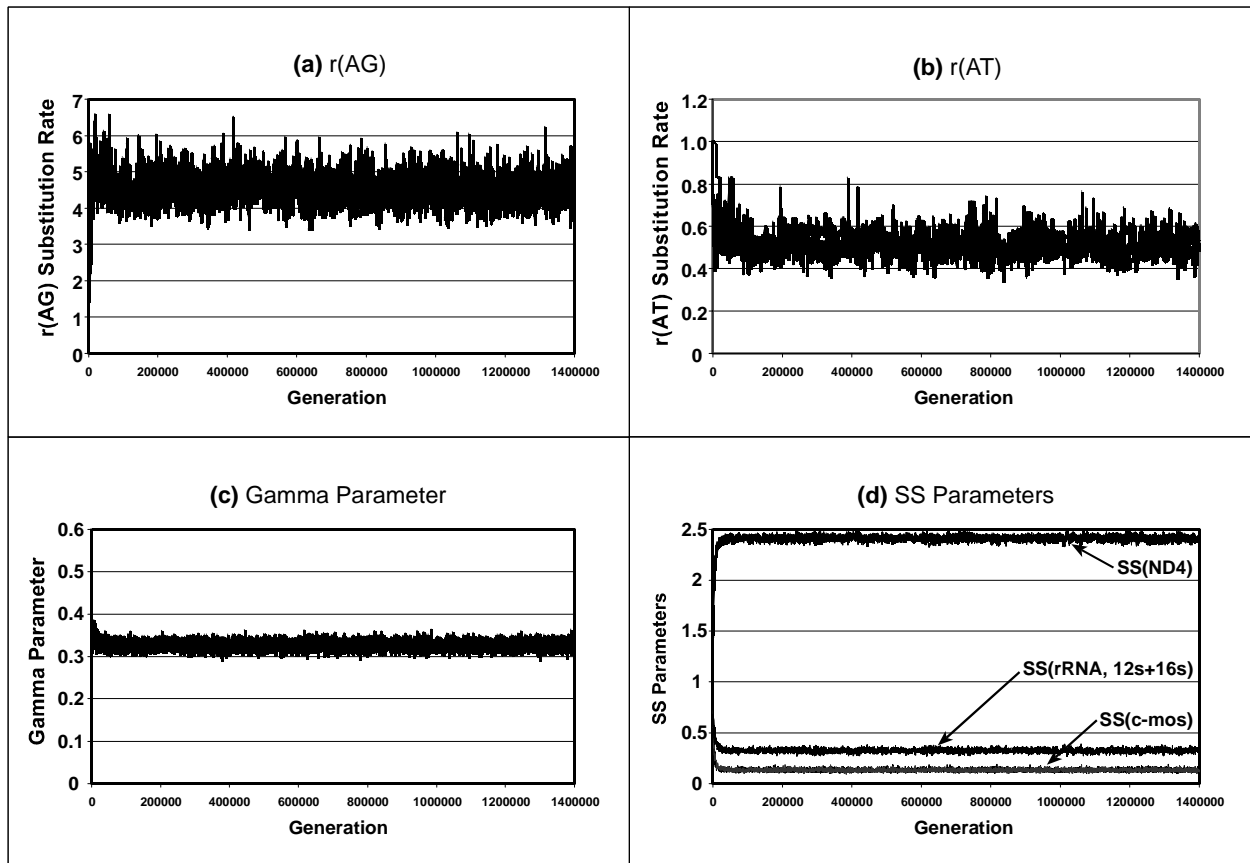


Figure 8. Plots of selected parameters of the CNR-SSG model through generations. All three independent runs are plotted per graph to show common burn-in rates and similar parameter estimates. (a) Plot of parametric estimates of $r(A-G)$ from the GTR rate matrix. (b) Plot of the parametric estimates of $r(A-T)$ from the GTR rate matrix. (c) Plot of the gamma parameter estimates. (d) Plot of the site-specific rate multiplier for the ND4, rRNA (12S + 16S), and c-mos site specific partitions of the gamma parameter

Table 5. Parameter estimates for CNR-SSG model MCMC runs summarized as means with 95% credibility interval in parentheses.

	CNR-SSG Run 1 (1.4 million generations)	CNR-SSG Run 2 (1.4 million generations)	CNR-SSG Run 3 (1.4 million generations)	CNR-SSG All Short Runs (1.4 million generations X 3)	CNR-SSG Long Run (33 million generations)
ln likelihood	-30318.7 (-30338 – -30300.6)	-30317.6 (-30336.3 – -30302.2)	-30319.7 (-30340.0 – -30301.8)	-30318.7 (-30338.2 – -30301.5)	-30319.5 (-30335.6 – -30283.4)
pi(A)	0.393 (0.377 – 0.409)	0.393 (0.378 – 0.409)	0.393 (0.379 – 0.410)	0.393 (0.378 – 0.410)	0.393 (0.362 – 0.385)
pi(C)	0.283 (0.270 – 0.295)	0.283 (0.270 – 0.295)	0.283 (0.271 – 0.294)	0.283 (0.270 – 0.295)	0.283 (0.257 – 0.284)
pi(G)	0.087 (0.081 – 0.094)	0.087 (0.081 – 0.094)	0.087 (0.081 – 0.094)	0.087 (0.081 – 0.094)	0.087 (0.075 – 0.088)
pi(T)	0.237 (0.227 – 0.248)	0.237 (0.226 – 0.248)	0.237 (0.225 – 0.248)	0.237 (0.226 – 0.248)	0.237 (0.214 – 0.243)
r(A-C)	0.409 (0.312 – 0.491)	0.412 (0.339 – 0.506)	0.412 (0.325 – 0.509)	0.411 (0.324 – 0.504)	0.411 (0.269 – 0.400)
r(A-G)	4.491 (3.843 – 5.257)	4.483 (3.86 – 5.203)	4.489 (3.744 – 5.210)	4.488 (3.81 – 5.221)	4.502 (3.163 – 4.324)
r(A-T)	0.504 (0.403 – 0.614)	0.508 (0.414 – 0.630)	0.504 (0.391 – 0.622)	0.506 (0.404 – 0.624)	0.507 (0.333 – 0.490)
r(C-G)	0.364 (0.257 – 0.474)	0.358 (0.259 – 0.488)	0.357 (0.262 – 0.483)	0.359 (0.259 – 0.481)	0.360 (0.178 – 0.340)
r(C-T)	3.745 (3.139 – 4.423)	3.771 (3.199 – 4.469)	3.759 (3.120 – 4.503)	3.758 (3.152 – 4.466)	3.767 (2.651 – 3.509)
r(G-T)	1	1	1	1	1
Gamma parameter	0.326 (0.308 – 0.344)	0.3254 (0.308 – 0.343)	0.325 (0.308 – 0.343)	0.325 (0.308 – 0.343)	0.325 (0.291 – 0.329)
SS1 (c-mos)	0.134 (0.113 – 0.162)	0.134 (0.114 – 0.158)	0.133 (0.112 – 0.157)	0.134 (0.113 – 0.158)	0.133 (0.095 – 0.133)
SS2 (ND4)	2.409 (2.359 – 2.454)	2.414 (2.365 – 2.456)	2.415 (2.370 – 2.459)	2.412 (2.364 – 2.457)	2.412 (2.312 – 2.407)
SS3 (rRNA genes)	0.326 (0.295 – 0.362)	0.323 (0.293 – 0.356)	0.323 (0.290 – 0.354)	0.324 (0.292 – 0.358)	0.324 (0.262 – 0.329)

Relative to the overall estimates of posterior probabilities (all 33 million generations minus 1 million burn-in), the deviation of posterior probability estimates at intervals of generations showed greater variance for the GTR+I+G model than did the CNR-SSG estimates (Fig. 9). The GTR+I+G model produced less precise point (intermediate interval) estimates than did the CNR-SSG model. In other words, as MCMC chains progressed through generations, the posterior probability estimates tended to vary more for the GTR+I+G than the CNR-SSG model. Although the GTR+I+G model included 20 nodes supported below 100%, while the CNR-SSG model included only 17, this bias was factored out by reporting deviations per interval after dividing by the number of nodes considered. This average nodal support deviation (from overall long run estimates) calculated from intervals of generations for the GTR+I+G model was more than twice that for the CNR-SSG model (Fig. 9). In comparisons of variance for nodes receiving similar levels of support (e.g., around 80% posterior probability) greater degrees of variation were evident in the GTR+I+G than the CNR-SSG model run (see Fig. 10 for detail), suggesting that elevated deviation observed in GTR+I+G estimates were not particularly biased by overall higher posterior probability estimates from the CNR-SSG model. Despite variance in posterior probability estimates for intervals of generations, however, no latent trends were observed in posterior probabilities that may indicate that new tree islands were sampled only late in runs (after many generations) or that chains were not completely burned-in after the inferred burn-in period (based on likelihood plateau). Instead, fluctuations in nodal support through generations appear to represent oscillating patterns (see Figure 10 for detail).

We re-examined the effect of model choice on the cumulative posterior probabilities for clades based on these two extended MCMC runs (for the GTR+I+G and CNR-SSG models).

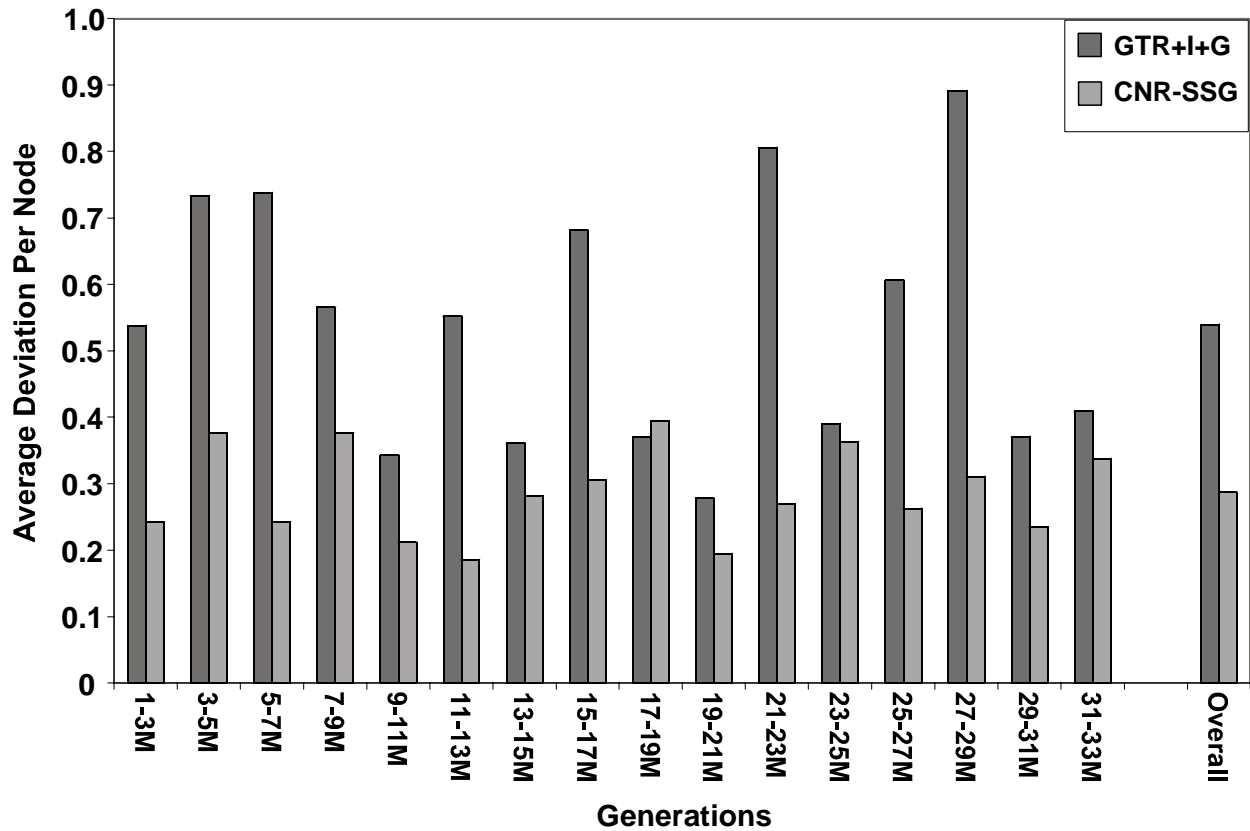


Figure 9. Comparison of deviation of posterior probability estimates at intervals of generations compared to overall means from long MCMC runs (33 million generations) for the GTR+I+G and CNR-SSG models. Values represent the absolute deviation of posterior probability estimates (relative to overall mean for long MCMC run) averaged across all nodes receiving less than 100% posterior probability.

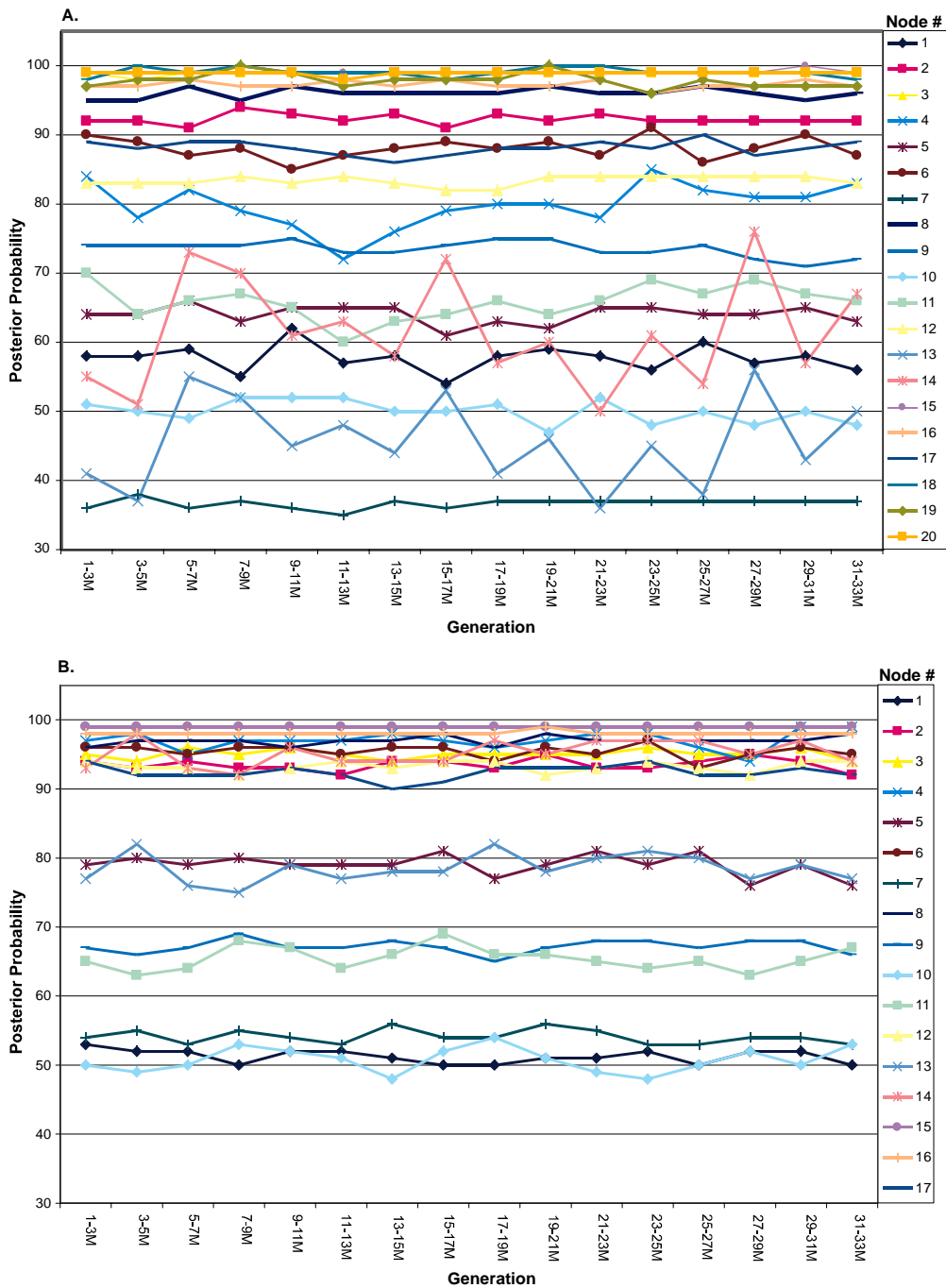


Figure 10. Plots of individual node posterior probability estimates over intervals of extended (33 million generation) MCMC runs of the GTR+I+G and CNR-SSG models. Numbers for nodes are given at the right; numbering of nodes is consistent between the two graphs for comparative purposes.

Results from a Wilcoxon signed rank test returned very similar results as previous estimates based on the three short MCMC runs suggesting the GTR+I+G model produced overall significantly lower estimates for posterior probabilities of clades than the preferred model (CNR-SSG; $z = 2.334$, $p = 0.019$).

The topology of the MCMC CNR-SSG tree (Fig. 6) has similarities with each of the other reconstructions but also differs from all aforementioned reconstructions in several ways. As with all of our phylogenetic reconstructions, *Alopoglossus* and *Ptychoglossus* form a well supported clade sister to the rest of the Gymnophthalmidae. The Cercosaurinae is polyphyletic. As in the c-mos reconstruction, the Heterodactylini is paraphyletic with respect to the Gymnophthalmini. Although the placement of *Rhachisaurus* as the sister taxon to the Gymnophthalminae is unique among the parsimony and data partitions (mitochondrial and nuclear) in this study, it is in the same position as that was recovered by Pellegrino et al. (2001; their Fig. 4). Additionally, *Bachia* is recovered (with weak support) as the sister taxon to the Cercosaurini, as was found by Pellegrino et al. (2001).

Comparison Among Phylogenetic Reconstructions

Not all individual gene data sets (Fig. 2, and not shown) were in agreement with the combined tree (Fig. 6). The c-mos data partition agreed with the combined tree on higher-level relationships except for the placement of *Rhachisaurus* and *Bachia*. Similar to the parsimony reconstruction (Fig. 1), *Rhachisaurus* and *Bachia* formed a clade instead of *Rhachisaurus* being related to the Gymnophthalminae and *Bachia* to the *Cercosaurini*. The ND4 data partition supported a monophyletic Heterodactylini. The 16S data partition resolved the same general

relationship as the combined data CNR-SSG tree except that one of the outgroups, *Tupinambis quadrilineatus*, was nested within the ingroup. The 12S data partition produced topologies most divergent from other genes, but nearly all of those relationships received poor posterior probability support.

Discussion

Model Selection and Evaluation

Bayesian methods have greatly improved our ability to estimate phylogenies using larger datasets and complex models of evolution. However, this creates a seemingly paradoxical dilemma with regard to model complexity and overparameterization. In general, it is assumed that more realistic models of evolution will yield more accurate trees and clade credibility (posterior probability) values, thus perhaps favoring parameter-rich models, because interpretations of posterior probabilities are contingent on model specifications (Huelsenbeck et al., 2002). However, a key assumption of Wald's (1949) proof of the consistency of maximum likelihood estimates is that all of the parameters of the likelihood function are identifiable from the true probability distribution of the data (Rogers, 2001). Even if a particular parameter may be intrinsic in the evolution of DNA sequences, we need to consider whether this parameter can be accurately estimated based on the data. This dilemma is manifest when attempting to construct and implement models that realistically describe DNA evolution, while avoiding

overparameterization, or using more parameters than can be meaningfully estimated from the data.

In a Bayesian analysis, the problem of identifying the best model may be condensed to two intertwined issues: evaluating model performance and fit and examining the sensitivity of posterior probability distributions to model specifications (Gelman et al., 1995; Huelsenbeck et al., 2002). Detecting overparameterized models, however, is not readily accomplished, especially in a Bayesian phylogenetic framework (Huelsenbeck et al., 2002; Rannala, 2002). Several authors have suggested features of MCMC analyses that may be monitored to identify overparameterization including: poor convergence of MCMC chains (Carlin and Louis, 1996), a strong correlation among parameters in the posterior density despite independence under the prior density (Rannala, 2002), delayed convergence of a MCMC chain to a stationary plateau relative to less parameterized models (Rannala, 2002), failure of multiple independent runs (chains) of the same model to converge on similar estimates of parameters and posterior probabilities (Huelsenbeck et al., 2002). We used these criteria, with the exception of testing among-parameter correlation, to guide the evaluation of what we tentatively identified as the best-fit model (CNR-SSG). Testing for among-parameter correlation, in our case, was not possible because the nature of the model causes inherent correlation of the parameters of interest (those which partition the among-site rate variation across partitions).

Different evolutionary rates and among-site rate patterns may be intrinsic evolutionary characteristics of different genes owing to their genomic origin (organellar vs. nuclear) or function (e.g., protein-coding vs. non-protein-coding). Sufficient evidence exists to suggest that drastically different evolutionary rates and distinct, gene-specific among-site rates can be

observed among different genes categorized within a particular class (e.g., among protein-coding mitochondrial genes; Miyata, 1982; Kelly and Rice, 1996). We used these observations to identify plausible alternative partitions across which to estimate specific rates of among-site rate variation. Comparison of post burn-in MCMC chain likelihood scores across alternative models for among-site rate variation showed that only the three partition (CNR-SSG and CNR-SAG) and four partition (4gene-SSG and 4gene-SAG) models fit the data better than the GTR+I+G model chosen by ModelTest. We found no evidence for substantial differences in burn-in time, chain likelihood score at stationarity, or overall clade posterior probability estimates across these models, yet we did detect a significant overall improvement in clade posterior probability estimates between one of these four models (CNR-SSG) and the GTR+I+G model. We found no evidence that this best-fit model (CNR-SSG) was parametrically over-fitted (excessively parameter rich). In fact, based on the analysis of the extended MCMC runs, we found this model to produce significantly more consistent posterior probabilities through generations than did the GTR+I+G model. Given available evidence, we concluded that the best-fit model of evolution, in keeping with our goal of practical improvement for the sake of phylogenetic inference, was the CNR-SSG model, upon which we base our preferred hypothesis for the phylogeny of *Gymnophthalmidae*.

Here, we summarize our approach to model construction and evaluation as an explicit hierarchical process:

Use hLRTs (e.g., ModelTest) to first identify best-fit conventional parameters (although other model choice criteria such as AIC [Akaike, 1974; also available in ModelTest], BIC [Schwarz, 1974], or DT (Minin et al., 2003) may be substituted)

Construct alternative models with data set partitions defined based on a priori expectations of potentially biologically relevant subsets of the data (e.g., protein-coding vs. non-protein-coding genes or mitochondrial vs. nuclear genes)

Examine model fit based on 95% CI of post burn-in MCMC chain likelihood values

Tentatively choose best-fit model

Examine this model for evidence of parameter identifiability or over-fitting

Compare relative burn-in period across alternative models

Check for topological consistency across multiple runs of tentative model

Examine consistency of parameter estimates across multiple independent runs of the tentatively optimal model

Check for consistency of clade posterior probabilities across independent runs

Check for consistency of posterior probability estimates across generations for extended MCMC runs (with large number of generations)

Test for significant differences in posterior probabilities between tentative model and those models of similar fit to the data

Given evidence for parameter identifiability and significant changes to posterior probabilities, accept model. If identifiability is questionable or no significant changes to posterior probability are observed, reduce model parameterization and repeat model evaluation.

Effects of Partitioning Gamma and Using Auto-correlated Rate Variation

Several studies based on simulated data have strongly supported the view that maximum likelihood estimates of phylogeny remain accurate and robust even when the model used to

estimate phylogeny differs markedly from that used to generate simulated data (Fukami-Kobayashi and Tateno, 1991; Yang et al., 1994; Sullivan and Swofford, 2001). Our results support this conclusion using empirical data, in that several different models for among-site rate variation support the same or very similar topologies. Many authors have underscored the importance of including estimates of among-site rate variation (e.g., Yang, 1993; Sullivan and Swofford, 2001; Buckley and Cunningham, 2002; Nylander et al., in press; see review in Yang, 1996b) in models of sequence evolution for increasing the consistency and accuracy of phylogenetic inference. Our results demonstrate that apparent inappropriate partitioning of gamma among loci (e.g., NM-SAG model) may lead to inconsistent and presumably inaccurate phylogenetic inferences. The fact that our preferred model (CNR-SSG) provided a significant increase in overall posterior probability estimates for clades over the GTR+I+G model suggests that well fitted partitioning of among-site rate variation appears to significantly affect the posterior probability distributions of MCMC analyses. These results parallel previous studies that have demonstrated the significant effects of substitution model on maximum likelihood bootstrap support (Yang et al., 1995; Sullivan et al., 1997; Buckley et al, 2001; Buckley and Cunningham, 2002).

An interesting, yet difficult to interpret, result was observed in comparisons of posterior probability monitored through intervals during extended MCMC runs. We found the CNR-SSG model to produce more consistent posterior probabilities through generations than did the GTR+I+G model. If we assume that the complexity of tree space remained relatively constant under the two models, this may suggest that the MCMC chains of the CNR-SSG model were more consistent over intervals of generations with respect to the regular visitation of tree islands.

Alternatively, it seems possible that implementing the partitioned gamma model (CNR-SSG) may have reduced the complexity of tree space by decreasing the number of optimal or near optimal peaks (reducing the number of major islands visited by MCMC chains over time), thereby reducing the variance through generations in trees sampled in the posterior distribution. This may have been accomplished by reducing the likelihood of certain peaks within tree space due to the different parameterization of among-site rates, thereby decreasing the number of near optimal tree islands. In general, the properties associated with the behavior of MCMC chains in tree space through generations has been essentially untouched in the literature, yet represents a significant gap in our understanding of Bayesian MCMC analyses. Future research is clearly necessary to answer questions about the number of generations and independent runs required for robust conclusions from MCMC and also how this may relate to model complexity and changes to the general topology of tree space.

While the results of this study favor use of models which partition among-site rate variation, they also highlight a potential pitfall of such parameter-rich models. Not all alternative models improved model fit relative to GTR+I+G. The GTR+AG, PR-SSG, PR-SAG, NM-SSG, and NM-SAG models all decreased the fit of the model to the data, relative to the GTR+I+G model (chosen initially by hLRT criteria). These results re-emphasize the need to test the fit of alternative models instead of choosing a particular model a priori (e.g., Huelsenbeck and Crandall, 1997; Posada and Crandall, 2001; Minin et al., 2003). The GTR+I+G model not only fit the data better than some partitioned models (e.g., NM-SSG, PR-SAG; Fig. 4), but also recovered the identical topology while (on average) underestimating posterior probability support for clades (Fig. 7). These findings support the utility of this conventionally employed

model and suggest that previous analyses using this model are likely to be as robust (with regard to topology) as more complex models, but provide more conservative estimates of posterior probability support for clades. However, our analysis of extended MCMC runs suggests that, for a reason which is not immediately clear, the GTR+I+G model appeared to take a large number of generations to undergo oscillation cycles with respect to estimates of posterior probabilities. This suggests that, for some models, at least one extended MCMC run (with a large number of generations) is desirable to precisely and accurately estimate posterior probabilities so that trees are sampled in the posterior distribution according to their posterior probability (Swofford, Warren, and Wilgenbusch, unpublished data). It is encouraging, from the standpoint of computational feasibility, that the estimates of model parameters, chain likelihood scores, and, particularly, posterior probabilities derived from the combination of three short (1.4 million generation) independent MCMC runs provided what appears to be, at least, a sufficient approximation of posterior probabilities derived from much longer MCMC runs.

Several authors have demonstrated the utility of employing a parameter to account for auto-correlated among-site rate variation in phylogenetic analyses (e.g., Yang, 1995; Penny et al., 2001; Huelsenbeck, 2002). While evidence for the occurrence of auto-correlated rates has been well documented (Yang, 1995; Nielsen, 1997; Penny et al., 2001), we found the addition of this parameter to alternative models to be of limited value for improving the fit of models to our data. As can be seen in Figure 4, models which fit the data poorly (PR-SSG and NM-SSG) did appear to be notably improved with the addition of a parameter for auto-correlation of among-site rates, although this increase in fit did not exceed that of the GTR+I+G model (or the CNR or 4gene models). Among models which showed the best-fit to the data (CNR and 4gene), the

addition of a parameter to account for among-site rate auto-correlation only slightly increased the likelihood scores for MCMC chains such that there was still broad overlap in the 95% credibility interval of likelihood scores (Fig. 4). Similarly, Wilcoxon signed rank tests comparing overall posterior probabilities for clades between SSG and SAG variants of the CNR and 4gene models found no significant differences attributable to the addition of the auto-correlation parameter.

In this study we have concentrated on accounting for one particular type of heterogeneity in among-site rate patterns in combined DNA sequence analysis, that which exists at or above the level of a gene or locus, ignoring potential partitions which may be prescribed within genes. Models that do examine and attempt to account for within gene heterogeneity by constructing partitions based on codon position (e.g., Yang, 1996a; Krajewski et al., 1999; Buckley et al., 2001), protein domain (Herron et al., in press), or secondary structure for rRNA or tRNA genes (e.g., Schoniger and von Haeseler, 1994; Savill et al., 2001) have also been implemented. These intra-locus partitions have yet to be thoroughly evaluated in a Bayesian framework and may potentially add additional realistic parameters to models of sequence evolution, especially in cases where very distant relationships are inferred (Penny et al., 2001) or where extreme accuracy of branch length estimates or model parameters are particularly critical to conclusions (Yang et al., 1994). Understanding and testing of parametric identifiability in complex models have been poorly studied and clearly requires additional attention. This issue, in addition to topology and posterior probability sensitivity to model choice, would benefit from future investigations using both simulated data and known phylogenies where more definitive conclusions about the effects of model choice may be drawn.

Taxonomic Considerations and Alterations

Much of our phylogeny reconstruction is consistent with that recovered by Pellegrino et al. (2001). However, our preferred phylogenetic hypothesis (combined data MCMC CNR-SSG reconstruction; Fig. 6) suggests four higher level taxonomic changes to the current classification (Pellegrino et al., 2001). The first change is that *Ptychoglossus* appears to be most closely related to *Alopoglossus* and not to the Cercosaurini. The placement of *Ptychoglossus* in the Cercosaurini by Pellegrino et al. (2001) was presumably the result of the swapping of taxon names between *Ptychoglossus* and *Neusticurus juruazensis*, as discussed in Appendix 2. This relationship was also inferred from the nuclear partition trees of Pellegrino et al. (2001) (and the c-mos reconstruction of Harris, 2003) in which *Ptychoglossus* was sister to the three *Alopoglossus* species. After making the correction to the Pellegrino et al. (2001) dataset, and adding our own sequences for this taxon, it seems clear that *Ptychoglossus brevifrontalis* is sister to *Alopoglossus*; therefore, we remove *Ptychoglossus* from the Cercosaurinae and place it in the *Alopoglossinae*. This relationship is also supported by the morphological synapomorphy (present in both *Ptychoglossus* and *Alopoglossus*) of infralingual plicae, unique in the family Gymnophthalmidae.

The second taxonomic alteration involves the tribe Heterodactylini. This tribe is paraphyletic with respect to the Gymnophthalmini in our combined tree (Fig. 6) and in the c-mos (Fig. 2) and 16S (not shown) reconstructions. The paraphyly of the tribe was also apparent in Pellegrino et al.'s (2001) maximum likelihood tree. Because there does not appear to be sufficient support for recognizing a separate tribe Heterodactylini, we remove both of the

Gymnophthalmini and Heterodactylini tribal names and refer all of the pertaining genera to subfamily Gymnophthalminae with no tribes.

The third taxonomic alteration involves species belonging to the cercosaurine tribe Ecleopini. The CNR-SSG tree (Fig. 6) suggests that the ecleopiines and the cercosauriines do not comprise a monophyletic Cercosaurinae. Although posterior probability support for intervening clades is low, monophyly of both groups is well supported. The Ecleopini appears to be distantly related to the Cercosaurini and we hereby raise the status of the former members of tribe Ecleopini (*Amapasaurus*, *Anotosaura*, *Arthrosaura*, *Colobosauroides*, *Ecleopus*, and *Leposoma*; Pellegrino et al., 2001) to subfamily status, the Ecleopinae Fitzinger.

The fourth taxonomic alteration involves the placement of *Bachia*. Pellegrino et al. (2001) recovered its placement as basal within the Cercosaurini. The node joining *Bachia* to the rest of the Cercosaurini was supported by bootstrap values less than 50% on their parsimony tree and by 81% on their maximum likelihood tree. We found conflict between our parsimony reconstructions and Bayesian reconstructions. In the parsimony trees (and 16S and c-mos individual Bayesian gene trees; 16S not shown) we found *Bachia* to be closely related to *Rhachisaurus brachylepis*, either joined with the Ecleopini or distantly related to the Cercosaurinae. In our CNR-SSG reconstruction *Bachia* appears to be the sister lineage to the rest of the Cercosaurini with low posterior probabilities supporting *Bachia* in that position. In addition, a large genetic distance separated *Bachia* from the other cercosauriines. Based on these data we are still unsure of the phylogenetic placement of *Bachia* within the family. However, we are confident that *Bachia* appears to be distantly related to all other sampled taxa. We believe the best course of action at the present time is to leave *Bachia* in the Cercosaurinae but elevate

the genus to tribe status, the Bachini. In this way the relationships of this genus with other genera of the Cercosaurinae are not confused.

A new phylogenetic classification for the family is presented in Table 6. The addition of *Pholidobolus macbrydei* provides preliminary support for a monophyletic *Pholidobolus*. The newly re-designated genus *Cercosaura*, which now includes all taxa formerly placed in *Pantodactylus* and *Prionodactylus* (Doan, 2003a), is supported in this study. The addition of *Neusticurus strangulatus* shows that this species forms a clade with two other members of its genus, *N. ecpleopus* and *N. juruazensis*, but overall the genus is polyphyletic. Additionally, *Anotosaura* is paraphyletic with respect to *Colobosauroides* and *Colobosaura* is paraphyletic with respect to *Iphisa*.

Proctoporus is the genus that was not included by Pellegrino et al. (2001). Contrary to the conclusions made by Doan (2003b) using morphological data, *Proctoporus* appears to be a polyphyletic member of the Cercosaurini. In the CNR-SSG reconstruction two separate *Proctoporus* clades are apparent, separated from each other by *Pholidobolus*, *Cercosaura*, and one clade of *Neusticurus*. One *Proctoporus* clade is composed of members from Ecuador, whereas the other includes members from Peru and Bolivia. In the parsimony reconstruction *Proctoporus ventrimaculatus* additionally forms a third lineage that appears to be most closely related to *Cercosaura*. This species (including an unidentified specimen designated as *P. cf. ventrimaculatus*) is the sole species from northern Peru, separated by a vast distance from the Ecuadorian clade to its north and the southern Peruvian and Bolivian clade to its south. It is clear that taxonomic rearrangement is necessary to rectify the taxonomy of this genus.

Table 6. Current phylogenetic classification of family Gymnophthalmidae.

Taxon
Gymnophthalmidae Merrem, 1820
Alopoglossinae Pellegrino, Rodrigues, Yonenaga-Yassuda, and Sites, 2001
<i>Alopoglossus</i> Boulenger, 1885
<i>Ptychoglossus</i> Boulenger, 1890
Cercosaurinae Gray, 1838
Tribe Bachini New Tribe
<i>Bachia</i> Gray, 1845
Tribe Cercosaurini Gray, 1838
<i>Anadia</i> Gray, 1845
<i>Cercosaura</i> Wagler, 1830
<i>Echinosaura</i> Boulenger, 1890
<i>Euspondylus</i> Tschudi, 1845
<i>Macropholidus</i> Noble, 1921
<i>Neusticurus</i> Duméril and Bibron, 1839
<i>Opipeuter</i> Uzzell, 1969
<i>Pholidobolus</i> Peters, 1862
<i>Placosoma</i> Tschudi, 1847
<i>Proctoporus</i> Tschudi, 1845
<i>Riolama</i> Uzzell, 1973
<i>Teuchocercus</i> Fritts and Smith, 1969
Ecleopinae Fitzinger, 1843
<i>Amapasaurus</i> Cunha, 1970
<i>Anotosaura</i> Amaral, 1933
<i>Arthrosaura</i> Boulenger, 1885
<i>Colobosauroides</i> Cunha and Lima Verde, 1991
<i>Ecleopus</i> Duméril and Bibron, 1839
<i>Leposoma</i> Spix, 1825
Gymnophthalminae Merrem, 1820
<i>Calyptommatus</i> Rodrigues, 1991
<i>Colobodactylus</i> Amaral, 1933
<i>Colobosaura</i> Boulenger, 1887
<i>Heterodactylus</i> Spix, 1825
<i>Iphisa</i> Gray, 1851
<i>Gymnophthalmus</i> Merrem, 1820
<i>Micrablepharus</i> Dunn, 1932
<i>Nothobachia</i> Rodrigues, 1984
<i>Procellosaurinus</i> Rodrigues, 1991
<i>Psilophthalmus</i> Rodrigues, 1991
<i>Stenolepis</i> Boulenger, 1888
<i>Tretioscincus</i> Cope, 1862
<i>Vanzosaura</i> Rodrigues, 1991
Rhachisaurinae Pellegrino, Rodrigues, Yonenaga-Yassuda, and Sites, 2001
<i>Rhachisaurus</i> Pellegrino, Rodrigues, Yonenaga-Yassuda, and Sites, 2001

References

- Akaike, H., 1974. A new look at statistical model identification. *IEE Trans. Autom. Contr.* 19, 716–723.
- Alfaro, M. E., Zoller, S., Lutzoni, F., 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and Bootstrapping in assessing phylogenetic confidence. *Mol. Biol. Evol.* 20, 255–266.
- Arévalo, E. S., S. K. Davis, J. W. Sites Jr., 1994. Mitochondrial DNA sequence divergence and phylogenetic relationships among eight chromosome races of the *Sceloporus grammicus* complex (Phrynosomatidae) in central Mexico. *Syst. Biol.* 43, 387–418.
- Buckley, T. R., 2002. Model misspecification and probabilistic tests of topology: evidence from empirical data sets. *Syst. Biol.* 51, 509–523.
- Buckley, T. R., C. W. Cunningham, 2002. The effects of nucleotide substitution model assumptions on estimates of nonparametric bootstrap support. *Mol. Biol. Evol.* 19, 394–405.
- Buckley, T. R., C. Simon, G. K. Chambers, 2001. Exploring among-site rate variation models in a maximum likelihood framework using empirical data: effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. *Syst. Biol.* 50, 67–86.
- Carlin, B. P., T. A. Louis, 1996. *Bayes and empirical Bayes methods for data analysis*. Chapman and Hall, New York.
- Caterino, M. S., R. D. Reed, M. M. Kuo, F. A. H. Sperling, 2001. A partitioned likelihood analysis of swallowtail butterfly phylogeny (Lepidoptera: Papilionidae). *Syst. Biol.* 50, 106–127.

- Cummings, M. P., S. A. Handley, D. S. Meyers, D. L. Reed, A. Rokas, K. Winka, 2003. Comparing bootstrap and posterior probability values in the four-taxon case. *Syst. Biol.* 52, 477–487.
- Dixon, M. T., D. M. Hillis, 1993. Ribosomal RNA secondary structure: compensatory mutations and implications for phylogenetic analysis. *Mol. Biol. Evol.* 10:256–267.
- Doan, T. M. 2003a. A new phylogenetic classification for the gymnophthalmid genera *Cercosaura*, *Pantodactylus*, and *Prionodactylus* (Reptilia: Squamata). *Zool. J. Linn. Soc.* 137, 101–115.
- Doan, T. M., 2003b. A south-to-north biogeographic hypothesis for Andean speciation: evidence from the lizard genus *Proctoporus* (Reptilia, Gymnophthalmidae). *J. Biogeography* 30, 361–374.
- Doan, T. M., T. A. Castoe, 2003. Using morphological and molecular evidence to infer species boundaries within *Proctoporus bolivianus* Werner (Squamata: Gymnophthalmidae). *Herpetologica* 59, 433–450.
- Douady, C. J., F. Delsuc, Y. Boucher, W. F. Doolittle, E. J. P. Douzery, 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Mol. Biol. Evol.* 20, 248–254.
- Erixon, S. P., B. Britton, B. Oxelman, 2003. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Syst. Biol.* 52, 665–673.
- Felsenstein, J., 1968. *Statistical inference and the estimation of phylogenies*. University of Chicago, Chicago.
- Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39, 83–791.

- Fukami-Kobayashi, K., Y. Tateno, 1991. Robustness of maximum likelihood tree estimation against different patterns of base substitutions. *J. Mol. Evol.* 32, 79–91.
- Gasser, R. B., X. Zhu, N. B. Chilton, L. A. Newton, T. Nedergaard, P. Guldborg, 1998. Analysis of sequence homogenization in rDNA arrays of *Haemonchus contortus* by denaturing gradient gel electrophoresis. *Electrophoresis* 19, 2391–2395.
- Gene Codes Corp. 1996. Sequencher, version 3.1. Gene Codes, Ann Arbor.
- Gonzales, I. L., J. E. Sylvester, 2001. Human rRNA: evolutionary patterns within genes and tandem arrays derived from multiple chromosomes. *Genomics* 73, 255–263.
- Gutell, R. R., 1994. Collection of small subunit (16S and 16S-like) ribosomal RNA structures. *Nucleic Acid Res.* 22, 3502–3507.
- Gutell, R. R., N. Larsen, C. R. Woese, 1994. Lessons from an evolving rRNA: 16S and 23S rRNA structures from a comparative perspective. *Microbiol. Rev.* 58, 10–26.
- Harris, D. J. 2003, Codon bias variation in C-mos between squamate families might distort phylogenetic inferences. *Mol. Phylogenet. Evol.* 27, 540–544.
- Herron, M. D., T. A. Castoe, C. L. Parkinson, 2004. Sciurid phylogeny and the paraphyly of Holarctic ground squirrels (*Spermophilus*). *Mol. Phylogenet. Evol.* 31, 1015–1030.
- Hickson, R., C. Simon, A. Cooper, G. Spicer, J. Sullivan, D. Penny, 1996. Conserved sequence motifs, alignment, and secondary structure for the third domain of animal 12s rRNA. *Mol. Biol. Evol.* 13, 150–169.
- Hillis, D. M., 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Syst. Biol.* 47, 3–8.

- Hillis, D. M., J. J. Bull, 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. *Syst. Biol.* 42, 182–192.
- Huelsenbeck, J. P., 1995. Performance of phylogenetic methods in simulation. *Syst. Biol.* 44, 17–48.
- Huelsenbeck, J. P., 1997. Is the Felsenstein zone a fly trap? *Syst. Biol.* 46, 69–74.
- Huelsenbeck, J. P., 2002. Testing a covariotide model of DNA substitution. *Mol. Biol. Evol.* 19, 698–707.
- Huelsenbeck, J. P., K. A. Crandall, 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. *Ann. Rev. Ecol. Syst.* 28, 437–466.
- Huelsenbeck, J. P., B. Larget, R. Miller, F. Ronquist, 2002. Potential applications and pitfalls of Bayesian Inference of Phylogeny. *Syst. Biol.* 51, 673–688.
- Huelsenbeck, J. P., F. Ronquist, 2001. MrBayes: Bayesian inference of phylogeny. *Bioinformatics* 17, 754–755.
- Huelsenbeck, J. P., R. Ronquist, R. Nielsen, J. Bollback, 2001. Bayesian inference of phylogeny and its impacts on evolutionary biology. *Science* 294, 2310–2314.
- International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. *Nature* 409, 860–921.
- Kelly, C., J. Rice, 1996. Modeling nucleotide evolution: a heterogeneous rate analysis. *Math. Biosci.* 133, 85–109.
- Kimura, M., 1986. The role of compensatory neutral mutations in molecular evolution. *J. Genetics* 64, 7–19

- Krajewski, C., M. G. Fain, L. Buckley, D. G. King, 1999. Dynamically heterogeneous partitions and phylogenetic inference: an evaluation of analytical strategies with cytochrome b and ND6 gene sequences in cranes. *Mol. Phylogenet. Evol.* 13, 302–313.
- Leaché, A. D., T. W. Reeder, 2002. Molecular systematics of the eastern fence lizard (*Sceloporus undulatus*): a comparison of parsimony, likelihood, and Bayesian approaches. *Syst. Biol.* 51, 44–68.
- Long, E. O., I. B. Dawid, 1980. Repeated genes in eukaryotes. *Ann. Rev. Biochem.* 49, 727–764.
- Minin, V., Z. Abdo, P. Joyce, J. Sullivan, 2003. Performance-based selection of likelihood models for phylogeny estimation. *Syst. Biol.* 52, 674–683.
- Miyata, T., 1982. Evolutionary changes and functional constraints in DNA sequences. Pages 233–266 In *Molecular evolution, protein polymorphism and the neutral allele theory* (M. Kimura, ed.). Japan Scientific Press, Tokyo.
- Moncalvo, J. M., D. Drehmel, R. Vilgalys, 2000. Variation in modes and rates of evolution in nuclear and mitochondrial ribosomal DNA in the mushroom genus *Aminita* (Agaricales, Basidiomycota): phylogenetic implications. *Mol. Phylogenet. Evol.* 16, 48–63.
- Muse, S. V., 1995. Evolutionary analyses when nucleotides do not evolve independently. Pages 115-124 In *Current topics on molecular evolution* (M. Nei and N. Takahata, eds.). Institute on Molecular Evolution and Genetics, University Park, Pennsylvania.
- Myers, C. W., M. A. Donnelly, 2001. Herpetofauna of the Yutajé-Corocoro Massif, Venezuela: second report from the Robert G. Goelet American Museum-Terramar expedition to the northwestern tepuis. *Bull. Am. Mus. Nat. Hist.* 261, 1–85.

- Nielsen, R., 1997. Site-by-site estimation of the rate of substitution and the correlation of rates in mitochondrial DNA. *Syst. Biol.* 46, 346–353.
- Nicholas, K. B., and H. B. Nicholas, Jr., 1997. GeneDoc: a tool for editing and annotating multiple sequence alignments. Distributed by the author at:
www.cris.com/~Ketchup/genedoc.shtml.
- Nylander, J. A. A., F. Ronquist, J. P. Huelsenbeck, J. L. Nieves-Aldrey, in press. Bayesian phylogenetic analysis of combined data. *Syst. Biol.*
- Pamilo, P., M. Nei, 1988. Relationships between gene trees and species trees. *Mol. Biol. Evol.* 5, 568–583.
- Pardue, M. L., 1974. Localization of repeated DNA sequences in *Xenopus* chromosomes. *Cold Spring Harbor Symposium on Quantitative Biology* 38, 475–482.
- Parkinson, C. L., 1999. Molecular systematics and biogeographical history of pitvipers as determined by mitochondrial ribosomal DNA sequences. *Copeia* 1999, 576–586.
- Parkinson, C. L., S. M. Moody, J. E. Alquist, 1997. Phylogenetic relationships of the ‘*Agkistrodon* Complex’ based on mitochondrial DNA sequence data. Pages 63–78. In *Venomous snakes: ecology, evolution, and snakebite* (R. S. Thorpe, W. Wüster, and A. Malhotra, eds.). *Symposia of the Zoological Society of London*. Clarendon Press, Oxford.
- Pellegrino, K. C. M., M. T. Rodrigues, Y. Yonenaga-Yassuda, J. W. Sites Jr., 2001. A molecular perspective on the evolution of microteiid lizards (Squamata: Gymnophthalmidae), and a new classification for the family. *Biol. J. Linnean Soc.* 74, 315–338.
- Penny, D., B. J. McComish, M. A. Carleston, M. D. Hendy, 2001. Mathematical elegance with biochemical realism: the covarion model of molecular evolution. *J. Mol. Evol.* 53, 711–723.

