The phylogeny of advanced snakes (Colubroidea), with discovery of a new subfamily and comparison of support methods for likelihood trees

R. Alexander Pyron [a,b,*], Frank T. Burbrink [b,f], Guarino R. Colli [c], Adrian Nieto Montes de Oca [d], Laurie J. Vitt [e], Caitlin A. Kuczynski [a], John J. Wiens [a]

[a] Dept. of Ecology and Evolution, Stony Brook University, Stony Brook, NY 11794-5245, United States
[b] Dept. of Biology, The Graduate School and University Center, The City University of New York, 365 Fifth Ave., NY, NY 10016, United States
[c] Departamento de Zoologia, Universidade de Brasília, Brasília, DF Brasil 70910-900, Brazil
[d] Museo de Zoología, Facultad de Ciencias, Universidad Nacional Autónoma de México, México, Distrito Federal 04510, Mexico
[e] Sam Noble Oklahoma Museum of Natural History and Zoology Dept., 2401 Chautauqua Avenue, University of Oklahoma, Norman, OK 73072, United States
[f] Dept. of Biology, The College of Staten Island, The City University of New York, 2800 Victory Blvd., Staten Island, NY 10314, USA

ARTICLE INFO

ABSTRACT

The superfamily Colubroidea (>2500 species) includes the majority of snake species and is one of the most conspicuous and well-known radiations of terrestrial vertebrates. However, many aspects of the phylogeny of the group remain contentious, and dozens of genera have yet to be included in molecular phylogenetic analyses. We present