- Posada, D., K.A. Crandall, 1998. Model-Test: testing the model of DNA substitution. *Bioinformatics* 14, 817–818.
- Posada, D., K.A. Crandall, 2001. Selecting the best-fit model of nucleotide substitution. *Syst. Biol.* 50, 580–601.
- Pupko, T., D. Huchon, Y. Cao, N. Okada, M. Hasegawa, 2002. Combining multiple data sets in a likelihood analysis: which models are the best? *Mol. Biol. Evol.* 19, 2294–2307.
- Rannala, B., 2002. Identifiability of parameters in MCMC Bayesian inference of phylogeny. *Syst. Biol.* 51, 754–760.
- Rogers, J. S., 2001. Maximum likelihood estimation of phylogenetic trees in consistent when substitution rates vary according to the invariable sites plus gamma distribution. *Syst. Biol.* 50, 713–722.
- Saint K. M., C. C. Austin, S. C. Donnellan, M. N. Hutchinson, 1998. C-mos, a nuclear marker useful for squamate phylogenetic analysis. *Mol. Phylogenet. Evol.* 10, 259–263.
- Savill, N. J., D. C. Hoyle, P. G. Higgs, 2001. RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods. *Genetics* 157, 399–411.
- Schöniger, M., and A. von Haeseler, 1994. A stochastic model and the evolution of auto-correlated DNA sequences. *Mol. Phylogenet. Evol.* 3, 240–247.
- Schwarz, G., 1974. Estimating the dimension of a model. *Ann. Stat.* 6, 461–464.
- Simon, C., F. Frati, A. Beckenbach, B. Brespi, H. Liu, P. Flook, 1994. Evolution, weighting, and phylogenetic utility of mitochondrial gene sequences and a compilation of conserved PCR primers. *Annals Entomol. Soc. Am.* 87, 651–701.

- StatSoft, Inc. 1993. Statistica for Windows, release 4.5. StatSoft, Tulsa.
- Sullivan, J., J. A. Markert, C. W. Kilpatrick, 1997. Phylogeography and molecular systematics of the *Peromyscus aztecus* species group (Rodentia: Muridae) inferred using parsimony and likelihood. *Syst. Biol.* 46, 426–440.
- Sullivan, J., D. L. Swofford, 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? *Syst. Biol.* 50, 723–729.
- Suzuki, Y., G. V. Glazko, M. Nei, 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *P.N.A.S.* 99, 16138–16143.
- Swofford, D. L., 2002. PAUP*. Phylogenetic Analysis Using Parsimony (*and Other Methods), version 4.0b10, Sinauer Associates, Sunderland, Massachusetts.
- Tavaré, S., 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Pages 57–86 In *Some mathematical questions in biology—DNA sequence analysis* (R.M. Miura, ed.). American Math Society, Providence, RI.
- Titus, T. A., D. R. Frost, 1996. Molecular homology assessment and phylogeny in the lizard family Opluridae (Squamata: Iguania). *Mol. Phylogenet. Evol.* 6, 49–62.
- Wald, A., 1949. Note on the consistency of the maximum likelihood estimate. *Ann. Math. Stat.* 20, 595–601.
- Wilcox, T. P., D. J. Zwickl, T. A. Heath, D. M. Hillis, 2002. Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. *Mol. Phylogenet. Evol.* 25, 361–371.

- Wu, C.I., 1991. Inferences of species phylogeny in relation to segregation of ancient polymorphisms. *Genetics* 127, 429–435.
- Yang, Z., 1993. Maximum likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites: approximate methods. *Mol. Biol. Evol.* 10, 1396–1401.
- Yang, Z., 1995. A space-time process model for the evolution of DNA sequences. *Genetics* 139, 993–1005.
- Yang, Z., 1996a. Maximum likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.* 42, 587–596.
- Yang, Z., 1996b. Among-site rate variation and its impacts on phylogenetic analyses. *Trends Ecol. Evol.* 11, 367–372.
- Yang, Z., N. Goldman, A. Friday, 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.* 11, 316–324.
- Yang, Z., N. Goldman, A. Friday, 1995. Maximum likelihood trees from DNA sequences: a peculiar statistical estimation problem. *Syst. Biol.* 44, 384–399.

CHAPTER 3 – MODELING NUCLEOTIDE EVOLUTION AT THE MESOSCALE: THE PHYLOGENY OF THE NEOTROPICAL PITVIPERS OF THE PORTHIDIUM GROUP (VIPERIDAE: CROTALINAE)

Introduction

Modeling nucleotide evolution at the mesoscale

Incorporating DNA sequence data from multiple genes to solve phylogenetic problems has essentially become a standard across contemporary molecular phylogenetic studies. Paralleling the increasing frequency of multi-locus datasets, model-based techniques have also become a standard in molecular phylogenetics. These methods are attractive because they effectively incorporate probabilistic models of DNA substitution and should, therefore, be less likely to be misled by the complexities of DNA evolution (Huelsenbeck and Crandall, 1997). Numerous empirical studies have demonstrated an array of molecular evolutionary patterns that vary across partitions of molecular datasets including mutation and base-compositional biases (e.g., Faith and Pollock, 2003; Reeder, 2003), and among-site rate variation (e.g., Castoe et al., 2004; Monclavo et al., 2000; Yang, 1996). Thus, an important concern arises when utilizing parametric model-based techniques: a single model with one set of parameters to account for molecular evolution over multiple heterogeneous partitions (e.g., multiple loci, codon positions, structural RNA vs. protein coding regions, etc.) in a combined analysis may fail to portray partition-specific evolutionary patterns.

The use of a single model of evolution for a dataset that is heterogeneous forces a compromise (or averaging) in parameter estimates that may introduce a major source of systematic error and mislead phylogenetic conclusions (Brandley et al., 2005; Reeder, 2003; Wilgenbusch and de Queiroz, 2000; see also Huelsenbeck and Rannala, 2004; Lemmon and Moriarty, 2004). This type of systematic error may be avoided by employing independent models of evolution (and parameter estimates) for subsets of a heterogeneous dataset within a combined analysis (Nylander et al., 2004; Ronquist and Huelsenbeck, 2003; Yang, 1996). Development of robust methods for fitting appropriately complex models of evolution to data partitions, however, has only recently been addressed directly (e.g., Brandley et al., 2005; Castoe et al., 2004; Nylander et al., 2004; Pupko et al., 2002; Yang, 1996).

Model choice has been shown to affect both phylogenetic topology (e.g., Huelsenbeck, 1995; Huelsenbeck, 1997; Sullivan and Swofford, 2001) as well as accurate estimation of posterior probabilities (e.g., Buckley, 2002; Castoe et al., 2004; Erixon et al., 2003; Huelsenbeck and Rannala, 2004; Suzuki et al., 2002). Because the accuracy of posterior probabilities in Bayesian phylogenetic methods relies (at least in part) on the model, models that may not affect the consensus topology may have notable effects on the posterior probability distribution of parameter estimates, and thus on confidence regarding phylogenetic conclusions. Based on this logic, employing complex models that more accurately portray DNA evolution should produce less-biased posterior-probability estimates as long as parameters can be accurately estimated from the data (Huelsenbeck et al., 2002; Huelsenbeck and Rannala, 2004). The benefits of constructing and employing more realistic evolutionary models of DNA substitution are challenged by the potential for imprecise and inaccurate parameter estimation (including

topology). This may result from overparameterization when the ratio of free parameters to data increases past a poorly characterized critical point (where parameters are no longer identifiable based on the data), beyond which a likelihood function may become unreliable (Huelsenbeck et al., 2002; Rannala, 2002; Rogers, 2001; Wald, 1949).

Fundamental differences in the process of optimization of Bayesian and maximum-likelihood methods (see reviews in Holder and Lewis, 2003; Huelsenbeck et al., 2001) have required reconsideration of methods and criteria for selection of best-fit models of evolution. Specific to Bayesian phylogenetics, analytical derivation of the marginal model likelihood is usually impossible when the number of parameters is large, although several estimators of the model likelihood have been proposed. Nylander et al. (2004) followed the proposal of Newton and Raftery (1994) by using the harmonic mean of the post burn-in likelihood values as a reasonable estimate of the marginal model likelihood (for details and justification see Nylander et al., 2004; see also Aris-Brosou and Yang, 2002; Suchard et al., 2001; Huelsenbeck et al., 2004). Here we take advantage of the harmonic mean estimation of Bayesian model likelihoods to employ Bayes factors (Nylander et al., 2004) and adapted version of Akaike weights (Buckley et al., 2002; based on Akaike Information Criteria: Akaike, 1973, 1974, 1983; Sakamoto et al., 1986) to identify the best-fit model of nucleotide substitution for our combined nucleotide data comprising two mitochondrial protein-coding gene fragments.

In this study, we analyze what we believe is representative of a mid-sized molecular phylogeny that ranges in sampling scope from intra-specific to inter-generic. The nucleotide data consist of two of the more common genes used in molecular phylogenetics, the mitochondrial NADH dehydrogenase subunit 4 (ND4) and cytochrome-*b* (cyt-*b*), from 61 terminal taxa. This

dataset provides a reasonably representative model of contemporary ‘mesoscale’ molecular phylogenetics. As such, understanding how phylogenetic hypotheses from this ‘mesoscale’ dataset are affected by analysis under various complex models of nucleotide evolution is an important concern relevant to a majority of contemporary analyses of similar molecular and taxon-sampling scope.

Systematics of the Neotropical pitvipers of the Porthidium group

Pitvipers (Viperidae: Crotalinae), comprise an extensive radiation of both Old and New World venomous snakes with over 180 species allocated to 29 genera (Campbell and Lamar, 2004; Malhotra and Thorpe, 2004; McDiarmid et al., 1999). This diverse radiation of highly venomous snakes has received substantial taxonomic and phylogenetic attention over the last several decades, yet many taxonomic and phylogenetic hypotheses remain unresolved. Recent studies examining molecular characters from a large number of taxa (Parkinson, 1999; Parkinson et al., 2002) have supported several higher-level relationships within Neotropical pitvipers. Within Neotropical pitvipers there appears to be: 1) several basal clades (genera: *Bothriechis*, *Lachesis*, and *Ophryacus*), 2) a primarily South American lineage (genera: *Bothrocophias*, *Bothriopsis*, and *Bothrops*), and 3) a primarily Middle American lineage (genera: *Atropoides*, *Cerrophidion*, and *Porthidium*). This study focuses on this third clade of Neotropical species, referred to as the ‘*Porthidium* group’ (Parkinson et al., 2002; Castoe et al., 2003; see Campbell and Lamar [2004] for detailed updated distribution maps of all *Porthidium* group species).

The *Porthidium* group radiation of Neotropical pitvipers contains three genera, each of which is morphologically and ecologically distinct. *Cerrophidion* (montane pitvipers) contains

four mid-sized species that inhabit mid-to-high elevation Middle American subtropical habitats. *Atropoides* (jumping pitvipers) contains five species of particularly stout-bodied pitvipers that inhabit low-to-middle elevation tropical and subtropical habitats in Middle America (ranging from rainforest and cloud forest to pine-oak forest). *Porthidium* (hognose pitvipers) contains nine more diminutive species that primarily inhabit low elevation wet and dry tropical and subtropical forests across Middle America and northern South America (Campbell and Lamar, 2004).

The *Porthidium* group has been the subject of a number of taxonomic rearrangements and specific additions over the last few decades (see detailed reviews in Campbell and Lamar, 1989, 2004; Castoe et al., 2003; Gutberlet and Harvey, 2004). Initially, all members of this group were recognized under the nominal genus *Porthidium* (Burger, 1971; Campbell and Lamar, 1989), and later were dissected into the three current genera (Werman, 1992; Campbell and Lamar, 1992). In addition to these revisions, two taxa that were once considered members of the *Porthidium* group have been subsequently reallocated to different genera (*Ophryacus melanurum*: Gutberlet, 1998; *Bothrocophias hyoprora*: Gutberlet and Campbell, 2001). At the level of alpha taxonomy, new species have been recently recognized in each of the three genera. Several of these new additions have suggested the taxonomic splitting of widely ranging species (*Atropoides* spp.: Campbell and Lamar, 2004; *P. porrasi*: Lamar and Sasa, 2003), while other recently described species represent previously unknown populations only recently discovered (e.g., *C. petlalcalensis*; López-Luna et al., 2000; *P. volcanicum*: Solórzano, 1995).

While no molecular phylogenetic analyses have inclusively examined relationships across the entire *Porthidium* group, several studies have provided insight into the phylogeny and systematics of the group. The most comprehensively sampled inter-generic molecular

phylogenetic study of pitvipers to date (Parkinson et al., 2002) resolved a monophyletic *Porthidium* group and the genus *Porthidium* as the sister taxon to *Atropoides* plus *Cerrophidion*. Castoe et al. (2003) did not find support for the monophyly of *Atropoides* and demonstrated the paraphyly of *A. nummifer* (later rectified by raising each subspecies to species status by Campbell and Lamar, 2004). Castoe et al. (2003) also demonstrated large divergences among populations of the widespread species *Cerrophidion godmani*. Similarly, Wüster et al. (2003) demonstrated paraphyly of the species *Porthidium nasutum* and *P. lansbergi* (each of which have also recently been taxonomically subdivided; Lamar and Sasa, 2003; Campbell and Lamar, 2004). In summary, the entirety of previous phylogenetic and systematic work conducted on the *Porthidium* group falls short of providing a united perspective on the relationships and taxonomy of these snakes as a result of conflicting or weakly resolved phylogenetic hypotheses, or because of limited taxonomic and geographic sampling. In this study we rectify these problems by reconstructing the phylogenetic relationships within this group including samples representing nearly all species, with many species represented by multiple samples from geographically distinct or isolated populations.

Theoretical and empirical scope of this study

The goals of this study incorporate a number of theoretical and empirical questions. We employ two different objective methods (Bayes factors and an adapted version of AIC) for identifying complex best-fit models of nucleotide evolution in a Bayesian phylogenetic context. In doing so, we address the question, “Is it practically important to consider complex models of

evolution for ‘mesoscale’ phylogenetic analyses?’” Given careful consideration of appropriate model choice, we apply the resulting phylogenetic hypotheses to outstanding questions regarding systematics of the *Porthidium* group. Specifically, we sought to address the following empirical questions: 1) Do we find evidence for the monophyly of *Atropoides*? 2) What are the relationships among the three *Porthidium* group genera? 3) Is there evidence of undescribed or non-monophyletic *Porthidium* group taxa?

Materials and methods

Taxon sampling

In total, 61 terminal taxa (OTU’s) were included in this study. The ingroup (members of the genera *Atropoides*, or *Cerrophidion*, and *Porthidium*) included 52 samples representing 15 of 18 nominal species. We included multiple representatives of nominal species where possible, Details of terminal-taxon sampling (along with voucher information) are provided in Table 7. Our sampling of recognized species included 5/5 *Atropoides* species, 3/4 *Cerrophidion* species

Table 7. Specimens used in this study including GenBank accession numbers.

Taxon	Specimen Reference ID	Voucher	Locality	ND4	cyt-b
Outgroups					
<i>Lachesis stenophrys</i>	Lachesis stenophrys		Costa Rica: Limón	U41885	AY223603
<i>Ophryacus melanurus</i>	Ophryacus melanurus	UTA-R-34605	Mexico	AY223634	AY223587
<i>Ophryacus undulatus</i>	Ophryacus undulatus	CLP-73	Mexico	AY223633	AY223586
<i>Bothriechis schlegelii</i>	Bothriechis schlegelii	MZUCR-11149	Costa Rica	AY223636	AY223590
<i>Bothriechis nigroviridis</i>	Bothriechis nigroviridis	MZUCR-11151	Costa Rica	AY223635	AY223589
<i>Bothriechis lateralis</i>	Bothriechis lateralis	MZUCR-11155	Costa Rica	U41873	AY223588
<i>Bothrocophias hyoprora</i>	Bothrocophias hyoprora		Colombia: Leticia	U41886	AY223593
<i>Bothriopsis taeniata</i>	Bothriopsis taeniata		Surinam	AY223637	AY223592
<i>Bothrops ammodytoides</i>	Bothrops ammodytoides	MVZ-223514	Argentina: Neuquén	AY223639	AY223595
Ingroup					
<i>Atropoides mexicanus</i>	A. mexicanus Costa Rica1	UTA-R-12943	Costa Rica: Cartago: Pavones de Turrialba	AY220335	AY220312
	A. mexicanus Costa Rica2	MSM	Costa Rica: Puntarenas: San Vito	AY220336	AY220313
	A. mexicanus Costa Rica3	CLP-168	Costa Rica: San José	U41871	AY223584
	A. mexicanus Guatemala1	UTA-R-35942	Guatemala: Baja Verapaz: Nino Perdido	AY220330	AY220037
	A. mexicanus Guatemala2	UTA-R-32746	Guatemala: Huehentanango: Finca Chiblac	AY220331	AY220308
	A. mexicanus Guatemala3	UTA-R-35944	Guatemala: Izabal: Puerto Barrios	AY220332	AY220309
	A. mexicanus Guatemala4	UTA-R-43592	Guatemala: Quiché: Mountains West of El Soch	AY220334	AY220311
	A. mexicanus Guatemala5	UTA-R-46616	Guatemala: Alta Verapaz: Finca San Juan	AY220329	AY220306
<i>Atropoides nummifer</i>	A. nummifer Guatemala6	UTA-R-32419	Guatemala: Petén: San José El Espinero	AY220333	AY220310
	A. nummifer Mexico1	UTA-R-24842	Mexico: Hidalgo: vic. Huejutla	AY220337	AY220314
	A. nummifer Mexico2	ENS-10515	Mexico: Puebla: San Andres Tziaulan	DQ061220	DQ061195
<i>Atropoides occiduus</i>	A. occiduus Guatemala1	UTA-R-29680	Guatemala: Escuintla: S. slope Volcán de Agua	AY220338	AY220315
	A. occiduus Guatemala2	UTA-R-46719	Guatemala: Sololá: San Lucas Tolimán	AY220340	AY220317
	A. occiduus Guatemala3	UTA-R-24763	Guatemala: Guatemala: Villa Nueva	AY220339	AY220316
	A. occiduus Honduras	ENS-10630	Honduras: Olancho: Sierra de Botaderos	DQ061219	DQ061194

Taxon	Specimen Reference ID	Voucher	Locality	ND4	cyt-b
<i>Atropoides olmec</i>	A. olmec Guatemala	UTA-R-34158	Guatemala: Baja Verapaz: Niño Perdido	AY220342	AY220319
	A. olmec Mexico1	ENS-10510	Mexico: Chiapas: Mapastepec	DQ061221	DQ061196
	A. olmec Mexico2	JAC-9745	Mexico: Oaxaca: Cerro El Baúl	AY220343	AY220320
	A. olmec Mexico3	UTA-R-25113	Mexico: Veracruz: Sierra de los Tuxtlas	AY220344	AY220321
	A. olmec Mexico4	UTA-R-14233	Mexico: Veracruz: Sierra de los Tuxtlas	AY220345	AY220322
<i>Atropoides picadoi</i>	A. picadoi Costa Rica1	CLP-45	Costa Rica: Alajuela: Varablanca	U41872	AY223593
	A. picadoi Costa Rica2	UTA-R-23837	Costa Rica: San José: Bajo la Hondura	AY220347	AY220324
	A. picadoi Costa Rica3	MSM-10350	Costa Rica: San José: Bajo la Hondura	DQ061222	DQ061197
<i>Cerrophidion godmani</i>	C. godmani Costa Rica1	MSM	Costa Rica: San José	AY220351	AY220328
	C. godmani Costa Rica2	MSM	Costa Rica: San José: Goicochea	DQ061224	DQ061199
	C. godmani Costa Rica3	MSM	Costa Rica: San José: Goicochea	DQ061225	DQ061200
	C. godmani Guatemala1	UTA-R-40008	Guatemala: Baja Verapaz: La Unión Barrios	AY220348	AY220325
	C. godmani Guatemala2	ENS-8195	Guatemala: Quiché	DQ061223	DQ061198
	C. godmani Honduras	ENS-10631	Honduras: Ocotepéque: Güisayote	DQ061226	DQ061201
	C. godmani Mexico	JAC-15709	Mexico: Oaxaca: Cerro El Baúl	AY220349	AY220326
	<i>Cerrophidion petlalcalensis</i>	C. petlalcalensis Mexico	ENS-10528	Mexico: Veracruz: Orizaba	DQ061227
<i>Cerrophidion tzotzilorum</i>	C. tzotzilorum Mexico1	ENS-10529	Mexico: Chiapas: Las Rosas	DQ061228	DQ061203
	C. tzotzilorum Mexico2	ENS-10530	Mexico: Chiapas: Zinacantán	DQ061229	DQ061204
<i>Porthidium arcoase</i>	P. arcosae Ecuador	WWW-750	Ecuador: Manabí: Salango	AY223631	AY223582
<i>Porthidium dunni</i>	P. dunni Mexico1	MS	Mexico: Chiapas: Guardiania	DQ061243	DQ061217
	P. dunni Mexico2	ENS-9705	Mexico: Oaxaca: near San Pedro Pochutla	AY223630	AY223581
<i>Porthidium lansbergii</i>	P. lansbergii Panama	MSM	Panama: Darién	DQ061231	DQ061206
	P. lansbergii Venezuela	WES	Venezuela: Isla Margarita	DQ061230	DQ061205
<i>Porthidium nasutum</i>	P. nasutum Costa Rica1	MSM	Costa Rica: Alajuela: Río Cuarto de Grecia	DQ061235	DQ061210
	P. nasutum Costa Rica2	MSM	Costa Rica: Cartago: Guayacán de Turrialba	DQ061233	DQ061208
	P. nasutum Costa Rica3	MSM	Costa Rica: Cartago: Guayacán de Turrialba	DQ061234	DQ061209
	P. nasutum Costa Rica4	MZUCR-11150	Costa Rica	U41887	AY223579

Taxon	Specimen Reference ID	Voucher	Locality	ND4	cyt-b
	P. nasutum Ecuador	FGO-live-517	Ecuador: Esmeraldas: Zapallo Grande	AF29574	AF292612
	P. nasutum Guatemala	UTA-R-44749	Guatemala: Alta Verapaz: Cobán	DQ061232	DQ061207
<i>Porthidium ophryomegas</i>	P. ophryomegas Costa Rica	UMMZ-210276	Costa Rica: Guanacaste	U41888	AY223580
	P. ophryomegas Guatemala	MSM-23	Guatemala: Zacapa	DQ061241	DQ061216
	P. ophryomegas Honduras	UTA-R-52580	Honduras: Gracias a Dios: Mocerón	DQ061240	
<i>Porthidium porrasi</i>	P. porrasi Costa Rica1	MSM	Costa Rica: Puntarenas	DQ061239	DQ061214
	P. porrasi Costa Rica2	MSM	Costa Rica: Puntarenas: Sierpe	DQ061236	DQ061211
	P. porrasi Costa Rica3	MSM	Costa Rica: Puntarenas: San Pedrillo	DQ061237	DQ061212
	P. porrasi Costa Rica4	MSM	Costa Rica: Puntarenas: Golfito	DQ061238	DQ061213
<i>Porthidium yucatanicum</i>	P. yucatanicum Mexico	JAC-24438	Mexico: Yucatán: Car. Yaxcabá-Tahdzibichen	DQ061244	DQ061215

(lacking *C. barbouri*), and 7/9 *Porthidium* species (lacking *P. hespere* and *P. volcanicum*).

Outgroup taxa were chosen based on results from recent large-scale pitviper phylogenetic studies (Parkinson, 1999; Parkinson et al., 2002; Parkinson and Castoe, unpublished). Additionally, we intentionally included two taxa (*Ophryacus melanurum* and *Bothrocophias hyoprora*) that were at one time considered members of the *Porthidium* group and later removed (Gutberlet, 1998; Gutberlet and Campbell, 2001). Our outgroup sampling strategy included multiple successive outgroups (Smith, 1994) based on the expectation that this approach would reduce potential biases imposed by rooting phylogenies with a single outgroup.

DNA sequencing and sequence alignment

In addition to novel sequences generated from this study, several sequences used in this study have been previously published (Parkinson 1999; Parkinson et al., 2002; Castoe et al., 2003; Wüster et al., 2002; see Table 1 for details). Laboratory methods for obtaining novel sequences used in this study are as follows. Genomic DNA was isolated from tissue samples (liver or skin preserved in ethanol) using the Qiagen DNeasy extraction kit and protocol (Qiagen Inc., Hilden, Germany). Two protein-coding mitochondrial gene fragments were amplified and sequenced per sample: the ND4 fragment (including the 3' region of the NADH dehydrogenase subunit 4 gene), and the *cyt-b* fragment (including the 3' region of the cytochrome-*b* gene).

The ND4 fragment was amplified via PCR using the primers ND4 and LEU or ND4 and HIS (Arévalo et al., 1994). The *cyt-b* fragment was PCR amplified using the primers Gludg and AtrCB3 (Parkinson et al., 2002). Genechoice or Sigma brand PCR reagents were used to conduct PCR in the following final concentrations: 1x standard PCR buffer, 1.5 units *Taq* polymerase,

0.1 μ M per primer, 1.0 mM dNTPs 2.0 mM MgCl₂ and 0.004% DMSO. Thermocycling conditions included initial denaturation at 95C for 3 min.; 35 cycles of 95C for 30 sec., 48C for 30 sec., 72C for 45 sec., and a final extension at 72C for 5 min. Positive PCR products were excised from agarose electrophoretic gels and purified using the GeneCleanIII kit (BIO101). Purified PCR products were sequenced in both directions with the amplification primers (and for ND4, an additional internal primer HIS; Arévalo et al., 1994). Purified PCR products were sequenced using the CEQ D Dye Terminator Cycle Sequencing (DTCS) Quick Start Kit (Beckman-Coulter) and run on a Beckman CEQ2000 automated sequencer according to the manufacturers' protocols. Raw sequence chromatographs for sequences generated in this study were edited using Sequencher 3.1 (Gene Codes Corp, 1996). Sequences of each fragment were aligned manually in GeneDoc (Nicholas and Nicholas, 1997). Alignment was unambiguous and contained no inferred indels within the ingroup but included the absence of a complete codon in the *cyt-b* fragment in several outgroup specimens. No internal stop codons were found in either fragment. The final alignment of both gene fragments concatenated comprised a total of 1405 aligned positions: 693 from ND4 and 712 from *cyt-b*. Novel sequences were deposited in GenBank (GenBank accession numbers for all sequences used are given in Table 7).

Phylogenetic reconstruction

Throughout all phylogenetic reconstructions, gaps in alignment were treated as missing data. Maximum parsimony (MP) and Bayesian Metropolis-Hastings coupled Markov chain Monte Carlo (MCMC) phylogenetic methods were used to reconstruct phylogenies. Both

methods were initially used to compare phylogenetic reconstructions based on each gene fragment independently to identify any instances where different gene fragments demonstrated strongly supported alternative phylogenetic arrangements. We expect that mitochondrial loci should all contain phylogenetic signal supporting a common phylogeny because mitochondrial haplotypes are inherited maternally as a single linkage unit. We tested this assumption (prior to combining data) by estimating individual gene fragment phylogenies and checking for bipartitions that differed between gene fragments and were well supported (e.g., Wiens, 1998) using both maximum parsimony and Bayesian MCMC analyses.

All MP phylogenetic analyses were conducted using PAUP* version 4.0b10 (Swofford, 2002). All characters were treated as equally-weighted in MP searches. We used the heuristic search option with inactive steepest descent option, tree bisection reconnection (TBR) branch-swapping option, and 10,000 random-taxon-addition sequences to search for optimal trees. Support for nodes in MP reconstructions was assessed using non-parametric bootstrapping (Felsenstein, 1985) with 1,000 full heuristic pseudo-replicates (10 random-taxon-addition sequence replicates per bootstrap pseudo-replicate).

ModelTest version 3.0 (Posada and Crandall, 1998, 2001) was used to select an appropriate model of evolution for MCMC analyses based on consideration of both available criteria, hLRT and AIC (with likelihoods for models estimated in PAUP*). In addition to the combined dataset, all putative partitions of the dataset were independently analyzed using ModelTest to determine best-fit models of nucleotide evolution. These estimates were used as a partial justification for partition-specific model choice during the construction of partitioned MCMC models, similar to the suggestions of Brandley et al. (2005).

All MCMC phylogenetic analyses were conducted in MrBayes 3.0b4 (Ronquist and Huelsenbeck, 2003) with vague priors and three heated chains in addition to the cold chain (as per the program's defaults). Each MCMC analysis was conducted in triplicate, with three independent runs initiated with random trees, and run for a total of 4.0×10^7 generations (sampling trees every 100 generations). Conservatively, the first 1.0×10^7 generations from each run were discarded as burn-in. Summary statistics and consensus phylograms with nodal posterior probability support were estimated from the combination of the triplicate set of runs per analysis.

An initial set of MCMC runs (for the individual and combined datasets) was run using the model estimated by ModelTest (considering both AIC and hLRT criteria) to fit each individual gene or combined dataset (or nearest model available in MrBayes 3.0, as explained below). In addition to the model selected by ModelTest, the combined dataset was subjected to five additional MCMC analyses under alternative evolutionary models. These five additional MCMC analyses were designed to allow independent models of evolution to be used for partitions of the combined dataset. This was accomplished by partitioning the dataset into what we assumed were biologically relevant partitions and specifying that an independent GTR+ Γ +I model, with independent base frequencies, be used for each identified partition (using the "unlink" command in MrBayes 3.0). For these complex models, only branch lengths and topology remained linked between partitions. The names and details of all models used to analyze the combined dataset are summarized in Table 8. These models partitioned the combined dataset based on combinations of codon position and/or gene fragment (ND4 vs. *cyt-b*).

Several methods are available for model selection in a Bayesian context. In this study we employ three statistics for the purposes of model selection: 1) Bayes factors (B_{10}), 2) relative Bayes factors (RBF), and 3) Akaike weights (A_w) to choose a best-fit model from among the alternative models outlined above. Each of these criteria allow testing of non-nested models (not allowed by hierarchical log-likelihood ratio tests: hLRTs), which is important here because two alternative models are non-nested (“2x-gene” and “2x-pos12,3” models). Also, each criteria allow accommodation of marginal model likelihoods (rather than maximum likelihoods) derived from Bayesian MCMC analyses (accommodation of marginal model likelihoods for AIC is described below).

Bayes factors were calculated following Nylander et al. (2004) and we report the results in the form of $2\ln B_{10}$. To compare two competing models, M_0 and M_1 , the Bayes factor supporting M_1 over M_0 is equal to the ratio of the model likelihoods. We considered $2\ln B_{10} > 10$ sufficient to support M_1 over M_0 (Kass and Raftery, 1995; see also Brandley et al., 2005; Nylander et al., 2004).

Relative Bayes factors (RBF) were used to quantify the average impact that each free model parameter had on increasing the fit of the model to the data. These values were also used qualitatively to estimate the ratio of parameters to posterior evidence (of prior modification by the data) of increasingly complex models. This statistic is a permutation of the Bayes factor between the simplest (best-fit unpartitioned) and the alternative partitioned model that is normalized to the difference in free model parameters between models. We calculated the RBF

Table 8. Best-fit models selected by ModelTest for various partitions of the dataset based on both hLTR and AIC criteria. P1-6 refer to the six independent partitions of the dataset under the 6x-gene,codon model.

Partition	hLTR	AIC
Entire dataset	GTR+ Γ +I	GTR+ Γ +I
ND4	TVM+ Γ +I	TrN+ Γ +I
cyt-b	TVM+ Γ +I	TrN+ Γ +I
codon position 1	TrN+ Γ +I	TVM+ Γ +I
codon position 2	HKY+ Γ +I	TIM+ Γ +I
codon position 3	TIM+ Γ +I	TIM+ Γ +I
P1 = (ND4,pos1)	TrNef+ Γ +I	GTR+ Γ +I
P2 = (ND4,pos2)	HKY+ Γ	TVM+ Γ +I
P3 = (ND4,pos3)	TrN+ Γ	GTR+ Γ +I
P4 = (cyt-b,pos1)	TrNef+ Γ +I	HKY+ Γ +I
P5 = (cyt-b,pos2)	HKY+ Γ +I	TrN+ Γ +I
P6 = (cyt-b,pos3)	HKY+ Γ +I	TrN+ Γ +I

of each complex model by calculating $2\ln B_{10}$ between the base model and each complex (partitioned) model and dividing this by the difference in the number of free model parameters between the base and complex model.

We used a statistic derived from Akaike Information Criteria (AIC) in addition to statistics based on Bayes factors. Specifically, we implemented an adapted version of Akaike weights to infer the best-fit model of nucleotide evolution. Instead of using the maximum likelihood value, we used the harmonic mean estimator of the $\ln L$ from MCMC analyses to incorporate an estimate of the marginalized likelihood of models to be compared using Akaike weights (A_w ; see also Kauermann, et al., 2004; Wager et al., in press). The estimation of A_w has been recently reviewed by Posada and Buckley (2004), and we provide a brief summary here. The AIC of each model is calculated as the $AIC = -2L + 2K$ where K is the number of estimatable parameters (model parameters plus branch lengths in our case; for unrooted bifurcating trees the total number of branches is equal to twice the number of taxa minus three). From this, we calculated the change in AIC across models by comparing the AIC of the i th model to the model with the highest likelihood (min AIC) using the equation $\Delta AIC_i = AIC_i - \min AIC$. Akaike (1983) suggested that the relative likelihood of the models given the data may be obtained using the formula $e^{(-\Delta AIC_i/2)}$, which may then be normalized over all models to obtain a set of positive Akaike weights (A_w). This is accomplished by dividing each $e^{(-\Delta AIC_i/2)}$ by the sum of all $e^{(-\Delta AIC_i/2)}$ values across all models. Thus, the higher the A_w for a model, the higher the relative support for that model.

In addition to employing Bayes factors and Akaike weights to identify best-fit models of nucleotide evolution, we secondarily evaluated the performance of alternative models to check

for problems with mixing and convergence indicative of model over-fitting (overparameterization). Once a tentative model was chosen, this model was rigorously examined to check for evidence of parameter identifiability, failed convergence, and unreliability (which would suggest the model may be parametrically over-fit). We investigated the performance of models (using Tracer; Rambout and Drummond, 2003) by examining features of model likelihood and parameter estimate burn-in, as well as the shapes and overlap of posterior distributions of parameters. Specifically, we looked for evidence that model likelihood and parameter estimates ascended directly and relatively rapidly to a stable plateau, and that independent runs converged on similar likelihood and parameter posterior distributions (considered evidence that a model was not over-fit). We also examined the model parameter estimates to confirm that the shape of their posterior distributions reflected a substantial modification of the priors (indicating their identifiability). As a secondary validation that the partitioning of the dataset was justified, we compared posterior distributions of parameter estimates across partitions (by inspecting posterior distributions using Tracer, and by comparing 95% credibility intervals of parameters) to confirm that, in fact, different partitions demonstrated unique posterior distributions of parameter estimates.

Results

Dataset characteristics and individual gene phylogenies

The concatenated alignment of 1405 characters contained 538 parsimony-informative characters and 713 constant characters. Nucleotide frequencies were similar between the two loci used, and the nucleotide frequencies of the combined dataset were G = 11.57%, A = 29.79%, T = 26.46%, and C = 32.18%. Individual gene phylogenetic reconstructions showed extremely similar, yet poorly resolved, phylogenetic estimates. Based on the apparent congruence in phylogenetic signal between the two gene fragments, we proceeded with combined data analyses.

The greatest pairwise sequence divergence among terminal taxa was between *Bothrops ammodytoides* and *Porthidium yucatanicum* (uncorrected divergence of 17.4%). Within ingroup genera, the highest sequence divergence within *Atropoides* was 11.6% (between “*A. picadoi* Costa Rica2” and “*A. mexicanus* Guatemala4”), within *Cerrophidion* was 9.4% (between “*C. tzotzilorum* Mexico2” and “*C. godmani* Costa Rica1”), and within *Porthidium* was 13.7% (between “*P. dunnii* Mexico2” and *P. lansbergii* Panama”).

Maximum parsimony phylogenetic analysis

The MP heuristic search on the combined dataset found 144 equally parsimonious trees of 2587 steps. A substantial degree of character-state homoplasy was inferred across these trees

based on the homoplasy index (HI = 0.6308) and rescaled consistency index (RCI = 0.2690). The 50% majority-rule consensus of these 144 MP trees, along with bootstrap support for nodes, is shown in Fig. 11.

The MP phylogenetic reconstruction did not infer a monophyletic *Atropoides*, placing *A. picadoi* in an unresolved clade with *Cerrophidion* and *Porthidium*. *Atropoides* minus *A. picadoi*, referred to as the *nummifer* complex (Castoe et al., 2003), was resolved as monophyletic with 100% bootstrap support (BS). All *Atropoides* and *Cerrophidion* species were estimated to be monophyletic, as were all species of *Porthidium* except *P. nasutum*. Samples of Central American *P. nasutum* formed a well-supported clade (BS = 100%) distantly related to South American (Ecuadorian) *P. nasutum*. The *P. nasutum* sample from Ecuador appears to be more closely related to South American and southern Central American *P. lansbergi*. A majority of MP phylogenetic results overlap broadly with those from MCMC analyses. For this reason, and our expectation that MCMC results should produce more accurate estimates of phylogeny, we limit our discussion to these results.

Bayesian MCMC model selection and evaluation

Both AIC and hLTR model selection criteria supported the GTR+ Γ +I model as the best fit for the combined dataset (Table 9). The TVM+ Γ +I (under hLTR criteria) and the TrN+ Γ +I (under AIC criteria) models were selected as best fitting the individual gene data sets. These models are restrictions of the GTR+ Γ +I model that are not available in MrBayes 3.0; instead we used a GTR+ Γ +I model as our base model for the analysis of both individual and combined data.

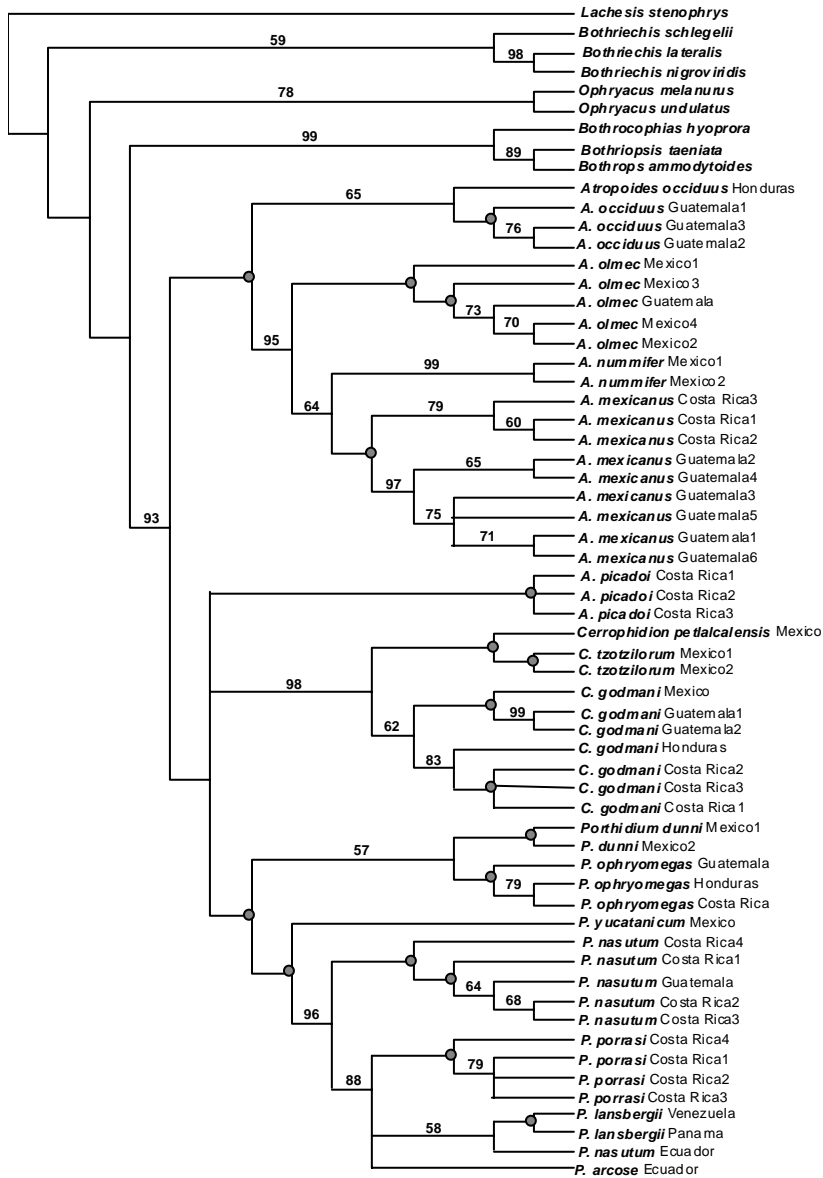


Figure 11. Majority-rule consensus of 144 equally-parsimonious trees (of 2587 steps) from heuristic maximum parsimony search based on 1405 bp. Bootstrap support for nodes is provided (values below 50% not shown). Bootstrap values of 100% are indicated with gray-filled circles.

Table 9. Description of complex partitioned models used in the analysis of the combined dataset.

Model	# Partitions	# Free Model Parameters	Description
1x-GTR+ Γ +I	1	10	base model employing a single GTR+ Γ +I model for the combined data
2x-gene	2	20	independent GTR+ Γ +I models for each of the two gene fragments
2x-pos12,3	2	20	one GTR+ Γ +I model for codon positions 1 and 2, and a second GTR+ Γ +I for position 3
3x-codon	3	30	one GTR+ Γ +I model per codon position
4x-gene12,3	4	40	each of the two gene fragments are allocated a set of two GTR+ Γ +I models, one for codon positions 1 and 2, a second for position 3
6x-gene,codon	6	60	each codon position of each of the two gene fragments are allocated an independent GTR+ Γ +I model

In addition to the GTR+ Γ +I model, we analyzed the combined dataset under five additional more complex models that employed multiple GTR+ Γ +I models assigned to specific partitions of the dataset (see Table 9). In MrBayes 3.0, available choices for modeling time-reversible nucleotide substitution include three possible substitution matrices including 1, 2, or 6 parameters. ModelTest results for all putative partitions indicated, in general, that there was evidence for the justification of nucleotide models including substitution matrices with greater than 2 parameters, as well as the parameters Γ and I (Table 8). Based on these results, we allocated independent GTR+ Γ +I models, per partition, in our partitioned MCMC analyses.

The evaluation of model fit for the complex models is visually depicted in Fig. 12. In comparing Bayes factors ($2\ln B_{10}$) between models, simple models were rejected in favor of more complex models that allowed parameters to be independently allocated to partitions of the dataset (Fig. 2). Ultimately, the most complex model tested, 6x-gene,codon, was supported as the best-fit model by $2\ln B_{10}$ estimates. Similarly, Akaike weights (A_w) placed nearly all relative weight ($A_w = 0.9998$) under the same 6x-gene,codon model as best fitting the data. Relative Bayes factors (RBF) demonstrate that, as model complexity and the number of free parameters increased, the relative improvements in model likelihood (per parameter added) decreased (Fig. 12). In summary, the RBF values suggest diminishing returns (in terms of likelihood) as more parameters were added to the model.

The best-fit complex model (6x-gene,codon) showed no evidence of parametric overfitting based on analysis of convergence and mixing. All independent MCMC runs of this model converged on nearly identical parameter and phylogenetic estimates. Model likelihoods and parameter estimates of all runs demonstrated effective mixing with burn-in characterized by a

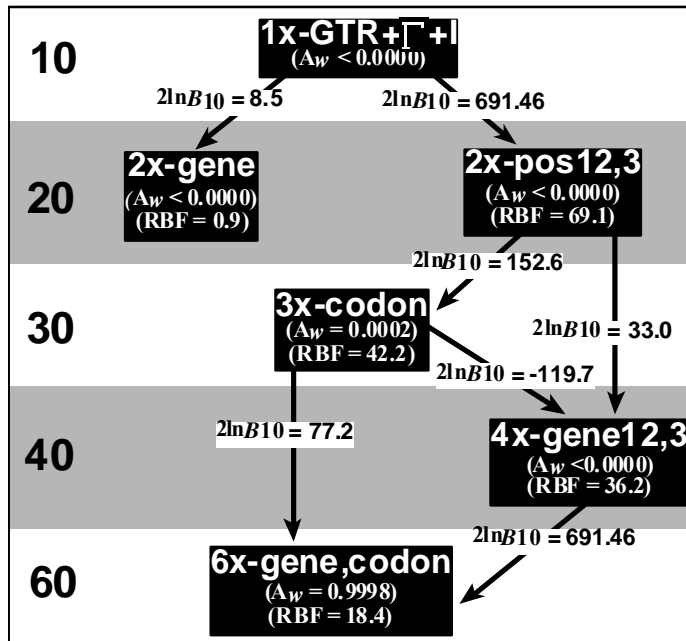


Figure 12. Flow chart illustrating the process of model selection among complex models tested for the analysis of the combined dataset. Statistics for models are given (A_w = Akaike weights, $2\ln B_{10}$ = $2\ln$ Bayes factor, RBF = Relative Bayes factor). For $2\ln B_{10}$ comparisons between models, M_i is represented by the model indicated by the arrowhead. See Table 9 for definitions of models.

direct rapid ascent to a stationary plateau (for model likelihood and parameters). Across all independent runs of the 6x-gene,codon model, likelihood values reached apparent stationarity (burned-in) prior to 1.5×10^6 generations, and parameter estimates reached apparent stationary by 2.0×10^6 generations. These observations confirm that our conservative *a-priori* choice of burn-in period at 1×10^7 effectively excluded non-stationary estimates.

Across partitions of the 6x-gene,codon model, base frequency, Γ , and I parameter estimates demonstrated posterior distributions with relatively low variance. In support of partitioning, these parameter-estimate distributions showed relatively little overlap between partitions (based on comparisons of the parameter distributions in Tracer and 95% confidence intervals; Table 10) and supported the distinctiveness of each partition. Posterior distributions of parameter estimates from the nucleotide substitution-rate matrix (i.e., GTR matrix parameters) of each partition showed higher degrees of overlap across partitions and greater variance compared with base frequencies, Γ , and I parameters (Table 10). While increasing parameter variance is expected when models are partitioned (because less data is available for estimation of each parameter), it was initially unclear if this increased variance may indicate that fitting each partition with a GTR substitution matrix over-fit the combined model. To test this we conducted a second set of partitioned runs in which we conducted MCMC analyses under an array of partitioned models where the substitution matrices were hierarchically re-linked (thereby reducing the number of free substitution matrix parameters overall). When we examined model fitting using A_w and $2\ln B_{10}$, we found that all tested restrictions of the 6x-gene,codon model were never favored by either statistic as being a better fit to the data than the 6x-gene,codon model (data not shown). Collectively, our *post-hoc* analyses of the 6x-gene,codon model support

Table 10. Mean and 95% credibility interval for each parameter sampled from the combined posterior distribution of three independent MCMC runs of the 6x-gene,codon model.

	P1 - (ND4,pos1)	P2 - (ND4,pos2)	P3 - (ND4,pos3)	P4 - (cyt-b,pos1)	P5 - (cyt-b,pos2)	P6 - (cyt-b,pos3)
r(G-T)	1	1	1	1	1	1
r(C-T)	32.32 (9.69 – 82.3)	33.36 (5.04 – 84.36)	17.46 (5.1 – 43.57)	16.7 (3.81 – 49.34)	2.57 (1.18 – 5.08)	13.39 (3.37 – 28.14)
r(C-G)	0.32 (0.02 – 1.18)	24.19 (3.11 – 67.99)	0.17 (0.01 – 0.94)	1.10 (0.14 – 4.02)	0.33 (0.01 – 1.22)	4.32 (0.92 – 9.79)
r(A-T)	3 (0.78 – 8.17)	4.64 (0.35 – 16.35)	1.47 (0.34 – 3.78)	2.05 (0.39 – 6.45)	0.25 (0.04 – 0.69)	1.41 (0.28 – 3.14)
r(A-G)	7.51 (2.13 – 20.28)	44.37 (7.06 – 94.96)	33.40 (10.18 – 82.42)	17.10 (4.37 – 48.85)	83.59 (53.96 – 99.42)	52.14 (13.85 – 97.19)
r(A-C)	0.78 (0.15 – 2.32)	6.88 (0.49 – 23.94)	0.92 (0.24 – 2.27)	1.71 (0.31 – 5.35)	0.32 (0.05 – 0.88)	0.50 (0.1 – 1.16)
pi(A)	0.361 (0.308 – 0.414)	0.161 (0.118 – 0.208)	0.408 (0.362 – 0.453)	0.338 (0.281 – 0.399)	0.295	0.313 (0.269 – 0.358)
pi(C)	0.306 (0.256 – 0.359)	0.32 (0.266 – 0.377)	0.367 (0.326 – 0.409)	0.254 (0.207 – 0.305)	0.257 (0.208 – 0.309)	0.469 (0.429 – 0.51)
pi(G)	0.178 (0.14 – 0.219)	0.128 (0.089 – 0.172)	0.065 (0.053 – 0.079)	0.158 (0.11 – 0.205)	0.104 (0.069 – 0.144)	0.036 (0.028 – 0.044)
pi(T)	0.155 (0.122 – 0.194)	0.392 (0.334 – 0.452)	0.16 (0.137 – 0.185)	0.249 (0.202 – 0.3)	0.399 (0.342 – 0.456)	0.182 (0.16 – 0.207)
Γ	0.218 (0.181 – 0.266)	0.098 (0.085 – 0.113)	3.836 (2.35 – 6.333)	0.306 (0.232 – 0.408)	0.264 (0.161 – 0.471)	4.958 (2.786 – 9.137)
I	0.170 (0.06 – 0.277)	0.599 (0.481 – 0.705)	0.054 (0.01 – 0.104)	0.400 (0.276 – 0.506)	0.549 (0.4 – 0.681)	0.039 (0.009 – 0.08)

this model as the superior best-fit model examined for our data. Hereafter, we consider the 6x-gene,codon model as our preferred model, and results based on analyses under this model as our preferred phylogenetic hypothesis.

Effects of model choice on Bayesian phylogenetic hypotheses

We present the majority-rule consensus topology of both the chosen model (6x-gene,codon) and the unpartitioned (1x-GTR+ Γ +I) model (Fig. 13) in order to compare the practical effects of model choice. No overall trend of increasing or decreasing posterior probability values for clades (Pp hereafter) is evident between the trees. Also, no relationships that were supported by 100% Pp changed more than a single percent across the two models. Instead, the majority of differences between consensus topologies and Pp support represented changes at weakly supported nodes ($Pp < 90\%$) that result in a change in the majority-rule consensus topology. The Pp support for basal relationships between *Porthidium* group genera becomes substantially stronger in the complex model (from $Pp = 64$ and 68 to $Pp = 81$ and 84 , respectively). Other deep nodes, including the resolution of relationships among outgroup taxa, showed substantial changes across the two models (Fig. 13). Also, the two models produce different consensus topologies affecting the resolution of members of *Atropoides* as well as *Porthidium* (although both relationships are weakly supported under either scenario).

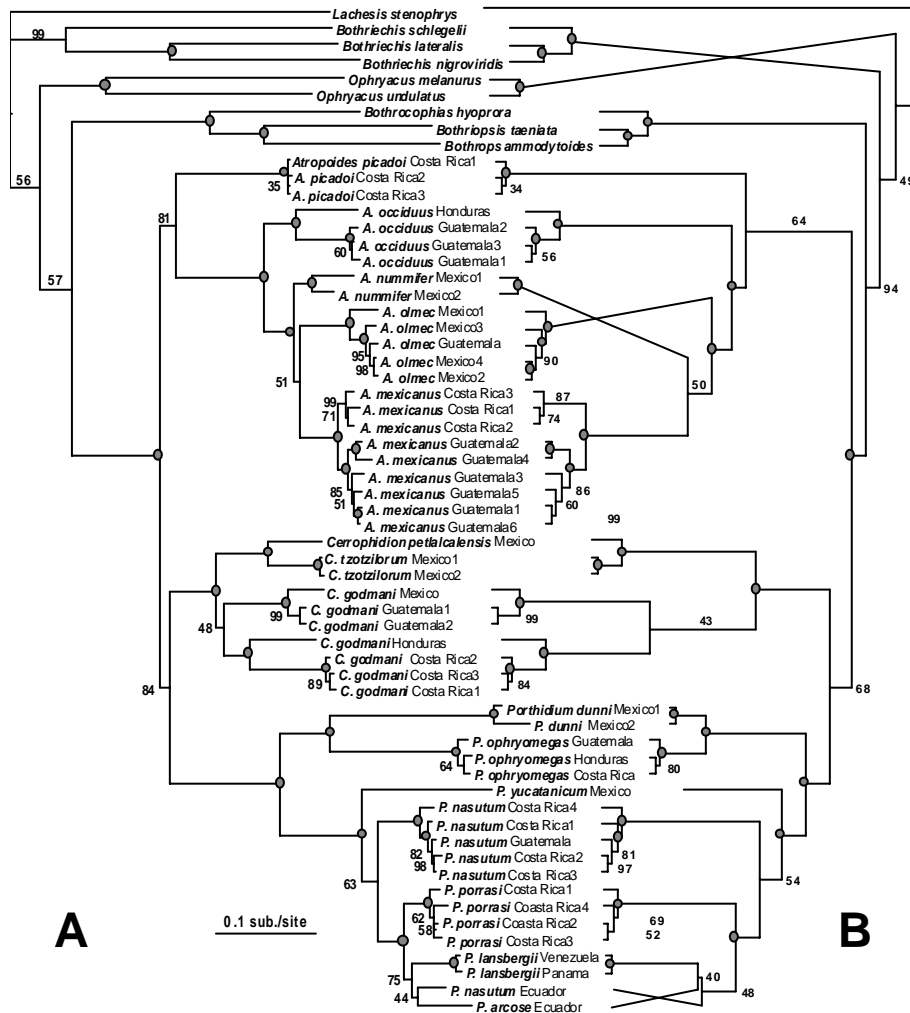


Figure 13. Majority rule consensus trees resulting from Bayesian MCMC phylogenetic reconstructions under two different models of nucleotide evolution (the favored partitioned model “6x-gene,codon” and the base unpartitioned 1x-GTR+ Γ +I). Nodal posterior probabilities are indicated; nodal posterior probabilities of 100% are indicated with a gray-filled circle. (A) Majority rule consensus phylogram based on a combined 9×10^7 post burn-in Bayesian MCMC generations of the favored “6x-gene,codon” partitioned model. (B) Majority rule consensus cladogram based on a combined 9×10^6 post burn-in Bayesian MCMC generations of the unpartitioned 1x-GTR+ Γ +I model (note: branch lengths are not informative in Fig. 13B).

Bayesian MCMC phylogenetic results under the best-fit model

The phylogenetic estimates for the *Porthidium* group derived from the MCMC analyses under the 6x-gene, codon model strongly support monophyly of the group ($Pp = 100\%$) and also inferred a clade comprising the primarily South American bothropoid lineages (genera *Bothrops*, *Bothriopsis*, and *Bothrocophias*). Monophyly is well supported for each of the genera *Cerrophidion* and *Porthidium* ($Pp = 100$), which are grouped ($Pp = 84$) as the sister taxon to a monophyletic *Atropoides* ($Pp = 81$). Within *Atropoides*, *A. picadoi* was inferred as the sister taxon to the remaining species ($Pp = 81\%$), which collectively form the *nummifer* complex. This group of *Atropoides* species was strongly supported as monophyletic, with a clade containing *A. mexicanus* and *A. olmec* ($Pp = 51$) forming the sister taxon to *A. nummifer*, and *A. occiduus* being the sister lineage to the remaining *nummifer* complex species ($Pp = 100$). Within *A. occiduus*, we found Honduran and Guatemalan populations to be well differentiated ($\sim 5.7\%$) compared to more shallow intraspecific divergences among populations of other *Atropoides* species.

Monophyly of the genus *Cerrophidion* received strong support ($Pp = 100$). The widespread species *C. godmani* was inferred with very weak support as monophyletic ($Pp = 48$), although a clade containing Honduran and Costa Rican populations received strong support ($Pp = 100$). Within this genus, we found evidence for an early phylogenetic split between a clade containing the two species restricted to Mexico (*C. tzotzilorum* and *C. petlalcalensis*) and *C. godmani*. Our sampling of *C. godmani* populations throughout Middle America highlights several cladogenetic divisions within this species (among northern, central, and southern Middle

American populations; divergences among the three lineages all > 7 %) that are deeper than those observed between the two other *Cerrophidion* species (< 6 %).

The first phylogenetic split within *Porthidium* separates a branch comprising *P. dumni* and *P. ophryomegas* ($Pp = 100$) from a branch comprising the remaining species ($Pp = 100$). All *Porthidium* species were resolved as monophyletic except *P. nasutum*. South American *P. nasutum* formed a weakly supported clade with *P. arcosae* ($Pp = 44$), the sister taxon to *P. lansbergii* ($Pp = 75$). This group of three South American lineages formed a clade with *P. porrasi* ($Pp = 100$). Central American populations of *P. nasutum* were found to represent a monophyletic group ($Pp = 100$) inferred to be the sister lineage ($Pp = 63$) to a clade comprising *P. porrasi* and the South American species.

Discussion

Model partitioning in Bayesian MCMC analyses

Our results support three important conclusions relevant to the use of complex partition-specific models in combined MCMC analyses. 1) Model choice may have important practical effects on phylogenetic conclusions even for mesoscale datasets such as the one used here. 2) The use of a complex partitioned model did not produce widespread increases or decreases in Pp nodal support. 3) A majority of differences in resolution resulting from model choice was

concentrated at deeper nodes. Also, a majority of these deeper nodes increased substantially in resolution (as measured by nodal Pp) with increasing model complexity.

Several studies have supported a direct relationship between accuracy of posterior probabilities and model complexity. In these studies, Bayesian analyses conducted with underparameterized models appear to experience elevated error rates, compared with parameter-rich models (Erixon et al., 2003; Huelsenbeck and Rannala, 2004; Lemmon and Moriarty, 2004; Suzuki et al., 2002). Also, simpler models have been shown to exhibit signs of poor mixing when compared to more complex partitioned models, based on the variance in Pp estimates through MCMC generations (Castoe et al., 2004). In addition to overall accuracy of results, this study (and Brandley et al., 2005) found that complex partitioned models may have important effects in the resolution of deeper nodes, a majority of which receive increased support under complex models. These results suggest that more complex models may be more effective at estimating patterns of molecular evolution when sequences are more divergent and phylogenetic signal is otherwise obscured by multiple substitutions or by homoplasy (see also discussion below). While not a panacea for resolving deep nodes, complex models that account for natural heterogeneity of molecular evolution within combined datasets appear to extract more phylogenetic signal than would a non-partitioned “compromise” model (see also Brandley et al., 2005; and analogous studies: Pupko et al., 2002; Voelker and Edwards, 1998; Yang, 1996).

Despite considerations favoring complex models, benefits of constructing and implementing more realistic evolutionary models of DNA substitution are challenged by the potential for imprecise and inaccurate model parameter and phylogeny estimation that may result from excess model complexity. Expanding computational power, increasing genomic resources,

and advances allowing broad flexibility in modeling evolutionary patterns in a Bayesian MCMC context collectively underscore the importance of developing accurate models and objective strategies for model testing.

As the implementation of complex models becomes more widespread in molecular phylogenetics, it may be useful to identify how reliant phylogenetic conclusions are on model specification. Reporting such details would provide an assessment of how much phylogenetic signal seems readily extracted from the data compared to that extracted through the implementation of more complex models (which may or may not ultimately contribute to the accuracy of phylogenetic results). In part, this is analogous to the common practice of providing results based on MP and likelihood-based phylogenetic methods. Also, advances with incorporating model averaging in phylogenetics (including reverse-jump Bayesian MCMC methods: Green, 1995; Suchard et al., 2001; Huelsenbeck et al., 2004) represent an attractive alternative to the common reliance on a single model for phylogeny estimation (see also Posada, 2003; Posada and Buckley, 2004).

Suggestions and prospects for complex Bayesian MCMC modeling and model testing

In accordance with previous empirical studies (e.g., Brandley et al., 2005; Castoe et al., 2004; Pupko et al., 2002), our results support the hypothesis that more complex models of evolution may have practical effects on phylogenetic inference. Furthermore, such models may more accurately portray heterogeneous patterns of evolution within a dataset, facilitating the extraction of more phylogenetic signal (i.e., at deep nodes) compared with simpler or non-

partitioned models. Support for the use of complex models has also been reiterated by simulation studies. With simulated data, Bayesian phylogenetic analyses conducted with oversimplified models suffer from inaccurate bipartition posterior probability estimates, whereas overly complex models do not appear to experience the same magnitude of inaccuracy (Alfaro et al., 2003; Huelsenbeck and Rannala, 2004; Lemmon and Moriarty, 2004). The potential utility of complex models, however, is balanced by potentially inaccurate or unreliable results that may be obtained from employing overly complex models. Resolving these opposing points requires robust and objective strategies for testing and evaluating such models.

In this study we exploited a three-part strategy for identifying, testing, and evaluating candidate complex models in a Bayesian MCMC context. We used standard methods implemented in ModelTest to examine potential models for biologically intuitive potential partitions of the dataset (as in Brandley et al., 2005), three statistics (A_w , $2\ln B_{10}$, and RBF) to examine model fit across partitioned Bayesian MCMC models, and *post-hoc* evaluation of model performance to check for proper mixing and convergence (including model parameter identifiability). We believe that these three steps represent a thorough strategy for the identification of best-fit models for partitioned Bayesian MCMC analyses that satisfy concerns (positive and negative) associated with employing complex models.

Several authors (Brandley et al., 2005; Nylander et al., 2004) have argued the efficacy of $2\ln B_{10}$ in Bayesian phylogenetic model selection. Here we find the results of A_w to support the same conclusions (picking the same model) as $2\ln B_{10}$, which is not entirely surprising given suggestions that AIC and Bayes factors are asymptotically equivalent (Akaike, 1983; see also Huelsenbeck et al., 2004). Through the use of the harmonic mean estimate of margin model

likelihood, both methods attractively incorporate parameter uncertainty into model choice (rather than maximum likelihood point estimates of model parameters and phylogeny). In terms of convenience, $2\ln B_{10}$ allows ready comparisons between two models, while A_w provides a useful perspective on model choice simultaneously over all models. Although the results of these two criteria were similar, they provide unique information and approaches to model selection (with different assumptions), and thus represent a desirable confirmatory approach to model selection when used together.

Although many interpretations exist, Bayes factors may be interpreted as the posterior evidence provided by the data for one model versus another being true (under uniform model priors) or as a comparison of the predictive likelihoods of the models (Gelfand and Dey, 1994; Kass and Raftery, 1995; Wasserman, 2000). Alternatively, Levine and Schervish (1999) suggested Bayes factors should be interpreted as measuring “the change in evidence in the odds in favor of the hypothesis when going from the prior to the posterior”, thus placing emphasis on the data modifying the priors as playing a primary role in determining the Bayes factor (see also Huelsenbeck et al., 2004; Wasserman, 2000). Unlike Bayes factors, AIC does not imply that the true model is contained in the set of candidate models (although the importance of this assumption for Bayes factors has been debated: e.g., Kass and Raftery, 1995; Posada and Buckley, 2004). Instead, AIC attempts to identify which model is most likely to be closest to the true model, or has the highest predictive accuracy, based on the Kullback-Liebler distance (Akaike, 1973; Forester, 2002; Sober, 2002). In comparing methods, some have suggested that Bayes factors may tend to favor simpler models than AIC (e.g., Barlett, 1957; Kass and Raftery, 1995; Lindley, 1957; Shibata, 1976). The AIC may also be less biased by specification of priors

(e.g., prior variance) whereas Bayes factors may become inaccurate if priors are too vague (diffuse and uninformative; Raferty and Zheng, 2003; see also Findley, 1991). However, AIC may only perform well when the dataset is large and when only ‘good’ models are compared (Burnham and Anderson, 1998). Neither method is clearly superior, but both have strengths, weaknesses, and potential biases. If methods agree, one can be more confident that biases or weaknesses of any one method have not misled model choice. If methods were to disagree regarding model choice, an investigator should weigh carefully the potential biases of each method in order to identify a preferred model; alternatively, one could evaluate multiple models and select the most complex that appears to not suffer from identifiability, mixing, and convergence problems (e.g., Huelsenbeck and Rannala, 2004).

In addition to Bayes factors and A_w , we also employed RBF (a rescaling of the Bayes factor) as a simple way to quantify the relative contribution of each added free parameter towards increasing overall model likelihood (starting from the base-unpartitioned model). As such, RBFs represent a simple *post-hoc* means of comparing the relative explanatory power of the added free parameters simultaneously across models. In general, as the number of free model parameters increase, we expect the RBF to decrease as the data to parameter ratio decreases. Thus, RBF values should generally decrease asymptotically with increasing model complexity. The rate of RBF decline should also be proportional to the size and heterogeneity of a dataset (assuming models are effectively portraying data heterogeneity).

These properties of the RBF make it a useful indicator that may help decide if model complexity is approaching the maximum justifiable complexity, or if the array of models tested still fall well below the maximum model complexity that may be warranted (e.g., through AIC or

Bayes factor model choice). If RBF values steadily decrease with model complexity, an investigator may be more convinced that they are approaching the higher end of model complexity justifiable by the data, as observed in this study. Contrastingly, if RBF values remain relatively constant across increasingly complex models, one may assume that the proportion of data to model parameters is high, which may suggest that even more complex models should be explored if possible. This latter pattern has been observed with large and more heterogeneous datasets (Castoe and Parkinson, unpublished manuscript).

Relationships and taxonomy of the Porthidium group

The intergeneric relationships among pitvipers have been investigated by numerous authors using either morphological or molecular data (recently reviewed by Gutberlet and Harvey, 2004). Despite this intensive systematic effort, a cohesive and robust hypothesis of relationships among genera has yet to be achieved. Many studies have supported a sister group relationship between the *Porthidium* group and South American bothropoid genera (*Bothrops*, *Bothriopsis*, *Bothrocophias*; e.g., Gutberlet and Harvey, 2002; Kraus et al., 1996; Parkinson, 1999; Parkinson et al., 2002). This relationship was supported in all our analyses, including MP and MCMC. As in previous molecular phylogenetic studies, we found strong support for the monophyly of the *Porthidium* group; this contrasts with previous studies based on morphology or morphology plus allozymes (Gutberlet and Harvey, 2002; Werman, 1992). Also, in accordance with previous studies (Parkinson, 1999; Parkinson et al., 2002; Gutberlet and Harvey, 2002), we found strong phylogenetic evidence supporting the previous removal of *Ophryacus*

melanurus and *Bothrocophias hyoproras* from the *Porthidium* group (Gutberlet, 1998; Gutberlet and Campbell, 2001).

Resolution of the basal relationships between the three genera of the *Porthidium* group appear to be a difficult phylogenetic problem to solve with either morphological or molecular data, as can be seen in our MP analyses (Fig. 12). Several molecular phylogenetic studies have either failed to resolve the relationships altogether or failed to resolve them with any substantial support (e.g., Castoe et al., 2003; Parkinson, 1999; Parkinson et al., 2002). In all cases, molecular phylogenies have inferred very short internodes connecting the three genera, implying a rapid radiation from a common ancestor and a difficult phylogenetic problem to solve. Parkinson et al. (2002) found weak support (BS = 68) for a clade containing *Cerrophidion* and *Atropoides*, as the sister taxon to *Porthidium*. Here, our partitioned MCMC analyses instead group *Cerrophidion* and *Porthidium* as a clade ($Pp = 84$) that is the sister lineage to *Atropoides*. It is important to note that resolution of these relationships appeared particularly dependent on MCMC model choice, with increasingly complex models recovering higher Pp for these relationships (Fig. 13). These different results across MCMC models would be reconciled under the hypothesis that complex models are, in fact, doing a better job extracting phylogenetic signal from the dataset which clearly does contain substantial homoplasy.

Despite the fact that species of *Atropoides* constitute a distinctive group of morphologically similar snakes, monophyly of this genus has not been well resolved based on molecular studies (Castoe et al., 2003; Parkinson, 1999; Parkinson et al., 2002). Our MP results also fail to resolve the question of monophyly. Similar to the resolution of the *Porthidium* group,

our MCMC analyses under the 6x-gene, codon model resolved monophyly of *Atropoides* with $Pp = 81$, compared to $Pp = 64$ in unpartitioned MCMC analyses.

Within the genus *Atropoides*, slight changes in the posterior distribution of trees under different MCMC models produced different majority-rule consensus trees of relationships among *Atropoides* species (in which *A. olmec* and *A. nummifer* exchanged positions). It is interesting to note that *A. olmec* and *A. mexicanus* share a presumed derived morphological feature (in having two or more subfoveal rows; Campbell and Lamar, 2004). Across all MP and MCMC analyses, *A. olmec* appears as the sister taxon to *A. mexicanus* only in the complex MCMC analysis (albeit with $Pp = 51$). These two species were resolved as the sister lineage to *A. nummifer*. These three species also all have nasorostral scales not present in the remaining species of *Atropoides*.

Previous molecular and morphological studies have supported *A. picadoi* as the sister lineage to all other *Atropoides*, and *A. occiduus* as the sister taxon to the remaining ‘*nummifer* complex’ species (Campbell and Lamar, 2004; Castoe et al., 2003; Parkinson et al., 2002). We also find strong evidence for these relationships based on MP (in part) and MCMC analyses.

Although not extensive, our intraspecific sampling within *Atropoides* illuminates several interesting patterns of phylogeography and undescribed taxonomic diversity. Castoe et al. (2003) demonstrated that the range of *A. olmec* included three closely-related disjunct populations in Veracruz and Oaxaca, Mexico, and Baja Verapaz, Guatemala. They concluded that in recent evolutionary time, the range of *A. olmec* may have been more continuous between these three known populations. Additional samples in this study include newly discovered populations in Chiapas, Mexico that further support the historical existence of a dispersal corridor spanning the Mexican Isthmus of Tehuantepec that facilitated relatively recent gene flow among these

populations. *Atropoides mexicanus* is the widest-ranging species in the genus and spans a majority of Middle America, although the occurrence of this species has not been confirmed throughout a large portion of Central America (in parts of Honduras and Nicaragua; Campbell and Lamar, 2004). We found evidence for phylogenetic structure within *A. mexicanus* whereby populations in northern Middle America form a clade, as do populations from Costa Rica. Shallow divergences between these clades indicate that gene flow across the large range of *A. mexicanus* has been prevalent at least within recent evolutionary time. These data support assertions that the ‘*nummifer* complex’ diversified in northern Middle America, and *A. mexicanus* later expanded its range southward (Castoe et al., 2003; Werman, 2005). Within *A. occiduus*, we found a Honduran sample to be substantially diverged from other Guatemalan populations. This and associated Honduran populations of *A. occiduus* may be candidates for species recognition if additional data support this distinction.

The genus *Cerrophidion* is composed of four species, three of which occupy small isolated ranges in Mexico. The two of these three range-restricted species sampled in this study, *C. tzotzilorum* and *C. petlalcalensis*, were recovered as a well-supported clade forming the sister lineage to the wide-ranging *C. godmani*. Although not sampled, the fourth *Cerrophidion* species, *C. barbouri*, shares a several presumably derived characters (low numbers of teeth and low numbers of middorsal scale rows) with *C. petlalcalensis*, suggesting these taxa may be sister species (Gutberlet and Harvey, 2004; although see Campbell, 1988).

The range of *C. godmani* extends from southern Mexico to northern Panama, although populations are patchily distributed across disjunct highland masses. Our results support for the existence of multiple divergent lineages within *C. godmani* that correspond to disjunct groups of

populations. We found strong support for three *C. godmani* lineages including: 1) populations in Mexico and Guatemala (BS = 100, *Pp* = 100); 2) populations in Honduras; 3) populations in Costa Rica (supported with BS = 83 and *Pp* = 100 as the sister lineage to Honduran *C. godmani*). These three lineages appear associated with three discrete geographic and geologic montane complexes that have been recognized as distinct biogeographic units in a number of studies (e.g.: Campbell, 1999; Savage, 1966, 1982; Stuart, 1966). Based on molecular evidence presented here, and on the allopatric distributions of these three lineages, additional work has been initiated to investigate the potential taxonomic recognition of these lineages of *C. godmani*.

Our results suggest a basal split within *Porthidium* between a clade including *P. dumni* and *P. ophryomegas* (both of which are restricted exclusively to tropical and subtropical dry habitats), and a clade comprising the remaining species, hereafter called the “*nasutum* group” (similar to Castoe et al., 2003; Parkinson, 1999; Parkinson et al., 2002). This basal split within *Porthidium* species is also supported by differences between clades in a dorsal-scale microstructural pattern (Estol, 1981; although not all *Porthidium* species were examined). The unsampled species *P. hespere* (of southwestern Mexico), like *P. ophryomegas* and *P. dumni*, is restricted to tropical dry forests and occurs geographically closest to *P. dumni*. While these facts suggest that *P. hespere* may be a member of the *P. ophryomegas* / *P. dumni* clade (see also Werman, 2005), no specific phylogenetic evidence is currently available to test this hypothesis. Within the widespread species *P. ophryomegas*, we observed shallow genetic structure across geographically distant populations, suggesting recent evolutionary genetic continuity across populations (Fig. 3, as inferred by Werman, 2005).

Porthidium yucatanicum has been hypothesized as being the sister taxon to all *Porthidium* species based on morphological data (Gutberlet and Harvey, 2002). We found strong support for this species to instead be the sister taxon to the remaining *nasutum* group species. This implies that early vicariance within the *nasutum*-group may have been centered in northern Middle America, which is not intuitive based on the lower Middle American and South American distribution of a majority of *nasutum* group taxa. We resolved *P. porrasi* as the sister lineage to this clade of South American lineages (*P. lansbergii*, *P. arcoase*, and Ecuadorian “*P. nasutum*”). *Porthidium porrasi* is restricted to the Osa Peninsula of southwestern Costa Rica (and immediately adjacent mainland), and was considered *P. nasutum* until recently (Lamar and Sasa, 2003). The close phylogenetic relationship of *P. porrasi* and South American *Porthidium* (rather than Central American lineages) seems to support a historical pattern of reticulating dispersal into and out of South America (see also Wüster et al., 2002).

We found strong evidence for paraphyly of *P. nasutum*, as reported by Wüster et al. (2002; see also Gutberlet and Harvey, 2004). Sampled populations of *P. nasutum* from Central America formed an evolutionarily shallow clade, distantly related to South American (Ecuadorian) “*P. nasutum*”. These results suggest that some taxonomic action may be required to rectify the phylogenetic relationships of South American “*P. nasutum*”, although the affinities of other populations allocated to *P. lansbergii* require further attention. We found Ecuadorian “*P. nasutum*” closely related to *P. lansbergii* and *P. arcosae* (both of which are geographically proximal and morphologically similar to South American populations of “*P. nasutum*”). Thus, decisive taxonomic treatment of *P. nasutum* may require a larger-scale reevaluation of the taxonomic status of *P. lansbergii* and *P. arcosae* (formerly considered a subspecies of *P.*

lansbergii; Campbell and Lamar, 2004). The unsampled species *P. volcanicum* (restricted to southwestern Costa Rica) has been suggested as a close relative of *P. lansbergii* by Solórzano (1995), which implies the potential for additional complications in clarifying the phylogeny and taxonomy of species related to *P. lansbergii*. *Porthidium* has historically been plagued with difficulties regarding taxonomic stability and correct species identification (reviewed by Campbell and Lamar, 2004). The taxonomic problems discussed here, and the likelihood of additional cryptic diversity among South American *Porthidium* populations (Campbell and Lamar, 2004) highlight future taxonomic activity for the genus.

References

- Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Second International Symposium on Information Theory. Akademiai Kiado, Budapest, pp. 673–681.
- Akaike, H., 1974. A new look at the statistical model identification. *IEEE Trans. Aut. Control* 19, 716–723.
- Akaike, H., 1983. Information measures and model selection. *Int. Stat. Inst.* 22, 277–291.
- Alfaro, M.E., Zoller, S., Lutzoni, F., 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. *Mol. Biol. Evol.* 20, 255–266.
- Arévalo, E.S., Davis, S.K., Sites Jr., J.W., 1994. Mitochondrial DNA sequence divergence and phylogenetic relationships among eight chromosome races of the *Sceloporus grammicus* complex (Phrynosomatidae) in central Mexico. *Syst. Biol.* 43, 387–418.
- Aris-Brosou, S., Yang, Z., 2002. Effects of models of rate evolution on estimation of divergence dates with special reference to the metazoan 18s ribosomal RNA phylogeny. *Syst. Biol.* 51, 703-714.
- Brandley, M.C., Schmitz, A., Reeder, T.W., 2005. Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of scincid lizards. *Syst. Biol.* in press.
- Buckley, T.R., 2002. Model misspecification and probabilistic tests of topology: evidence from empirical data sets. *Syst. Biol.* 51, 509–523.
- Burger, W.L. 1971. Genera of pitvipers. Ph.D. dissertation, University of Kansas, Lawrence, KS.

- Burnham, K.P., Anderson, D.R., 1998. Model Selection and Inference: A Practical Information Theoretic Approach. Springer, New York.
- Campbell, J.A., 1988. The distribution, variation, and natural history of *Porthidium barbouri*. *Acta Zoologica Mexicana, nueva serie* 26, 1–32.
- Campbell, J.A., 1999. Distribution patterns of amphibians in Middle America. In: Duellman, W.E. (Ed.), *Distribution Patterns of Amphibians: A Global Perspective*. Johns Hopkins University Press, Baltimore, MD, pp. 111–209.
- Campbell, J.A., Lamar, W.W., 1989. *The Venomous Reptiles of Latin America*. Cornell University Press, Ithaca, NY.
- Campbell, J.A., Lamar, W.W., 1992. Taxonomic status of miscellaneous neotropical viperids with the description of a new genus. *Occ. Papers Texas Tech. Univ.* 153, 1–31.
- Campbell, J.A., Lamar, W.W., 2004. *The Venomous Reptiles of the Western Hemisphere*. Cornell University Press, Ithaca, NY.
- Castoe, T.A., Chippindale, P.T., Campbell, J.A., Ammerman, L.A., Parkinson, C.L., 2003. The evolution and phylogeography of the Middle American jumping pitvipers, genus *Atropoides*, based on mtDNA sequences. *Herpetologica* 59, 421–432.
- 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.
- Erixon, S.P., Britton, B., Oxelman, B., 2003. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Syst. Biol.* 52, 665–673.

- Estol, C.O., 1981. Scale microdermatoglyphics of the viperid snake genera *Bothrops* and *Trimeresurus*: taxonomic relationships. Ph.D. dissertation, New York University, New York, NY.
- Faith, J.J., Pollock, D.D., 2003. Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. *Genetics* 165, 735–745.
- Farris, J.S., Källersjö, M., Kluge, A.G., 1994. Testing significance of incongruence. *Cladistics* 10, 315–319.
- Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39, 783–791.
- Findley, D.F., 1991. Counterexamples to parsimony and BIC. *Ann. Inst. Statist. Math.* 43, 505–514.
- Forster, M.R., 2002. Predictive accuracy as an achievable goal in science. *Phil. Sci.* 69, S124–S134.
- Gelfand, A.E., Dey, D.K., 1994. Bayesian model choice: Asymptotics and exact calculations. *J. R. Stat. Soc. B.* 56, 501–514.
- Gene Codes Corp., 1996. Sequencher, version 3.1. Gene Codes, Ann Arbor, MI.
- Green, P.J., 1995. Reversible jump MCMC computation and Bayesian model determination. *Biometrika* 92, 711–732.
- Gutberlet Jr., R.L., 1998. The phylogenetic position of the Mexican black-tailed pitviper (*Squamata*: *Viperidae*: *Crotalinae*). *Herpetologica* 54, 184–206.

- Gutberlet Jr., R.L., Campbell, J.A., 2001. Generic recognition for a neglected lineage of South American pitvipers (Squamata: Viperidae: Crotalinae), with the description of a new species from the Colombian Chocó. *Am. Mus. Novit.* 3316, 1–15.
- Gutberlet Jr., R.L., Jr., Harvey, M.B., 2002. Phylogenetic relationships of New World pitvipers as inferred from anatomical evidence. In: Schuett, G.W., Höggren, M., Douglas, M.E., Greene, H.W. (Eds.), *Biology of the Vipers*. Eagle Mountain Publishing, Salt Lake City, UT, pp. 51–68.
- Gutberlet Jr., R.L., Harvey, M.B., 2004. The evolution of New World venomous snakes. In: Campbell, J.A., Lamar, W.W., *The Venomous Reptiles of the Western Hemisphere*. Cornell University Press, Ithaca, NY, pp. 634–682.
- Holder, M., Lewis, P.O., 2003. Phylogeny estimation: traditional and Bayesian approaches. *Nat. Rev. Genet.* 4, 275–284.
- Huelsenbeck, J.P., Crandall, K.A., 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. *Ann. Rev. Ecol. Syst.* 28, 437–466.
- Huelsenbeck, J.P., Larget, B., Alfaro, M.E., 2004. Bayesian phylogenetic model selection using reverse jump Markov chain Monte Carlo. *Mol. Biol. Evol.* 21, 1123–1133.
- Huelsenbeck, J.P., Larget, B., Miller, R., Ronquist, F., 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. *Syst. Biol.* 51, 673–688.
- Huelsenbeck, J.P., Rannala, B., 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Syst. Biol.* 53, 904–913.
- Huelsenbeck, J.P., Ronquist, F., Nielsen, R., Bollback, J.P., 2001. Bayesian inference and its impact on evolutionary biology. *Science* 294, 2310–2314.

- Kass, R.E., Raferty, A.E., 1995. Bayes factors. *J. Am. Stat. Assoc.* 90, 773–795.
- Kauermann, G., Xu, R., Vaida, F., 2004. Smoothing, random effects and generalized linear mixed models in survival analysis. Technical report, University of Bielefeld, Bielefeld, Germany.
- Kraus, F., Mink, D.G., Brown, W.M., 1996. Crotaline intergeneric relationships based on mitochondrial DNA sequence data. *Copeia* 1996, 763–773.
- Lamar, W.W., Sasa, M., 2003. A new species of hognose pitviper, genus *Porthidium*, from the southwestern Pacific of Costa Rica (Serpentes: Viperidae). *Rev. Biol. Trop.* 51, 797–804.
- Lavine, M., Schervish, M.J., 1999. Bayes factors: What they are and what they are not. *Am. Stat.* 53, 119–122.
- Lemmon, A.R., Moriarty, E.C., 2004. The importance of proper model assumption in Bayesian phylogenetics. *Syst. Biol.* 53, 265–277.
- Leviton, A.E., Gibbs Jr., R.H., Heal, E., Dawson, C.E., 1985. Standards in herpetology and ichthyology. Part I. Standard symbolic codes for institutional resource collections in herpetology and ichthyology. *Copeia* 1985, 805–832.
- López-Luna, M.A., Vogt, R.C., de la Torre-Loranca, M.A., 2000. A new species of montane pitviper from Veracruz, Mexico. *Herpetologica* 55, 382–389.
- Malhotra, A., Thorpe, R.S., 2004. A phylogeny of four mitochondrial gene regions suggests a revised taxonomy for Asian pitvipers (*Trimeresurus* and *Ovophis*). *Mol. Phylogenet. Evol.* 32, 83–100.

- McDiarmid, R.W., Campbell, J.A., Touré, T.A., 1999. Snake Species of the World: A Taxonomic and Geographical Reference, Vol. 1. The Herpetologists' League, Washington, D.C.
- Monclavo, J.M., Drehmel, D., Vilgalys, R., 2000. Variation in modes and rates of evolution in nuclear and mitochondrial ribosomal DNA in the mushroom genus *Aminita* (Agaricales, Basidiomycota): phylogenetic implications. *Mol. Phylogenet. Evol.* 12, 48–63.
- Newton, M.A., Raftery, A.E., 1994. Approximate Bayesian inference with the weighted likelihood bootstrap. *J. R. Stat. Soc. B* 56, 3–48.
- Nicholas, K.B., Nicholas Jr., H.B., 1997. GeneDoc: a tool for editing and annotating multiple sequence alignments. Distributed by the authors at <http://www.cris.com/~Ketchup/genedoc.shtml>.
- 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.
- Parkinson, C.L., 1999. Molecular systematics and biogeographical history of pitvipers as determined by mitochondrial ribosomal DNA sequences. *Copeia* 1999, 576–586.
- Parkinson, C.L., Campbell, J.A., Chippindale, P.T., 2002. Multigene phylogenetic analyses of pitvipers; with comments on the biogeographical history of the group. In: Schuett, G.W., Höggren, M., Douglas, M.E., Greene, H.W. (Eds.), *Biology of the Vipers*. Eagle Mountain Publishing, Salt Lake City, UT, pp. 93–110.

- Posada, D., 2003. Using Modeltest and PAUP* to select a model of nucleotide substitution. In: Baxevanis, A.D., Davidson, D.B., Page, R.D.M., Petsko, L.D., Stein, L.D., Stormo, G.D. (Eds.), *Current protocols in Bioinformatics*. John Wiley & Sons, Inc., Hoboken, NJ, pp. 6.5.1–6.5.14.
- Posada, D., Buckley, T.R., 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.A., 1998. Modeltest: testing the model of DNA substitution. *Bioinformatics* 14, 817–818.
- Posada, D., Crandall, K.A., 2001. Selecting the best-fit model of nucleotide substitution. *Syst. Biol.* 50, 580–601.
- Pupko, T., Huchon, D., Cao, Y., Okada, N., Hasegawa, M., 2002. Combining multiple data sets in a likelihood analysis: which models are the best? *Mol. Biol. Evol.* 19, 2294–2307.
- Rambaut, A., Drummond, A.J., 2003. Tracer, version 1.0.1. Distributed by the authors at <http://evolve.zoo.ox.ac.uk/>.
- Rannala, B., 2002. Identifiability of parameters in MCMC Bayesian inference of phylogeny. *Syst. Biol.* 51, 754–760.
- Reeder, T.W., 2003. A phylogeny of the Australian Sphenomorphus group (Scincidae: Squamata) and the phylogenetic placement of the crocodile skinks (Tribolonotus): Bayesian approaches to assessing congruence and obtaining confidence in maximum likelihood inferred relationships. *Mol. Phylogenet. Evol.* 4, 203–222.

- Rogers, J.S., 2001. Maximum likelihood estimation of phylogenetic trees is consistent when substitution rates vary according to the invariable sites plus gamma distribution. *Syst. Biol.* 50, 713–722.
- Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572–1574.
- Sakamoto, Y., Ishiguro, M., Kitagawa, G., 1986. Akaike Information Criterion Statistics. Springer, NY.
- Savage, J.M., 1966. The origins and history of the Central American herpetofauna. *Copeia* 1966, 719–766.
- Savage, J.M., 1982. The enigma of the Central American herpetofauna: dispersal or vicariance? *Ann. Missouri Bot. Gard.* 69, 464–549.
- Shibata, R., 1978. Selection of the order of an autoregressive model by Akaike's Information Criteria. *Biometrika* 63, 117–126.
- Smith, A.B., 1994. Rooting molecular trees: problems and strategies. *Biol. J. Linnean Soc.* 51, 279–292.
- Sober, E., 2000. instrumentalism, parsimony, and the Akaike framework. *Phil. Sci.* 69, S112–S123.
- Solórzano, A., 1995. Una nueva especie de serpiente venenosa terrestre del género *Porthidium* (Serpentes: Viperidae), del Suroeste de Costa Rica. *Rev. Biol. Trop.* 42, 695–701.
- StatSoft Inc., 1993. Statistica for Windows, release 4.5. StatSoft, Tulsa, OK.
- Stuart, L.C., 1966. The environment of the Central American cold-blooded vertebrate fauna. *Copeia* 1966, 684–699.

- Suchard, M.A., Weiss, R.E., Sinsheimer, J.S., 2001. Bayesian selection of continuous-time Markov chain evolutionary models. *Mol. Biol. Evol.* 18, 1001–1013.
- Swofford, D.L., 2002. PAUP*: Phylogenetic Analysis Using Parsimony (* and Other Methods), version 4.0b10. Sinauer Associates, Sunderland, MA.
- Suzuki, Y., Glazko, G.V., Nei, M., 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *Proc. Nat. Acad. Sci. U.S.A.* 99, 16138–16143.
- Voelker, G., Edwards, S.V., 1998. Can weighting improve bushy trees? Models of cytochrome b evolution and the molecular systematics of pipits and wagtails (Aves: Montacillidae). *Syst. Biol.* 47, 589–603.
- Wager, C., Vaida, F. Kauermann, G., in press. Model selection for P-spline smoothing using Akaike information criteria. *J. Comput. Graph. Stat.*
- Wald, A., 1949. Note on the consistency of the maximum likelihood estimate. *Ann. Math. Stat.* 20, 595–601.
- Wasserman, L., 2000. Bayesian model selection and model averaging. *J. Math. Psychol.* 44, 92–107.
- Werman, S., 1992. Phylogenetic relationships of Central and South American pitvipers of the genus *Bothrops* (sensu lato): cladistic analyses of biochemical and anatomical characters. In: Campbell, J.A., Brodie, E.D., Jr. (Eds.), *Biology of the Pitvipers*. Selva, Tyler, TX. pp. 21–40.
- Werman, S., 2005. Hypotheses on the historical biogeography of bothropoid pitvipers and related genera of the Neotropics. In: Donnelly, M.A., Crother, B.I., Guyer, C., Wake, M.H., White, M.E. (Eds.), *Ecology and Evolution in the Tropics*. University of Chicago Press, Chicago, IL. In Press.

- Weins, J.J., 1998. Combining data sets with different phylogenetic histories. *Syst. Biol.* 47, 568–581.
- Wilgenbusch, J., de Queiroz, K. 2000. Phylogenetic relationships among the phrynosomatid sand lizards inferred from mitochondrial DNA sequences generated by heterogeneous evolutionary processes. *Syst. Biol.* 49, 592–612.
- Wüster, W., da Graca Salomão, M., Quijada-Mascareñas, J.A., Thorpe, R.S., Butantan-British Bothrops Systematics Project, 2002. Origin and evolution of the South American pitviper fauna: evidence from mitochondrial DNA sequence data. In: Schuett, G.W., Höggren, M., Douglas, M.E., Greene, H.W. (Eds.), *Biology of the Vipers*. Eagle Mountain Publishing, Salt Lake City, UT, pp. 111–128.
- Yang, Z., 1996. Maximum likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.* 42, 587–596.

CHAPTER 4 – BAYESIAN MIXED MODELS AND THE PHYLOGENY OF PITVIPERS (VIPERIDAE: SERPENTES)

Introduction

Pitvipers and their contemporary systematics

The venomous snake family Viperidae (asps, moccasins, rattlesnakes, and true vipers) includes about 260 species in four subfamilies: Azemiopinae, Causinae, Crotalinae and Viperinae (McDiarmid et al., 1999). The Crotalinae (pitvipers) is the most species rich of the four subfamilies, containing over 190 species (approximately 75% of viperid species) allocated to 29 genera (Gutberlet and Campbell, 2001; Malhotra and Thorpe, 2004; McDiarmid et al., 1999; Zhang, 1998; Ziegler et al., 2000). Among viperid groups, pitvipers are also the most widely distributed subfamily, with major radiations of species in the Old World and the New World (Campbell and Lamar, 2004; Gloyd and Conant, 1990; McDiarmid et al., 1999).

Pitviper species produce a wide diversity of proteinaceous venom toxins, and many species are capable of inflicting fatal bites to humans (e.g., Russell, 1980). Accordingly, a valid taxonomy and a robust understanding of relationships among these venomous species are important for systematics, in addition to the fields of medicine, pharmacology, and toxicology (e.g., > 3000 citations on PubMed [National Center for Biotechnical Information] for “pit viper venom”). The phylogeny and taxonomy of this group has received substantial research attention

that has led to many revisions to make taxonomy consistent with estimates of phylogeny (see reviews in Campbell and Lamar, 2004; Gutberlet and Harvey, 2004; Malhotra and Thorpe, 2004; Parkinson et al., 2002). Of the 29 generic names in use, 19 have been recognized in the last three decades (Burger, 1971; Campbell and Lamar, 1989; Campbell and Lamar, 1992; Gutberlet and Campbell, 2001; Hoge and Romano-Hoge, 1981; Hoge and Romano-Hoge, 1983; Werman, 1992; Zhang, 1998; Ziegler et al., 2000; Malhotra and Thorpe, 2004).

The deepest phylogenetic divergences among pitvipers have yet to be resolved with strong support. Current evidence indicates either: 1) a clade containing *Hypnale*, *Calloselasma*, *Deinagkistrodon*, and *Tropidolaemus* as the sister group to the remaining pitvipers (Malhotra and Thorpe, 2004; Parkinson et al., 2002) or, 2) a clade comprised of *Deinagkistrodon* and *Tropidolaemus* as the sister group to the remaining pitvipers (Knight et al., 1992; Parkinson, 1999; Parkinson et al., 2002; Vidal et al., 1998).

The Old World genus *Trimeresurus* (*sensu lato*; e.g., Burger, 1971) was found to be polyphyletic by a number of studies (e.g., Malhotra and Thorpe, 2000; Parkinson, 1999), and was subsequently dissected into a total of 11 genera, including: *Protobothrops* (Hoge and Romano-Hoge, 1983), *Ovophis* (Burger, 1971; Hoge and Romano-Hoge, 1981), *Zhaoermia* (described as *Ermia* by Zhang, 1993, changed to *Zhaoermia* by Gumprecht and Tillac, 2004), *Triceratolepidophis* (Ziegler et al., 2000), and *Cryptelytrops*, *Garthius*, *Himalayophis*, *Parias*, *Peltopelor*, *Popeia*, and *Viridovipera* (Malhotra and Thorpe, 2004). Despite these changes, recent pitviper phylogenetic estimates suggest that *Ovophis* and *Trimeresurus* (*sensu stricto*) remain polyphyletic (e.g., Malhotra and Thorpe, 2000, 2004; Parkinson et al., 2002).

Kraus et al. (1996) hypothesized that New World pitvipers are monophyletic, and recent molecular studies have shown increasing support for this clade (e.g., Malhotra and Thorpe, 2004; Parkinson, 1999; Parkinson et al., 2002). This contradicts all morphology-based phylogenetic hypotheses (not constraining New World pitviper monophyly) which find a polyphyletic origin of New World pitvipers (Brattstrom, 1964; Burger, 1971; Gloyd and Conant, 1990). Currently there are twelve genera of New World pitvipers recognized (Campbell and Lamar, 2004) and the relationships among these remain poorly understood and inconsistent across studies. Certain molecular studies (Parkinson, 1999; Parkinson et al., 2002), and the morphological data set of Gutberlet and Harvey (2002), support the earliest New World divergence as being between a temperate North American clade and a Neotropical clade. Within this temperate clade, rattlesnakes (*Crotalus* and *Sistrurus*) have been consistently inferred to be monophyletic, and to be the sister group to a clade containing the cantils/copperheads/moccasins (*Agkistrodon*; Knight et al., 1992; Murphy et al., 2002; Parkinson, 1999; Parkinson et al., 2002; Vidal et al., 1999).

Few relationships among the tropical New World genera are supported by multiple studies, although several notable relationships have been repeatedly identified. A primarily South American bothropoid clade, with *Bothrocophias* inferred as the sister group to *Bothrops* plus *Bothriopsis*, has been found by both morphological and molecular-based studies (Castoe et al., 2005; Gutberlet and Campbell, 2001; Parkinson et al., 2002). Results of several studies have agreed on the paraphyly of *Bothrops* (*sensu stricto*) with respect to *Bothriopsis* (Gutberlet and Campbell, 2001; Knight et al., 1992; Parkinson, 1999; Parkinson et al., 2002; Salomão et al., 1997, 1999; Vidal et al., 1997, 1999; Wüster et al., 2002). Although studies incorporating morphological data disagree (Gutberlet and Harvey, 2002; Werman, 1992), several molecular

studies have inferred a clade comprising the primarily Middle American genera *Porthidium*, *Atropoides*, and *Cerrophidion* (Castoe et al., 2003, 2005; Parkinson, 1999; Parkinson et al., 2002).

Challenges and strategies for resolving pitviper phylogeny

Despite the efforts of numerous authors, phylogenetic relationships within the subfamily Crotalinae remain controversial, particularly at the intergeneric level (e.g. Gutberlet and Harvey, 2004; Malhotra and Thorpe, 2004; Parkinson et al., 2002). Three issues have likely played major roles in the generation of inconsistent conclusions or poor resolution across studies: 1) Only four (Kraus et al., 1996; Malhotra and Thorpe, 2004; Parkinson, 1999; Parkinson et al., 2002) of nearly twenty inter-generic molecular-based studies have included most of the proposed crotaline genera. No study has included a large representation of both Old World and New World genera and species. Limited taxonomic sampling can be problematic in phylogenetic analyses (Hillis, 1998; Poe, 1998; Poe and Swofford, 1999; Salisbury and Kim, 2001), and when only a few representatives of a diverse group are sampled, the resulting phylogenies may represent sampling artifacts (e.g., due to long-branch attraction) rather than accurate and objective phylogenetic reconstructions (Graybeal, 1998; Hillis, 1996, 1998). 2) Many studies (particularly earlier studies) employed only a small gene region to infer inter-generic relationships providing few informative characters. 3) Most DNA-based studies to date have analyzed relationships based on mitochondrial gene sequences. Mitochondrial-based phylogenetics has proven very successful largely because of the rapid rate of sequence evolution characteristic of this genome (Brown et al., 1979; Caccone et al., 1997; Vidal et al., 1999), yielding large proportions of potentially

informative (variable) sites. This strength becomes problematic, however, because the probability of continued sequence turnover at sites increases with phylogeny depth. Confident estimation of deeper relationships becomes increasingly difficult as the phylogenetic signal-to-noise ratio becomes unfavorable. This problematic feature of molecular evolution, combined with limited taxon sampling and limited character sampling has synergistically weighed against previous attempts to reconstruct crotaline phylogeny.

Here, we use DNA sequences from four mitochondrial gene regions sampled from a large array of pitviper taxa (including 28 of 29 genera) to estimate pitviper phylogeny. Our extensive taxonomic sampling design targets difficulties that limited taxon sampling may impose on recovering accurate phylogenetic estimates. Our sampling of gene regions (mitochondrial genes), however, remains potentially susceptible to problems associated with the high rate of sequence evolution characteristic of mitochondrial genes, leading to excessive homoplasy and obscured phylogenetic signal at deeper nodes. We target this latter problem analytically through complex-partitioned modeling of nucleotide evolution during phylogenetic analyses.

Model-based phylogenetic methods (including Bayesian phylogenetic techniques) are particularly useful for reconstructing phylogenies from divergent sequences because they incorporate probabilistic models of DNA substitution that should be less likely to be misled by complexities of DNA evolution (Huelsenbeck, 1995; Huelsenbeck and Crandall, 1997). Multigene datasets, as in this study, may contain partitions (e.g., multiple genes, rRNA vs. protein coding genes, codon positions, types of RNA secondary structures) that evolve under different models (or patterns) of evolution. In these cases, using a single likelihood model for the entire dataset forces a compromise in parameter estimates that must (under a single model) be

averaged over the entire dataset. This compromise may lead to systematic error and mislead phylogenetic conclusions (Brandley et al., 2005; Huelsenbeck and Rannala, 2004; Lemmon and Moriarty, 2004; Reeder, 2003; Wilgenbusch and de Queiroz, 2000). Important for our phylogenetic problem, a single compromise model may not capture the range of complexities in nucleotide substitution across the entire mixed dataset. In turn, this compromise may result in increased error identifying substitutions with high likelihoods of change (and homoplasy), versus substitutions with low likelihoods of change (with higher probabilities of containing phylogenetic signal). This type of modeling compromise may also increase the error in reconstructing ancestral states. This problematic compromise may be avoided by allocating independent models of nucleotide evolution to partitions of a heterogeneous dataset (e.g., Nylander et al., 2004; Pagel and Meade, 2004; Yang, 1996).

Model choice may affect both phylogenetic topology (e.g., Huelsenbeck, 1995; Huelsenbeck, 1997; Sullivan and Swofford, 2001) and posterior probability estimation (e.g., Buckley, 2002; Castoe et al., 2004; Suzuki et al., 2002; Erixon et al., 2003). Complex partitioned models may have important effects in the resolution of deeper nodes, a majority of which receive increased support under complex models (Brandley et al., 2005; Castoe et al., 2004, 2005). Complex models appear to be more effective at estimating patterns of molecular evolution when sequences are highly divergent and phylogenetic signal is otherwise obscured by multiple substitutions (Brandley et al., 2005; Castoe et al., 2005; see also Huelsenbeck and Rannala, 2004; Lemmon and Moriarty, 2004).

In this study we combine taxon sampling and analytical strategies to estimate a robust hypothesis for the phylogeny of pitvipers. Along with maximum parsimony analyses, we

implement complex partitioned models of nucleotide evolution (in a Bayesian MCMC framework) to help counter problems likely to have biased previous analyses of pitviper phylogeny. We compare phylogeny and parameter estimates between simple and complex models to identify the impacts that complex models have on phylogenetic inference and on modeling patterns of nucleotide evolution. Based on our estimates of pitviper phylogeny we evaluate the current genus-level taxonomy and discuss the relevance of our estimates to previous phylogenetic and taxonomic hypotheses.

Materials and methods

Taxon sampling

A total of 167 terminals were included in this study. We base our taxonomic assignment of species and genera on McDiarmid et al. (1999), Malhotra and Thorpe (2004) and Campbell and Lamar (2004), unless specifically noted (see Table 11). The ingroup, members of the subfamily Crotalinae (pitvipers), were represented by 157 terminals comprising 116 currently recognized species, including 45 Old World and 71 New World species (Table 11). Collectively, our sampling included representatives of 28 of 29 genera, excluding only the monotypic Old World genus *Peltopelor*. Outgroup taxa including representatives of the three other subfamilies

Table 11. Taxon sampling with voucher information, locality data, and Genbank accession numbers for gene fragments. An asterisk is used to indicate novel sequences generated in this study.

Taxon and sample identifier	Voucher	Locality	Genbank numbers (12s, 16s, cyt-b, ND4)
<i>Causus rhombeatus</i>		Africa	<u>DO305409*</u> , <u>DO305432*</u> , <u>DO305455*</u> , <u>DO305473*</u>
<i>Causus resimus</i>	Moody 515	Africa	<u>AY223649</u> , <u>AY223662</u> , <u>AY223555</u> , <u>AY223616</u>
<i>Causus defilippi</i>	CLP154	Tanzania	<u>AF057186</u> , <u>AF057233</u> , <u>AY223556</u> , <u>AY223617</u>
<i>Atheris ceratophora</i>			<u>DQ305410*</u> , <u>DQ305433*</u> , <u>DQ305456*</u> , <u>DQ305474*</u>
<i>Atheris nitchei</i>	CAS201653	Tanzania	<u>AY223650</u> , <u>AY223663</u> , <u>AY223557</u> , <u>AY223618</u>
<i>Bitis nasicornis</i>	CAS207874		<u>DQ305411*</u> , <u>DQ305434*</u> , <u>DQ305457*</u> , <u>DQ305475*</u>
<i>Bitis peringueyi</i>	CAS193863		<u>DQ305412*</u> , <u>DQ305435*</u> , <u>DQ305458*</u> , <u>DQ305476*</u>
<i>Bitis arietans</i>		Togo	<u>AF057185</u> , <u>AF57232</u> , <u>AY223558</u> , <u>AY223619</u>
<i>Daboia russelii</i>	CAS205253		<u>DO305413*</u> , <u>DO305436*</u> , <u>DO305459*</u> , <u>DO305477*</u>
<i>Azemiops feae</i>	CLP-157	China	<u>AF057187</u> , <u>AF057234</u> , <u>AY223559</u> , <u>AFU41865</u>
<i>Calloselasma rhodostoma</i>	UTA-R22247		<u>AF057190</u> , <u>AF057237</u> , <u>AY223562</u> , <u>U41878</u>
<i>Cryptelytrops albolabris</i> (A165)	AM A165	Thailand, Loei Prov.	<u>AF517169</u> , <u>AF517182</u> , <u>AF517185</u> , <u>AF517214</u>
<i>Cryptelytrops albolabris</i> (A229)	AM A229	Thailand, Pha Yao Prov.	<u>AY059544</u> , <u>AY059560</u> , <u>AY059566</u> , <u>AY059583</u>
<i>Cryptelytrops albolabris</i> (B22)	AM B22	Thailand, Nonthaburi	<u>AF517165</u> , <u>AF517178</u> , <u>AF517189</u> , <u>AF517221</u>
<i>Cryptelytrops albolabris</i> (B47)	AM B47	Thailand, Phetburi Prov.	<u>AF517160</u> , <u>AF517173</u> , <u>AF517187</u> , <u>AF517216</u>
<i>Cryptelytrops albolabris</i> (B6)	AM B6	Indonesia, Java, Cilacap	<u>AF517158</u> , <u>AF517171</u> , <u>AF517186</u> , <u>AF517213</u>
<i>Cryptelytrops albolabris</i> (MCZR)	MCZR-177966	Hong Kong, Port Shelter Is., Yim Tin Tsi	<u>AF057195</u> , <u>AF057242</u> , <u>AY223567</u> , <u>U41890</u>
<i>Cryptelytrops andersonii</i>	AM A77	India, Andaman Is.	<u>AY352801</u> , <u>AY352740</u> , <u>AF171922</u> , <u>AY352835</u>
<i>Cryptelytrops cantori</i> (A85)	AM A85	India, Nicobar Is.	<u>AY352802</u> , <u>AY352741</u> , <u>AF171889</u> , <u>AY352836</u>
<i>Cryptelytrops cantori</i> (CLP)		India, Nicobar Is., Kamurta	<u>AF057196</u> , <u>AF057243</u> , <u>-AY223568</u> , <u>U41891</u>
<i>Cryptelytrops erythrurus</i> (A209)	AM A209	Myanmar, Rangoon	<u>AF517161</u> , <u>AF517174</u> , <u>AF171900</u> , <u>AF517217</u>
<i>Cryptelytrops erythrurus</i> (B220)	AM B220	Bangladesh, Chittagong	<u>AY352800</u> , <u>AY352739</u> , <u>AY352768</u> , <u>AY352834</u>
<i>Cryptelytrops insularis</i> (A109)	AM A109	Indonesia, Java	<u>AY352799</u> , <u>AY352738</u> , <u>AY352767</u> , <u>AY352833</u>
<i>Cryptelytrops insularis</i> (B7)	AM B7	Indonesia, Timor	<u>AY059534</u> , <u>AY059550</u> , <u>AY059568</u> , <u>AY059586</u>
<i>Cryptelytrops macrops</i>	AM B27	Thailand, Bangkok	<u>AF517163</u> , <u>AF517176</u> , <u>AF517184</u> , <u>AF517219</u>
<i>Cryptelytrops purpureomaculatus</i> (A83)	AM A83	Thailand, Satun Prov.	<u>AF517162</u> , <u>AF517175</u> , <u>AF517188</u> , <u>AF517218</u>
<i>Cryptelytrops purpureomaculatus</i> (B418)	CAS212246	Myanmar, Ayeyarwade	<u>AY352807</u> , <u>AY352746</u> , <u>AY352772</u> , <u>AY352746</u>
<i>Cryptelytrops septentrionalis</i> (A100)	AM A100	Nepal, Mahattari Dist.	<u>AY059543</u> , <u>AY059559</u> , <u>AF171909</u> , <u>AY059592</u>
<i>Cryptelytrops septentrionalis</i> (B487)	AM B487	Nepal, Kathmandu Dist.	<u>AY352784</u> , <u>AY352724</u> , <u>AY352755</u> , <u>AY352818</u>
<i>Cryptelytrops venustus</i>	AM A241	Thailand, Thammarat Prov.	<u>AY293931</u> , <u>AY352723</u> , <u>AF171914</u> , <u>AY293930</u>
<i>Deinagkistrodon acutus</i>	CLP-28	China	<u>AF057188</u> , <u>AF057235</u> , <u>AY223560</u> , <u>U41883</u>
<i>Garthius chaseni</i>	AM B306	Malaysia, Sabah	<u>AY352791</u> , <u>AY352729</u> , <u>AY352760</u> , <u>AY352825</u>

Taxon and sample identifier	Voucher	Locality	Genbank numbers (12s, 16s, cyt-b, ND4)
<i>Gloydus halys</i>		Kazakhstan	AF057191 , AF057238 , AY223564 , AY223621
<i>Gloydus shedaoensis</i>	ROM-20468	China, Liaoning	AF057194 , AF057241 , AY223566 , AY223623
<i>Gloydus strauschi</i>	ROM-20473	China, Jilin, Waqie Sichuan	AF057192 , AF057239 , AY223563 , AY223620
<i>Gloydus ussuriensis</i>	ROM-20452	China, Jilin, Kouqian	AF057193 , AF057240 , AY223565 , AY223622
<i>Himalayophis tibetanus</i>	ZMB-65641	Nepal, Helambu Prov.	AY352776 , AY352715 , AY352749 , AY352810
<i>Hypnale hypnale</i>	CLP-164	Sri Lanka, Columbo	AF057189 , AF057236 , AY223561 , U41884
<i>Ovophis monticola</i> (A87)	AM A87	Taiwan	AY059545 , AY059561 , AF171907 , AY059582
<i>Ovophis monticola</i> (JBS)	CAS215050	China, Yunnan Prov., Nu Jiang Prefecture	DQ305416* , DQ305439* , DQ305462* , DQ305480*
<i>Ovophis monticola</i> (MAK)	NTNUB200800		DQ305417* , DQ305440* , DQ305463* , DQ305481*
<i>Ovophis okinavensis</i> (162)	CLP-162	Japan, Okinawa	AF057199 , AF057246 , AY223573 , U41895
<i>Ovophis okinavensis</i> (FK)	FK		DQ305418* , DQ305441* , DQ305464* , U41895
<i>Parias flavomaculatus</i> (B289)	AM B289	Philippines, Batan Is.	AY371756 , AY371795 , AY371831 , AY371858
<i>Parias flavomaculatus</i> (B3)	AM B3	Philippines, Luzon	AY059535 , AY059551 , AF171916 , AY059584
<i>Parias flavomaculatus</i> (B4)	AM B4	Philippines, Mindanao	AY352796 , AY352734 , AY352764 , AY352830
<i>Parias hageni</i> (B33)	AM B33	Thailand, Songkhla Prov.	AY059536 , AY059552 , AY059567 , AY059585
<i>Parias hageni</i> (B364)	AM B364	Indonesia, Sumatra, Bengkulu Prov.	AY371763 , AY371790 , AY371825 , AY371863
<i>Parias malcomi</i>	AM B349	Malaysia, Sabah	AY371757 , AY371786 , AY371832 , AY371861
<i>Parias schultzei</i>	AM B210	Philippines, Palawan	AY352785 , AY352725 , AY352756 , AY352819
<i>Parias sumatranus</i> (B347)	AM B347	Malaysia, Sabah	AY371759 , AY371788 , AY371823 , AY371859
<i>Parias sumatranus</i> (B367)	AM B367	Indonesia, Sumatra, Bengkulu Prov.	AY371765 , AY371791 , AY371824 , AY371864
<i>Popeia popeiorum</i> (A203)	AM A203	Thailand, Thammarat Prov.	AY059537 , AY059553 , AY371796 , AY059588
<i>Popeia popeiorum</i> (B196)	FMNH-258950	Laos, Phongsaly Prov.	AY059538 , AY059554 , AY059571 , AY059590
<i>Popeia popeiorum</i> (B246)	AM B246	Malaysia, Selangor	AY059540 , AY059556 , AY059570 , AY059589
<i>Popeia popeiorum</i> (B34)	AM B34	Thailand, Phetburi Prov.	AY059542 , AY059558 , AY059572 , AY059591
<i>Protobothrops cornutus</i>	ZFMK75067	Vietnam, Phong Nha- Ke NP	AY294272 , AY294262 , AY294276 , AY294267
<i>Protobothrops elegans</i>	UMMZ-199970	Japan, Ryuku Is., Ishigaki	AF057201 , AF057248 , AY223575 , U41893
<i>Protobothrops falvoviridis</i>	UMMZ-199973	Japan, Ryuku Is., Tokunoshima	AF057200 , AF057247 , AY223574 , U41894
<i>Protobothrops jerdonii</i>	CAS215051	China, Nu Jiang, Yunnan	AY294278 , AY294269 , AY294274 , AY294264
<i>Protobothrops mucrosquamatus</i> (2717)	ROM-2717	Vietnam	AY223653 , AY223666 , AY223577 , AY223629
<i>Protobothrops mucrosquamatus</i> (B106)	AM B106	Vietnam, Vin Phuc Prov.	AY294280 , AY294271 , AY294275 , AY294266
<i>Protobothrops tokarensis</i>	FK-1997	Japan, Ryuku Is., Takarajima	AF057202 , AF057249 , AY223576 , AY223628
<i>Triceratolepidophis sieversorum</i> (B162)	AM B162	Vietnam	AY352782 , AY352721 , AY352753 , AY352816
<i>Triceratolepidophis sieversorum</i> (CLP)	ZFMK 75066	Vietnam, Phong Nha- Quang Ping Province	DQ305414* , DQ305437* , DQ305460* , DQ305478*

Taxon and sample identifier	Voucher	Locality	Genbank numbers (12s, 16s, cyt-b, ND4)
<i>Trimeresurus borneensis</i>	AM B301	Malaysia, Sabah	AY352783 , AY352722 , AY352754 , AY352817
<i>Trimeresurus gracilis</i> (A86)	AM A86	Taiwan	AY352789 , AY352728 , AF171913 , AY352823
<i>Trimeresurus gracilis</i> (NTUB)	NTNUB 200515		DQ305415* , DQ305438* , DQ305460* , DQ305478*
<i>Trimeresurus gramineus</i> (A220)	AM A220	India, Tamil Nadu	AY352793 , AY352731 , AY352761 , AY352827
<i>Trimeresurus gramineus</i> (B261)	AM B261	India, Maharashtra	AY352794 , AY352732 , AY352762 , AY352828
<i>Trimeresurus malabaricus</i> (A218)	AM A218	India, Tamil Nadu	AY059548 , AY059564 , AY059569 , AY059587
<i>Trimeresurus malabaricus</i> (B260)	AM B260	India, Maharashtra	AY352795 , AY352733 , AY352763 , AY352829
<i>Trimeresurus puniceus</i>	AM B213	Indonesia	AF517164 , AF517177 , AF517192 , AF517220
<i>Trimeresurus trigonocephalus</i>	AM A58	Sri Lanka, Balangoda	AY059549 , AY059565 , AF171890 , AY059597
<i>Tropidolaemus wagleri</i> (B132)	AM B132	Malaysia, Perak	AF517167 , AF517180 , AF517191 , AF517223
<i>Tropidolaemus wagleri</i> (B311)	AM B311	Malaysia, Sabah	AY352788 , AY352727 , AY352759 , AY352822
<i>Tropidolaemus wagleri</i> (I41)	CLP-141	Indonesia, West Kalimantan	AF057198 , AF057245 , AY223571 , AY223625
<i>Viridovipera gumprechtii</i> (A164)	AM A164	Thailand, Loei Prov.	AF517168 , AF517181 , AY352766 , AF157224
<i>Viridovipera gumprechtii</i> (B15)	NMNS-3113	China, Yunnan Prov.	AY352798 , AY352736 , AY3521487 , AY352736
<i>Viridovipera gumprechtii</i> (B174)	FMNH-255579	Vietnam, Nghe An Prov.	AY059547 , AY059563 , AY059573 , AY059595
<i>Viridovipera medoensis</i>	CAS 221528	Myanmar, Kachin	AY352797 , AY352735 , AY352765 , AY352831
<i>Viridovipera stejnegeri</i> (A160)	AM A160	Taiwan, Taipei	AY059539 , AY059555 , AF171896 , AY059593
<i>Viridovipera stejnegeri</i> (A222)	NMNS-3651	China, Fujian Prov.	AY059541 , AY059557 , AF277677 , AY059594
<i>Viridovipera stejnegeri</i> (UMMZ)	UMMZ-190532	Taiwan, Taipei	AF057197 , AF057244 , AY223570 , U41892
<i>Viridovipera vogeli</i> (B97)	AM B97	Thailand, Ratchasima Prov.	AY059546 , AY059562 , AY059574 , AY059596
<i>Viridovipera vogeli</i>	ROM-7234		AY223651 , AY223664 , AY223569 , AY223624
<i>Zhaoermia mangshanensis</i>	AM B300	China, Hunan Prov.	AY352787 , AY352726 , AY352758 , AY352821
<i>Agkistrodon bilineatus</i>	WWL	Costa Rica, Guanacaste	AF156593 , AF156572 , AY223613 , AF156585
<i>Agkistrodon contortrix</i>	Moody 338	USA, Ohio, Athens Co.	AF057229 , AF057276 , AY223612 , AF156576
<i>Agkistrodon piscivorus</i>	CLP-30	USA, South Carolina	AF057231 , AF057278 , AY223615 , AF156578
<i>Agkistrodon taylori</i>	CLP-140	Mexico, Tamaulipas	AF057230 , AF057230 , AY223614 , AF156580
<i>Atropoides mexicanus</i>	CLP-168	Costa Rica	AF057207 , AF057254 , AY223584 , U41871
<i>Atropoides nummifer</i>	ENS-10515	Mexico, Puebla, San Andres Tziaulan	DQ305422* , DQ305445* , DQ061195 , DQ061220
<i>Atropoides occiduus</i>	UTA-R29680	Guatemala, Escuintla	DQ305423* , DQ305446* , AY220315 , AY220338
<i>Atropoides olmec</i>	JAC-16021	Mexico, Veracruz	AY223656 , AY223669 , AY220321 , AY220344
<i>Atropoides picadoi</i>	CLP-45	Costa Rica, Alajuela	AF057208 , AF057255 , AY223593 , U41872
<i>Bothriechis aurifer</i>	UTA-R35031	Guatemala	DQ305425* , DQ305448* , DQ305466* , DQ305483*
<i>Bothriechis bicolor</i>	UTA-R34156		DQ305426* , DQ305449* , DQ305467* , DQ305484*
<i>Bothriechis lateralis</i>	MZUCR-11155	Costa Rica, Acosta	AF057211 , AF057258 , AY223588 , U41873

Taxon and sample identifier	Voucher	Locality	Genbank numbers (12s, 16s, cyt-b, ND4)
<i>Bothriechis marchi</i>	UTA-R52959	Guatemala: Zacapa: Cerro del Mono	<u>DO305428*, DQ305451*, DO305469*, DO305486*</u>
<i>Bothriechis nigroviridis</i>	MZUCR-11151	Costa Rica, San Gerondo de Dota	<u>AF057212, AF057259, AY223589, AY223635</u>
<i>Bothriechis rowleyi</i>	JAC 13295	Mexico: Cerro Baúl	<u>DQ305427*, DQ305450*, DQ305468*, DQ305485*</u>
<i>Bothriechis schlegelii</i>	MZUCR-11149	Costa Rica, Cariblanco de Sarapiquí	<u>AF057213, AF057260, AY223590, AY223636</u>
<i>Bothriechis superciliaris</i>		San Vito, Costa Rica	<u>DQ305429*, DQ305452*, DQ305470*, DQ305487*</u>
<i>Bothriechis thalassinus</i>	UTA-R52958	Guatemala: Zacapa	<u>DQ305424*, DQ305447*, DQ305465*, DQ305482*</u>
<i>Bothriopsis bilineata</i>		Colombia, Leticia	<u>AF057214, AF057261, AY223591, U41875,</u>
<i>Bothriopsis chloromelas</i>	LSUMZ 41037	Peru, Pasco Dept.	<u>DQ305430*, DQ305453*, DQ305471*, DQ305488*</u>
<i>Bothriopsis taeniata</i>		Suriname	<u>AF057215, AF057262, AY223592, AY223637</u>
<i>Bothrocophias hyoprora</i>		Colombia, Leticia	<u>AF057206, AF057253, AY223593, U41886</u>
<i>Bothrocophias microphthalmus</i>	LSUMZ H-9372	Peru, Pasco Dept.	<u>AY223657, AY223670, AY223594, AY223638</u>
<i>Bothrops alternatus</i>	DLP-2879		<u>AY223660, AY223673, AY223601, AY223642</u>
<i>Bothrops ammodytoides</i>	MVZ-223514	Argentina, Neuguen	<u>AY223658, AY223671, AY223595, AY223639</u>
<i>Bothrops asper</i>	MZUCR-11152	Costa Rica	<u>AF057218, AF057265, AY223599, U41876</u>
<i>Bothrops atrox</i>	WWW-743		<u>AY223659, AY223672, AY223598, AY223641</u>
<i>Bothrops cotiara</i>	WWW	Brazil	<u>AF057217, AF057264, AY223597, AY223640</u>
<i>Bothrops diporus</i>	PT3404	Depto. Castro Barros, Prov. La Rioja, Argentina	<u>DQ305431*, DQ305454*, DQ305472*, DQ305489*</u>
<i>Bothrops erythromelas</i>	RG-829	Brazil, Algóóas, Piranhas	<u>AF057219, AF057266, -AY223600, U41877</u>
<i>Bothrops insularis</i>	WWW	Brazil, São Palo, Iiha Queimada Grande	<u>AF057216, AF057263, AY223596, AF188705</u>
<i>Bothrops jararacussu</i>	DPL-104		<u>AY223661, AY223674, AY223602, AY223643</u>
<i>Cerrophidion godman</i> (CR)	MZUCR-11153	Costa Rica, San Jose	<u>AF057203, AF057250, AY223578, U41879</u>
<i>Cerrophidion godmani</i> (GM)	UTAR-40008	Guatemala: Baja Verapaz	<u>DQ305419*, DQ305442*, AY220348, AY220325</u>
<i>Cerrophidion petalcalensis</i>	ENS-10528	Mexico, Veracruz, Orizaba	<u>DQ305420*, DQ305443*, DQ061202, DQ061227</u>
<i>Crotalus adamanteus</i>	CLP-4	USA, Florida, St. Johns Co.	<u>AF057222, AF057269, AY223605, U41880</u>
<i>Crotalus aquilus</i>	ROM-18117	Mexico, San Luis Potosi	<u>AF259232, AF259125, AF259162, ----</u>
<i>Crotalus atrox</i>	CLP-64	USA, Texas, Jeff Davis Co.	<u>AF0572225, AF057272, AY223608, AY223646</u>
<i>Crotalus basiliscus</i>	ROM-18188	Mexico, Nyarit	<u>AF259244, AF259136, AF259174, ----</u>
<i>Crotalus catalinensis</i>	ROM-18250, BYU-34641-42	Mexico, Baja California Sur, Isla Santa Catalina	<u>AF259259, AF259151, AF259189, ----</u>
<i>Crotalus cerastes</i>	ROM-FC-20099, ROM-19745	USA, California, Riverside Co.	<u>AF259235, AF259128, AF259165, ----</u>
<i>Crotalus durissus</i>	ROM-18138	Venezuala	<u>AF259248, AF259140, AF259178, ----</u>

Taxon and sample identifier	Voucher	Locality	Genbank numbers (12s, 16s, cyt-b, ND4)
<i>Crotalus enyo</i>	ROM-FC411, ROM13648	Mexico, Baja California Sur	AF259245 , AF259137 , AF259175 , ----
¹ <i>Crotalus</i> “ <i>exsul</i> ”	BYU-34753-54	Mexico, Baja California, Isla de Cedros	AF259260 , AF259152 , AF259190 , ----
<i>Crotalus horridus</i> (AR)	UTA-R14697	USA, Arkansas	AF259252 , AF259144 , AF259182 , ----
<i>Crotalus horridus</i> (NY)	ROM-18132-33	USA, New York	AF259251 , AF259143 , AF259181 , ----
<i>Crotalus intermedius</i>	ROM-FC223, ROM-18164	Mexico, Veracruz	AF259238 , AF259131 , AF2589205 , ----
<i>Crotalus lepidus</i>	ROM-18128	Mexico, Chihuahua	AF259230 , AF259123 , AF259160 , ----
<i>Crotalus mitchelli</i>	ROM-18178	USA, California, Imperial Co.	AF259250 , AF259142 , AF259180 , ----
<i>Crotalus molossus</i>	CLP-66	USA, Texas, El Paso Co.	AF057224 , AF057271 , AY223607 , AY223645
<i>Crotalus oreganus</i>	ROM-19656	USA, California, Los Angeles Co.	AF259253 , AF259145 , AF259183 , ----
<i>Crotalus polystictus</i>	ROM-FC263, ROM-18139	Mexico, Distrito Federal	AF259236 , AF259129 , AF259166 , ----
<i>Crotalus pricei</i>	ROM-FC2144, ROM-18158	Mexico, Nuevo Leon	AF259237 , AF259130 , AF259167 , ----
<i>Crotalus pusillus</i>	ROM-FC271	Mexico, Michoacan	AF259229 , AF259122 , AF259159 , ----
<i>Crotalus ravus</i>	UTA-live	Mexico, Puebla, Zapotitlán	AF057226 , AF057273 , AY223609 , AY223647
<i>Crotalus ruber</i>	ROM-18197-98, ROM18207	USA, California, Riverside CO.	AF259261 , AF259153 , AF259191
<i>Crotalus scutulatus</i>	ROM-18210, ROM-18218	USA, Arizona, Mojave Co.	AF259254 , AF259146 , AF259184 , ----
<i>Crotalus tigris</i>	CLP169	USA, Arizona, Pima Co.	AF057223 , AF057270 , AY223606 , AF156574
<i>Crotalus tortugensis</i>	ROM-18192, ROM-18195	Mexico, Baja California Sur, Isla Tortuga	AF259257 , AF259149 , AF259187 , ----
<i>Crotalus transversus</i>	KZ-shed skin	Mexico	AF259239 , AF259206 , AF259169 , ----
<i>Crotalus triseriatus</i> (LG)	ROM-18114	Mexico, Distrito Federal, Llano Grande	AF259231 , AF259124 , AF259161 , ----
<i>Crotalus triseriatus</i> (TO)	ROM-18121	Mexico, Distrito Federal, Toluca	AF259233 , AF259126 , AF259163 , ----
<i>Crotalus triseriatus</i> (XO)	ROM-18120	Mexico, Distrito Federal, Xochochomiko	AF259234 , AF259127 , AF259164 , ----
<i>Crotalus unicolor</i>	ROM-18150	Aruba Island	AF259246 , AF259138 , AF259176 , ----
² <i>Crotalus</i> “ <i>vegrandis</i> ”	ROM-18261	Venezuela	AF259247 , AF259139 , AF259177 , ----
<i>Crotalus willardi</i> (2575)	HWG-2575	USA, Arizona, Coshise Co.	AF259242 , AF259134 , AF259172 , ----
<i>Crotalus willardi</i> (413)	ROM-FC363, KZ-413	USA, Arizona, Santa Cruz Co.	AF259241 , AF259133 , AF259171 , ----
<i>Crotalus willardi</i> (ROM)	ROM-18183, ROM-18185	Mexico, Sonora	AF259240 , AF259132 , AF259170 , ----
<i>Lachesis muta</i>	Cadle 135	Peru	AF057221 , AF057268 , AY223604 , AY223644
<i>Lachesis stenophrys</i>		Costa Rica, Limón	AF057220 , AF057267 , AY223603 , U41885
<i>Ophryacus melanurus</i>	UTA-R34605	Mexico	AF057210 , AF057257 , AY223587 , AY223634
<i>Ophryacus undulatus</i>	CLP-73	Mexico	AF057209 , AF057256 , AY223586 , AY223633
<i>Porthidium arcose</i>	WWW-750	Ecuador	AY223655 , AY223668 , AY223582 , AY223631
<i>Porthidium dummi</i>	ENS-9705	Mexico, Oaxaca	AY223654 , AY223667 , AY223581 , AY223630

Taxon and sample identifier	Voucher	Locality	Genbank numbers (12s, 16s, cyt-b, ND4)
<i>Porthidium nasutum</i>	MZUCR-11150	Costa Rica	AF057204 , AF057251 , AY223579 , U41887
<i>Porthidium ophryomegas</i>	UMMZ-210276	Costa Rica, Guanacaste Prov.	AF057205 , AF057252 , AY223580 , U41888
<i>Porthidium porrasii</i>	MSM	Costa Rica, Puntarenas	DQ305421* , DQ305444* , DQ061214 , DQ061239
<i>Sistrurus catenatus</i>	Moody-502	USA, Texas, Haskell Co.	AF057227 , AF057274 , AY223610 , AY223648
<i>Sistrurus miliaris</i>	UTA-live	USA, Florida, Lee Co.	AF057228 , AF057275 , AY223611 , U41889

of viperids (Causinae, Viperinae, Azemiopinae) were also included so that the monophyly of the Crotalinae could be assessed. We rooted phylogenies with members of the genus *Causus* based on previous suggestions that the Causinae is the sister group to all other viperids (McDiarmid et al., 1999).

DNA sequencing and sequence alignment

A majority of sequences used in this study have been published previously (Castoe et al., 2003, 2005; Kraus et al., 1996; Malhotra and Thorpe, 2004; Murphy et al., 2002; Parkinson, 1999; Parkinson et al., 1997, 2000, 2002). Laboratory methods for novel sequences generated for this study are provided below. Genomic DNA was isolated from tissue samples (liver or skin preserved in ethanol) using the Qiagen DNeasy extraction kit and protocol. Four mitochondrial gene fragments were independently PCR amplified and sequenced per sample. The 12s gene was amplified using the primers L1091 and H1557, and the 16s gene was amplified using the primers L2510 and H3059 (described in Parkinson et al., 1997; Parkinson, 1999). The *cyt-b* fragment was PCR amplified using the primers Gludg and AtrCB3 (described in Parkinson et al., 2002) and the ND4 fragment was amplified via PCR using the primers ND4 and LEU or ND4 and HIS as described in Arévalo et al. (1994). Positive PCR products were excised from agarose electrophoretic gels and purified using the GeneCleanIII kit (BIO101). Purified PCR products were sequenced in both directions with the amplification primers (and for ND4, an additional internal primer HIS; Arévalo et al., 1994). In cases where PCR products were too weak to sequence directly, they were cloned using the Topo TA cloning kit (Invitrogen). Plasmids were isolated from multiple clones per individual using the Qiaquick spin miniprep kit (Qiagen) and

sequenced using M13 primers. All sequencing was accomplished using the CEQ Dye Terminator Cycle Sequencing Quick Start Kit (Beckman-Coulter) and run on a Beckman CEQ8000 automated sequencer. Raw sequence chromatographs were edited using Sequencher 4.2 (Gene Codes Corp.). Sequences of each fragment were aligned manually in GeneDoc (Nicholas and Nicholas, 1997). Alignment of protein-coding genes was straightforward and included several indels that represented deletions or insertions of complete codons. No internal stop codons were found in either protein coding fragment. Alignment of rRNA genes was based on models of secondary structure for snake mitochondrial rRNAs (Parkinson, 1999). A total of 24 sites were excluded because positional homology was not obvious (all occurred in loop structural regions of rRNA genes), including 10 sites from 12s and 14 sites from 16s. Novel sequences were deposited in GenBank (accession numbers DQ305409 – DQ305489; Table 12).

Phylogenetic reconstruction

Gaps in alignment were treated as missing data for all phylogenetic reconstructions. Maximum parsimony (MP) and Bayesian Metropolis-Hastings coupled Markov chain Monte Carlo (MCMC) phylogenetic methods were used to reconstruct phylogenies. Both methods were initially used to compare phylogenetic reconstructions based on each gene fragment independently. In general, we expect that mitochondrial loci should all contain phylogenetic signal supporting a common phylogeny because mitochondrial haplotypes are inherited maternally as a single linkage unit. We verified this assumption, prior to combining data, by reconstructing phylogenies of each gene independently and searching for strongly supported incongruent relationships across gene trees (e.g., Wiens, 1998).

Table 12. Description of complex partitioned models used in the analysis of the combined dataset. Each partition identified below was allocated the model selected by AIC criteria estimated in MrModeltest.

Model	Partitions	Free Model Parameters	Description of Partitions	Harmonic Mean of Marginal Likelihood	Akaike weight (A_w)	Relative Bayes Factor (RBF)
1x	1	11	single model for the entire dataset	-66557.76	0.0000	----
2x	2	22	protein coding genes; rRNA genes	-66405.69	0.0000	27.65
3x	3	33	codon positions 1+2; codon position 3; rRNA genes	-66337.62	0.0000	20.01
4xA	4	44	12s; 16s; codon positions 1+2; codon position 3	-66300.39	0.0000	15.60
4xB	4	44	12s; 16s; ND4; <i>cyt-b</i>	-66342.22	0.0000	13.06
5xA	5	51	rRNA stems, rRNA loops, codon position 1; codon position 2; codon position 3	-66195.33	0.0000	18.12
5xB	5	55	12s; 16s; codon position 1; codon position 2; codon position 3	-66255.71	0.0000	13.73
5xC	5	55	rRNA genes; ND4 position 1+2; ND4 position 3; <i>cyt-b</i> position 1+2; <i>cyt-b</i> codon position 3	-66043.64	0.0000	23.37
8x	8	84	12s; 16s; ND4 position 1; ND4 position 2; ND4 position 3; <i>cyt-b</i> position 1; <i>cyt-b</i> position 2; <i>cyt-b</i> position 3	-65842.18	0.0000	19.60
10x	10	94	all codon positions or stem and loop regions of each gene allocated independent model (labeled $P1-10$ in Table 2)	-65737.02	1.0000	19.78

All MP phylogenetic analyses were conducted using PAUP* version 4.0b10 (Swofford, 2002). All characters were treated as equally-weighted in MP searches. We used the heuristic search option with tree bisection reconnection (TBR) branch-swapping option, and 1,000 random-taxon-addition sequences to search for optimal trees. Support for nodes in MP reconstructions was assessed using non-parametric bootstrapping (Felsenstein, 1985) with 1,000 full heuristic pseudo-replicates (10 random-taxon-addition sequence replicates per bootstrap pseudo-replicate).

MrModeltest v.2.2 (Nylander, 2004) was used to select an appropriate model of evolution for MCMC analyses because this program only considers nucleotide substitution models that are currently available in MrBayes v3.04b (Ronquist and Huelsenbeck, 2003). PAUP* was used to calculate model likelihoods for use in MrModeltest. Based on arguments presented by Posada and Buckley (2005), we used AIC (Akaike, 1973, 1974; Sakamoto et al., 1986) to select best-fit models in MrModeltest. In addition to the combined dataset, putative *a priori* partitions of the dataset were independently analyzed using MrModeltest to estimate best-fit models of nucleotide evolution. These best-fit models for each partition were implemented as partition-specific models within partitioned-model analyses of the combined dataset, similar to the suggestions of Brandley et al. (2005).

All MCMC phylogenetic analyses were conducted in MrBayes 3.0b4 (Ronquist and Huelsenbeck, 2003) with vague priors and three incrementally heated chains in addition to the cold chain (as per the program's defaults). Each MCMC analysis was conducted in triplicate, with three independent runs initiated with random trees, and run for a total of 4.0×10^6 generations (sampling trees every 100 generations). Conservatively, the first 1.0×10^6

generations from each run were discarded as burn-in. Summary statistics and consensus phylograms with nodal posterior probability support were estimated from the combination of the triplicate set of runs per analysis.

An initial set of MCMC runs (for the individual and combined datasets) was conducted using the model estimated by AIC in MrModeltest for each dataset. In addition to the unpartitioned model selected by AIC for the entire dataset, the combined dataset was subjected to additional MCMC analyses under nine alternative evolutionary models. These additional MCMC analyses were designed to allow independent models of nucleotide evolution to be applied to partitions of the combined dataset. This was accomplished by dividing the dataset into *a priori* assumed biologically relevant partitions and specifying that an independent (partition-specific) model be used for each partition (using the “unlink” command in MrBayes). For these complex-partitioned models, only branch lengths and topology remained linked between partitions. These mixed models partitioned the combined dataset based on gene fragment type (protein coding or rRNA), gene, codon position (for protein encoding genes), and stem and loop secondary structure (for rRNA genes). The names and details of all models used to analyze the combined dataset are summarized in Table 13. MrBayes blocks containing the settings for various MCMC analyses are available from the authors upon request.

We used three statistics to choose the best-fit partitioned model for analysis of the combined data: 1) Bayes factors (B_{10}), 2) relative Bayes factors (RBF), and 3) Akaike weights (A_w) (as in Castoe et al., 2005). Each of these criteria allow objective evaluation of non-nested partitioned models, which is important here because several alternative models are non-nested. Bayes factors were calculated using the harmonic mean approximation of the marginal model

Table 13. Results of AIC model selection conducted in MrModeltest for partitions of the dataset.

Partition	AIC Model
all data	GTR+ΓI
all rRNA	GTR+ΓI
all rRNA, stems	SYM+ΓI
all rRNA, loops	GTR+ΓI
12S	GTR+ΓI
12s, stems (=P1)	SYM+ΓI
12s, loops (=P2)	HKY+ΓI
16s	GTR+ΓI
16s, loops (=P3)	GTR+ΓI
16s, stems (=P4)	SYM+ΓI
all protein coding	GTR+ΓI
positions 1+2	GTR+ΓI
position 1	GTR+ΓI
position 2	GTR+ΓI
position 3	GTR+ΓI
cyt-b	GTR+ΓI
cyt-b, positions 1+2	GTR+ΓI
cyt-b, position 1 (=P5)	GTR+ΓI
cyt-b, position 2 (=P6)	HKY+ΓI
cyt-b, position 3 (=P7)	GTR+ΓI
Nd4	GTR+ΓI
Nd4, positions 1+2	GTR+ΓI
Nd4, position 1 (=P8)	GTR+ΓI
Nd4, position 2 (=P9)	GTR+ΓI
Nd4, position 3 (=P10)	GTR+ΓI

likelihood following Nylander et al. (2004; see also Kass and Raftery, 1995), and we report the results in the form of $2\ln B_{10}$. Evidence for model M_I over M_0 was considered very strong (and considered sufficient for our purposes) if $2\ln B_{10} > 10$ (Kass and Raftery, 1995, see also Nylander et al., 2004).

Relative Bayes factors (RBF; Castoe et al., 2005) were used to quantify the average impact that each free model parameter had on increasing the fit of the model to the data. These values were also used to estimate the ratio of parameters to posterior evidence (of prior modification by the data) of increasingly complex partitioned models. This may provide a simple means of determining the parameter richness of candidate models tested in relation to how complex a model may be justified by the size and heterogeneity of a dataset (Castoe et al., 2005). We calculated the RBF of each complex model by calculating $2\ln B_{10}$ between the base model and each complex (partitioned) model and dividing this by the difference in the number of free model parameters between the base and complex model (Castoe et al., 2005).

Akaike weights (A_w) were employed as a means of confirming model choice, together with $2\ln B_{10}$ estimates. To estimate A_w , we used the harmonic mean estimator of the model likelihood from MCMC analyses to incorporate an estimate of the marginalized likelihood of models (following Castoe et al., 2005). The higher the A_w for a model, the higher the relative support for that model.

Once a tentative best-fit model was chosen for the combined data, this model was checked for evidence of parameter identifiability, failed convergence, and unreliability (which would suggest the model may be parametrically over-fit; e.g., Castoe et al., 2004; Huelsenbeck et al., 2002; Rannala, 2002). We investigated the performance of models (using Tracer; Rambout

and Drummond, 2003) by examining features of model likelihood and parameter estimate burn-in, as well as the shapes and overlap of posterior distributions of parameters. We looked for evidence that model likelihood and parameter estimates ascended directly and rapidly to a stable plateau, and that independent runs converged on similar likelihood and parameter posterior distributions (considered evidence that a model was not over-fit). We also examined the model parameter estimates to confirm that the shape of their posterior distributions reflected a substantial modification of the priors (indicating their identifiability based on the data). As a secondary validation that the partitioning of the dataset was justified, we graphically compared posterior distributions of parameter estimates across partitions to confirm that, in fact, different partitions demonstrated unique posterior distributions of parameter estimates.

Results

Properties of the dataset

The final alignment of all four gene fragments concatenated consisted of a total of 2306 aligned positions: 417 from 12s, 503 from 16s, 717 from *cyt-b*, and 669 from ND4. This alignment contained 1105 parsimony-informative characters and 906 invariant characters.

The greatest pairwise sequence divergence (uncorrected percent divergence) across all taxa was 20.8% (*Causus resimus* and *Bothrops atrox*), and 17.7% among crotaline taxa (*Calloselasma rhodostoma* and *Sistrurus miliaris*). The maximum divergence among Old World

pitvipers was 16.4% (*C. rhodostoma* and *Cryptelytrops venustus*), and 16.2% among New World pitvipers (*Porthidium porrasi* and *Crotalus transverses*). The mean divergence between Old and New World pitvipers was 12.9%.

Individual gene phylogenies generally suffered from poor resolution and low support under MP and MCMC analyses. No instances of strongly supported differences across individual gene trees were observed, providing evidence for the assumption that individual genes supported a common phylogeny and are appropriate for combined data analysis. Previous studies that have analyzed many of the sequences used in this study have come to the same general conclusion supporting the combinability of these four gene fragments (e.g., Castoe et al., 2005; Malhotra and Thorpe, 2004; Murphy et al., 2002; Parkinson, 1999; Parkinson et al., 2002).

Maximum Parsimony phylogenetic analyses

The MP heuristic search found 12 equally-parsimonious trees, each with 14,816 steps. These trees had a consistency index of 0.162, a retention index of 0.568, and a homoplasy index of 0.838. The strict consensus of these 12 trees, along with nodal bootstrap support (BS hereafter) values, is provided (Fig. 14).

Maximum parsimony phylogenetic estimates (Fig. 14) show strong support for a clade containing the monotypic Azeimopinae (*Azemiops fea*) and the Crotalinae (BS = 100), as well as the sister-group relationship of these two subfamilies (BS = 89). Three ancient clades of pitvipers are inferred by MP analyses: two exclusively Old World clades, and a third containing both Old and New World species, although support for these clades is low. The deepest phylogenetic split among pitvipers is estimated as being between a clade including *Hypnale* and *Calloselasma* and



Figure 14. Strict consensus cladogram of 12 equally-parsimonious trees obtained from maximum parsimony analysis of 2306 bp of mitochondrial DNA sequences (14816 steps, consistency index = 0.162, retention index = 0.568, homoplasy index = 0.838). Bootstrap support for nodes above 50% is given adjacent to nodes; nodes receiving bootstrap support of 100% are indicated by gray-filled circles.

the remaining Crotalinae. Following this divergence, a clade including *Deinagkistrodon*, *Garthius*, and *Tropidolaemus* is estimated to be the sister group to the third ancient pitviper clade comprising the remaining Asiatic and New World species (Fig. 14).

A large clade containing nearly all members of *Trimeresurus sensu lato* was strongly supported (BS = 89), as were a majority of intra and intergeneric relationships within this clade (Fig. 14). *Trimeresurus sensu stricto* is inferred to be polyphyletic, with *T. gracilis* distantly related to the remaining members. Monophyly of *Popeia*, *Viridovipera*, and *Parias* received moderate to strong (BS > 74) support, although *Cryptelytrops* was found to be polyphyletic, with a clade containing *C. venustus* and *C. macrops* distantly related to the remaining *Cryptelytrops* species (Fig. 14). *Ovophis* was found to be polyphyletic, with *O. monticola* estimated to be the sister lineage to a clade containing *Triceratolepidophis*, *Zhaoermia*, and *Protobothrops* (Fig. 14). The other representative of this genus included in this study, *O. okinavensis*, was strongly supported as the sister taxon to *Trimeresurus gracilis*, both forming the sister clade to *Gloydius*. This clade was weakly supported as the sister taxon to a moderately supported (BS = 76) clade including all New World genera (Fig. 14).

The deepest phylogenetic relationships among New World genera were poorly resolved by MP analyses (Fig. 14). The temperate New World genera (*Agkistrodon*, *Sistrurus*, and *Crotalus*) did not form a clade (Fig. 14). *Ophryacus* and *Lachesis* formed a weakly supported clade, inferred as the sister group to *Agkistrodon*. Monophyly of *Ophryacus*, *Lachesis*, and *Agkistrodon* were all strongly supported (BS > 96), and monophyly of *Bothriechis* received weak support (BS = 58). The primarily Middle American genera *Atropoides*, *Cerrophidion*, and *Porthidium* formed a strongly supported (BS = 95) clade inferred to be the sister group to a clade

(BS = 100) containing the primarily South American genera *Bothrocophias*, *Bothrops*, and *Bothriopsis*. Within the Middle American group, monophyly of *Porthidium* was well supported (BS = 100). *Atropoides* was inferred to be paraphyletic (BS = 72) with respect to *Cerrophidion* and *Porthidium*, with *A. picadoi* distantly related to other *Atropoides* species. Within the South American group, a *Bothrocophias* clade (BS = 100) was inferred to be the sister taxon to a clade containing a *Bothriopsis* clade (BS = 100) and paraphyletic clustering of *Bothrops* species. Monophyly of the rattlesnakes, *Sistrurus* and *Crotalus*, was strongly supported (BS = 100), with a monophyletic (BS = 89) *Sistrurus* forming the sister taxon to a weakly supported (BS = 57) monophyletic *Crotalus*. Deep phylogenetic relationships among *Crotalus* species generally received weak support (Fig. 14).

Selection, evaluation, and comparison of Bayesian MCMC models

The single (unpartitioned) best-fit model for the combined dataset identified by AIC criteria was the GTR+ Γ I model (Tavaré, 1996; Table 13; “1x” model in Table 12). In addition to this unpartitioned model, nine other models that allocated an independent model of nucleotide evolution to various partitions of the dataset within a combined data analysis were examined (Table 12). Partition-specific best-fit models selected using AIC criteria in MrModeltest are shown in Table 13, and included one of three different models selected for various partitions: the GTR+ Γ I (11 free model parameters), the HKY+ Γ I (Hasegawa et al., 1985; 7 free parameters), and SYM+ Γ I (a GTR model with fixed equal base frequencies; 7 free parameters). Across all models for the combined dataset, Akaike weights ($A_w = 1.0000$; Table 12) and Bayes factors ($2\ln B_{10} > 210$; Table 3) provided extremely strong support for the most complex partitioned

model examined, 10x, as the best-fit to the combined data. Relative Bayes factors demonstrate that, despite the large number of free model parameters in the 10x model, the average contribution of each parameter to increasing the overall likelihood remains high (RBF = 19.78), compared across other partitioned models (Table 12). Only one model, the 2x model in which protein-coding and rRNA genes were allocated separate models, had a RBF (27.65; Table 12) substantially higher than the 10x model.

The best-fit 10x model showed no indications of being parametrically overfitted, or of poor mixing or convergence. The three independent runs of the 10x model produced identical tree topologies, extremely similar posterior probability estimates (all values within three percentage points, most less than three), and model likelihoods and parameter estimates that were nearly identical. Plots of the model likelihoods through generations from independent runs all show a rapid and direct ascent to a stationary plateau by no later than 200,000 generations (suggesting that burn-in occurred by this period), implying that our exclusion of the first 10^6 generations (as “burn-in”) was conservative. Similar to plots of model likelihoods through time, plots of parameter estimates all demonstrated a direct approach to a stationary range, occurring at approximately the same number of generations as likelihood values appeared to reach stationarity (as visualized using Tracer). Based on our model-selection criteria, combined with our inability to identify any problems indicating that the 10x model is excessively parameter rich, we treat phylogenetic estimates based on the 10x model as our favored phylogenetic hypothesis hereafter.

Substantial differences in parameter estimates were observed between the 1x model and the parameters of the 10x partitions, as well as among different partitions of the 10x model

(based on parameter means and 95% credibility intervals, CI hereafter; Table 14). A subset of parameter estimates is shown in Fig. 2. For each of the five parameters plotted across models and partitions, at least two partition-specific parameter estimates (based on CIs) from the 10x model do not overlap with the CI of the analogous parameter from the 1x model (Fig. 15). Among parameter CIs that do overlap between the 1x and 10x partitions, many partitions have parameter estimates in which a majority of posterior density is concentrated outside the 95% CI of the 1x model estimates (Fig. 15). Among model parameters, estimates of the gamma shape parameter (and I parameter, pInvar.) show the least overlap between 10x partitions and the 1x model, followed in magnitude by nucleotide frequencies, and then by parameters of the GTR substitution matrix (Fig. 15; Table 14).

Bayesian phylogenetic hypotheses based on 10x partitioned model

Bayesian phylogenetic estimates under the 10x partitioned model inferred a strongly supported clade ($Pp = 100$) comprising the Azemiopinae (*Azemiops*) and the Crotalinae, with the Crotalinae forming its own monophyletic group ($Pp = 100$; Fig. 16). This MCMC phylogeny implied the same three early phylogenetic splits among pitvipers as did MP, although the relationships between the three were unresolved (Fig. 16). The first of these clades ($Pp = 100$) includes *Hypnale* and *Calloselasma*. The second of these clades ($Pp = 92$) includes *Deinagkistrodon*, *Garthius*, and *Tropidolaemus*. The third basal pitviper clade ($Pp = 100$) includes all remaining Old World and New World genera (Fig. 16).

Table 14. Mean and 95% credibility interval (in parentheses) of model parameters from Bayesian phylogenetic analyses of the combined data set conducted under the 1x and 10x models. Parameter estimates for each model are based on a total of 9 x 10⁶ generations combined from three independent MCMC runs. Partitions of the 10x model (P1 – P10) are defined in Table 2.

Model - Partition	Ti:Tv	r(C–T)	r(C–G)	r(A–T)	r(A–G)	r(A–C)
1x	---	7.21 (6.12–8.61)	0.77 (0.60–0.96)	0.83 (0.68–1.01)	11.63 (9.63–13.70)	0.57 (0.47–0.70)
10x-P1	---	70.32 (34.36–98.54)	1.50 (0.47–3.24)	4.70 (2.25–7.94)	19.61 (10.35–30.26)	6.33 (2.99–10.82)
10x-P2	11.44 (9.89–13.18)	---	---	---	---	---
10x-P3	---	10.16 (5.97–16.51)	1.37 (0.59–2.69)	3.26 (1.74–5.58)	10.93 (6.02–18.78)	1.68 (0.75–3.17)
10x-P4	---	12.90 (6.78–26.99)	1.18 (0.47–2.67)	1.40 (0.71–2.95)	8.14 (4.41–16.01)	0.99 (0.48–2.07)
10x-P5	---	11.46 (6.01–21.67)	0.70 (0.25–1.52)	0.95 (0.45–1.82)	16.93 (9.51–30.88)	0.54 (0.29–1.03)
10x-P6	---	4.19 (2.59–6.86)	0.05 (0.01–0.14)	0.66 (0.35–1.18)	5.47 (3.33–8.88)	0.30 (0.17–0.52)
10x-P7	---	17.08 (4.12–60.87)	20.04 (6.09–65.38)	1.71 (0.34–6.17)	28.82 (8.15–82.77)	3.85 (0.65–14.12)
10x-P8	---	3.40 (2.27–5.09)	0.25 (0.12–0.44)	0.50 (0.31–0.78)	3.63 (2.44–5.29)	0.21 (0.12–0.35)
10x-P9	7.27 (5.50–9.48)	---	---	---	---	---
10x-P10	---	6.05 (3.78–9.73)	1.70 (0.90–2.99)	0.63 (0.35–1.08)	15.74 (9.38–26.20)	0.36 (0.21–0.59)

Model - Partition	pi(A)	pi(C)	pi(G)	pi(T)	Γ	pInvar.
1x	0.35 (0.34–0.37)	0.36 (0.35–0.37)	0.07 (0.06–0.07)	0.22 (0.21–0.23)	0.63 (0.60–0.66)	0.31 (0.28–0.33)
10x-P1	---	---	---	---	0.39 (0.36–0.42)	0.29 (0.22–0.38)
10x-P2	0.39 (0.36–0.42)	0.31 (0.29–0.34)	0.08 (0.07–0.09)	0.22 (0.20–0.24)	0.30 (0.27–0.32)	0.08 (0.03–0.14)
10x-P3	---	---	---	---	0.21 (0.19–0.22)	0.17 (0.08–0.25)
10x-P4	0.47 (0.43–0.52)	0.24 (0.21–0.27)	0.07 (0.05–0.09)	0.22 (0.19–0.25)	0.46 (0.41–0.5)	0.32 (0.26–0.37)
10x-P5	0.43 (0.40–0.47)	0.35 (0.32–0.38)	0.06 (0.05–0.06)	0.16 (0.15–0.18)	3.63 (2.76–4.62)	0.03 (0.00–0.07)
10x-P6	0.35 (0.30–0.40)	0.38 (0.34–0.43)	0.10 (0.08–0.13)	0.17 (0.14–0.20)	0.33 (0.29–0.37)	0.21 (0.15–0.28)
10x-P7	0.16 (0.12–0.20)	0.28 (0.23–0.34)	0.11 (0.06–0.15)	0.46 (0.40–0.52)	0.22 (0.20–0.25)	0.41 (0.33–0.49)
10x-P8	0.37 (0.33–0.42)	0.33 (0.29–0.38)	0.09 (0.07–0.11)	0.20 (0.18–0.23)	0.48 (0.44–0.51)	0.34 (0.28–0.39)
10x-P9	0.24 (0.20–0.29)	0.26 (0.22–0.30)	0.11 (0.09–0.14)	0.38 (0.34–0.43)	0.21 (0.20–0.23)	0.31 (0.23–0.38)
10x-P10	0.32 (0.29–0.35)	0.43 (0.40–0.46)	0.04 (0.03–0.04)	0.21 (0.20–0.23)	2.89 (2.31–3.62)	0.03 (0.00–0.07)

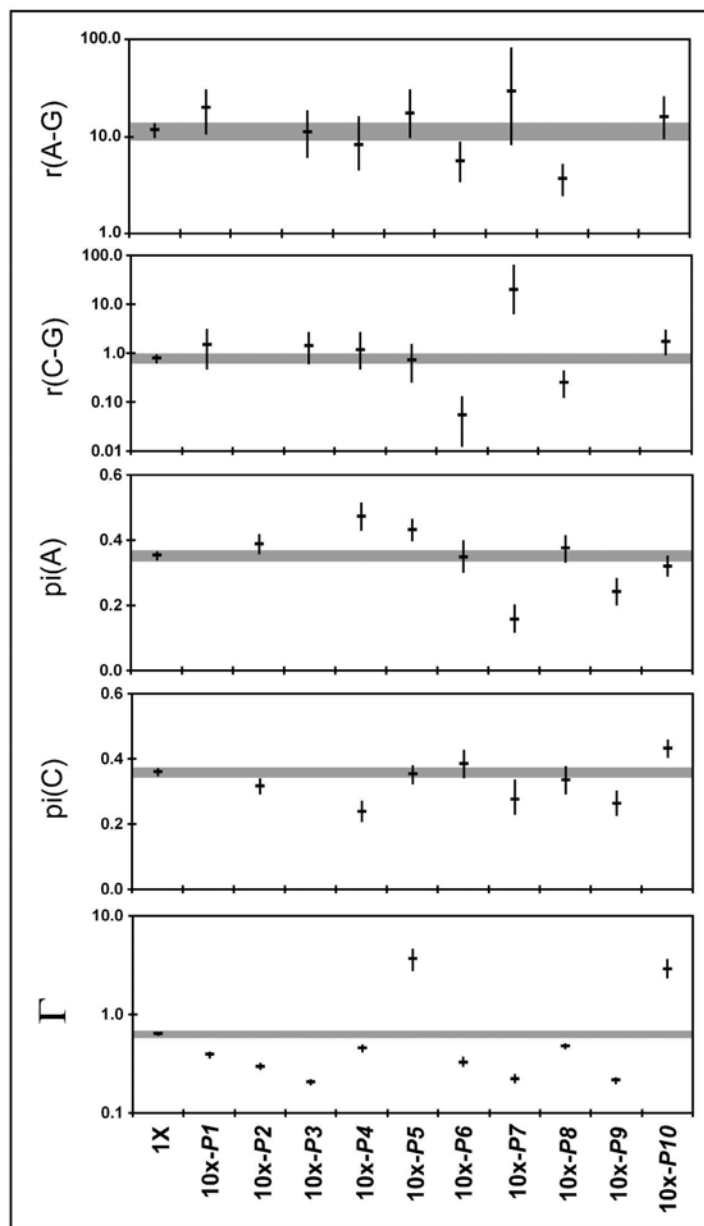


Figure 15. Comparisons of means and 95% credibility intervals (CI) of selected nucleotide model parameters estimated from Bayesian MCMC analyses conducted under the 1x (unpartitioned) and the 10x (partitioned) models. Partitions of the 10x model are designated P1–P10 and correspond with Table 2. Gray-shaded bands indicate the 95% CI of parameters estimated under the 1x model.

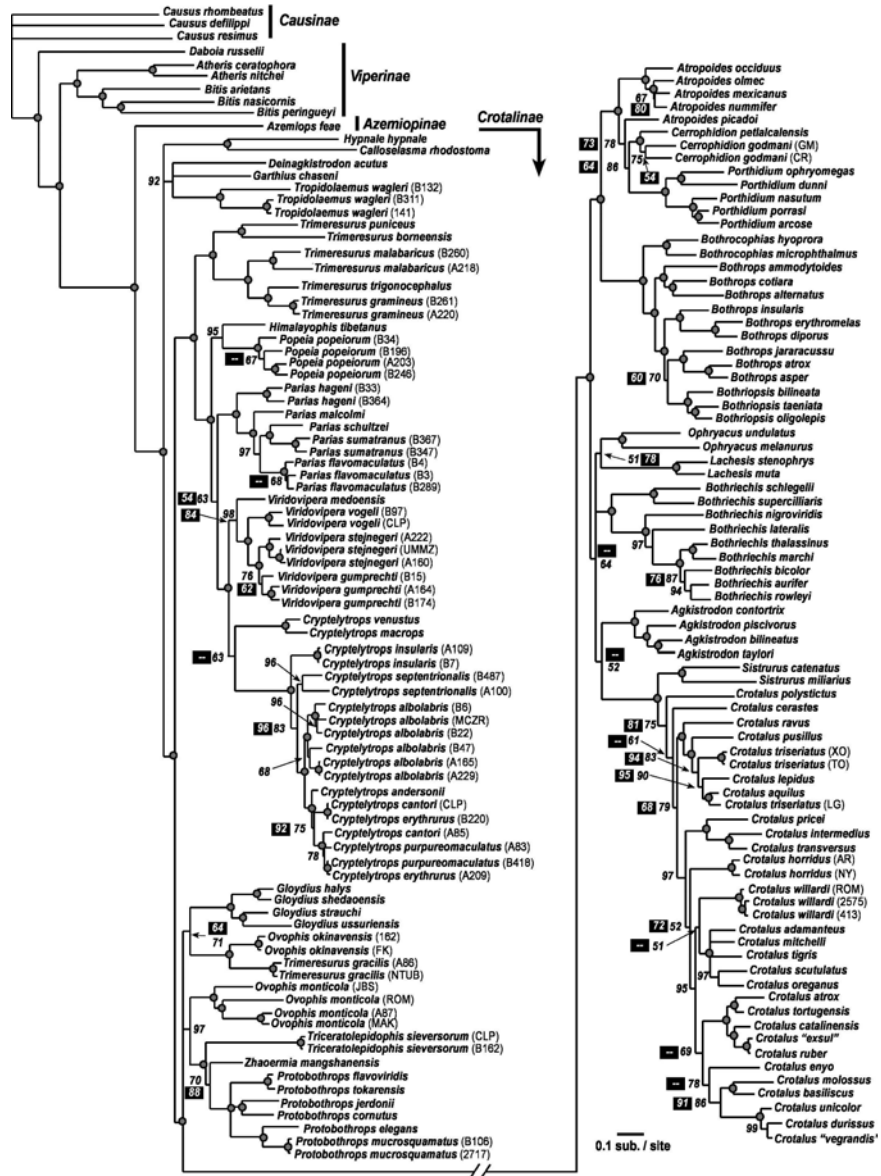


Figure 16. Bayesian MCMC fifty-percent majority-rule consensus phylogram compiled from analyses of 2306 bp of mitochondrial DNA sequences analyzed under the best-fit “10x” partitioned model (see text for model definition and selection). Consensus phylogram and posterior probabilities (shown adjacent to nodes) were estimated from a total of 9 x 10⁶ post-burn-in generations (from three independent MCMC runs). Nodes receiving posterior probability support of 100% are indicated by grey-filled circles; otherwise, posterior probability support for nodes based on the 10x model is shown in black print. Posterior probability estimates based on the unpartitioned 1x model that differed notably from those from the 10x model are shown in black rectangles with white print (black boxes with dashes indicate clades that were not present in the consensus topology of the 1x tree).

A large clade containing almost all members of *Trimeresurus sensu lato* is strongly supported ($Pp = 100$). *Trimeresurus sensu stricto* was inferred to be polyphyletic (with strong support across several intervening nodes), with *T. gracilis* distantly related to a strongly supported clade ($Pp = 100$) containing the remaining members of *Trimeresurus* (Fig. 16). Monophyly of *Popeia* ($Pp = 100$), *Viridovipera* ($Pp = 98$), and *Parias* ($Pp = 100$) received strong support. *Cryptelytrops* was found to be monophyletic, unlike in the MP tree, but with low support ($Pp = 63$). *Ovophis* was estimated to be polyphyletic, with *O. monticola* placed as the sister lineage ($Pp = 97$) to a clade containing *Triceratolepidophis*, *Zhaoermia*, and *Protobothrops*. Within this clade, *Zhaoermia* was inferred as the sister lineage ($Pp = 70$) to a monophyletic ($Pp = 100$) *Protobothrops* clade. *Ovophis okinavensis* was strongly supported ($Pp = 100$) as the sister lineage to *Trimeresurus gracilis* (both taxa placed far from congeneric species); collectively, this clade formed the sister group to a monophyletic ($Pp = 100$) *Gloydus* (Fig. 16). The sister group to all New World genera was not resolved, with a polytomy uniting three clades ($Pp = 100$) including: a *Gloydus*, *O. okinavensis*, *T. gracilis* clade; an *O. monticola*, *Triceratolepidophis*, *Zhaoermia*, *Protobothrops* clade; and a third clade ($Pp = 100$) including all New World genera (Fig. 16).

The earliest phylogenetic divisions among New World pitvipers were generally inferred with weak support and poor resolution. The earliest divergence within New World genera was estimated between a clade ($Pp = 100$) including Middle and South American bothropoid genera (*Atropoides*, *Cerrophidion*, *Porthidium*, *Bothrocophias*, *Bothrops*, *Bothriopsis*) and a weakly supported clade ($Pp = 64$) containing the remaining temperate and tropical New World genera (Fig. 16). The Middle American genera *Atropoides*, *Cerrophidion*, and *Porthidium* formed a

clade inferred to be the sister group to a clade comprising the South American genera *Bothrocophias*, *Bothrops*, and *Bothriopsis* ($Pp = 100$). Within the Middle American clade, the monophyly of *Porthidium* received strong support ($Pp = 100$). *Atropoides* was estimated to be paraphyletic ($Pp = 78$) with respect to *Cerrophidion* and *Porthidium*, due to *A. picadoi* not being grouped with other *Atropoides* species (Fig. 16). Among South American bothropoids, a monophyletic ($Pp = 100$) *Bothrocophias* formed the sister group to a clade containing a monophyletic ($Pp = 100$) *Bothriopsis* and a paraphyletic *Bothrops* group.

Relationships among members of the second basal clade of New World genera (including tropical and temperate genera) were unresolved, with a polytomy between three clades: a clade ($Pp = 51$) containing a monophyletic *Ophryacus* ($Pp = 100$) and a monophyletic *Lachesis* ($Pp = 100$), a clade ($Pp = 100$) including all *Bothriechis* species, and a clade ($Pp = 52$) containing the temperate New World genera (*Agkistrodon*, *Sistrurus*, and *Crotalus*). Monophyly of *Agkistrodon* and *Sistrurus* received strong support (both $Pp = 100$) and *Crotalus* monophyly received weak support ($Pp = 75$). *Agkistrodon* was weakly inferred to be the sister taxon ($Pp = 52$) to a clade including *Crotalus* and *Sistrurus* ($Pp = 100$). Deep phylogenetic relationships among *Crotalus* species received poor support (Fig. 16).

Differences in MCMC phylogenetic estimates between 1x and 10x partitioned analyses

Consensus topology and nodal posterior probabilities from the 1x model analyses that differed notably (Pp difference > 5 for weakly supported clades, > 3 for Pp values above 90) from that of the 10x model are indicated in Fig. 16. A majority of the differences between the MCMC phylogeny based on the unpartitioned 1x model, compared to the partitioned 10x model,

represented changes in the posterior probability for moderately or weakly supported nodes. No nodes receiving 100% Pp under one model received less than 97% Pp support under the other model. Posterior probabilities that differed notably between the 1x and 10x estimates tended to show higher Pp estimates in the 10x model, although examples to the contrary were observed. This trend of increased Pp support under the 10x model was more pronounced at deeper nodes (Fig. 16).

There were no major changes in the tree topology between the 1x and 10x analyses (considering moderate to well supported clades). The 50% majority rule consensus topology, however, did show several differences in resolution of poorly supported clades between estimates. The only important difference in the majority-rule consensus topology among Old World pitvipers was the collapse of the internode supporting *Cryptelytrops venustus* plus *C. macrops* as sister to the remaining members of the genus, hence the failure of the 1x model to infer/resolve the monophyly of *Cryptelytrops* (1x- $Pp < 50$, 10x- $Pp = 63$). Deep phylogenetic relationships among New World pitvipers, based on the 50% majority-rule consensus of the 1x analyses, suggest a different (yet poorly supported) topology with a primary phylogenetic division occurring between a clade containing *Sistrurus* and *Crotalus* (the rattlesnakes; $Pp = 100$), and the remaining New World genera ($Pp = 51$), similar to that seen in the MP tree. Within this second large New World clade, there was a polytomy of three lineages in the 1x tree including the following clades: 1) an *Agkistrodon* clade, 2) a *Lachesis* and *Ophryacus* clade, and 3) a clade containing *Bothriechis* as the sister group ($Pp = 56$) to Middle and South American bothropoid genera. Relationships among several *Crotalus* species also show alternative consensus topology between models, largely resulting from the placement of *C. enyo* shifting

from the sister taxon to *C. willardi* in the 1x tree ($Pp = 59$), to the sister lineage ($Pp = 78$) of a clade containing *C. molossus*, *C. basiliscus*, *C. unicolor*, *C. durissus*, and *C. “vergrandis”* in the 10x tree.

Discussion

Strengths and limitations of complex partitioned models

Model specification in Bayesian MCMC analyses is inherently critical to the accuracy of phylogeny estimates since Bayesian Pps represent estimates of bipartition support that are dependent on the model (and priors) and the data (Huelsenbeck et al., 2002; Larget and Simon, 1999; also see Huelsenbeck and Rannala, 2004). In general, Pps have been shown to be less conservative than bootstrap values (Douady et al., 2003; Erixon et al., 2003; Leaché and Reeder, 2002; see also Cummings et al., 2003). Nonetheless, broad claims that bipartition Pps represent over-inflated estimates of phylogenetic confidence (e.g., Simmons et al., 2004; Suzuki et al., 2002) are not necessarily justifiable. Available evidence suggests, instead, that Pp values provide a more powerful estimate of phylogenetic structure present in aligned sequences than do BS values (Alfaro et al., 2003; Wilcox et al., 2002), provided major assumptions of the method are not violated (e.g., Suzuki et al., 2002). Many studies agree that Bayesian analyses conducted using overly simplistic models suffer from decreased Pp accuracy (e.g., Erixon et al., 2003; Huelsenbeck and Rannala, 2004; Suzuki et al., 2002; Wilcox et al., 2002). In contrast, simulation

studies have shown that when Bayesian analyses are conducted using models more complex than that used to generate simulated data, Pp accuracy remains high (Huelsenbeck and Rannala, 2004; Lemmon and Moriarty, 2004). Collectively, these conclusions suggest that using a “compromise” model, in which multiple unique patterns of evolution are modeled using a single set of parameters, appears to be a major concern for phylogenetic estimation. Partitioning models of evolution across portions of a dataset provides a straightforward means of reducing the biases inherent with oversimplified modeling in Bayesian phylogenetic analyses. Generally, favoring the use of more complex models offers the best chance of recovering an accurate Bayesian phylogenetic estimate, as long as parameters can be accurately identified from the data (see also Huelsenbeck and Rannala, 2004). The upper limit of model complexity imposed by the need for parameters to be estimatable (or identifiable; see Castoe et al., 2004; Huelsenbeck et al., 2002; Rannala, 2002) is the primary justification for employing methods of model selection (e.g., Bayes factors, Akaike weights) and *post hoc* MCMC run evaluation in Bayesian phylogenetic analyses.

To what extent is an unpartitioned model forced to compromise estimates of model parameters in the analysis of a combined multi-gene dataset (as in our case), versus a model like the 10x that contains several partitions? Our results suggest that this compromise is extreme in some cases, and is evident across different classes of model parameters. Comparisons of the 95% CI of parameter estimates derived from the 1x, versus partitions of the 10x model (Fig. 15, Table 14), show many instances where 95% CIs of partitions do not overlap those based on the 1x model. Furthermore, many CIs that do overlap do not coincide for a majority of their posterior densities. These findings point directly at the elevated potential for an unpartitioned model to fall

into the trap identified in simulation studies where an oversimplified model suffers from decreased posterior probability accuracy. Collectively, available evidence supports not only the use of complex models (including partitioned models), but implies that these may be crucial for accurate phylogenetic estimates (see also Huelsenbeck and Rannala, 2004).

Across the models we tested for the combined data, all model-selection criteria supported the most complex partitioned model by a large margin (the 10x model). A majority of Bayes factors provided extremely strong support for increasingly complex models (data not shown). Relative Bayes factors (RBF) for increasingly complex models remained high, suggesting high returns on parameter addition even with increasing model complexity (Castoe et al., 2005). Collectively, these results seem to suggest that even more complex models than those tested here are likely to have been favored by model-selection criteria. Our most complex candidate model exhausted our *a priori* conceptions of biologically meaningful partitions of the data, placing an upper limit on the models examined. Future studies that investigate additional partitioning schemes (e.g., identify heterogeneous patterns within genes not examined here) may provide additional suggestions for partitioning heterogeneous datasets (Faith and Pollock, 2003; Huelsenbeck et al., 2004).

How should the differences in phylogenetic hypotheses between simple and complex models be interpreted? We found complex models to result in changes in *Pps* of clades that, in some instances, altered the Bayesian consensus topology. These changes tended to provide higher *Pps* in the complex (10x) model, with a majority of changes concentrated at deeper nodes (e.g., Brandley et al., 2005; Castoe et al., 2004, 2005; see also Alfaro et al., 2003). This observation raises two possibilities, either complex models result in over-inflated *Pp* support, or

they provide (at least on average) more accurate estimates of nodal support. Three points of evidence suggest that complex models do generally provide to more accurate, rather than over-inflated, posterior probability estimates: 1) the results of simulation studies discussed above, 2) empirical studies, including this one, demonstrating that even though a majority of nodes may increase, some decrease under complex model analyses (see also Brandley et al., 2005; Castoe et al., 2004, 2005; Nylander et al., 2004), and 3) results that show a coincidence between clades that show increased *Pp* support under complex-model analyses and are also supported by other independent data (noted below; see also examples in Castoe et al., 2005).

Phylogeny and systematics of pitvipers

In agreement with previous studies (e.g., Kraus et al., 1996; Malhotra and Thorpe, 2004; Parkinson et al., 2002), our results provide strong support for the monophyly of the Crotalinae (BS = 100, *Pp* = 100) and the Azemiopinae as its sister lineage (BS = 89, *Pp* = 100). We found evidence of three early-diverging lineages of pitvipers, two exclusively Old World clades, and a third containing both Old and New World species, although the branching pattern and order among these three clades was poorly resolved (Figs. 14, 16). Strong support for two exclusively Old World clades, *Hypnale* plus *Calloselasma*, and *Deinagkistrodon*, *Garthius*, and *Tropidolaemus*, was found by MP and MCMC analyses, although it remains unclear whether these two clades are sister groups (Figs. 14, 16). The third early-diverging pitviper group included all other Old and New World genera (Fig. 16), including a clade containing all members of *Trimeresurus sensu lato* (except *T. gracilis*) inferred to be the sister lineage to the remaining Old and New World genera.

The recent generic subdivision of *Trimeresurus* (Malhotra and Thorpe, 2004) is supported by our results. Monophyly of *Popeia*, *Viridovipera*, and *Parias* received strong support under MCMC ($Pp > 97$) and MP ($BS > 74$) analyses. Although *Cryptelytrops* was paraphyletic under MP (Fig. 14) and unresolved in the 1x MCMC tree, the 10x MCMC tree weakly supported the monophyly of this new genus ($Pp = 63$; Fig. 16). Monophyly of *Cryptelytrops* is additionally supported by the presence of long, slender, deeply-bifurcated papillose hemipenes (and other external morphological characters) in members of this genus (Malhotra and Thorpe, 2004). Interestingly, the monophyly of *Viridovipera*, united by the possession of spinose “type 2” hemipenes (Malhotra and Thorpe, 2004), also received increased support under the 10x ($Pp = 98$) versus the 1x model ($Pp = 84$; Fig. 16). We found strong support for the validity of two newly described monotypic genera, *Triceratolepidophis* (Ziegler et al., 2000) and *Zhaoermia* (Zhang, 1993; Gumprecht and Tillac, 2004), which formed a clade with *Protobothrops* ($BS < 50$, $Pp = 100$). *Zhaoermia* was inferred with weak to moderate support ($BS = 73$, $1x-Pp = 88$, $10x-Pp = 70$) as the sister lineage to a clade ($BS = 97$, $Pp = 100$) comprising *Protobothrops* species.

All analyses provided strong evidence that *Trimeresurus sensu stricto* is rendered polyphyletic by *T. gracilis* being placed distantly from remaining members of *Trimeresurus*. Similarly, the placement of *O. okinavensis* (distant from the type species *O. monticola*) renders the genus *Ovophis* polyphyletic. These two enigmatic species, *O. okinavensis* and *T. gracilis*, formed a strongly supported clade in all analyses ($BS = 100$, $Pp = 100$). Our results supporting the close relationship of *T. gracilis* and *O. okinavensis*, and the distant relationship of these taxa to congeneric species, is in agreement with previous studies based on mitochondrial gene

sequences (Malhotra and Thorpe, 2000, 2004) as well as sequences of a nuclear intron (Giannasi et al., 2001). The close relationship of these two species is particularly surprising because *T. gracilis* (like a majority of pitvipers) gives live birth to offspring, whereas *O. okinavensis* is among the few egg-laying species. Malhotra and Thorpe (2004) discussed possible actions to rectify the current generic allocation of *O. okinavensis* and *T. gracilis* (i.e., recognition of these species as a new genus versus allocating them to the genus *Gloydus*). These authors deferred taxonomic action until they could amass additional hemipenial and other morphological characters (work in progress by Malhotra and Thorpe), and we follow their decision.

Which lineage is the sister group to the New World pitvipers is an important question, with numerous ramifications relative to biogeography and trait evolution, yet no two studies have yielded identical results. Among molecular-based hypotheses, four Old World genera (*Protobothrops*, *Ovophis*, *Trimeresurus* and *Gloydus*) have been variously estimated as the sister group to the New World clade (Knight et al., 1992; Malhotra and Thorpe, 2004; Parkinson, 1999; Parkinson et al., 2002). Although support was weak, our MP tree inferred a clade containing *Gloydus*, *O. okinavensis*, and *T. gracilis* as the sister group to all New World genera (Fig. 14). Bayesian estimates did not resolve this relationship (based on the 50% majority-rule consensus), and yielded a polytomy between three clades: 1) a clade including all New World genera, 2) a *Gloydus*, *O. okinavensis*, *T. gracilis* clade, and 3) a clade containing *Protobothrops*, *Zhaoermia*, *Triceratolepidophis*, and *O. monticola*.

Early pitviper systematic studies suggested a close relationship between terrestrial pitvipers with large head shields (rather than many small head scales) in the Old World and New World, recognizing a trans-continental genus *Agkistrodon* (e.g., Gloyd and Conant, 1990).

Several studies, including our results, indicate that New World and Old World *Agkistrodon* (*sensu lato*) do not form a clade exclusive of other New World pitvipers (e.g., Knight et al., 1992; Kraus et al., 1996; Parkinson et al., 1997, 2002), supporting the recognition of *Gloydus* (Hoge and Romano-Hoge, 1981) for the Asiatic members of *Agkistrodon sensu lato*. Despite the polyphyly of *Agkistrodon sensu lato*, *Gloydus* is relatively close phylogenetically to New World pitvipers (Figs. 14, 16).

All non-crotaline members of the Viperidae are distributed exclusively in the Old World. Here, as in other studies (Kraus et al., 1996; Malhotra and Thorpe, 2004; Parkinson, 1999; Parkinson et al., 2002), we find strong evidence for multiple early-diverging lineages of Old World pitvipers, and the relatively recent origin of a monophyletic clade of New World pitvipers. Kraus et al. (1996) were the first to provide molecular evidence for the monophyly of all New World pitvipers and suggest a historical biogeographic scenario for pitvipers including a single dispersal event from the Old World into the New World, and subsequent studies have supported this hypothesis (Kraus et al., 1996; Parkinson, 1999; Parkinson et al., 2002; see also Gutberlet and Harvey, 2002, 2004).

Phylogenetic estimates based on both MP and MCMC did not resolve the deep phylogenetic relationships among New World genera with any decisive levels of support (Figs. 14, 16). We did not find evidence for a temperate (*Agkistrodon*, *Sistrurus*, *Crotalus*) clade as the sister group to the remaining New World (Neotropical) genera, as has been suggested by several studies (e.g., Gutberlet and Harvey, 2002; Parkinson et al., 2002). The Bayesian 10x tree placed the earliest New World phylogenetic split between a clade ($Pp = 100$) including the Middle and South American bothropoid genera (*Atropoides*, *Cerrophidion*, *Porthidium*, *Bothrocophias*,

Bothrops, *Bothriopsis*) and a weakly supported clade ($Pp = 64$) containing the remaining temperate and tropical New World genera (Fig. 16).

Morphological and molecular studies have found strong support for the monophyly of the primarily temperate genera (*Agkistrodon*, *Sistrurus*, and *Crotalus*; e.g., Gutberlet and Harvey, 2002; Parkinson et al., 2002). Although MP and Bayesian analyses under the 1x model did not resolve this temperate clade, this clade was weakly supported ($Pp = 52$) under the 10x MCMC model. Monophyly of *Agkistrodon* and the rattlesnakes (*Sistrurus* and *Crotalus*) was strongly supported by both MP and MCMC analyses. The monophyly of the rattlesnake genera was supported by both MP and MCMC, although *Crotalus* monophyly received weak support (BS = 57, 1x- $Pp = 81$, 10x- $Pp = 75$). Our estimates of *Crotalus* phylogeny differ notably from estimates of Murphy et al. (2002, based only on MP including many of the same sequences as this study), although many deep phylogenetic relationships among *Crotalus* species received weak support under MP and MCMC analyses (Figs. 14, 16). Both MP and MCMC inferred *C. polystictus* to be the sister taxon to the remaining *Crotalus* species, instead of *C. ravus* as suggested by Murphy et al. (2002). Other novel relationships in our trees include the early divergence of *C. cerastes*, and the placement of *C. enyo* as the sister taxon to a clade containing *C. molossus*, *C. basiliscus*, *C. unicolor*, *C. durissus*, and *C. "vegrandus"* (Fig. 16; rather than nested within it). Despite the inclusion of nearly all *Crotalus* species by Murphy et al. (2002), and in this study, our understanding of relationships among rattlesnakes remains incomplete.

Several molecular studies have supported a clade comprising the primarily Middle American genera *Porthidium*, *Atropoides*, and *Cerrophidion* (Castoe et al., 2003, 2005; Parkinson, 1999; Parkinson et al., 2002), although studies incorporating morphological data

disagree (Gutberlet and Harvey, 2002; Werman, 1992; see also Gutberlet and Harvey, 2004). These Middle American genera formed a strongly supported clade (BS = 96, Pp = 100) inferred as the sister group to a clade comprising the South American genera *Bothrocophias*, *Bothrops*, and *Bothriopsis* (as in Castoe et al., 2005; Parkinson et al., 2002). Within the Middle American group, *Atropoides* appeared paraphyletic (BS = 72, $1x-Pp$ = 73, Pp = 78) with respect to *Cerrophidion* and *Porthidium*, with *A. picadoi* distantly related to other *Atropoides* species (Fig. 16). Based on results of several studies, the phylogenetic status of *Atropoides* appears to be a difficult problem to solve with molecular data (Castoe et al., 2003, 2005; Kraus et al., 1996; Parkinson, 1999; Parkinson et al., 2002). A recent study using two mitochondrial gene sequences (ND4 and cyt-b) for a large sample of Middle American pitvipers did resolve *Atropoides* monophyly with moderate support (Castoe et al., 2005), as had been found by studies based on morphology (Gutberlet and Harvey, 2002) and morphology plus allozymes (Werman, 1992). This example demonstrates the potential impact of taxon sampling and inclusion of morphological characters on estimating pitviper phylogeny.

As the sister group to Middle American pitvipers in all analyses, the South American bothropoid genera formed a strongly supported clade (BS = 100, Pp = 100) with *Bothrocophias* estimated to be the sister taxon to a clade containing a monophyletic (BS = 100, Pp = 100) *Bothriopsis* and a paraphyletic *Bothrops* grouping. The problem of the recognition of *Bothriopsis*, rendering *Bothrops* paraphyletic, has been noted by many studies (e.g., Gutberlet and Harvey, 2002; Gutberlet and Campbell, 2001; Parkinson, 1999; Salomão et al., 1997; Wüster et al., 2002), with some suggesting that *Bothriopsis* should not be recognized (e.g., Salomão et al., 1997; Wüster et al., 2002). Currently, *Bothrops* contains a large and diverse assemblage

(around 40 species; Campbell and Lamar, 2004) of primarily South American pitvipers, and some have argued that the genus *Bothriopsis* should be retained and *Bothrops* be subdivided to rectify the current paraphyly of the genus. The subdivision of *Bothrops* is most consistent with recent trends in pitviper systematics characterized by the recognition of genera that include restricted numbers of ecologically and morphologically similar species, rather than recognition of genera including a broad diversity and large number of species (e.g., Gutberlet and Campbell, 2001; Campbell and Lamar, 1992; Malhotra and Thorpe, 2004). Neither this study, nor previous studies, have sufficiently sampled *Bothrops* species to the extent that new generic allocations from within *Bothrops* are obvious. Our results do suggest, however, that subdivision of *Bothrops* may be accomplished by recognition of at least the three major groups receiving strong support throughout our analyses, including: 1) *B. ammodytoides*, *B. cotiara*, and *B. alternatus*, 2) *B. jararacussu*, *B. atrox*, and *B. asper*, and 3) *B. insularis*, *B. erythromelas*, and *B. diporus* clades (see also Salomão et al., 1997, 1999; Parkinson, 1999; Parkinson et al., 2002; Werman, 1992; Wüster et al., 2002). The challenge of placing unsampled species within these groups, and confirming that these three groups are monophyletic, needs to be confronted before a valid taxonomy can be proposed (see also Gutberlet and Harvey, 2004).

Studies incorporating morphological data have inferred *Ophryacus* to be the sister taxon to *Bothriechis* (Gutberlet, 1998; Gutberlet and Harvey, 2002; Werman, 1992), although no DNA-sequence-based evidence has supported this relationship (Kraus et al., 1996; Parkinson, 1999; Parkinson et al., 2002; see discussion in Gutberlet and Harvey, 2004). Our phylogenies place *Ophryacus* in a clade with *Lachesis* with weak support (BS < 50, 10x model *Pp* = 51). It is interesting to note that the 10x MCMC analyses showed decreased *Pp* support for this

relationship compared to the 1x model ($Pp = 78$), vaguely suggesting convergence of the 10x model on trees that are more in agreement with morphological studies (that reject the existence of this clade). Neither MP nor MCMC results resolved the sister lineage to *Bothriechis*, but both supported monophyly of the genus ($BS = 66$, $Pp = 100$).

Future directions for pitviper systematics

Over thirty years of intense research on pitviper systematics, including works by numerous authors, have produced a phylogeny that is nearing resolution and a current taxonomy that is approaching stability. Sampling of molecular phylogenetic characters has, to date, been largely restricted to mitochondrial gene data, except for studies restricted to particular groups (Giannasi et al., 2001; Creer et al., 2003). Although mitochondrial gene sequences provide a large number of variable characters, homoplasy due to the high divergence of mitochondrial sequences probably substantially hinders estimates of deep relationships among pitvipers. Sequences of nuclear genes may hold valuable synapomorphies required to solidify estimates of relationships at deeper nodes that are not confidently resolved in this study. Additionally, no studies have combined morphological and molecular data to estimate pitviper relationships. These future directions have the potential for establishing robust synapomorphic evidence for relationships, particularly at the inter-generic level, that comprise a majority of the currently outstanding questions in pitviper phylogeny and systematics.

References

- Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Second International Symposium on Information Theory. Akademiai Kiado, Budapest, pp. 673–681.
- Akaike, H., 1974. A new look at the statistical model identification. *IEEE Trans. Aut. Control* 19, 716–723.
- Alfaro, M.E., Zoller, S., Lutzoni, F., 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. *Mol. Biol. Evol.* 20, 255–266.
- Arévalo, E.S., Davis, S.K., Sites, J.W. Jr., 1994. Mitochondrial DNA sequence divergence and phylogenetic relationships among eight chromosome races of the *Sceloporus grammicus* complex (Phrynosomatidae) in central Mexico. *Syst. Biol.* 43, 387–418.
- Brandley, M.C., Schmitz, A., Reeder, T.W., 2005. Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of scincid lizards. *Syst. Biol.* 54, 373–390.
- Brattstrom, B.H., 1964. Evolution of pit vipers. *Trans. San Diego Soc. Nat. Hist.* 13, 185–268.
- Brown, W.M., George M. Jr., Wilson, A.C., 1979. Rapid evolution of animal mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* 76, 1967–1971.
- Buckley, T.R., 2002. Model misspecification and probabilistic tests of topology: evidence from empirical data sets. *Syst. Biol.* 51, 509–523.
- Burger, W.L., 1971. Genera of pitvipers. Ph.D. Dissertation, University of Kansas, Lawrence, KS.

- Caccone, A., Milinkovitch, M.C., Sbordoni, V., Powell, J.R., 1997. Mitochondrial DNA rates and biogeography in European newts (genus *Euproctus*). *Syst. Biol.* 46, 126–144.
- Campbell, J.A., Lamar, W.W., 1989. *The Venomous Reptiles of Latin America*. Cornell University Press, Ithaca, NY.
- Campbell, J.A., Lamar, W.W., 1992. Taxonomic status of miscellaneous neotropical viperids with the description of a new genus. *Occ. Papers Texas Tech. Univ.* 153, 1–31.
- Campbell, J.A., Lamar, W.W., 2004. *The Venomous Reptiles of the Western Hemisphere*. Cornell University Press, Ithaca, NY.
- Castoe, T.A., Chippindale, P.T., Campbell, J.A., Ammerman, L.A., Parkinson, C.L., 2003. The evolution and phylogeography of the Middle American jumping pitvipers, genus *Atropoides*, based on mtDNA sequences. *Herpetologica* 59, 421–432.
- Castoe, T.C., 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.C., Sasa, M., Parkinson, C.L., 2005. Modeling nucleotide evolution at the mesoscale: The phylogeny of the Neotropical pitvipers of the *Porthidium* group (Viperidae: Crotalinae). *Mol. Phylogent. Evol.* 37, 881–898.
- Creer, S., Malhotra, A., Thorpe, R., 2003. Assessing the phylogenetic utility of four mitochondrial genes and a nuclear intron in the Asian pit viper genus, *Trimeresurus*: separate, simultaneous, and conditional data combination analyses. *Mol. Biol. Evol.* 20, 1240–1251.
- 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.

- Douady, C.J., Delsuc, F., Boucher, Y., Doolittle, W.F., Douzery, E.J.P., 2003. Comparison of Bayesian and maximum likelihood bootstrap measures of phylogenetic reliability. *Bol. Biol. Evol.* 20, 248–254.
- Erixon, S.P., Britton, B., Oxelman, B., 2003. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Syst. Biol.* 52, 665–673.
- Faith, J.J., Pollock, D.D., 2003. Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. *Genetics* 165, 735–745.
- Felsenstein, J., 1985. Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* 39, 783–791.
- Giannasi, N., Malhotra, A., Thorpe, R., 2001. Nuclear and mtDNA phylogenies of the *Trimeresurus* complex: implications for the gene versus species tree debate. *Mol. Phylogenet. Evol.* 19, 57–66.
- Gloyd, H.K., Conant, R., 1990. Snakes of the *Agkistrodon* complex: a monographic review. *SSAR Contributions to Herpetology* 6, Ithaca, NY.
- Graybeal, A., 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst. Biol.* 47, 9–17.
- Gumprecht, A., Tillack, F., 2004. Proposal for a replacement name of the snake genus *Ernia* Zhang, 1993. *Russ. J. Herpetol.*, 11, 73–76.
- Gutbierlet, R.L. Jr., 1998. The phylogenetic position of the Mexican black-tailed pitviper (*Squamata: Viperidae: Crotalinae*). *Herpetologica* 54, 184–206.

- Gutberlet, R.L. Jr., Campbell, J.A., 2001. Generic recognition for a neglected lineage of South American pitvipers (Squamata: Viperidae: Crotalinae), with the description of a new species from the Colombian Chocó. *Am. Mus. Novit.* 3316, 1–15.
- Gutberlet, R.L. Jr., Harvey, M.B., 2002. Phylogenetic relationships of New World pitvipers as inferred from anatomical evidence. In: Schuett, G.W., Höggren, M., Douglas, M.E., Greene, H.W. (Eds.), *Biology of the Vipers*. Eagle Mountain Publishing, Salt Lake City, UT, pp. 51–68.
- Gutberlet, R.L. Jr., Harvey, M.B., 2004. The evolution of New World venomous snakes. In: Campbell, J.A., Lamar, W.W., *The Venomous Reptiles of the Western Hemisphere*. Cornell University Press, Ithaca, NY, pp. 634–682.
- Hasegawa, M., Kishino, H., Yano, T., 1985. Dating of the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* 21, 160–174.
- Hillis, D.M., 1996. Inferring complex phylogenies. *Nature* 383, 130–131.
- Hillis, D.M., 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Syst. Biol.* 47, 3–8.
- Hoge, A.R., Romano-Hoge, S.A.R.W., 1981 [dated 1979]. Poisonous snakes of the World. Part I. Check list of the pitvipers: Viperioidea, Viperidae, Crotalinae. *Mem. Inst. Butantan.* 42/43, 179–310.
- Hoge, A.R., Romano-Hoge, S.A.R.W., 1983 [dated 1981]. Notes on micro and ultrastructure of "oberhauschen" in Viperioidea. *Mem. Inst. Butantan.* 44/45, 81–118.
- Huelsenbeck, J.P., 1995. The performance of phylogenetic methods in simulation. *Syst. Biol.* 44, 17–48.

- Huelsenbeck, J.P., 1997. Is the Felsenstein zone a fly trap? *Syst. Biol.* 46, 69–74.
- Huelsenbeck, J.P., Crandall, K.A., 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. *Ann. Rev. Ecol. Syst.* 28, 437–466.
- Huelsenbeck, J.P., Larget, B., Miller, R., Ronquist, F., 2002. Potential applications and pitfalls of Bayesian inference of phylogeny. *Syst. Biol.* 51, 673–688.
- Huelsenbeck, J.P., Larget, B., Alfaro, M.E., 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. *Mol. Biol. Evol.* 21, 1123–1133.
- Huelsenbeck, J.P., Rannala, B., 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Syst. Biol.* 53, 904–913.
- Kass, R.E., Raftery, A.E., 1995. Bayes factors. *J. Am. Stat. Assoc.* 90, 773–795.
- Knight, A., Densmore, L.D., Rael, E.D., 1992. Molecular systematics of the *Agkistrodon* complex. In: Campbell, J.A., Brodie, E.D. Jr. (Eds.), *Biology of the Pitvipers*. Selva, Tyler, TX. pp. 49–70.
- Kraus, F., Mink, D.G., Brown, W.M., 1996. Crotaline intergeneric relationships based on mitochondrial DNA sequence data. *Copeia* 1996, 763–773.
- Larget, B., Simon, D.L., 1999. Markov chain Monte Carlo algorithms for the Bayesian analysis of phylogenetic trees. *Mol. Biol. Evol.* 16, 750–759.
- Lemmon, A.R., Moriarty, E.C., 2004. The importance of proper model assumption in Bayesian phylogenetics. *Syst. Biol.* 53, 265–277.
- Leaché, A.D., Reeder, T.W., 2002. Molecular systematics of the eastern fence lizard (*Sceloporus undulatus*): a comparison of parsimony, likelihood, and Bayesian approaches. *Syst. Biol.* 51, 44–68.

- Malhotra, A., Thorpe, R.S., 2000. A phylogeny of the *Trimeresurus* group of pit vipers: New evidence from a mitochondrial gene tree. *Mol. Phylogenet. Evol.* 16, 199–211.
- Malhotra, A., Thorpe, R.S., 2004. A phylogeny of four mitochondrial gene regions suggests a revised taxonomy for Asian pitvipers (*Trimeresurus* and *Ovophis*). *Mol. Phylogenet. Evol.* 32, 83–100.
- McDiarmid, R.W., Campbell, J.A., Touré, T.A., 1999. *Snake Species of the World: A Taxonomic and Geographical Reference*, Vol. 1. The Herpetologists' League, Washington, D.C.
- Murphy, R.W., Fu, J., Lathrop, A., Feltham, J.V., Kovac, V., 2002. Phylogeny of the rattlesnakes (*Crotalus* and *Sistrurus*) inferred from sequences of five mitochondrial genes. In: Schuett, G.W., Höggren, M., Douglas, M.E., Greene, H.W. (Eds.), *Biology of the Vipers*. Eagle Mountain Publishing, Salt Lake City, UT, pp. 69–92.
- Nicholas, K.B., Nicholas, H.B. Jr., 1997. GeneDoc: a tool for editing and annotating multiple sequence alignments. Available from <http://www.cris.com/~Ketchup/genedoc.shtml>.
- Nylander, J.A.A., 2004. MrModeltest v2. Program distributed by the author. Evolutionary Biology Centre, Uppsala 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.
- Pagel, M., Meade, A., 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character data. *Syst. Biol.* 53, 571–581.
- Parkinson, C.L., 1999. Molecular systematics and biogeographical history of pitvipers as determined by mitochondrial ribosomal DNA sequences. *Copeia* 1999, 576–586.

- Parkinson, C.L., Campbell, J.A., Chippindale, P.T., 2002. Multigene phylogenetic analyses of pitvipers; with comments on the biogeographical history of the group. In: Schuett, G.W., Höggren, M., Douglas, M.E., Greene, H.W. (Eds.), *Biology of the Vipers*. Eagle Mountain Publishing, Salt Lake City, UT, pp. 93–110.
- Parkinson, C.L., Moody, S.M., Ahlquist, J.E., 1997. Phylogenetic relationships of the “Agkistrodon Complex” based on mitochondrial DNA sequence data. In: Thorpe, R.S., Wüster, W., Malhotra, A. (Eds.), *Venomous snakes: ecology, evolution and snakebite*. The Zoological Society of London, Clarendon Press, Oxford, UK, pp. 63–78.
- Poe, S., 1998. Sensitivity of phylogeny estimation to taxonomic sampling. *Syst. Biol.* 47, 18–31.
- Poe, S., Swofford, D.L., 1999. Taxon sampling revisited. *Nature* 398, 299–300.
- Posada, D., Buckley, T.R., 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.
- Rambaut, A., Drummond, A.J., 2003. Tracer. Version 1.0.1. Available from <http://evolve.zoo.ox.ac.uk/>.
- Rannala, B., 2002. 2002. Identifiability of parameters in MCMC Bayesian inference of phylogeny. *Syst. Biol.* 51, 754–760.
- Reeder, T.W., 2003. A phylogeny of the Australian Sphenomorphus group (Scincidae: Squamata) and the phylogenetic placement of the crocodile skinks (Tribolonotus): Bayesian approaches to assessing congruence and obtaining confidence in maximum likelihood inferred relationships. *Mol. Phylogenet. Evol.* 4, 203–222.

- Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572–1574.
- Russell, F.E., 1980. Snake Venom Poisoning. Lippincott, Philadelphia, PA.
- Sakamoto, Y., Ishiguro, M., Kitagawa, G., 1986. Akaike Information Criterion Statistics. Springer, NY.
- Salisbury, B.A., Kim, J., 2001. Ancestral state estimation and taxon sampling density. *Syst. Biol.* 50, 557–564.
- Salomão, M.G., Wüster, W., Thorpe, R.S., Butantan-British Bothrops Systematics Project, 1999. MtDNA phylogeny of Neotropical pitvipers of the genus *Bothrops* (Squamata: Serpentes: Viperidae). *Kaupia* 8, 127–134.
- Salomão, M.G., Wüster, W., Thorpe, R.S., Touzet, J.M., Butantan-British Bothrops Systematics Project, 1997. DNA evolution of South American pitvipers of the genus *Bothrops*. In: Thorpe, R.S., Wüster, W., Malhotra, A. (Eds.), *Venomous snakes: ecology, evolution and snakebite*. The Zoological Society of London, Clarendon Press, Oxford, UK, pp. 89–98.
- Simmons, M.P., Pickett, K.M., Miya, M., 2004. How meaningful are Bayesian support values? *Mol. Biol. Evol.* 21, 188–199.
- Swofford, D.L., 2002. PAUP*: Phylogenetic Analysis Using Parsimony (and Other Methods). Version 4.0b10. Sinauer, Sunderland, MA.
- Sullivan, J., Swofford, D.L., 2001. Should we use model-based methods for phylogenetic inference when we know assumptions about among-site rate variation and nucleotide substitution pattern are violated? *Syst. Biol.* 50, 723–729.

- Suzuki, Y., Glazko, G.V., Nei, M., 2002. Overcredibility of molecular phylogenies obtained by Bayesian phylogenetics. *Proc. Nat. Acad. Sci. U.S.A.* 99, 16138–16143.
- Tavaré, S., 1996. Some probabilistic and statistical problems on the analysis of DNA sequences. In: Miura, R.M. (Ed.), *Some mathematical questions in biology—DNA sequence analysis*. American Math Society, Providence, RI, pp. 57–86.
- Vidal, N., Lecointre, G., 1998. Weighting and congruence: a case study based on three mitochondrial genes in pitvipers. *Mol. Phylogenet. Evol.* 9, 366–374.
- Vidal, N., Lecointre, G., Vié, J.C., Gasc, J.P., 1997. Molecular systematics of pitvipers: Paraphyly of the Bothrops complex. *C. R. Acad. Sci. Paris, Science de la vie* 320, 95–101.
- Vidal, N., Lecointre, G., Vié, J.C., Gasc, J.P., 1999. What can mitochondrial gene sequences tell us about intergeneric relationships of pitvipers? *Kaupia* 8, 113–126.
- Werman, S., 1992. Phylogenetic relationships of Central and South American pitvipers of the genus *Bothrops* (sensu lato): cladistic analyses of biochemical and anatomical characters. In: Campbell, J.A., Brodie, E.D. Jr. (Eds.), *Biology of the Pitvipers*. Selva, Tyler, TX. pp. 21–40.
- Wiens, J.J., 1998. Combining data sets with different phylogenetic histories. *Syst. Biol.* 47, 568–581.
- Wilcox, T.P., Zwickl, D.J., Heath, T.A., Hillis, D.M., 2002. Phylogenetic relationships of the dwarf boas and a comparison of Bayesian and bootstrap measures of phylogenetic support. *Mol. Phylogenet. Evol.* 25, 361–271.
- Wilgenbusch, J., de Queiroz, K., 2000. Phylogenetic relationships among the phrynosomatid sand lizards inferred from mitochondrial DNA sequences generated by heterogeneous evolutionary processes. *Syst. Biol.* 49, 592–612.

- Wüster, W., da Graca Salomão, M., Quijada-Mascareñas, J.A., Thorpe, R.S., Butantan-British Bothrops Systematics Project, 2002. Origin and evolution of the South American pitviper fauna: evidence from mitochondrial DNA sequence data. In: Schuett, G.W., Höggren, M., Douglas, M.E., Greene, H.W. (Eds.), *Biology of the Vipers*. Eagle Mountain Publishing, Salt Lake City, UT, pp. 111–128.
- Yang, Z., 1996. Maximum likelihood models for combined analyses of multiple sequence data. *J. Mol. Evol.* 42, 587–596.
- Zhang, F.J., 1998. Description of the distinct pitviper of genus *Ermia* (Serpentes: Viperidae) of China. *Russ. J. Herpetol.* 5, 83–84.
- Ziegler, T., Herrmann, H.-W., David, P., Orlov, N.L., Pauwels, S.G., 2000. *Triceratolepidophis sieversorum*, a new genus and species of pitviper (Reptilia: Serpentes: Viperidae: Crotalinae) from Vietnam. *Russ. J. Herpetol.* 7, 199–214.

CHAPTER 5 – COMPARATIVE MITOCHONDRIAL GENOMICS OF SNAKES: EXTRAORDINARY SUBSTITUTION RATE DYNAMICS AND FUNCTIONALITY OF THE DUPLICATE CONTROL REGION

Introduction

The vertebrate mitochondrial genome has been an important model system for studying molecular evolution, organismal phylogeny, and genome structure. The versatility and prominence of vertebrate mitochondrial genomes stems from their compactness and manageable size for sequencing and analysis, well-characterized replication and transcription processes (e.g., (Clayton, 1982; Fernandez-Silva et al., 2003; Shadel and Clayton, 1997; Szczesny et al., 2003); see also (Holt and Jacobs, 2003; Reyes et al., 2005; Yang et al., 2002)), and the diversity of protein and structural RNA genes that they encode. Vertebrate mitochondrial genomes generally lack recombination and have a conserved genome structure, although instances of intramolecular recombination have been proposed (Piganeau et al., 2004; Tsaousis et al., 2005), and there are numerous examples of structural rearrangements (Cooper et al., 2001; Mindell et al., 1998; Sankoff et al., 1992). Despite extensive molecular studies, little is known regarding the ways in which genome architecture might affect the various aspects of genome function and evolution (including replication, transcription, and function of proteins and RNAs). Nevertheless, patterns linking mitochondrial genome structure, function, and nucleotide evolution have begun to emerge (Krishnan et al., 2004a; Krishnan et al., 2004b; Raina et al., 2005).

The mitochondrial genome (mtDNA) has long been believed to replicate asymmetrically (Clayton, 1982), which creates a substantial difference in mutation rates and nucleotide composition biases between strands (Bielawski and Gold, 2002; Jermiin et al., 1995; Perna and Kocher, 1995a, b; Tanaka and Ozawa, 1994). During replication under the classical model, the synthesis of the nascent heavy strand initiates at the origin of heavy strand replication (O_H), within the control region (CR). This has been extensively reviewed elsewhere (Bielawski and Gold, 2002; Faith and Pollock, 2003), but in brief, after two thirds of the nascent heavy strand is synthesized, the synthesis of the nascent light strand starts at the origin of light strand replication (O_L), a short secondary structure forming segment located within the tRNA cluster (the WANCY region) between the NADH dehydrogenase subunit 2 (ND2) and Cytochrome C oxidase subunit 1 (COX1) genes. The strand-asymmetric replication mechanism has been thought to expose different regions of the parental heavy strand to varying amounts of time in the single-stranded state during replication (D_{ssH} ; (Tanaka and Ozawa, 1994)), depending on the distances of the regions from the O_H and O_L . Variation in this strand-asymmetric mutation processes appears to have contributed substantially to variation in substitution rates among genes (Bielawski and Gold, 2002; Faith and Pollock, 2003; Raina et al., 2005).

Controversy has recently arisen concerning the classical mitochondrial replication mechanism, mostly concerning the asymmetry of the process, the role of the putative origin of light strand replication, and whether the replicating DNA spends substantial amounts of time single-stranded (Reyes et al., 2005; Yang et al., 2002; Yasukawa et al., 2005). Although the newly proposed models of replication are directly at odds with the genetic data, one of us has hypothesized (Pollock, *in review*) that most of the biochemical and genetic data is compatible

with a reconciled model of mitochondrial replication, which retains most critical features of the classical model except for single strandedness. Regardless of the final reconciliation, to take a neutral position on the biochemical issue of single-strandedness we will refer to the time that a gene or nucleotide is predicted to spend in an asymmetric mutagenic state (T_{AMS}), rather than the predicted duration of time that the heavy strand spends single-stranded (D_{SSH}); the calculation is, however, identical to that for D_{SSH} (Faith and Pollock, 2003; Reyes et al., 1998; Tanaka and Ozawa, 1994).

Cytosine →Uracil deaminations are common in single-stranded DNA, while Adenine → Hypoxanthine deaminations are less common (Frederico et al., 1990; Impellizzeri et al., 1991). These two deaminations lead to mutations (Cytosine→Thymine and Adenine→Guanine, or C→T and A→G) that appear to account for most of the asymmetry in synonymous substitutions found in vertebrate mtDNA (Bielawski and Gold, 1996; Faith and Pollock, 2003; Frank and Lobry, 1999; Krishnan et al., 2004a; Krishnan et al., 2004b; Raina et al., 2005; Rand and Kann, 1998; Reyes et al., 1998). C→T and A→G mutations on the heavy strand during replication apparently lead respectively to G→A and T→C substitutions (and G and T deficiencies) on the light strand. Most protein-coding genes (all but ND6) use the heavy strand as a template; thus, the mutation biases observed in the light strand parallel the biases in most protein-coding gene transcripts. Faith and Pollock (Faith and Pollock, 2003) found that, in vertebrates, T→C light strand substitutions at four-fold and two-fold redundant 3rd codon positions increase linearly with increasing T_{AMS} . In contrast, G→A light strand substitutions increase rapidly but quickly reach a

maximal level. Consequently, T→C substitutions and the resultant C/T nucleotide frequency gradient are good predictors of T_{AMS} .

The mitochondrial genomes of snakes contain a number of qualities and structural features that are unusual among the vertebrates. Snake mitochondrial genomes have elevated evolutionary rates and contain truncated tRNAs (Dong and Kumazawa, 2005; Kumazawa et al., 1998). All snake species sampled to date, except the scolecophidian snake *Leptotyphlops dulcis*, have a duplicated control region (CR2) between NADH dehydrogenase subunit 1 (ND1) and subunit 2 (ND2), in addition to a control region (CR1) adjacent to 5'-end of the 12s rRNA, as it is in other vertebrates. These two control regions appear to undergo concerted evolution that acts to homogenize the nucleotide sequence of each duplicate copy within a given genome (Dong and Kumazawa, 2005; Kumazawa et al., 1996, 1998). The functionality of these two control regions in transcription and initiation of heavy strand replication is not clear, but since the nucleotide sequence of each is nearly identical, any functional features that are not dependent on surrounding sequences should be similar. In contrast, recent evidence suggest that initiation of heavy strand replication may be distributed across a broad zone, including cytochrome b (CytB) and NADH dehydrogenase subunit 6 (ND6) (Reyes et al., 2005), indicating that CR2 may not function as effectively in this role.

A number of interesting questions arise that might be addressed through comparative analysis, including: (1) does one or the other, or do both control regions function as origins of heavy strand DNA synthesis? (2) does the altered genome structure affect patterns of snake mtDNA molecular evolution? (3) when during snake evolution did various features arise, and do particular features appear to coincide? (4) do patterns of molecular evolution vary at different

depths of phylogeny? and (5) is there any evidence or plausible rationale for selection as a causative agent in generating these differences in genomic structure and molecular evolutionary patterns?

To investigate outstanding questions regarding snake mitochondrial genome evolution, structure, and function, we analyzed a dataset consisting of three new complete snake mitochondrial genomes together with eight previously published snake mitochondrial genomes, and 42 other vertebrate mitochondrial genomes for comparative purposes. The new snake genomes were obtained from *Pantherophis slowinskii* (a corn snake from Louisiana; previously *Elaphe guttata*), and from *Agkistrodon piscivorus* (the cottonmouth or water moccasin; one specimen from Florida and the other from Louisiana). These genomes were targeted in order to increase the phylogenetic density of sampling in alethinophidian snakes, which appear to show among the most interesting mitochondrial genome evolutionary patterns based on previous studies (Kumazawa et al., 1996, 1998).

The research presented here constitutes an exploratory comparative study of genomic architecture and substitution rate variation among genes and among lineages. Given the large amount and diversity of data in this study, we have deferred to a future study all analysis of site-specific selection via dN/dS ratios and its relation to details of protein structure and function. Although this dataset does not (and was not designed to) resolve any major questions in squamate phylogeny, we were able to map onto the phylogeny changes in genome size, gene organization, tRNA size and structure, and dynamics of gene-specific evolutionary rates, and to conduct detailed comparisons of mtDNA evolution at the intraspecific level with the two *A. piscivorus* samples. We also used predictions based on the asymmetrical pattern of mitochondrial

genome replication (and corresponding nucleotide substitution and frequency biases) to make a preliminary assessment of control region functionality.

Material and Methods

Sampling, sequencing and annotation

Several complete mitochondrial genomes of snakes have been published, and previous snake mtDNA sampling has targeted divergent lineages (e.g., no family of snakes is represented by multiple examples). To complement this broader sampling, we sequenced complete mtDNAs of two species, each of which representing the second taxon within a family from which a complete mtDNA was already available. Also, we sequenced two mtDNAs from divergent populations of a single species. Thus, our taxonomic sampling was designed to complement existing snake mtDNA sequences by providing comparative genomic data at shallower levels of phylogenetic divergence. Such sampling is essential to more accurately assess details concerning the process of evolution.

DNA was extracted from vouchered specimens available at the Louisiana State University Museum of Natural Science (LSUMZ) and the University of Central Florida (CLP). The *A. piscivorus* (cottonmouth or water moccasin; Viperidae) specimens were from Louisiana, USA (LSUMZ-17943) and from Florida, USA (CLP-73). We will refer to these as Api1 (Louisiana specimen) and Api2 (Florida specimen). The *P. slowinskii* (corn snake; Colubridae)

specimen was from Louisiana, USA (LSUMZ- H-2036). The genus *Pantherophis* (Utiger et al., 2002) was recently erected to contain a clade of species formerly allocated to *Elaphe*. The species *P. slowinskii* was formerly considered *Pantherophis (Elaphe) guttatus*, and was recently recognized as a distinct species (Burbrink, 2002). The *P. slowinskii* specimen used as a source of DNA in this study is the type specimen for the species. Since no genera in this study are represented by multiple species, for mnemonic convenience we will hereafter primarily use the names of genera to identify sources of mtDNA genomes. Details of molecular laboratory methods (e.g., PCR, cloning, sequencing), genome annotation (Slack et al., 2003), and accession numbers are provided below.

Total DNA was isolated from frozen (-80C) liver tissue of *Api2* using the Qiagen DNeasy extraction kit and protocol (Qiagen Inc.). Using the Expand Long Template PCR system (Roche Molecular Biochemicals), the mitochondrial genome was amplified in six overlapping fragments with 12 primers (Table S1). In addition, several smaller fragments were also amplified using the BIO-X-ACT Short PCR kit (Bioline) to fill-in otherwise inadequately sequenced regions. Cycling conditions followed the manufacturers' suggestions, with annealing temperatures between 50°C and 55°C, and for 35 cycles.

Positive PCR products were electrophoretically separated and excised from agarose gels, followed by purification using the GeneCleanIII kit (BIO101). Purified PCR products were cloned using either the TopoTA or TopoXL cloning kits (Invitrogen). Plasmids containing amplification fragments were isolated and purified using QIAprep Spin Miniprep kits (Qiagen) and sequenced using M13 primers (flanking the cloning site in the Topo vectors), an array of internal primers (details available upon request), and the CEQ Dye Terminator Cycle Sequencing

Quick Start Kit (Beckman-Coulter), and were run on a Beckman CEQ8000 automated sequencer according to the manufacturers' protocols.

Total DNA was extracted from *Api1* using a High Pure PCR Template Preparation Kit (Roche), and amplified into two long overlapping fragments, 8kb and 9kb, using the Expand Long Template PCR Amplification System (Roche) and 4 primers (Table 15). These two fragments overlap in the 16s RNA and COIII genes. Conditions followed the manufacturer's recommendations, with annealing temperatures of 58.4°C (9kb fragment), and 52.2°C (8kb fragment). After electrophoresis as above, PCR products were purified using the Agarose Gel DNA Purification kit (Mo Bio Laboratory), followed by end phosphorylation, ligation, and shearing in a nebulizer (Invitrogen). Fragments ranging from 1.5-3kb were purified from 0.8% agarose gels using QIAquick Gel Extraction Kit (Qiagen), cloned into pPCR-Script Amp SK(+) vector (Stratagene PCR-Script Amp Cloning Kit), and transformed into XL-10 Gold Kan ultracompetent cells (Stratagene). Bacterial clones containing plasmids with snake mitochondrial inserts were amplified using M13 primers, and the products were purified by QIAquick PCR Purification Kit and sequenced using T3 primer and Big Dye Terminator Sequence Master (PE Biosystems) using standard protocols. The reactions were purified on DyeEx columns (Qiagen), and the DNA sequence was determined using an ABI 3700 automated sequencer.

Total DNA from *Pantherophis* was extracted and amplified using the same protocol and reagents as for *Api1*, but with a different set of four primers (Table 15) yielding 12.5 Kb and 4.5 Kb fragments. These two fragments overlap in the CytB and 16s rRNA genes, and were sequenced following the same protocol as used for *Api1*, with additional internal primers.

Table 15. Table S1 - Primer sets used to amplify mitochondrial genome fragments in this study.

Primer Name	Primer sequence (5' – 3')	Source
<i>Agkistrodon piscivorus</i> - <i>Api2</i> amplification primers		
L2932	MYTGGTGCCAGCCGCCGCGG	This study
tRNATrpR	GGCTTTGAAGGCTMCTAGTTT	R. Lawson, unpub.
ND1L	CTATCCCCATCATAGCMC	This study
ND2H	TCGGGGTATGGGCCCC	This study
LRattle	ACTCTAACGCTCCTAACCTGAC	K. Zamudio, unpub.
Leu	CCAACACCTVTTCTGATT	Arévalo et al. 1994
L6929	CCAACACCTVTTCTGATT	This study
ND4CP200	ARATTGYRGCTRCTACTARGCC	This study
ND4	CACCTATGACTACCAAAAGCTCATGTAGAAGC	Arévalo et al. 1994
AtrCB3	TGAGAAGTTTTTCYGGGTCRTT	Parkinson et al. 2002
Gludg	TGACTTGAARAACCAAYCGTTG	Parkinson et al. 2002
H3059	CCGGTCTGAACTCAGATCACGT	This study
<i>Agkistrodon piscivorus</i> - <i>Api1</i> amplification primers		
DPFB002R	AGTGGTCAWGGGCTKGGGACTA	This study
DPFB0013F	CGGCCGCGGTATYCTAACCGTGCAAAG	This study
DPFB001F	TAGTAGACCCMAGCCCWTGACCACT	This study
DPFB0021R	CTGATCCAACATCGAGGTCGTAAACC	This study
<i>Pantherophis slowinskii</i> amplification primers		
DPAL007	CTACGTGATCTGAGTTCAGACC	This study
DPFB007	CTCAGAAKGATATYTGCCYCATGG	This study
DPFB006	CCATGRGGACARATATCMTTCTGAG	This study
DPAL006	CTCCGGTCTGAACTCAGATCAC	This study

Most tRNAs in the raw genome sequences were detected using tRNAscan (Lowe et al. 1997), followed by manual verification. The tRNAs not identified by tRNAscan were identified by their position in the genome and folded manually based on homology. The tRNAs were then used to identify approximate boundaries of protein coding genes, control region, and ribosomal RNAs. Final boundaries of protein coding genes were set based on position of the most plausible first start and last stop codons in each region, including non-canonical signal codons known to operate in vertebrate mitochondrial genome (Slack et al. 2003). Proteins were also translated to their amino acid sequence, and all amino acid and DNA sequences were compared to the corresponding genes or regions from published snake genomes to verify the annotation.

Phylogenetic and sliding-window analyses

In addition to the three new snake mitochondrial genome sequences, the sequence dataset used included all eight available snake mtDNAs, and 42 additional taxa for comparative purposes, including heavy sampling of birds, mammals (mostly primates), and lizards (species scientific names and access numbers are in Table 16). We limited our sampling of mammalian mtDNAs almost exclusively to primates (and *Bos taurus*) because we were particularly interested in obtaining precise comparative estimates of mutation rates that may otherwise become unreliable when sampling is overly sparse, due to the high rates of mitochondrial genome evolution. Also, focused sampling of primates was incorporated to keep the total number of sequences low enough to facilitate complex likelihood analyses (which would otherwise be computationally unfeasible), and to facilitate comparisons in rates and patterns between snakes and primates (Raina et al., 2005).

Table 16. Table S2 - Complete mitochondrial genomes used in this study, and associated Genbank accession numbers.

Vertebrate Group	Genbank Accession	Taxon	Vertebrate Group	Genbank Accession	Taxon
Amphibians	NC_002756	<i>Mertensiella luschani</i>	Birds	NC_002782	<i>Apteryx haastii</i>
	NC_001573	<i>Xenopus laevis</i>		NC_003128	<i>Buteo buteo</i>
Turtles	NC_000886	<i>Chelonia mydas</i>		NC_002196	<i>Ciconia boyciana</i>
	NC_002073	<i>Chrysemys picta</i>		NC_002197	<i>Ciconia ciconia</i>
	NC_002780	<i>Dogania subplana</i>		NC_002069	<i>Corvus frugilegus</i>
	NC_001947	<i>Pelomedusa subrufa</i>		NC_002784	<i>Dromaius novaehollandiae</i>
Tuatara	NC_004815	<i>Sphenodon punctatus</i>		NC_000878	<i>Falco peregrinus</i>
Lizards	NC_005958	<i>Abronia graminea</i>		NC_001323	<i>Gallus gallus</i>
	NC_005962	<i>Cordylus warreni</i>		NC_000846	<i>Rhea americana</i>
	NC_000888	<i>Eumeces egregius</i>		NC_000879	<i>Smithornis sharpei</i>
	NC_002793	<i>Iguana iguana</i>		NC_002785	<i>Struthio camelus</i>
	NC_005960	<i>Sceloporus occidentalis</i>		NC_002781	<i>Tinamus major</i>
	NC_005959	<i>Shinisaurus crocodilurus</i>		NC_000880	<i>Vidua chalybeata</i>
	AB080275-6	<i>Varanus komodoensis</i>	Mammals	NC_001567	<i>Bos taurus</i>
Snakes	NC_007400	<i>Acrochordus granulatus</i>		NC_002763	<i>Cebus albifrons</i>
	GB_#####	<i>Agkistrodon piscivorus (Api1)</i>		NC_002082	<i>Hylobates lar</i>
	GB_#####	<i>Agkistrodon piscivorus (Api2)</i>		NC_001646	<i>Pongo pygmaeus</i>
	NC_007398	<i>Boa constrictor</i>		NC_001644	<i>Pan paniscus</i>
	NC_007401	<i>Cylindrophis ruffus</i>		NC_001645	<i>Gorilla gorilla</i>
	NC_001945	<i>Dinodon semicarinatus</i>		NC_001807	<i>Homo sapiens</i>
	NC_005961	<i>Leptotyphlops dulcis</i>		NC_001992	<i>Papio hamadryas</i>
	NC_007397	<i>Ovophis okinavensis</i>		NC_002764	<i>Macaca sylvanus</i>
	GB_#####	<i>Pantherophis slowinskii</i>		NC_002811	<i>Tarsius bancanus</i>
	NC_007399	<i>Python regius</i>		NC_004025	<i>Lemur catta</i>
	NC_007402	<i>Xenopeltis unicolor</i>		NC_002765	<i>Nycticebus coucang</i>

Sequences of protein-coding and rRNA genes were aligned using ClustalX (Thompson et al., 1997), followed by manual adjustment. Protein-coding genes were first aligned at the amino acid level, and then the nucleotide sequences were aligned according to the corresponding amino acid alignment. The alignment of rRNAs contained a small number of sites (corresponding to the loop-forming structures of the rRNAs) with ambiguous alignments only among major tetrapod lineages. Since we wanted to compare estimates of mitochondrial gene evolutionary rates and patterns, we chose not to exclude any sites of the alignment. This was also justified by preliminary phylogenetic estimates that suggested the incorporation of these few potentially ambiguous sites did not effect phylogenetic results. The main phylogeny used and presented here was inferred using the concatenated nucleotide sequence of all 13 protein-coding and two rRNA genes by maximum-likelihood (ML) analysis in PAUP 4.0 beta10 (Swofford, 1997). This analysis incorporated the GTR+ Γ +I model of evolution, which was the best-fit model under all criteria in ModelTest (Posada and Crandall, 1998). Estimated ML model parameters were as follows: $r_{AC} = 1.51278$, $r_{AG} = 2.46909$, $r_{AT} = 0.90191$, $r_{CG} = 0.2503$, $r_{CT} = 4.56723$, Γ (alpha shape) = 0.997413, and I (proportion of invariable sites) = 0.19647.

Support for this topology was evaluated in two ways: (1) based on 1000 NJ bootstraps (in PAUP) with ML distances calculated under the same model as above, but with down-weighted synonymous sites to avoid saturation problems (rRNAs relative weight = 5 and 1st, 2nd, and 3rd codon positions relative weights = 4, 5, and 1) and (2) based on Bayesian posterior probability support estimated by conducting two simultaneous independent MCMC runs conducted for 10⁶ generations (with the first 400,000 generations of each run discarded as burn-in) using a GTR+ Γ +I model of evolution (in MrBayes 3.1 (Ronquist and Huelsenbeck, 2003)). The burnin period

was determined by visual assessment of stationarity and convergence of likelihood values between the chains. To analyze nucleotide substitution rate variation in different lineages and different genes, branch length estimates were separately calculated under the GTR+ Γ +I model for different genes (COX1, ND1, ND2, ND4, ND5, CytB) and gene clusters (COX2 + ATP8 + ATP6, and COX3 + ND3 + ND4L; each comprising groups of individually short genes adjacent along the mtDNA) using the ML topology and PAML (Yang, 1997). We also calculated the length of the internal branch (ancestral branch) leading to each of three nominal clades (mammals, snakes, and lizards), and the total branch lengths within each of these clades (species cluster length).

To further analyze fluctuations in nucleotide substitution rates, we conducted sliding window analyses (SWA) on the phylogenetic dataset. The program Hyphy (Pond et al., 2005) was used to estimate branch lengths (estimated numbers of substitutions) for 1000 bp windows. SWA was conducted using the GTR model with global parameter estimation and topological relationships specified based on the ML tree estimate, with a window slide of 200 bp. Based on preliminary trials, the size of the window and slide length were chosen to minimize noise observed with shorter windows, but to allow differentiation of patterns in different regions. To compare patterns of substitution across the mitochondrial genome for select branches or groups of branches, we first divided substitution estimates for each window by the median substitution rate across all windows. Since branch lengths are estimates of $\delta_b t_b$ (the branch-specific substitution rate times divergence time) this procedure estimates a ratio of substitution rates, $\delta_b^w / \delta_b^{\bar{w}}$, where δ_b^w is the branch- and window-specific substitution rate, and $\delta_b^{\bar{w}}$ is the branch-specific substitution rate in the median window. To evaluate whether the windows had

relative rates that were slower or faster than expected, we took the substitution rate ratio from the set of all branches in the non-snakes (NS) as a standard. This was then subtracted from the branch-specific ratio to obtain a “standardized substitution rate”, $\delta_b^w / \delta_b^e - \delta_{NS}^w / \delta_{NS}^e$. When relative rates of substitution are distributed similarly across the mtDNA, in comparison with NS, this standardized rate comparison approaches zero.

tRNA structure

To compare predicted tRNA stabilities, the secondary structures of squamate (snake and lizard) tRNAs were determined under the guidance of the mammalian tRNA cloverleaf structures (Helm et al., 2000) and the tRNAscan program (Lowe and Eddy, 1997), and then used to modify tRNA alignments by hand (tRNA^{Ser} [AGY] was not included in these analyses since it does not form a cloverleaf structure). To determine the relative stabilities of the tRNA secondary structures, we calculated the energy (ΔG) of the cloverleaf structure using the Vienna Package version 1.4 (Hofacker et al., 1994). The minimum energy (ΔG) is the predicted amount of energy (in calories) required to destroy the structure: the lower the energy of the molecules, the more stable its secondary structure.

Analysis of control region functionality

The calculation of T_{AMS} differs depending on whether CR1 or CR2 is functional, but only for the genes that are in between the two control regions, the two rRNAs and ND1 (Table 17).

Table 17. Estimated T_{AMS} values of genes for squamates. Two T_{AMS} values are given for each species of alethinophidian snakes; T_{AMS}^1 is estimated based on the assumption of exclusive CR1 usage, whereas T_{AMS}^2 is estimated based on exclusive CR2 usage. Genes that have alternative T_{AMS} estimates under different CR usage scenarios in alethinophidian mtDNAs are indicated in bold.

Snakes																			
Genes	<i>Agkistrodon</i>		<i>Ovophis</i>		<i>Pantherophis</i>		<i>Dinodon</i>		<i>Acrochordus</i>		<i>Boa</i>		<i>Cylindrophis</i>		<i>Python</i>		<i>Xenopeltis</i>		<i>Leptotyphlops</i>
	T_{AMS}^1	T_{AMS}^2	T_{AMS}^1	T_{AMS}^2	T_{AMS}^1	T_{AMS}^2	T_{AMS}^1	T_{AMS}^2	T_{AMS}^1	T_{AMS}^2	T_{AMS}^1	T_{AMS}^2	T_{AMS}^1	T_{AMS}^2	T_{AMS}^1	T_{AMS}^2	T_{AMS}^1	T_{AMS}^2	T_{AMS}
12s	0.35	1.36	0.34	1.34	0.35	1.35	0.35	1.35	0.35	1.35	0.33	1.33	0.35	1.35	0.36	1.36	0.32	1.32	0.45
16s	0.50	1.51	0.48	1.48	0.50	1.50	0.50	1.49	0.50	1.49	0.47	1.46	0.50	1.49	0.51	1.50	0.46	1.45	0.61
ATP6	0.36	0.36	0.36	0.36	0.36	0.36	0.36	0.36	0.35	0.35	0.33	0.33	0.35	0.35	0.36	0.36	0.33	0.33	0.39
ATP8	0.31	0.31	0.31	0.31	0.31	0.31	0.31	0.31	0.31	0.31	0.29	0.29	0.31	0.31	0.31	0.31	0.29	0.29	0.34
COX1	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.11	0.10	0.10	0.10	0.10	0.11	0.11	0.10	0.10	0.12
COX2	0.26	0.26	0.25	0.25	0.26	0.26	0.26	0.26	0.25	0.25	0.23	0.23	0.25	0.25	0.26	0.26	0.23	0.23	0.28
COX3	0.45	0.45	0.44	0.44	0.45	0.45	0.45	0.45	0.44	0.44	0.41	0.41	0.44	0.44	0.45	0.45	0.41	0.41	0.48
CytB	1.10	1.10	1.08	1.08	1.09	1.09	1.09	1.09	1.07	1.07	1.00	1.00	1.08	1.08	1.10	1.10	1.01	1.01	1.17
ND1	0.64	1.65	0.62	1.62	0.64	1.64	0.64	1.63	0.64	1.64	0.60	1.60	0.64	1.64	0.66	1.66	0.59	1.59	0.77
ND2	0.91	0.91	0.92	0.92	0.91	0.91	0.91	0.91	0.91	0.92	0.92	0.92	0.92	0.92	0.91	0.91	0.92	0.92	0.91
ND3	0.52	0.52	0.51	0.51	0.52	0.52	0.52	0.52	0.51	0.51	0.47	0.47	0.51	0.51	0.52	0.52	0.47	0.47	0.55
ND4	0.66	0.66	0.65	0.65	0.66	0.66	0.66	0.66	0.64	0.64	0.60	0.60	0.65	0.65	0.66	0.66	0.60	0.60	0.70
ND4L	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.56	0.55	0.55	0.51	0.51	0.55	0.55	0.56	0.56	0.52	0.52	0.60
ND5	0.86	0.86	0.85	0.85	0.86	0.86	0.86	0.86	0.84	0.84	0.79	0.79	0.85	0.85	0.86	0.86	0.79	0.79	0.92
ND6	0.99	0.99	0.98	0.98	0.99	0.99	0.99	0.99	0.97	0.97	0.91	0.91	0.98	0.98	1.00	1.00	0.91	0.91	1.06

Lizards						
Genes	<i>Iguana</i>	<i>Eumeces</i>	<i>Sceloporus</i>	<i>Cordylus</i>	<i>Abronia</i>	<i>Shinisaurus</i>
	T_{AMS}	T_{AMS}	T_{AMS}	T_{AMS}	T_{AMS}	T_{AMS}
12s	0.44	0.47	0.46	0.47	0.43	0.45
16s	0.60	0.62	0.62	0.62	0.59	0.60
ATP6	0.37	0.35	0.36	0.36	0.39	0.37
ATP8	0.32	0.31	0.31	0.31	0.33	0.32
COX1	0.11	0.11	0.11	0.11	0.12	0.11
COX2	0.26	0.25	0.26	0.25	0.27	0.26
COX3	0.46	0.44	0.45	0.45	0.48	0.46
CytB	1.15	1.10	1.12	1.11	1.19	1.15
ND1	0.76	0.78	0.77	0.77	0.76	0.76
ND2	0.91	0.91	0.91	0.91	0.91	0.91
ND3	0.54	0.51	0.52	0.52	0.56	0.54
ND4	0.68	0.65	0.67	0.66	0.71	0.68
ND4L	0.58	0.56	0.57	0.56	0.60	0.58
ND5	0.90	0.86	0.88	0.87	0.93	0.90
ND6	1.04	0.99	1.01	1.01	1.08	1.04

Based on previous work, the light strand C/T ratio at synonymous two-fold and fourfold redundant 3rd codon positions is expected to increase linearly with T_{AMS} , so we used this prediction to determine whether there was any evidence for activity of CR1 or CR2 in initiating heavy strand replication. We implemented a slightly modified version of the MCMC approach in (Raina et al., 2005) to estimate the most likely slope and intercept of the C/T ratio gradient depending on the calculated T_{AMS} at every site. We applied these calculations using T_{AMS} from CR1 and CR2, and also separately calculated the slope and intercept for the most likely weighted average T_{AMS} for the two control regions. Other than the addition of the weighting parameter, all details of the Markov chain were as in (Raina et al., 2005).

Results

Brief summary of the new complete snake mitochondrial genomes

The gene contents of *A. piscivorus* and *P. slowinskii* mtDNAs are similar to other snakes (Figure 17; detailed genome annotation in Tables 18 and 19). There is a duplicated control region (CR2) between ND1 and ND2, in addition to the original control region (CR1) present in all vertebrates adjacent to the 5' end of the 12s rRNA gene (Dong and Kumazawa, 2005; Kumazawa et al., 1996, 1998). These genomes also possess the translocated tRNA^{Leu} common to

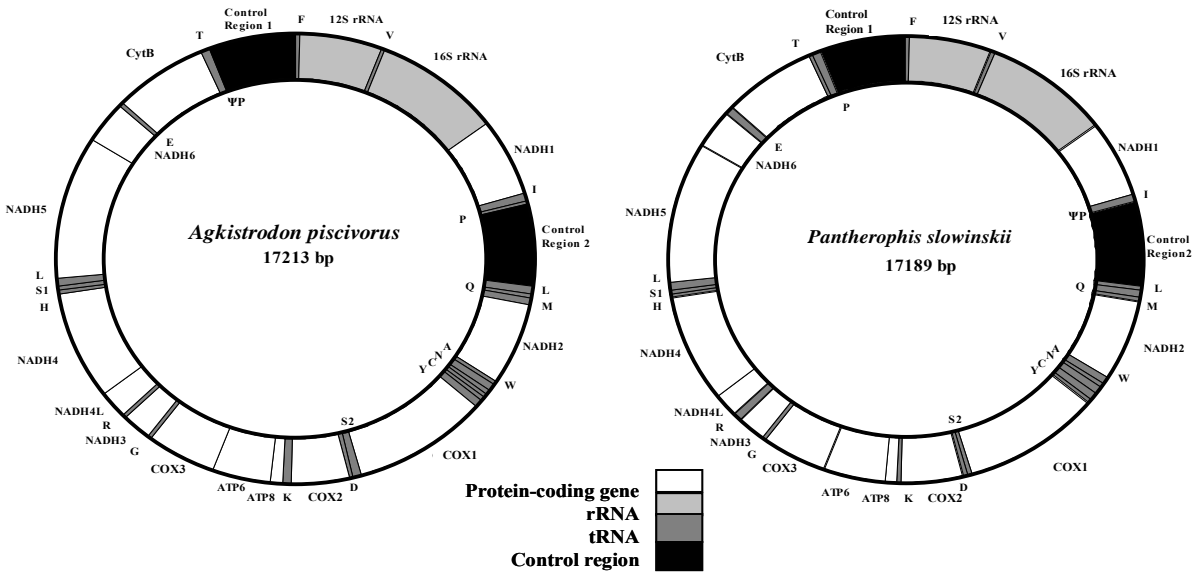


Figure 17. Annotated mitochondrial genome maps of *Agkistrodon piscivorus* and *Pantherophis slowinskii*. The two *Agkistrodon* samples (*Api1* and *Api2*) have identical annotations except for minor variations in gene length. Labels of genes outside the circle refer to genes transcribed from the light strand, and names within the circle represent genes transcribed from the heavy strand.

Table 18. Detailed genome annotation of *Agkistrodon piscivorus*.

	From	To	Size	Strand	Codon	StartCodon	StopCodon
Phe	1	65	65	L	TTC		
12sRNA	62	976	915	-			
Val	977	1040	64	L	GTA		
16sRNA	1041	2527	1487	-			
ND1	2528	3488	961	L		ATC	T
Ile	3489	3556	68	L	ATC		
Pro	3560	3622	63	H	CCA		
CR1	3623	4642	1020	-			
Leu	4643	4715	73	L	TTA		
Gln	4716	4785	70	H	CAA		
Met	4786	4848	63	L	ATG		
ND2	4849	5878	1030	L		ATA	T
Trp	5879	5944	66	L	TGA		
Ala	5945	6009	65	H	GCA		
Asn	6010	6081	72	H	AAC		
O_L	6084	6117	34	-			
Cys	6116	6175	60	H	TGC		
Tyr	6176	6236	61	H	TAC		
COX1	6238	7839	1602	L		GTG	AGA
Ser4	7830	7897	68	H	TCA		
Asp	7898	7960	63	L	GAC		
COX2	7962	8646	685	L		ATG	T
Lys	8647	8710	64	L	AAA		
ATP8	8711	8875	165	L		ATG	TAA
ATP6	8866	9546	681	L		ATG	TAA
COX3	9546	10329	784	L		ATG	T
Gly	10330	10390	61	L	GGA		
ND3	10391	10733	343	L		ATC	T
Arg	10734	10797	64	L	CGA		
ND4L	10798	11087	290	L		ATG	TA
ND4	11088	12425	1338	L		ATG	AGA
His	12426	12487	62	L	CAC		
Ser2	12488	12542	55	L	AGC		
Leu4	12543	12614	72	L	CTA		
ND5	12616	14403	1788	L		ATG	TAA
ND6	14399	14908	510	H		GTG	AGG
Glu	14918	14980	63	H	GAA		
CytB	14981	16094	1114	L		ATG	T
Thr	16095	16159	65	L	ACA		
Pseudo-Pro	16160	16190	31	-			
CR2	16191	17213	1019	-			

Table 19. Detailed genome annotation of *Pantherophis slowinskii*.

	From	To	Size (bp)	Strand	Codon	StartCodon	StopCodon
Phe*	1	60	60	L	TTC		
12sRNA	59	991	933	-			
Val	992	1054	63	L	GTA		
16sRNA	1055	2531	1477	-			
ND1	2532	3495	964	L		ATA	T
Ile	3496	3561	66	L	ATC		
Pseudo-Pro	3558	3592	35				
CR1	3593	4613	1021	-			
Leu2	4614	4686	73	L	TTA		
Gln	4689	4759	71	H	CAA		
Met	4761	4822	62	L	ATG		
ND2	4823	5852	1030	L		ATT	T
Trp	5853	5917	65	L	TGA		
Ala	5919	5981	63	H	GCA		
Asn	5983	6055	73	H	AAC		
O_L	6058	6093	36	-			
Cys	6092	6152	61	H	TGC		
Tyr	6153	6214	62	H	TAC		
COX1	6216	7817	1602	L		GTG	AGA
Ser4	7808	7874	67	H	TCA		
Asp	7875	7938	64	L	GAC		
COX2	7940	8624	685	L		ATG	T
Lys	8625	8688	64	L	AAA		
ATP8	8690	8848	159	L		ATG	TAA
ATP6	8839	9519	681	L		ATG	TAA
COX3	9519	10302	784	L		ATG	T
Gly	10303	10363	61	L	GGA		
ND3	10364	10706	343	L		GTG	T
Arg	10707	10771	65	L	CGA		
ND4L	10772	11061	290	L		ATG	TA
ND4	11062	12399	1338	L		ATG	TAA
His	12400	12464	65	L	CAC		
Ser2	12465	12521	57	L	AGC		
Leu4	12519	12589	71	L	CTA		
ND5	12590	14536	1947	L		ATG	ATT
ND6	14353	14853	501	H		ATG	TAG
Glu	14863	14924	62	H	GAA		
CytB	14923	16039	1117	L		ATG	T
Thr	16040	16103	64	L	ACA		
Pro	16104	16164	61	H	CCA		
CR2	16165	17189	1025	-			

all alethinophidian snakes (3' of CR2). In addition to an intact tRNA^{Pro} between CytB and CR1, *Pantherophis* has an apparent pseudo-tRNA^{Pro} gene (Ψ -tRNA^{Pro}) between ND1 and CR2 (as does the previously sequenced colubrid, *Dinodon*). This Ψ -tRNA^{Pro} exactly matches the first 35 bases of tRNA^{Pro}. In contrast, the intact tRNA^{Pro} of *Agkistrodon* (and the previously sequenced viperid, *Ovophis*) is located between ND1 and CR2 (exactly the location of Ψ -tRNA^{Pro} in the colubrids), and there is a 31 bp non-coding fragment between tRNA^{Thr} and CR1, where tRNA^{Pro} is usually located. In *Ovophis*, this is clearly a Ψ -tRNA^{Pro} as these 31 bp are an exact match the CR1-proximal end of the complete tRNA^{Pro}, but in *Agkistrodon* the homology is much less clear (see below for further detail). These alternative positions of tRNA^{Pro}, Ψ -tRNA^{Pro}, and a previously noted (Dong and Kumazawa, 2005) duplication of tRNA^{Phe} in *Ovophis* (see below) are the only notable mtDNA gene rearrangements identified within the alethinophidian snakes.

Comparison of A. piscovorus genomes

Polymorphisms were observed between the two *Agkistrodon* genomes, *Api1* and *Api2*, for all protein and rRNA genes (Table S6) and for 14 of 22 tRNAs (Table S7). The 12s and 16s rRNAs were the most conserved genes between the two *Agkistrodon* individuals, with 2% and 3% sequence divergence respectively (Figure 18A; Table S6). Protein-coding genes differed more, up to 6.2% for ND3 (Figure 18A; Table 20). Most differences occurred at 3rd codon positions (Figure 18A; Table 20), as expected under predominantly neutral patterns of divergence (for example, 57/58 substitutions in COX1 were at 3rd codon positions). Within a

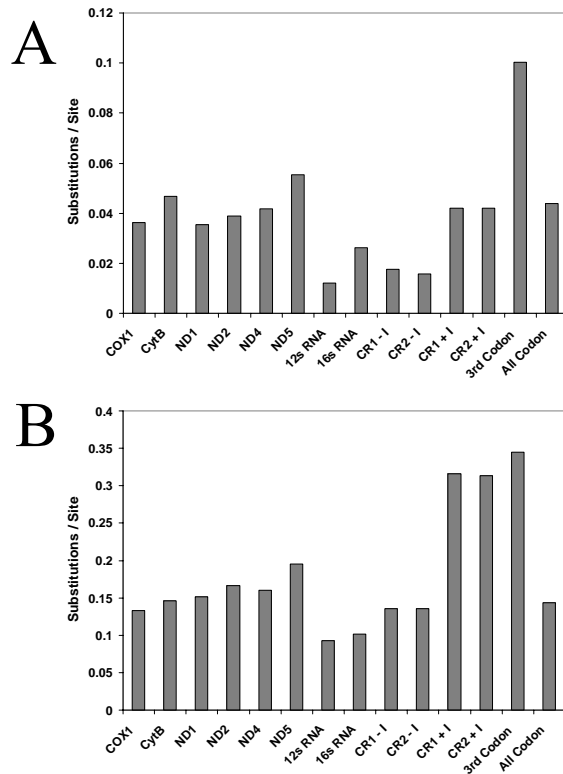


Figure 18. Differences per site for homologous genes or groups of sites in the two *Agkistrodon* genomes and in the two viperid genomes. The differences per site are shown for a comparison of *Api1* and *Api2* (A), and for *Agkistrodon* (mean of *Api1* and *Api2*) and *Ovophis* (B). Differences are shown only for the longer protein-coding genes. For the control regions only (shaded black), differences are shown for each aligned site including indels (e.g., CR1+I), or excluding indels (e.g., CR1-I). For all other genes, indels are not included in the difference measure. The bars for 3rd codon positions (3rd Codon) and for all codon positions (All Codon) are summed over all protein-coding genes.

Table 20. Gene-specific polymorphisms observed between the two *Agkistrodon piscivorus* genomes (*Api1* and *Api2*).

Genes	Length	Similarity	Substitutions				
			all	1st	2nd	3rd	AA
12s RNA	915	98.80%	11	-	-	-	-
16s RNA	1487	97.40%	39	-	-	-	-
ATP6	681	95.00%	32	5	2	25	4
ATP8	165	93.94%	11	3	1	7	3
COX1	1602	96.38%	58	0	1	57	2
COX2	685	96.50%	24	6	0	18	3
COX3	786	96.40%	28	6	1	21	5
CytB	1114	95.33%	52	10	3	39	10
ND1	960	96.46%	34	8	1	25	3
ND2	1030	96.12%	40	6	4	30	8
ND3	343	93.88%	21	2	6	20	8
ND4	1338	95.81%	56	9	3	44	5
ND4L	290	97.93%	6	2	0	4	2
ND5	1788	94.46%	96	21	9	69	28
ND6	510	95.00%	26	3	4	19	5
CR1	1021	98.20%	19	-	-	-	-
CR2	1022	98.40%	18	-	-	-	-

mtDNA, the duplicated CRs of each newly sequenced species are nearly identical, as is typical for alethinophidian snakes (Dong and Kumazawa, 2005; Kumazawa et al., 1998). In *Pantherophis* there is a single point mutation and four extra nucleotides at one end of CR1, in *Api1* there is one indel plus 14 extra nucleotides on one end of CR1, and in *Api2* there are seven indels and two base changes between the two control regions. Comparing within a species between *Api1* and *Api2*, CR1 differs by five indels and 19 point mutations, whereas CR2 differs by three indels (two at the 5' end) and 18 point mutations. Within *Agkistrodon*, the control regions (e.g., CR1 in *Api1* vs. CR1 in *Api2*) are as similar to each other as rRNAs and more similar than the protein coding genes (Figure 18A). This is in strong contrast to the normal pattern of divergence between vertebrate species, for which control region similarity is far less than that of protein-coding or rRNA genes. Between *Agkistrodon* and the other viperid *Ovophis*, the control regions have 30% more differences (with indels included) than the rRNAs, and are on par with divergence in the protein-coding genes (Figure 18B). If indels are included, the control regions between these two species are nearly as different as the average 3rd codon position (Figure 18B). The high degree of similarity (low divergence) observed between the CRs of the two *Agkistrodon* individuals (e.g., CR1 of *Api1* vs. CR1 of *Api2*) is surprising, and contrasts sharply with the high relative divergence of CRs between *Ovophis* and *Agkistrodon* (Fig. 18).

Phylogenetics

Taxonomic sampling in this study was designed to include multiple groups to compare with the snakes. We included all available snakes, crocodylians and turtles with complete

mitochondrial genomes, as well as a sampling of birds and mammals (mostly primates), and all lizards with an unambiguous evolutionary relationship to snakes (including the tuatara Rest et al. 2003). The phylogenetic tree obtained by ML is shown, with NJ bootstrap values (BS) and posterior probabilities (PP) for nodal support, which were generally high (Figure 19). Our phylogeny estimate provides a well-resolved and, in many cases, strongly-supported amniote phylogeny that is consistent with previous molecular studies. Differences between the ML topology (Figure 19), and the topology based on Bayesian analysis (not shown) were minor, and included an alternative placement of *Bos* among mammals, and alternative placements of *Gallus* and *Rhea* among birds. Additionally, relationships among lizard taxa varied, with *Cordylus* estimated to be the sister lineage to all other lizards, and an alternative placement of *Varanus* in the Bayesian estimate.

All phylogenetic estimates provided an identical well-supported topology for relationships among snakes (Figure 19), and a summary of results concerning snake relationships is shown in Figure 20. The Scolecophidia (Typhlopoidea), represented here by *Leptotyphlops*, formed the sister group to the remaining snakes. Rather than finding support for a sister-group relationship between Henophidia and Caenophidia (*Acrochordus* plus Colubroidea (Dong and Kumazawa, 2005; Gower et al., 2005)), we find strong support for *Acrochordus* as the sister lineage to the Henophidia. Hereafter we will therefore operationally refer to Henophidia as including *Acrochordus*, and we will refer to the sister clade of the Henophidia as the Colubroidea (Lawson et al., 2005).

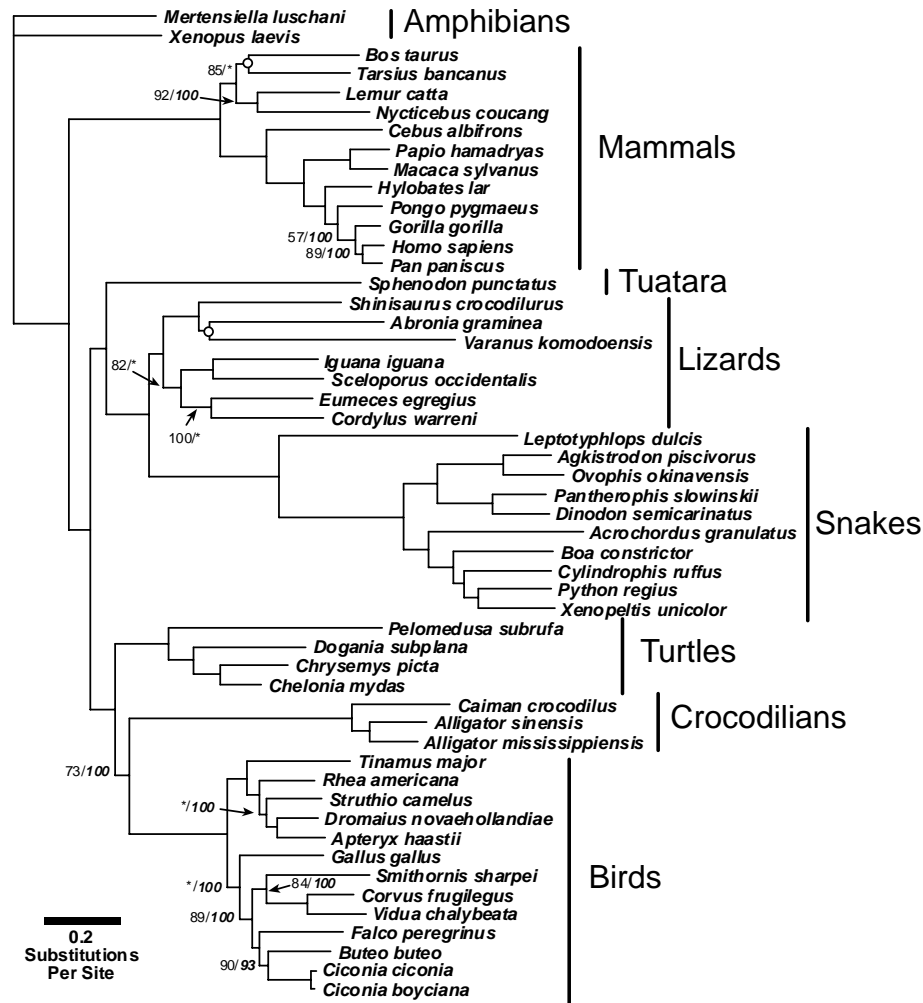
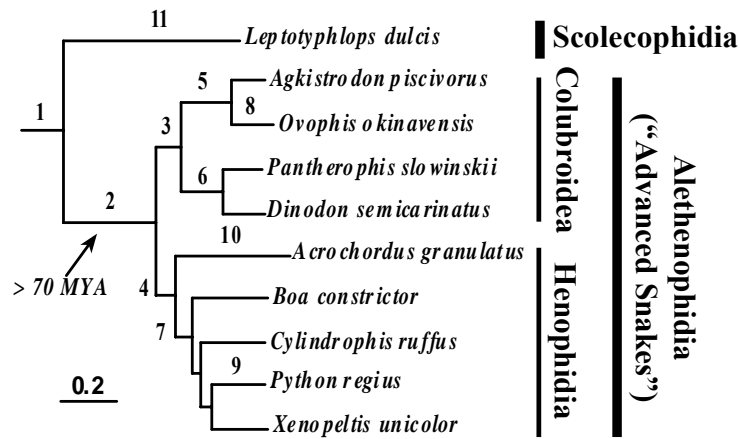


Figure 19. Maximum likelihood phylogeny for vertebrate taxa included in this study. This phylogeny is based on all protein-coding and rRNA genes. Most branches have greater than 95% support for both NJ ML distance bootstrap and Bayesian posterior probability support (see Methods), and are not annotated with support values. Where support from either measure is less than 95%, the support values are indicated by ratios, with the ML bootstrap support on top and the Bayesian posterior probability support below in italics, except for two nodes with less than 50% support by either measure, which are indicated by a hollow circle. Other than for these two nodes, support values less than 50% are indicated with an asterisk (*).



Branch	Major Genomic and Molecular Evolutionary Events
1	Length reduction in all genes; Simplification of the tRNA T-arms Acceleration of ATP6, ATP8, COX1, COX2, CytB, ND1, ND2, and ND5
2	Duplication of CR; Transposition of tRNA ^{Leu} Acceleration of ATP6, ATP8, COX1, COX2, CytB, and ND6
3	Duplication of tRNA ^{Pro} ; Length reduction in tRNA and rRNA genes Acceleration of ND5, ND6, and 12s, 16s rRNAs Rate of CR concerted evolution increases Increase in ND1 C/T ratio indicating CR2 function
4	Length increase in rRNA genes Acceleration of ATP6, COX3, ND3, ND4L, ND6, 16s rRNA
5	Degradation/loss of tRNA ^{Pro} duplicate (3' of CR1)
6	Degradation of tRNA ^{Pro} duplicate (3' of CR2)
7	Decrease in C/T ratio of ND1 indicating CR1 preference
8	Duplication/translocation of tRNA ^{Phe} Concerted evolution of tRNA ^{Phe} copies along with CRs Acceleration of 16s rRNA
9	Increase in C/T ratio of ND1 indicating CR2 preference
10	Acceleration of ATP6, ATP8, and COX2
11	Loss of light strand origin; Translocation of tRNA ^{Gln}

Figure 20. Hypotheses for the relative timing of alterations in mitochondrial genome architecture and molecular evolution throughout snake phylogeny. The topological relationships among snakes and branch lengths shown are the same as in Figure 3. Major groups of snakes are indicated along with the approximate diversification time of the Alethinophidia.

Since both the snake and the overall amniote phylogeny are strongly supported by our analysis of this dataset, we will henceforth treat this phylogeny as though it is accurate. We wish to emphasize, however, that the consistency of the phylogenetic results do not guarantee that they are, in fact, accurate. Some difficult questions were avoided (amphisbaenian lizards were not included because their placement in relation to snakes is uncertain), and we used a single nucleotide substitution model for the entire dataset rather than a complex set of partitioned models. We have, however, analyzed an expanded version of this dataset (with additional mtDNAs) using complex partitioned models for each gene and codon position, and the resulting phylogeny estimates were essentially identical to those presented here. We provide evidence below for extremely complex non-stationary patterns of nucleotide substitution across branches and mtDNA regions, and have previously identified asymmetric substitution gradients in mtDNA (Faith and Pollock, 2003) that may vary among species (e.g., primates (Raina et al., 2005)). These latter patterns cannot be modeled using available phylogenetic programs (e.g., MrBayes (Ronquist and Huelsenbeck, 2003)). Some of us are currently developing new analytical strategies to accommodate these spatial and temporal nucleotide substitution dynamics, but the subject of improved phylogenetic reconstruction using such methods is a complicated topic that is outside the scope of this study, and we will reserve it for future research. We expect our phylogenetic estimates here to represent a good estimate of the relationships among mtDNAs sampled, and if minor inaccuracies in the topology have occurred in our estimates, these changes should not substantially impact the qualitative conclusions of further analyses (e.g., sliding window analysis, SWA) because a majority of these later estimates are averaged over many

branches of the tree, and the dynamics we concentrate on are quite dramatic and are likely to be obvious and qualitatively similar even with slight changes in the topology estimate.

Nucleotide frequencies and control region functionality

In *Agkistrodon* and *Pantherophis* mtDNA, as in other vertebrates (Reyes et al., 1998), nucleotides A and C are favored on the light strand, particularly at 3rd codon positions. This bias is probably related to elevated rates of deamination mutations on the heavy strand incurred during replication (see Background), and is not systematically different between lizards and snakes, although there is considerable variation among individual mtDNAs.

Due to the simple linear relationship in most vertebrate mtDNAs between C/T ratios and T_{AMS} predicted based on the location of the (functional) control region, it is of interest to determine whether there has been any clear genetic effect of the duplicated control region in alethinophidians. Exclusive use of one control region or the other would be most strongly observable in ND1, the only protein-coding gene located between the two control regions in alethinophidian snake mtDNAs. Since the nucleotide sequence of duplicate control regions is nearly identical within each genome, however, it is also reasonable to consider the possibility that both control regions are functional.

To test these predictions, we applied our MCMC analysis (Raina et al., 2005) to fit alternative models of exclusive CR1 or CR2 usage, or mixed control region effect (Table 21). The Akaike weights for the alternative individual models provide a prediction of the degree to which a control region is exclusively functional, while the weight parameter in the mixed model

Table 21. Negative log likelihood values and Akaike weights (in parentheses) for individual origin of replication models and the mixed model, along with the most likely CR2 preference parameter in the mixed model, for alethinophidian snakes.

Species	Individual model		Mixed model	
	O_H^{CR1}	O_H^{CR2}	$O_H^{CR1} + O_H^{CR2}$	% O_H^{CR2}
<i>Agkistrodon piscivorus</i>	1179.2 (18%)	1178.0 (60%)	1179.0 (22%)	99%
<i>Pantherophis slowinskii</i>	1164.6 (29%)	1164.1 (47%)	1164.8 (24%)	54%
<i>Dinodon semicarinatus</i>	1167.1 (21%)	1166.2 (57%)	1167.1 (22%)	78%
<i>Ovophis okinavensis</i>	1252.7 (38%)	1252.6 (45%)	1253.5 (17%)	59%
<i>Boa constrictor</i>	854.5 (29%)	853.9 (50%)	854.8 (21%)	64%
<i>Acrochordus granulatus</i>	1245.0 (2%)	1241.5 (72%)	1242.5 (26%)	100%
<i>Xenopeltis unicolor</i>	1159.4 (31%)	1159.0 (45%)	1159.6 (24%)	50%
<i>Python regius</i>	1133.0 (1%)	1128.9 (72%)	1130.0 (26%)	100%
<i>Cylindrophis ruffus</i>	1129.8 (70%)	1132.6 (4%)	1130.8 (26%)	<1%

represents the time-averaged effect of mixed control region usage on the C/T ratios. There is evidence for at least mixed CR2 usage in all but one species (*Cylindrophis*). The evidence is good for exclusive or nearly exclusive CR2 functionality in two species (*Acrochordus* and *Python*), and for a strong CR2 preference in *Agkistrodon*. The patterns appear to be species-specific (strong preferences for a particular control region are widely dispersed on the tree), which may indicate rapid evolution of the strength of the gradient (as suggested in primates (Raina et al., 2005)) or rapid evolution of differential usage of the two control regions. Species with ambiguous control region preferences may have mixed usage, may not have a strong enough gradient to differentiate, or may have previously switched usage and thus have not reached mutational equilibrium. A potentially relevant observation is that three of the five henophidians have both strong control region preferences and also greater divergence between their CR sequences than do colubroids (Dong and Kumazawa, 2005).

Gene length and stability of truncated tRNAs in snakes

In snakes, all protein-coding genes (except COX1), ribosomal RNAs, tRNAs, and individual CRs are shorter than their counterparts in most lizards and most other vertebrates (Figure 21). An exception to this is *Sphenodon*, for which the control region, ATP8 (ATP synthase subunit 8) and the 12s rRNA are all shorter than in snakes. With the increased sampling in this study, it appears that while the tRNAs and proteins became shorter prior to the divergence of all snakes, the tRNAs became shorter still in the Colubroidea (Figures 20 and 21).

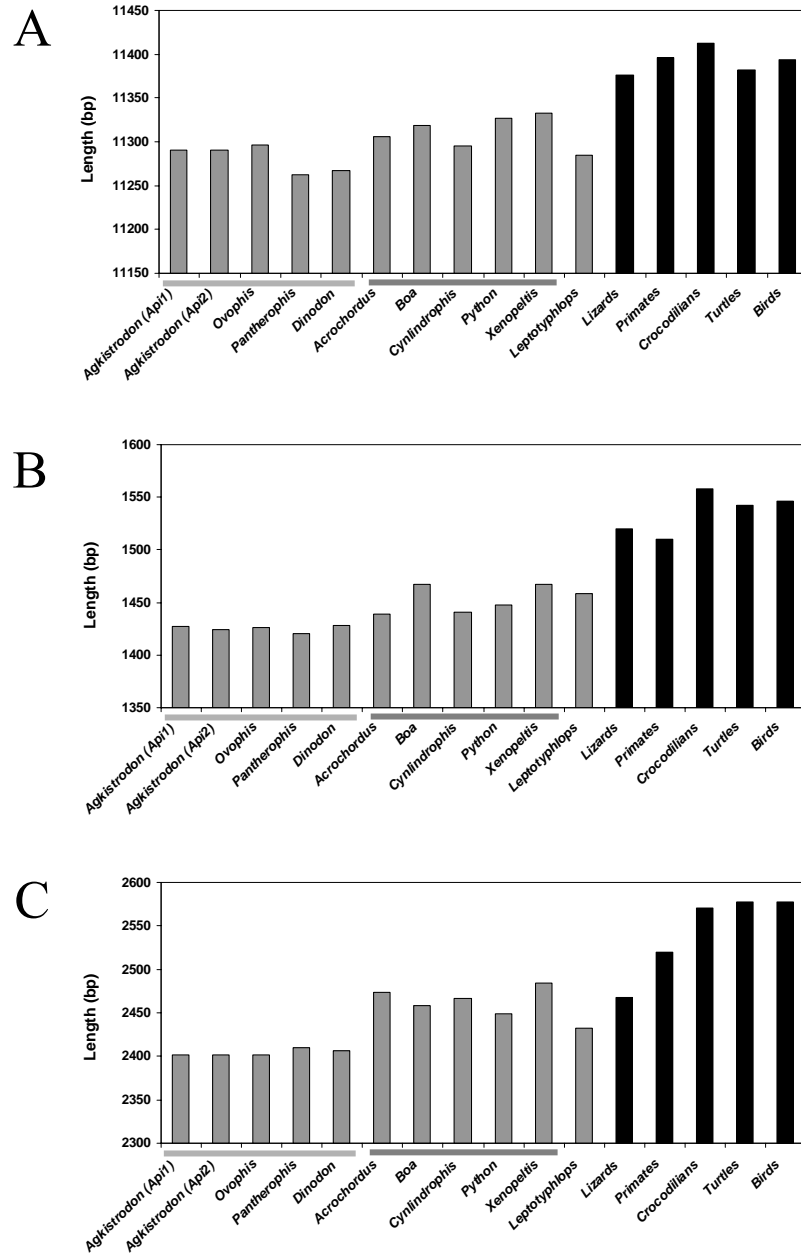


Figure 21. Comparison of gene lengths in snakes and other squamates. The total length is shown for all protein coding regions (A), tRNAs (B), and rRNAs (C). All snakes are in gray, while other squamates (lizards) are in black, and light gray and dark gray bars are drawn under snake species to indicate membership in the Colubroidea or Henophidia, respectively.

Additionally, the rRNAs did not become shorter in *Leptotyphlops* or the Henophidia, but are dramatically shorter in the Colubroidea (Figures 20 and 21).

The shorter length of tRNAs in snakes results mainly from a truncated T-arm in the secondary structure (see also (Kumazawa et al., 1996, 1998)). In some tRNAs, the D-arm is also shorter, but to a lesser extent than the T-arms. Although short tRNAs are typically less stable than long ones, there is only a minor effect of sequence length on secondary structure stability (ΔG) in snake tRNAs. The cloverleaf structures of most snake tRNAs are slightly less stable than their lizard counterparts (Table S8), but two tRNAs (tRNA^{Ile}, tRNA^{Met}) are actually more structurally stable in snakes than in other squamates with longer tRNAs.

Spatio-temporal substitution rate dynamics across mtDNA genes and regions

Although the mitochondrial genomes of snakes (as well as crocodylians) have been identified as evolving faster than other tetrapods (Hughes and Mouchiroud, 2001; Janke et al., 2001; Kumazawa and Nishida, 1999), the details and uniformity of such rate dynamics have not been investigated. To assess the difference in substitution rates among genes, we fixed the topology (Figure 19) and calculated branch lengths based on rRNAs and on all protein-coding genes (Figure 22). Along the branches leading to modern snake taxa there was a slight increase in the rate of molecular evolution of rRNAs and a dramatic increase in protein-coding gene rates. For the rRNAs, most other major amniote groups have experienced similar amounts of total evolution from their common ancestor with the amphibians, and the snake lineages stand out as unusual in their accelerated evolution (Figure 22A). For protein-coding genes, there is much

A. rRNA Genes

B. Protein-Coding Genes

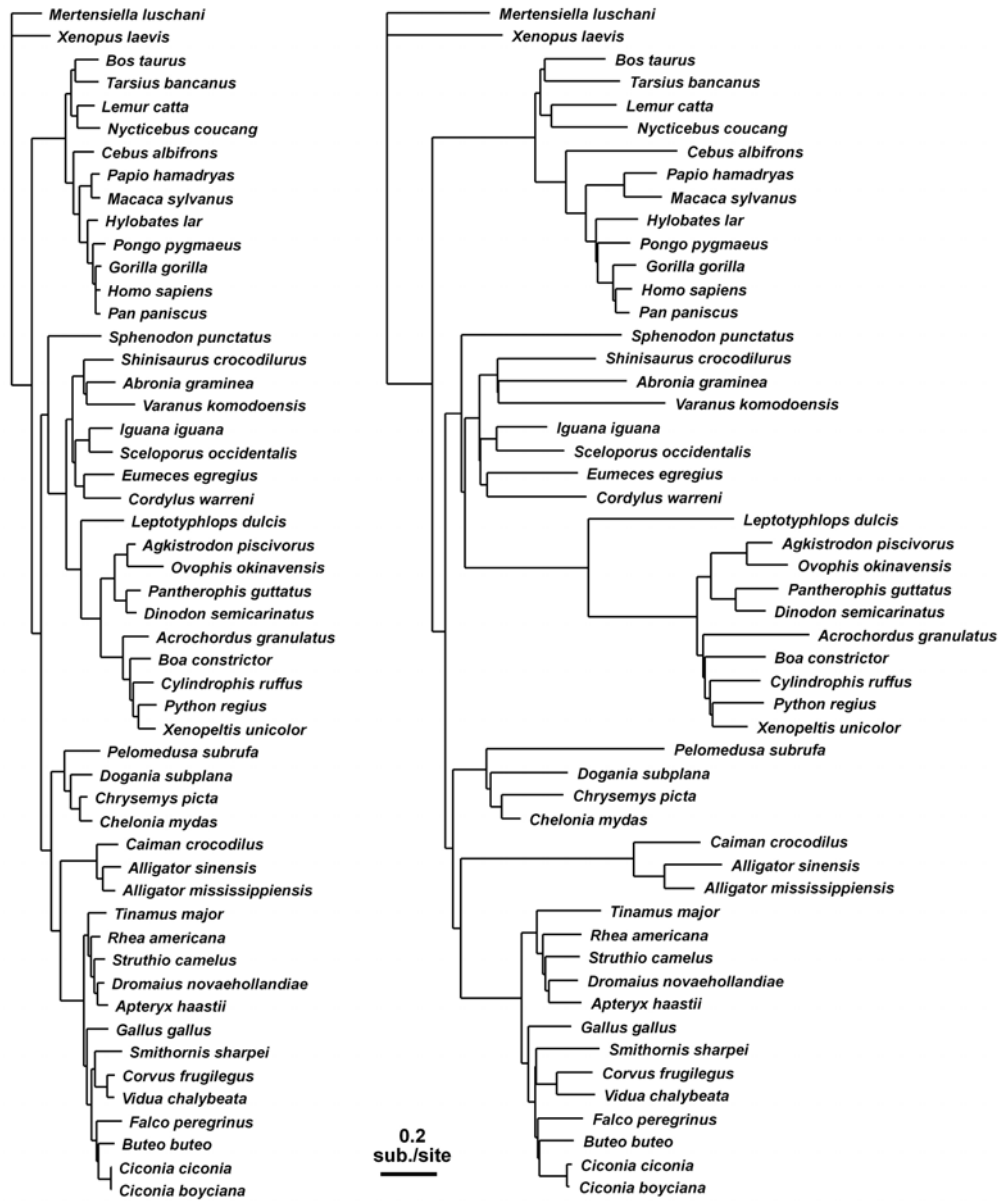


Figure 22. Phylograms based on the relative branch lengths for rRNA and protein-coding genes, topologically constrained based on the ML phylogeny (Figure 3). Branch lengths on this constrained topology were estimated using all rRNA genes (A) or all protein-coding genes (B). The substitution rate scale is the same in both trees.

more variation, and mammals, some lizards, crocodylians, and one turtle have longer branches than the other turtles, lizards, and all birds (Figure 22B). The snake lineage has, comparatively, even longer branches than any of these groups, and certain branches (e.g., the ancestor of all snakes and the ancestor of Alethinophidia) are disproportionately long compared to branch lengths based on rRNAs (Figure 22). To evaluate this further, branch lengths were calculated for different genes and gene clusters. There was considerable variation among genes with respect to relative branch lengths in the ancestral snake lineages (data not shown). As an example, for each gene or gene cluster we compared cumulative branch lengths within three clades (mammals, snakes, or lizards) and among the lineages leading to their common ancestors (Figure 23). There is a remarkable degree of consistency in the total and relative amounts of evolution between the mammal clade and the lizard clade (Figure 23A). In contrast, four genes and gene clusters (COX1, CytB, the COX2+ATP6+ATP8 cluster, and the COX3+ND3+ND4L cluster) have relatively longer branch lengths (indicating higher substitution rates) in snakes than in lizards and mammals. For the remaining genes (ND1, ND2, ND4, and ND5) the total branch lengths for snakes are either intermediate or similar to that of mammals and lizards. There is more variation for the ancestral branches (Figure 23B), which is not surprising given that it is a single branch with shorter total length, but a few details stand out. First, the snake ancestral branch length is similar to the mammal ancestral branch length for a majority of genes, but is considerably shorter for the rRNAs and ND2, and is obviously far longer for COX1. Combining evidence from Figure 23 with the tree-based evidence (Figure 22), we interpret these patterns as indicating that there has been accelerated evolution in many mitochondrially-encoded proteins along ancestral

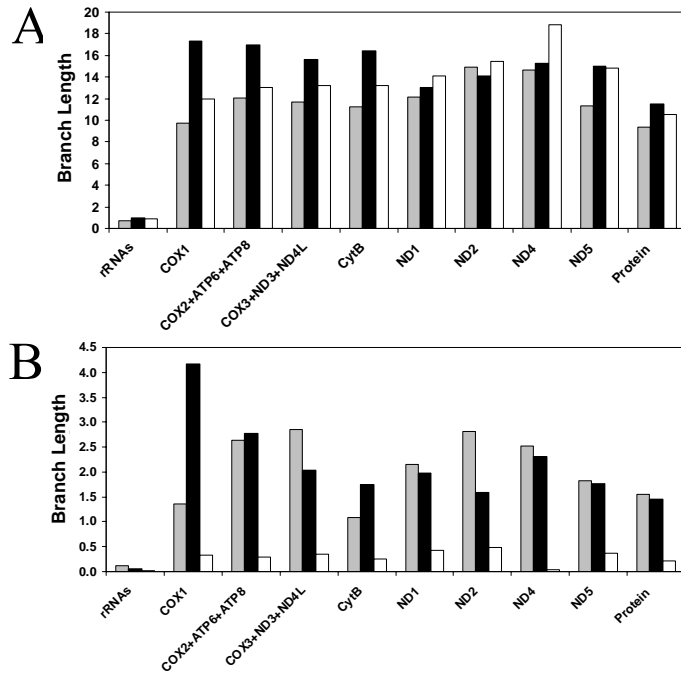


Figure 23. Comparison of branch lengths from different genes and gene clusters for mammals, snakes, and lizards. Branch lengths for each gene or gene cluster are shown based on the cumulative branch lengths within each clade (A), or based on the gene or gene cluster branch length estimated along the ancestral branch leading to each nominal clade (B). Mammals are shown in gray, snakes in black, and lizards in white fill. rRNA branch lengths have been multiplied by ten to make them visible in this figure compared to protein branch lengths.

branches of the snake phylogeny, but that most ND subunits have experienced minimal acceleration, similar to the rRNAs.

To qualitatively elucidate the spatio-temporal dynamics in rates of substitution between gene regions that occur across branches, we plotted the branch lengths derived from rRNAs (which appear to have had only minimal acceleration; e.g., Figure 22A) versus the branch lengths of various genes and gene clusters (Figure 24). All gene pairs generally appear to have highly correlated branch lengths (Figure 24), but some branches are outside the main distribution. These are of the greatest interest since they may indicate unusual molecular evolutionary dynamics in these genes, including possible accelerated evolution.

Two branches consistently below the main distribution in most comparisons are the terminal branch leading to *Ovophis* and the ancestral branch leading to the henophidians (Figure 24). Looking back (Figure 22), it is apparent that these two branches are disproportionately longer in the rRNA trees than in the protein trees. These two lineages (the ancestor of Henophida, and *Ovophis*) appear to have experienced acceleration of rRNA genes well beyond the mild accelerated evolution of rRNA that occurred along the ancestral lineages leading to all snakes and to the Alethinophidia.

The ancestral branches leading to all snakes and to the alethinophidians are well above the main distribution in comparisons of COX1 (Figure 24A), CytB (Figure 24B), and COX2+ATP6+ATP8 (Figure 24C). Notably, these clusters include nearly all mitochondrially-encoded protein-coding genes except those from ND (although ND6 does show some dramatic acceleration; Figure 24H). This suggests that the acceleration was targeted at certain functional groups of genes, and was not ubiquitous or evenly distributed across all mitochondrial genes.

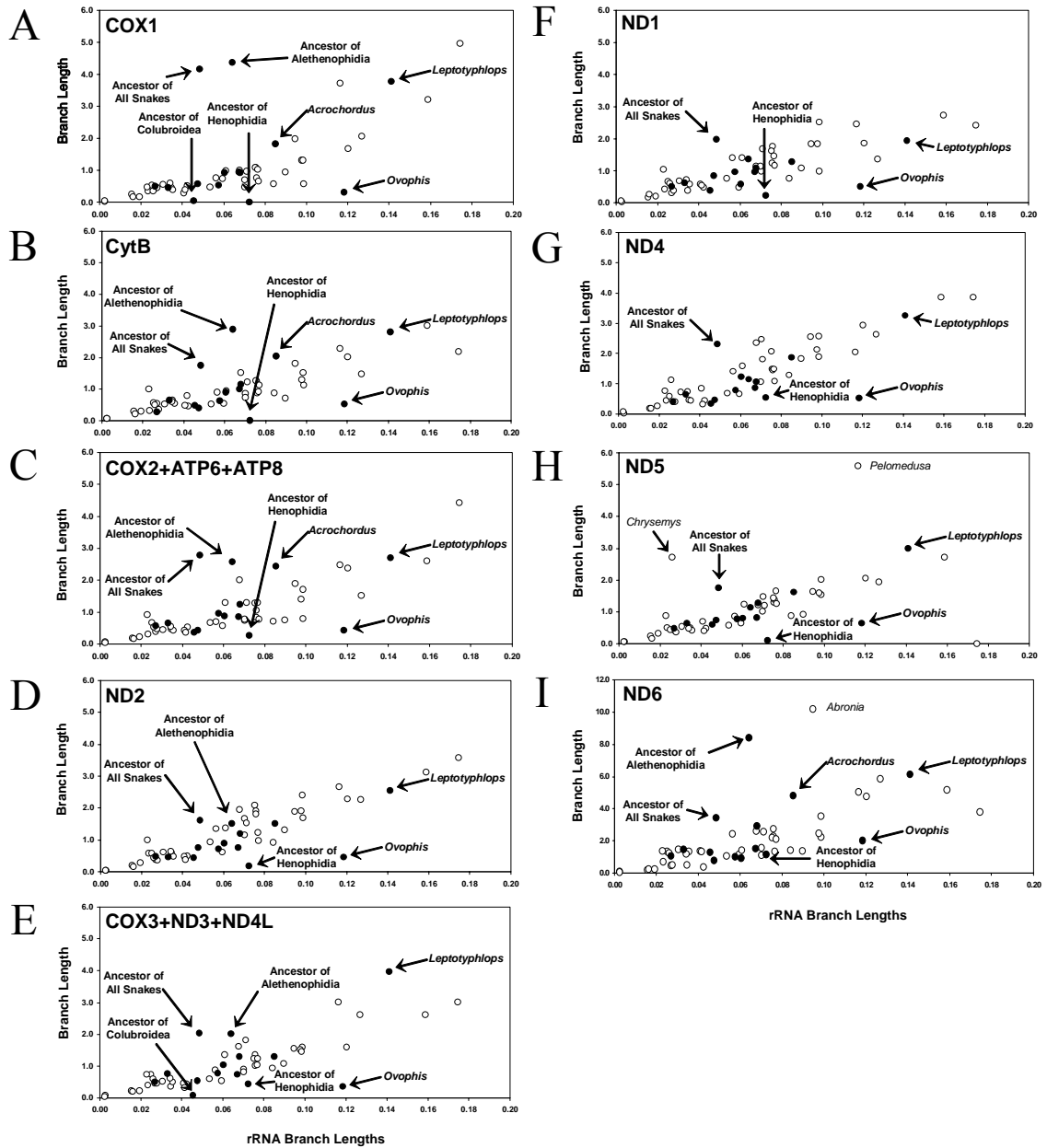


Figure 24. Plot of branch lengths obtained from rRNA versus various genes and gene clusters. Snake branches are indicated with filled circles, and non-snake tetrapod branches are indicated with an unfilled circle. The locations of selected snake branches are labeled (in bold) with arrows. Outlying non-snake branches are indicated and labeled in normal type. Genes and gene clusters shown are (A) COX1, (B) CytB, (C) COX2 + ATP6 + ATP8, (D) ND2, and (E) COX3 + ND3 + ND4L, (F) ND1, (G) ND4, (H) ND5, (I) ND6.

The ancestor of the Colubroidea does not stand out as having had experienced notable accelerated evolution in these comparisons, which could mean that it did not, or that acceleration across various genes is balanced by acceleration of rRNA evolution. We also observed several non-snake tetrapod tip branches that were outliers on these plots (Figure 24), indicating that differential selection on a single gene has occasionally occurred in taxa other than snakes.

The branch leading to *Leptotyphlops* is not detectably accelerated in any comparison in this analysis (Figure 24), and generally falls amidst the distribution of non-snake vertebrates. The branch leading to *Acrochordus* (the most divergent henophidian, as described earlier) is outstanding only in the COII+ATP6+ATP8 comparison (and slightly in CytB; Figure 24). All other branches in the snakes (unlabelled filled circles in Figure 24) are consistently in the midst of the distribution, indicating either that any accelerated evolution in their proteins is proportionally matched by acceleration in their rRNAs (which is somewhat inconsistent with Figure 22A), or that genome-wide evolutionary rates conform to average relative rates in tetrapods (Figure 24).

To further evaluate the variation in spatio-temporal dynamics of substitution rates across the mitochondrial genome, we used SWA of branch-specific and group-specific patterns of relative substitution. Only one of these comparisons, that of the henophidian terminal branches, shows little variation of standardized substitution rates across the genome (Figure 25C). This suggests that the distribution of substitutions across the mtDNA of contemporary henophidians is nearly identical to the distribution across the mtDNA of other tetrapods, and thus that contemporary henophidians are not undergoing atypical gene-specific selection. The terminal colubroid branches are also fairly flat except for the downstream half of the 16s rRNA (Figure

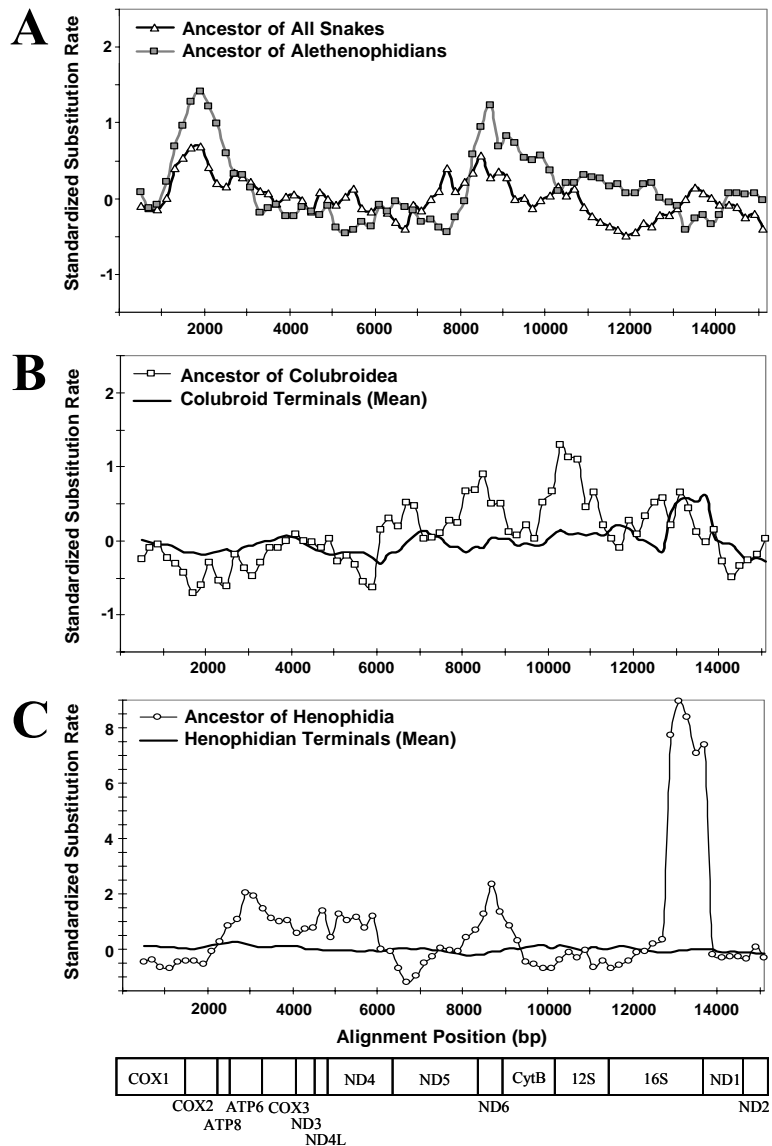


Figure 25. Standardized substitution rates across the mitochondrial genome for selected branches or clusters. For each 1000 bp window applied to a set of branches, standardized substitution rates were obtained by first dividing by the median window value for that branch, and then subtracting this value from the average across all non-snake branches. This helps to visualize regions of the genome that are evolving at slower or faster rates, with the average tetrapod relative rate being zero. Branches or branch sets shown are (A) the ancestor of all snakes and the ancestor of the Alethinophidia; (B) the ancestor of the Colubroidea and the sum of all colubroid terminal branches; and (C) the ancestor of the Henophidia and the sum of all henophidian terminal branches.

25B), which may be entirely attributable to acceleration of the 16s rRNA in *Ovophis*, as discussed earlier. The patterns in the ancestors of henophidians, colubroids, alethinophidians (henophidians plus colubroids), and of all snakes contrast sharply with this background, and instead have distinctive atypical gene-specific patterns (Figure 25). In the ancestor of alethinophidians, there is a strong peak coinciding with the end of COX1, and covering COX2, ATP6, and ATP8, and there is another peak in ND6 and CytB (Figure 25A). In the ancestor of all snakes, there are less distinctive rises in the same areas. In contrast, the ancestor of the Colubroidea has low relative rates in the region from COX1 to ND4, but has rate peaks in the beginning of ND5, in ND6, in the 12s rRNA, and somewhat of a peak in the middle of the 16s rRNA (Figure 25B). The ancestor of the Henophidia has a broad low peak from ATP6 to ND4 (including COX3, ND3, and ND4L), another peak in ND6, and an extremely large peak in the end of the 16s rRNA (Figure 25C). It is notable that the henophidian ancestral 16s peak closely matches the *Ovophis* peak in the same region

In summary, the ancestor of all snakes appears to have had moderately accelerated evolution in the region starting near the end of COX1 thru COX2, ATP8, and somewhat into ATP6, and also in the separate region including the end of ND5, ND6, and CytB (and a rise in ND1). The COX1, COX2, ATP8, and ND6 accelerations increased and were stronger in the ancestor of the Alethinophidia, while the ND5 acceleration decreased, and a notable acceleration of CytB also occurred. In the ancestor of the Colubroidea, only the ND6 acceleration continued, but new rate peaks arose in ND5, 12s rRNA, and the first part of the 16s rRNA, followed by a strong dropoff in all gene-specific acceleration in modern colubroid lineages, except in the end of 16s rRNA in *Ovophis*. In the ancestor of the Henophidia, the accelerated rates of evolution (in

COX1, COX2, ATP8, and ND5 genes) observed along the branch leading to the alethinophidians diminished (except for ND6 as in the Colubroidea), but new rate peaks arose in ATP6, COX3, ND3, ND4L, and the latter half of the 16s rRNA. These punctuated gene-specific accelerations were followed by the complete elimination of all atypical gene-specific signals of rate differentiation in contemporary henophidian lineages. We find no evidence for a constant accelerated rate of snake mtDNA evolution. Instead, our analyses of rates and patterns of substitution underscore both the spatial (gene-specific) and temporal (branch-specific) nature of molecular evolutionary rate dynamics in snake mtDNA.

Discussion

In this exploratory comparative analysis, we have investigated the potential causes and molecular evolutionary consequences of the unique mitochondrial genomic architecture of snakes. The three new complete snake mitochondrial genomes presented here, together with previously existing vertebrate genomes, compose an intriguing dataset that provides a preliminary perspective on a complex history of potentially adaptive genomic change in snakes. Unusual changes in gene size and nucleotide substitution rates have accompanied or followed the change in genomic architecture (Figure 20), but despite evidence for variable among-lineage functionality of the duplicate control region in snakes, the changes in substitution dynamics cannot be directly explained by the changes in genome architecture. Collectively, the patterns we

have identified over the course of snake mitochondrial genome evolution are most consistent with some type of broad selective pressure on the efficiency and function of oxidative metabolism in snakes.

Gene size reduction and control region functionality

All vertebrate mitochondrial genomes are compact, but nevertheless there is a strong trend for genes to be smaller in snakes than in other vertebrate mitochondrial genomes. Most of the reductions in gene lengths are evident in all snakes, including *Leptotyphlops* (Figures 20 and 21), but there are large further reductions in rRNA genes in the Colubroidea, and more moderate further reductions in tRNAs and some proteins. We do not have a direct measure of how this gene shortening affects the function of mitochondrial genes, but in the case of tRNAs, stability (presumably related to functionality) was only slightly affected by reduced length in snakes. It is interesting that the genomic size reduction due to gene shortening in alethinophidians is more than offset by the retention of duplicate control regions in alethinophidians, maintained by concerted evolution. This suggests that these dual CRs are maintained because they provide some selective advantage potentially including enhancement of mitochondrial genome replication and/or transcription, perhaps allowing these processes to occur more quickly (Sessions and Larson, 1987), or facilitating increased transcriptional control (see below).

Based on the genetic evidence of C/T gradients on the light strand, the duplicate control region appears to function in heavy strand replication in at least some snakes, although there is evidence for considerable variation in CR usage across snake lineages (Table 21). It is difficult to

extrapolate from the genetic data, however, a precise molecular model to explain the mechanism of dual control region function, and the mixed model weight cannot be directly interpreted as measuring control region functionality. For example, if the control regions usually function simultaneously and equally well in the same replication event, then it is possible that (due to their relative positions) the T_{AMS} of ND1 would be higher than the average of the two individual T_{AMS} , perhaps close to the value predicted if only CR2 were functional. In other words, strong evidence for a T_{AMS} consistent with CR2 function may indicate that CR2 functions alone during replication, but may also be indicative of dual CR function in each replication event. Future analyses with increased taxon sampling (especially with more closely related snake taxa) should help clarify patterns resulting from recent replication activity, and may be able to discern between potential molecular models.

Despite some uncertainty regarding the details of how dual control regions may be involved in genome replication, our data provide considerable evidence that all but one species (*Cylindrophis*) of alethinophidian snakes utilize CR2, to some extent, to initiate genome replication. A number of apparently evolutionarily independent origins of CR duplication, coupled with CR concerted evolution, have been recently identified in several divergent vertebrate lineages, including eels (Inoue et al., 2003), frogs (Sano et al., 2005), birds (Abbott et al., 2005; Eberhard et al., 2001), and lizards (Amer and Kumazawa, 2005; Kumazawa and Endo, 2004), although no examples are known from mammalian taxa. It seems reasonable to expect that these other vertebrates with dual CRs (homogenized by concerted evolution) may also use the duplicate CR or both CRs as origins of genome replication. Each of these examples is associated with unique rearrangements of genome architecture, and it would be interesting to search for

potential mutational effects of these rearrangements and evidence of differential or dual CR usage. In contrast, however, our results (and additional unpublished data) suggest that the dramatic shifts in rates and patterns of molecular evolution in snakes represent a unique phenomenon that we do not expect to be necessarily associated with CR duplication, but rather more likely associated with selection for mitochondrial function. As an example, the *Sphenodon* and *Varanus* samples included both have duplicated CRs, and the *Varanus* CRs are homogenized via concerted evolution, but no indications of dramatic rate dynamics were observed for either of these lineages.

Concerted evolution in and around the duplicate control regions

The control region appears to have duplicated only once in the ancestor of alethinophidian snakes over 70 MYA (Dong and Kumazawa, 2005; Kumazawa et al., 1996, 1998) (based on the fossil record of snakes (Rage, 1987)), and this duplication has been maintained in all alethinophidians sequenced to date (Figure 20). The two control regions clearly undergo concerted evolution to maintain reciprocal homogeneity between control regions within a genome (Dong and Kumazawa, 2005; Kumazawa et al., 1996, 1998), presumably through gene conversion. Two interesting points arise from the greater sampling of the relatively closely-related viperids and colubrids presented here.

First, there is an apparently nonfunctional partial (or pseudo) proline tRNA (Ψ -tRNA^{Pro}) in the colubrids that appears to be maintained by concerted evolution (Figure 17). In

Pantherophis, Ψ -tRNA^{Pro} is identical to the first 35 bp of tRNA^{Pro}, and in *Dinodon* the Ψ -

tRNA^{Pro} differs from tRNA^{Pro} by only a single insertion; thus, the Ψ -tRNA^{Pro} closely reflects the divergence patterns of functional tRNAs (there is only one indel between the tRNA^{Pro} from *Pantherphis* and *Dinodon*) rather than the pattern expected from nonfunctional DNA in a genome selected for reduction in gene size. In colubrids and most other snakes, tRNA^{Pro} is located between CR1 and tRNA^{Thr}, and the colubrid Ψ -tRNA^{Pro} is located in the same relative position next to CR2 and adjacent to tRNA^{Ile} (Figure 17).

The concerted evolution of these tRNAs could be explained by a tendency for gene conversion events involving the duplicate control regions to extend into the homologous tRNA regions. If this is correct, the Ψ -tRNA^{Pro} may be only slowly lost as differences accumulate at the end distal to CR2. It is possible that the pseudogene is a leftover remnant from the original duplication that created the duplicate control region.

The location of tRNA^{Pro} in *Agkistrodon* (and other viperids) between CR2 and tRNA^{Ile}, precisely where the Ψ -tRNA^{Pro} is located in colubrids (Figure 17), could also be explained as a remnant from the original CR duplication. Under this hypothesis, the functional tRNA^{Pro} of viperids would have been retained adjacent to the duplicate control region (CR2), and the original tRNA^{Pro} (adjacent to CR1) was eliminated or became a pseudogene. Both *Ovophis* and *Agkistrodon* have a 31 bp sequence between tRNA^{Thr} and CR1, but in *Ovophis* these 31 bp are identical to the CR2-proximal portion of the intact tRNA^{Pro}, while in *Agkistrodon* this 31 bp segment shares only 12 bp with the canonical tRNA^{Pro}, and is thus only marginally identifiable as homologous. Although this is not definitive proof of concerted evolution, it is suggestive that

there was only one duplication, and that concerted evolution has occurred recently in *Ovophis* and the colubrids, but that the Ψ -tRNA^{Pro} in *Agkistrodon* (Figure 17) has diverged too much, and is no longer capable of concerted evolution.

The time span during which both duplicate tRNA^{Pro} genes would have had to remain functional is long (i.e., tens of millions of years). If this is a remnant of the original CR duplication, it is surprising that the functional tRNA^{Pro} is almost always in the same location as in the colubrids. A simple alternative explanation is that a tRNA^{Pro} duplication occurred in some common ancestor of the Colubridae and Viperidae, and was resolved differently in different lineages. The gene conversion process that homogenizes the control region may occasionally pick up extra DNA, making tRNA^{Pro}, or part of it, prone to duplication at this location. Alternatively, gene duplications adjacent to the control region may simply be more likely to be preserved for long periods of time by concerted evolution. The existence of a duplicate tRNA^{Phe} between CR2 and tRNA^{Leu} in *Ovophis* (Dong and Kumazawa, 2005) makes repeated duplication seem a more likely possibility (these two tRNA^{Phe} differ by only 3 of 64 bp; implying either concerted evolution or recent duplication).

The second point of interest concerning gene conversion that arises from this study is a preliminary indication of differential evolutionary processes operating on the CRs within versus between species. Vertebrate mitochondrial control regions typically evolve very rapidly, and this is the case in a comparison of the two viperid species (*Ovophis* and *Agkistrodon*) in which CRs from these species are approximately as divergent as the fastest positions within the mtDNA, third codon positions (Figure 18B). In contrast, the two *Agkistrodon pisvictorus* genomes, *Api1*

and *Api2*, have surprisingly similar CRs between individuals (Figure 18A; Table 20), comparable to the similarity between rRNA genes, among the slowest regions in the mtDNA. A previous study on viperid snakes also showed slow within-species CR evolutionary rates (Ashton and de Queiroz, 2001), and other studies have demonstrated alternative rates of CR evolution operating within versus between species in fish (Tang et al., 2006).

In this study we have found a great deal of rate heterogeneity among genes, so it is certainly possible that the normally unconserved control regions have become suddenly critical and conserved in *Agkistrodon*. Alternatively, it is plausible that the complex (and poorly understood) process of gene conversion of CRs within a genome may also alter rates of CR evolution within species through a yet unknown process of gene conversion that may involve intragenomic (or even intergenomic) recombination. Although occasional cases of recombination between mitochondria have been proposed (Piganeau et al., 2004; Tsaousis et al., 2005), there is still very little evidence for a molecular mechanism to explain how concerted evolution in mitochondrial genomes may operate. A densely sampled collection (with intra and interspecific examples) of snake mtDNAs may eventually be able to directly address such questions.

Potential impacts of genome architecture on genome replication and transcription

In mitochondrial genomes (particularly in vertebrates), the processes of replication and transcription are not entirely functionally independent, and genome structural organization plays a prominent role in both processes. The CR acts as the origin of heavy strand replication, in addition to its role as the promoter for both heavy and light strand transcription (Fernandez-Silva et al., 2003). Genome replication also depends on the processing of light strand transcripts to

produce short primers required for heavy strand initiation of genome replication (originating from the CR (Clayton, 1982)). The regular distribution of the tRNA genes throughout the mtDNA is functionally significant, and these play an important role in RNA processing of polycistrons to yield mature RNAs, transcription initiation and termination, as well as initiation of light strand replication (Fernandez-Silva et al., 2003). Collectively, many functional ramifications are linked tightly to genome architecture in vertebrate mitochondria.

The possession of two functional control regions in most snake mtDNA could be advantageous by increasing the rate at which genome replication proceeds, and/or increasing the overall number of mtDNA copies per mitochondrion. It is also possible that dual control regions could alter patterns of transcription, since either could potentially serve as an origin of light or heavy strand transcripts.

Since the dual CRs essentially flank the rRNA genes, they (along with adjacent tRNAs) could also plausibly function to independently control rates of protein-coding and rRNA gene transcription. Across snake species, there are several alterations of the tRNAs flanking the CRs, including the translocation of tRNA^{Leu} (3' of CR2) and the duplication / translocation / truncation of tRNA^{Pro}. In vertebrates, tRNA^{Leu} has been shown to decouple rates of rRNA and mRNA transcription by acting as a terminator of ~95% of heavy strand transcripts (leading to ~20-fold higher rRNA vs. mRNA levels; (Fernandez-Silva et al., 2003)). Considering the ectothermy of snakes, transcriptional decoupling via independent control regions could provide a more direct means of countering thermodynamic depression of enzymatic rates at low temperatures.

The role of the tRNA^{Pro} in genome regulation is not entirely clear, but it is adjacent to the promoter site for light strand transcription (for some tRNAs and ND6), and is also adjacent to the initiation site for heavy strand replication. It is therefore plausible that tRNA^{Pro} plays roles in initiation or attenuation of both processes. Despite considerable progress in deciphering the molecular mechanisms involved in vertebrate mitochondrial replication and transcription, many intriguing questions remain regarding these processes. Vertebrate mtDNAs with unique mitochondrial genome architectures, such as alethinophidian snakes, represent an ideal comparative model for future research examining the impacts of genome architecture on mitochondrial function.

Comparative rates of molecular evolution

Previous studies have suggested that snake mitochondrial genomes have an accelerated rate of evolution (Dong and Kumazawa, 2005; Kumazawa et al., 1998). Our results suggest this general conclusion is actually an oversimplification of a much more complex scenario, and that rates of snake mtDNA evolution incorporate broad temporal (branch-specific) and spatial (gene and gene region-specific) dynamics. Ancestral branches early in snake evolution appear to be associated with dramatically elevated evolutionary rates and rate dynamics across the mitochondrial genome (Figure 20). In contrast, terminal snake lineages (branches) appear to have patterns of mtDNA evolution that are strikingly similar to other (non-snake) vertebrate mtDNAs. Our analyses here have concentrated on relative rates of evolution across the mtDNA, and future studies that incorporate a greater diversity of snake mtDNA together with estimates of absolute

rates of evolution (by calibrating nodes with divergence times) will be required to further characterize the absolute rate dynamics that have occurred.

There is no obvious reason why the existence of duplicate control regions or the usage of CR2 as an origin of heavy strand replication should result in genome-wide acceleration of protein evolutionary rates. Among protein-coding genes, only ND1 might be expected to experience relatively higher rates of evolution in genomes with duplicate CRs, due to higher rates of mutation (based on increased T_{AMS}), yet it and other ND genes are among the least accelerated of the mitochondrial protein-coding genes. Although it is possible that the usage of dual CRs leads to decreased accuracy of DNA synthesis (Kumazawa et al., 1998), we were unable to find evidence for an increased neutral transversion rate (data not shown), nor would this hypothesis explain the rate dynamics observed among genes.

Our results suggest that terminal alethinophidian branches have not experienced particularly accelerated rates of molecular evolution (except for rRNA in *Ovophis*), but that the early branches in snake evolution did experience highly differential rate acceleration that varied along lineages and among genes (Figure 20). The punctuated nature of this phenomenon suggests that the evolution of two CRs, gene shortening, and the variable molecular evolutionary rate dynamics may be collectively related by a larger pattern of selection for functionality (perhaps correlating with a shift in metabolic function).

In support of a hypothesis involving selection for overall oxidative metabolic function, the accelerated rates of molecular evolution in snakes appears to depend greatly on gene function, with most ND subunits accelerating only slightly and occasionally, while the COX, ATP, CytB, and rRNA evolutionary accelerations are dramatic and punctuated. The roles of

these accelerated proteins (and the mitochondria in general) in energetics via oxidative phosphorylation are well known, and it may be that a single causative agent accompanying the diversification of snakes that dramatically altered metabolic demand, or led to a fluctuation in metabolic demand, was responsible for large-scale changes in selective pressure on these proteins. If so, it may eventually be possible to find evidence for similar adaptive pressure on related nuclear-encoded snake proteins. It is worth noting that other cases have recently been identified in which mitochondrial proteins appear to have undergone bursts of selection in response to fluctuating energetic demands (McClellan et al., 2005).

We are undertaking a detailed analysis of coevolutionary interactions (Pollock et al., 1999; Wang and Pollock, 2005), three-dimensional structure, and site-specific selection events in snake mitochondrial proteins in an attempt to understand this acceleration in greater functional detail. This requires further sampling of snake genomes to obtain sufficient accuracy and statistical power, and is complicated by the ancient nature of the evolutionary acceleration; the most dramatic evidence for acceleration exists at the base of the Serpentes clade rather than in modern snake lineages (Figure 20).

Conclusions

Snake mitochondrial genomes present a rare opportunity to observe, and investigate, the evolutionary interactions and functional ramifications that link genome architecture, molecular evolution, and multi-level molecular function. Available evidence points to selective pressures acting at many hierarchical levels of snake mitochondrial genomes, and at different times during

snake evolution, leading to diverse, dramatic, and broad-scale changes in the genome. Interestingly, some consequences of this adaptive shift appear to have diminished over time (e.g., accelerated rates of COX and other gene evolution), whereas others appear to continue in modern snakes (i.e., the effects of control region duplication on mutation gradients, replication, and potentially transcription, and remnant functional consequences of short and highly substituted genes). Although the precise cause is unknown, this outstanding example of an apparent punctuated adaptive shift involving multiple aspects of genome architecture evolution provides an important comparative tool for the study of vertebrate mitochondrial genome evolution. Overall, this highlights the need for further comparative genomic research in snakes to provide more accurate resolution of evolutionary patterns and possible site-specific effects of mutational dynamics.

References

- Abbott, C.L., Double, M.C., Trueman, J.W.H., Robinson, A., Cockburn, A., 2005. An unusual source of apparent mitochondrial heteroplasmy: duplicate mitochondrial control regions in *Thalassarche* albatrosses. *Mol. Ecol.* 14, 3605-3613.
- Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov, B.N., Csake, F. (Eds.), *Second International Symposium on Information Theory*. Akademia Kiado, Budapest, pp. 673-681.
- Akaike, H., 1983. Information measures and model selection. *Int. Stat. Inst.* 22, 277-291.
- Amer, S.A.M., Kumazawa, Y., 2005. Mitochondrial genome of *Pogona vitticeps* (Reptilia; Agamidae): control region duplication and the origin of Australasian agamids. *Gene* 346, 249-256.
- Ashton, K.G., de Queiroz, A., 2001. Molecular systematics of the western rattlesnake, *Crotalus viridis* (Viperidae), with comments on the utility of the D-loop in phylogenetic studies of snakes. *Mol. Phylogenet. Evol.* 21, 176-189.
- Bielawski, J.P., Gold, J.R., 1996. Unequal synonymous substitution rates within and between two protein-coding mitochondrial genes. *Mol. Biol. Evol.* 13, 889-892.
- Bielawski, J.P., Gold, J.R., 2002. Mutation patterns of mitochondrial H- and L-Strand DNA in closely related cyprinid fishes. *Genetics* 161, 1589-1597.
- Burbrink, F.T., 2002. Phylogeographic analysis of the cornsnake (*Elaphe guttata*) complex as inferred from maximum likelihood and Bayesian analyses. *Mol. Phylogenet. Evol.* 25, 465-476.

- Clayton, D.A., 1982. Replication of animal mitochondrial DNA. *Cell* 28, 693-705.
- Cooper, A., Lalueza-Fox, C., Anderson, S., Rambaut, A., Austin, J., Ward, R., 2001. Complete mitochondrial genome sequences of two extinct moas clarify ratite evolution. *Nature* 409, 704-707.
- Dong, S., Kumazawa, Y., 2005. Complete mitochondrial DNA sequences of six snakes: Phylogenetic relationships and molecular evolution of genomic features. *J. Mol. Evol.* 61, 12-22.
- Eberhard, J.R., Wright, T.F., Bermingham, E., 2001. Duplication and concerted evolution of the mitochondrial control region in the parrot genus *Amazona*. *Mol. Biol. Evol.* 18, 1330-1342.
- Faith, J.J., Pollock, D.D., 2003. Likelihood analysis of asymmetrical mutation bias gradients in vertebrate mitochondrial genomes. *Genetics* 165, 735-745.
- Fernandez-Silva, P., Enriquez, J.A., Montoya, J., 2003. Replication and transcription of mammalian mitochondrial DNA. *Exp. Physiol.* 88, 41-56.
- Frank, A.C., Lobry, J.R., 1999. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene* 238, 65-77.
- Frederico, L.A., Kunkel, T.A., Shaw, B.R., 1990. A sensitive genetic assay for the detection of cytosine deamination - determination of rate constants and the activation energy. *Biochemistry (Mosc)*. 29, 2532-2537.
- Gower, D.J., Vidal, N., Spinks, J.N., McCarthy, C.J., 2005. The phylogenetic position of Anomochilidae (Reptilia: Serpentes): first evidence from DNA sequences. *Journal of Zoological Systematics and Evolutionary Research* 43, 315-320.

- Helm, M., Brule, H., Friede, D., Giege, R., Putz, D., Florentz, C., 2000. Search for characteristic structural features of mammalian mitochondrial tRNAs. *RNA* 6, 1356-1379.
- Hofacker, I.L., Fontana, W., Stadler, P.F., Bonhoeffer, L.S., Tacker, M., Schuster, P., 1994. Fast folding and comparison of RNA secondary structures. *Monatshefte Fur Chemie* 125, 167-188.
- Holt, I.J., Jacobs, H.T., 2003. Response: The mitochondrial DNA replication bubble has not burst. *Trends Biochem. Sci.* 28, 355-356.
- Hughes, S., Mouchiroud, D., 2001. High evolutionary rates in nuclear genes of squamates. *J. Mol. Evol.* 53, 70-76.
- Impellizzeri, K.J., Anderson, B., Burgers, P.M.J., 1991. The spectrum of spontaneous mutations in a *Saccharomyces cerevisiae* Uracil-DNA-Glycosylase mutant limits the function of this enzyme to cytosine deamination repair. *J. Bacteriol.* 173, 6807-6810.
- Inoue, J.G., Miya, M., Tsukamoto, K., Nishida, M., 2003. Evolution of the deep-sea gulper eel mitochondrial genomes: Large-scale gene rearrangements originated within the eels. *Mol. Biol. Evol.* 20, 1917-1924.
- Janke, A., Erpenbeck, D., Nilsson, M., Arnason, U., 2001. The mitochondrial genomes of the iguana (*Iguana iguana*) and the caiman (*Caiman crocodylus*): implications for amniote phylogeny. *Proc Biol Sci* 268, 623-631.
- Jermiin, L.S., Graur, D., Crozier, R.H., 1995. Evidence from analyses of intergenic regions for strand-specific directional mutation pressure in metazoan mitochondrial DNA. *Mol. Biol. Evol.* 12, 558-563.
- Krishnan, N.M., Raina, S.Z., Pollock, D.D., 2004a. Analysis of among-site variation in substitution patterns. *Biological Procedures Online* 6, 180-188.

- Krishnan, N.M., Seligmann, H., Raina, S.Z., Pollock, D.D., 2004b. Detecting gradients of asymmetry in site-specific substitutions in mitochondrial genomes. *DNA Cell Biol.* 23, 707-714.
- Kumazawa, Y., Endo, H., 2004. Mitochondrial genome of the Komodo dragon: Efficient sequencing method with reptile-oriented primers and novel gene rearrangements. *DNA Res.* 11, 115-125.
- Kumazawa, Y., Nishida, M., 1999. Complete mitochondrial DNA sequences of the green turtle and blue-tailed mole skink: Statistical evidence for Archosaurian affinity of turtles. *Mol. Biol. Evol.* 16, 784-792.
- Kumazawa, Y., Ota, H., Nishida, M., Ozawa, T., 1996. Gene rearrangements in snake mitochondrial genomes: Highly concerted evolution of control-region-like sequences duplicated and inserted into a tRNA gene cluster. *Mol. Biol. Evol.* 13, 1242-1254.
- Kumazawa, Y., Ota, H., Nishida, M., Ozawa, T., 1998. The complete nucleotide sequence of a snake (*Dinodon semicarinatus*) mitochondrial genome with two identical control regions. *Genetics* 150, 313-329.
- Lawson, R., Slowinski, J.B., Crother, B.I., Burbrink, F.T., 2005. Phylogeny of the Colubroidea (Serpentes): New evidence from mitochondrial and nuclear genes. *Mol. Phylogenet. Evol.* 37, 581-601.
- Lowe, T.M., Eddy, S.R., 1997. tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25, 955-964.

- McClellan, D.A., Palfreyman, E.J., Smith, M.J., Moss, J.L., Christensen, R.G., Sailsbery, A.K., 2005. Physicochemical evolution and molecular adaptation of the cetacean and artiodactyl cytochrome b proteins. *Mol. Biol. Evol.* 22, 437-455.
- Mindell, D.P., Sorenson, M.D., Dimcheff, D.E., 1998. Multiple independent origins of mitochondrial gene order in birds. *Proc. Natl. Acad. Sci. U. S. A.* 95, 10693-10697.
- Perna, N.T., Kocher, T.D., 1995a. Patterns of nucleotide composition at fourfold degenerate sites of animal mitochondrial genomes. *J. Mol. Evol.* 41, 353-358.
- Perna, N.T., Kocher, T.D., 1995b. Unequal base frequencies and estimation of substitution rates. *Mol. Biol. Evol.* 12, 359-361.
- Piganeau, G., Gardner, M., Eyre-Walker, A., 2004. A broad survey of recombination in animal mitochondria. *Mol. Biol. Evol.* 21, 2319-2325.
- Pollock, D.D., Taylor, W.R., Goldman, N., 1999. Coevolving protein residues: Maximum likelihood identification and relationship to structure. *J. Mol. Biol.* 287, 187-198.
- Pond, S.L.K., Frost, S.D.W., Muse, S.V., 2005. HyPhy: hypothesis testing using phylogenies. *Bioinformatics* 21, 676-679.
- Posada, D., Crandall, K.A., 1998. ModelTest: testing the model of DNA substitution. *Bioinformatics* 14, 817-818.
- Rage, J.C., 1987. Fossil history. In: R. A. Seigel, J.T.C., S. S. Novak (Ed.), *Snakes: Ecology and Evolutionary Biology*. Macmillian, New York, pp. 51-76.
- Raina, S.Z., Faith, J.J., Disotell, T.R., Seligmann, H., Stewart, C.B., Pollock, D.D., 2005. Evolution of base-substitution gradients in primate mitochondrial genomes. *Genome Res.* 15, 665-673.

- Rand, D.M., Kann, L.M., 1998. Mutation and selection at silent and replacement sites in the evolution of animal mitochondrial DNA. *Genetica* 103, 393-407.
- Reyes, A., Gissi, C., Pesole, G., Saccone, C., 1998. Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Mol. Biol. Evol.* 15, 957-966.
- Reyes, A., Yang, M.Y., Bowmaker, M., Holt, I.J., 2005. Bidirectional replication initiates at sites throughout the mitochondrial genome of birds. *J. Biol. Chem.* 280, 3242-3250.
- Ronquist, F., Huelsenbeck, J.P., 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* 19, 1572-1574.
- Sankoff, D., Leduc, G., Antoine, N., Paquin, B., Lang, B.F., Cedergren, R., 1992. Gene order comparisons for phylogenetic inference - evolution of the mitochondrial genome. *Proc. Natl. Acad. Sci. U. S. A.* 89, 6575-6579.
- Sano, N., Kurabayashi, A., Fujii, T., Yonekawa, H., Sumida, M., 2005. Complete nucleotide sequence of the mitochondrial genome of Schlegel's tree frog *Rhacophorus schlegelii* (family Rhacophoridae): duplicated control regions and gene rearrangements. *Genes Genet. Syst.* 80, 213-224.
- Sessions, S.K., Larson, A., 1987. Developmental correlates of genome size in plethodontid salamanders and their implications for genome evolution. *Evolution* 41, 1239-1251.
- Shadel, G.S., Clayton, D.A., 1997. Mitochondrial DNA maintenance in vertebrates. *Annu. Rev. Biochem.* 66, 409-435.
- Slack, K.E., Janke, A., Penny, D., Arnason, U., 2003. Two new avian mitochondrial genomes (penguin and goose) and a summary of bird and reptile mitogenomic features. *Gene* 302, 43-52.

- Swofford, D.L., 1997. PAUP*. Phylogenetic Analysis Using Parsimony (* and Other Methods). Sinauer Associate, Sunderland, Massachusetts.
- Szczesny, B., Hazra, T.K., Papaconstantinou, J., Mitra, S., Boldogh, I., 2003. Age-dependent deficiency in import of mitochondrial DNA glycosylases required for repair of oxidatively damaged bases. *Proc. Natl. Acad. Sci. U. S. A.* 100, 10670-10675.
- Tanaka, M., Ozawa, T., 1994. Strand asymmetry in human mitochondrial-DNA mutations. *Genomics* 22, 327-335.
- Tang, Q., Liu, H., Mayden, R., Xiong, B., 2006. Comparison of evolutionary rates in the mitochondrial DNA cytochrome b gene and control region and their implications for phylogeny of the Cobitoidea (Teleostei: Cypriniformes). *Mol Phylogenet Evol.* 39, 347-357. Epub 2005 Oct 2004.
- Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., Higgins, D.G., 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res.* 25, 4876-4882.
- Tsaousis, A.D., Martin, D.P., Ladoukakis, E.D., Posada, D., Zouros, E., 2005. Widespread recombination in published animal mtDNA sequences. *Mol. Biol. Evol.* 22, 925-933.
- Utiger, U., Helfenberger, N., Schätti, B., Schmidt, C., Ruf, M., Ziswiler, V., 2002. Molecular systematics and phylogeny of Old World and New World ratsnakes, *Elaphe Auct.*, and related genera (Reptilia, Squamata, Colubridae). *Russ. J. Herpetol.* 9, 105-124.
- Wang, Z.O., Pollock, D.D., 2005. Context dependence and coevolution among amino acid residues in proteins. *Methods Enzymol.* 395, 779-790.

- Yang, M.Y., Bowmaker, M., Reyes, A., Vergani, L., Angeli, P., Gringeri, E., Jacobs, H.T., Holt, I.J., 2002. Biased incorporation of ribonucleotides on the mitochondrial L-strand accounts for apparent strand-asymmetric DNA replication. *Cell* 111, 495-505.
- Yang, Z.H., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* 13, 555-556.
- Yasukawa, T., Yang, M.Y., Jacobs, H.T., Holt, I.J., 2005. A bidirectional origin of replication maps to the major noncoding region of human mitochondrial DNA. *Mol. Cell* 18, 651-662.

CHAPTER 6 – CONCLUSION

Advancing the Framework for Functional Comparative Genomics

The primary limitation imposed on the exploitation of vast genomic resources currently available is our limited abilities to distill meaningful comparative and functionally relevant patterns from these practically infinite arrays of four nucleotides. Given that essentially the entirety of biological diversity is encoded via the patterns of nucleotide occurrence in genomic sequences, developing our abilities to understand and extract information about these patterns is absolutely crucial to advancing our understanding of the function and diversity of biological systems. Only through the development of a robust comparative framework, whereby information about the evolutionary relationships among compared units may be synthesized together with structural and functional information, may these vast genomics resources yield meaningful insight into the biological relevance of the variation and conservation across genes and genomes.

To advance our ability to understand and draw conclusions from these patterns via a comparative framework, we conducted several studies that examined 1) methodologies for complex modeling of nucleotide evolution, including the impact these methods have on increasing the power and accuracy of phylogenetic inference, 2) exploratory analyses that examine novel potential links and interrelationships between genome structure, function, and nucleotide evolution, and 3) an truly extreme example of a massive genome-wide adaptive shift

that appears to have altered nucleotide evolution, genome architecture, and overall molecular function of the genome itself and of the gene products (both RNAs and proteins) encoded by the mitochondrial genome. Collectively, these studies significantly contribute a critical foundation required for functional comparative genomics by overcoming previous practical, theoretical, and methodological limitations, while also providing crucial examples that demonstrates coordinated changes in genome structure, function, and nucleotide evolution may collectively contribute substantially to system-wide biological functional change.

Complex Modeling of the Nucleotide Evolutionary Process

Likelihood-based methods, including Bayesian Markov-chain Monte Carlo (MCMC) methods, have greatly improved our ability to estimate evolutionary patterns using larger datasets and complex models of evolution. However, this also has lead to a seemingly paradoxical dilemma with regard to evolutionary model complexity. In general, it is assumed that more realistic models of evolution will yield more accurate phylogenetic estimates and clade credibility (posterior probability) values, thus perhaps favoring parameter-rich models, since interpretations of posterior probabilities are contingent on model specifications (Huelsenbeck et al., 2002). A key assumption of Wald's (1949) proof of the consistency of maximum likelihood estimates, however, is that all of the parameters of the likelihood function are identifiable from the true probability distribution of the data (Rogers, 2001). Even if a particular parameter may be intrinsic in the evolution of DNA sequences, we need to consider whether this parameter can be

accurately estimated based on the data. This dilemma is manifest when attempting to construct and implement models that realistically describe DNA evolution, while avoiding overparameterization, or using more parameters than can be meaningfully estimated from the data. Thus, despite considerations favoring complex models, benefits of constructing and implementing more realistic evolutionary models of DNA substitution are challenged by the potential for imprecise and inaccurate model parameter and phylogeny estimation that may result from excess model complexity. Expanding computational power, increasing genomic resources, and advances allowing broad flexibility in modeling evolutionary patterns in a Bayesian MCMC context collectively underscore the importance of developing accurate models and objective strategies for model testing.

We have developed a three-part strategy for identifying, testing, and evaluating candidate complex models in a Bayesian MCMC context. 1) We have employed simple Akaike Information Criteria (Akaike, 1973, 1974, 1983; Sakamoto et al., 1986) to identify best-fit models for independent biologically intuitive (potential) partitions of the dataset (see also Brandley et al., 2005). 2) To identify the best partitioning strategy for heterogeneous datasets, we have developed a three-part means of cross-validation using marginalized Akaike weights (novel; see also Buckley et al., 2002), Bayes factors (Nylander et al., 2004), and the Relative Bayes Factors (novel) to examine model fit across alternatively partitioned Bayesian MCMC models. 3). To test the important assumption that models are not excessively parameter rich (which may lead to serious problems in accuracy), we have developed an array of *post-hoc* model evaluation methods to evaluate the performance of complex models and to check for proper mixing and convergence. Collectively this novel and robust strategy facilitates the use of

more realistic complex models of nucleotide evolution (which we have shown are necessary for accurate evolutionary inference) while ensuring that issues that may lead to problems with employing overly complex models are avoided.

Across the multiple studies on complex modeling that we have conducted (including manuscripts not included in this dissertation listed in references: Castoe et al., 2007; Castoe et al., in press; Doan et al., 2005; Herron et al., 2004) we have found that complex models of nucleotide evolution are extremely critical for accurate phylogenetic inference, and thus essential for meaningful comparative genomic analyses. Our results support four important conclusions relevant to the use of complex partition-specific models in combined MCMC analyses. 1) Model choice may have important practical effects on phylogenetic conclusions even for smaller datasets. 2) The use of complex partitioned models does not produce widespread increases or decreases in inferred support for phylogenetic relationships. 3) A majority of differences in resolution resulting from model choice is concentrated at deeper nodes, thus complex models become more critical as sequence divergence increases. Also, a majority of these deeper nodes increased substantially in resolution (as measured by nodal posterior probability support) with increasing model complexity. 4) Appropriately complex models appear to facilitate superior exploration of tree and parameter space, thus increasing the speed and effectiveness of evaluation of all possible estimates to determine the most optimal and accurate set of likely possibilities.

Since we observed substantial differences between estimates based on simple versus complex models, two important questions arise with regard to these differences. The first is, to what extent is an unpartitioned model forced to compromise estimates of model parameters in the analysis of a complex heterogeneous dataset, versus a complex model that contains several

distinct partitions of evolutionary patterns for portions of the data? In other words, how misleading may a single model really be (in terms of nucleotide substitution model parameters) if used for a complex dataset? Our results suggest that this compromise is extreme in some cases, and is evident across different classes of model parameters. Comparisons of the 95% confidence intervals (CIs) of parameter estimates derived from simple models show many instances where 95% CIs of partitions do not overlap those based on the unpartitioned simple model. Furthermore, many CIs that do overlap between simple and complex models do not coincide for a majority of their posterior densities. These findings point directly at the elevated potential for an unpartitioned model to fall into the trap identified in simulation studies where an oversimplified model suffers from decreased accuracy. Collectively, available evidence supports not only the use of complex models (including partitioned models), but implies that these may be crucial for accurate phylogenetic estimates (see also Huelsenbeck and Rannala, 2004). Consequently, these results suggest that accounting for the realistic heterogeneity of nucleotide evolution using complex models is essential for accurate, meaningful comparative genomic inference.

Given the differences between the estimates based on simple and complex models, the second question arises: how should the differences in phylogenetic hypotheses between simple and complex models be interpreted? We found complex models to result in changes in phylogenetic support (posterior probabilities) for clades that, in some instances, altered the estimate of the consensus topology. These changes tended to provide higher support in complex models, with a majority of changes concentrated at deeper nodes (e.g., Brandley et al., 2005; see also Alfaro et al., 2003). This observation raises two possibilities, either complex models result

in over-inflated support estimates, or they provide (at least on average) more accurate estimates of nodal support. Three points of evidence suggest that complex models do generally provide to more accurate, rather than over-inflated, posterior probability estimates: 1) the results of simulation studies discussed above, 2) empirical studies, including the studies included in this dissertation, demonstrating that even though a majority of nodes may increase, some decrease under complex model analyses (see also Brandley et al., 2005; Nylander et al., 2004), and 3) results described here that show a coincidence between clades that show increased support under complex-model analyses and are also supported by other independent data (see also Doan et al., 2005; Castoe et al., 2007; Castoe et al., in press).

It may not be immediately obvious how these studies that utilize organismal phylogeny examples are relevant to the broad field of functional comparative genomics and ultimately to human biology and disease. To illustrate this direct connection, I present several examples from our recent work (not included in this dissertation) that build upon these examples of nucleotide modeling to make important inferences about broader biological questions.

Although not completed, we have initiated collaborative work on the functional evolution of the Rho GTPase family of proteins across eukaryotes aimed at understanding the evolutionary functional context of Rho diversification, and also at definitively identifying orthologous and paralogous members of this gene family across model systems (yeast, flies, worms, mammals, and humans). The importance of understanding the evolution and relationships among RhoGTPases is illustrated by the massive differential expansion of some members of this protein family across eukaryotes (e.g., 5 members in *C. elegans* to over 20 members in mammals). This large and diverse family of proteins, ubiquitous among eukaryotes, is so named based on their

high homology with the small GTPase Ras, the first oncogene identified. Accordingly, Rho-family GTPases are of extreme interest based on their roles in directing cellular differentiation, development, and cancer. Unlike organismal phylogeny questions, deciphering genealogical relationships among members of this gene family is particularly difficult because of the small size of the protein involved (typically < 200 codons) which only permits a small number of aligned homologous nucleotides to be analyzed. Our results clearly show that all previous estimates of Rho family phylogeny are significantly inaccurate (e.g., Boureux et al., 2006; Wherlock and Mellor, 2002), and that using complex nucleotide modeling strategies to estimate the phylogeny of Rho GTPases provides a drastically novel perspective, whereby the current views of which members are homologs across model systems is very incorrect. Our comparative genomic perspective suggests a completely new model of which functional types of Rho GTPases evolved first, and also which single Rho proteins in some model systems have numerous paralogous sister proteins in humans, while other proteins in model systems have essentially no homolog in humans are less interesting for human health.

Similarly, our recent work on the eukaryotic type II polyketide and fatty acid synthase (PKS/FAS) gene family demonstrated significantly different estimates of the evolutionary relationships, grouping very unique homologous clusters of genes across animals (Castoe et al., 2007). These results also provided strong evidence for several novel groups of non-FAS PKS genes in some animal genomes that may play key roles in primitive innate immunity. We have also been able to show (only using these advanced models) that there has been a strong trend of gene loss of these novel PKS genes in many animal lineages that suggests that the function of

these genes may be either strongly favored or disfavored, depending on the unique physiology and immune system function of particular animal groups.

Lastly, in ongoing studies to understand in more functional detail the dramatic patterns of evolutionary change that we have discovered in vertebrate mitochondrial genomes discussed above (in Chapter 5), these advances in nucleotide modeling have been critical to the accuracy and ultimate inference of functional patterns of genomic change. Using these complex models has allowed us to add extensive power to our ability to detect even subtle changes in the patterns of nucleotide change, facilitating our identification of differential activity of mitochondrial control region function (in initiation of genome replication) and also apparent changes in the activity of the gamma DNA polymerase activity of different lineages of animals. Understanding these types of dynamics has many functional ramifications, including our enhanced ability to accurately and meaningfully compare different animal models used for investigating mitochondria-related diseases. Additionally, using different models of nucleotide evolution provide very different estimates of how selection may have driven the major remodeling of the mitochondrial genomes of snakes, and these more accurate complex models yield phylogeny and character-state change estimates that drastically revise earlier estimates of where, when, and through correlations with other co-occurring phenomenon, why these changes may have occurred.

Collectively, the work in this dissertation that focuses on modeling the nucleotide evolutionary process has made a succinct and significant contribution that directly satisfies the initial goal of removing limitations on the functional utilization of genomic data by constructing a cohesive and robust framework for comparative genomic analyses. Our work demonstrates

that, given an infrastructure of careful model selection and evaluation, complex modeling of the nucleotide evolutionary process not only contributes to, but is required for accurate inference of phylogenetic and evolutionary patterns estimated from comparative genomic data.

Insight into the Evolutionary Process at the Genome Scale

In this exploratory comparative genomic analysis, we have focused on the extreme example of snake mitochondrial genomes, and investigated the potential causes and molecular evolutionary consequences of the unique mitochondrial genomic architecture of snakes. The novel complete snake mitochondrial genomes presented here, together with previously existing vertebrate genomes, compose an intriguing dataset that provides a preliminary perspective on a complex history of potentially adaptive genomic change in snakes that involved a coordinated change in genome structure, nucleotide evolution, and molecular function. Unusual changes in gene size and nucleotide substitution rates have accompanied changes in genomic architecture, but despite evidence for variable among-lineage functionality of the duplicate control region in snakes, the changes in substitution dynamics cannot be directly explained by the changes in genome architecture alone. Collectively, the patterns we have identified over the course of snake mitochondrial genome evolution are most consistent with some type of broad selective pressure on the efficiency and function of oxidative metabolism in snakes.

Previous to the work here, studies have suggested that snake mitochondrial genomes (mtDNA) have an accelerated rate of evolution (Dong and Kumazawa, 2005; Kumazawa et al.,

1998). Our results suggest this general conclusion is actually a drastic oversimplification of a much more complex scenario, and that rates of snake mtDNA evolution incorporate broad temporal (branch-specific) and spatial (gene and gene region-specific) dynamics. Ancestral branches early in snake evolution appear to be associated with dramatically elevated evolutionary rates and rate dynamics across the mitochondrial genome. In contrast, terminal (recent) snake lineages appear to have patterns of mtDNA evolution that are strikingly similar to other (non-snake) vertebrate mtDNAs.

The punctuated nature of these drastic changes in evolutionary rates suggests that the evolution of two mitochondrial control regions, gene shortening, and the variable molecular evolutionary rate dynamics may be collectively related by a larger pattern of selection for functionality (perhaps correlating with a shift in metabolic function). In support of a hypothesis involving selection for overall oxidative metabolic function, the accelerated rates of molecular evolution in snakes appears to depend greatly on gene function, with most NADH-dehydrogenase subunits accelerating only slightly and occasionally, while the Cytochrome C Oxidase (COX), Cytochrome-B (Cyt-B), ATPase, and rRNA evolutionary accelerations are dramatic and extremely punctuated. The roles of these accelerated proteins (and the mitochondria in general) in energetics via oxidative phosphorylation are well known, and it may be that a single causative agent accompanying the diversification of snakes that dramatically altered metabolic demand, or led to a fluctuation in metabolic demand, was responsible for large-scale changes in selective pressure on these proteins. It is also interesting that the two most accelerated protein complexes in snake mtDNAs also happen to be involved in mitochondrially-induced apoptosis as they bind Cytochrome-C, which when released is the primary trigger for

initiation of the intrinsic apoptotic pathway. Presently, however, there is no other data which obviously link the abnormal patterns of snake mtDNA evolution with the control of programmed cell death.

Snake mitochondrial genomes present a rare opportunity to observe and investigate the evolutionary interactions and functional ramifications that link genome architecture, molecular evolution, and multi-level molecular function. Available evidence points to selective pressures acting at many hierarchical levels of snake mitochondrial genomes, and at different times during snake evolution, leading to diverse, dramatic, and broad-scale changes in the genome. Interestingly, some consequences of this adaptive shift appear to have diminished over time (e.g., accelerated rates of COX and other gene evolution), whereas others appear to continue in modern snakes (i.e., the effects of control region duplication on mutation gradients, replication, and potentially transcription, and remnant functional consequences of short and highly substituted genes). Our ongoing work on comparative mitochondrial genomics of snakes is aimed at testing hypotheses for the cause of these changes, and also investigating the spectrum of potential functional ramifications these changes may have lead to. Although the precise cause is unknown, this outstanding example of an apparent punctuated adaptive shift involving multiple aspects of genome architecture evolution provides an important comparative tool for the study of vertebrate mitochondrial genome evolution, and a novel example of coordinated structural, molecular and functional evolution for the field of comparative genomics in general.

References

- Akaike, H., 1973. Information theory and an extension of the maximum likelihood principle. In: Second International Symposium on Information Theory. Akademiai Kiado, Budapest, pp. 673–681.
- Akaike, H., 1974. A new look at the statistical model identification. *IEEE Trans. Aut. Control* 19, 716–723.
- Akaike, H., 1983. Information measures and model selection. *Int. Stat. Inst.* 22, 277–291.
- Alfaro, M.E., Zoller, S., Lutzoni, F., 2003. Bayes or bootstrap? A simulation study comparing the performance of Bayesian Markov chain Monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. *Mol. Biol. Evol.* 20, 255–266.
- Brandley, M.C., Schmitz, A., Reeder, T.W., 2005. Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of scincid lizards. *Syst. Biol.* 54, 373–390.
- Boureux, A., Vignal, E., Faure, S., Fort, P., 2006. Evolution of the Rho family of Ras-like GTPases in eukaryotes. *Mol. Biol. Evol.* 24, 203–216.
- Buckley, T.R., 2002. Model misspecification and probabilistic tests of topology: evidence from empirical data sets. *Syst. Biol.* 51, 509–523.
- Castoe, T.A., Stephens, T.S., Noonan, B.P., Calestani, C.L., 2007. A novel group of type I polyketide synthases (PKS) in animals and the complex phylogenomics of PKSs. *Gene*.

- Castoe, T.A., Smith, E.N., Brown, R.M., Parkinson, C.L., In press. Higher-level phylogeny of Asian and American coralsnakes, their placement within the Elapidae (Squamata), and the systematic affinities of the enigmatic Asian coralsnake *Hemibungarus calligaster*. *Zool. J. Linnean Soc.*
- Doan, T.M., Castoe, T.A., Arizabal Arriaga, W., 2005. Phylogenetic relationships of the genus *Proctoporus* sensu stricto (Squamata: Gymnophthalmidae), with a new species from Puno, southeastern Peru. *Herpetologica* 61, 325–336.
- Dong, S., Kumazawa, Y., 2005. Complete mitochondrial DNA sequences of six snakes: Phylogenetic relationships and molecular evolution of genomic features. *J. Mol. Evol.* 61, 12–22.
- Herron, M.D., Castoe, T.A., Parkinson, C.L., 2004. Sciurid phylogeny and the paraphyly of Holarctic ground squirrels (*Spermophilus*). *Mol. Phylogenet. Evol.* 31, 1015–1030
- Huelsenbeck, J.P., 2002. Testing a covariotide model of DNA substitution. *Mol. Biol. Evol.* 19, 698–707.
- Huelsenbeck, J.P., Rannala, B., 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Syst. Biol.* 53, 904–913.
- Kumazawa, Y., Ota, H., Nishida, M., Ozawa, T., 1998. The complete nucleotide sequence of a snake (*Dinodon semicarinatus*) mitochondrial genome with two identical control regions. *Genetics* 150, 313–329.
- McClellan, D.A., Palfreyman, E.J., Smith, M.J., Moss, J.L., Christensen, R.G., Sailsbery, A.K., 2005. Physicochemical evolution and molecular adaptation of the cetacean and artiodactyl cytochrome b proteins. *Mol. Biol. Evol.* 22, 437–455.

- 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.
- Rogers, J. S., 2001. Maximum likelihood estimation of phylogenetic trees in consistent when substitution rates vary according to the invariable sites plus gamma distribution. *Syst. Biol.* 50, 713–722.
- Sakamoto, Y., Ishiguro, M., Kitagawa, G., 1986. Akaike Information Criterion Statistics. Springer, NY.
- Wald, A., 1949. Note on the consistency of the maximum likelihood estimate. *Ann. Math. Stat.* 20, 595–601.
- Wherlock, M., Mellor, H., 2002. The Rho GTPase family: a Racs to Wrchs story. *J. Cell Sci.* 115, 239–240.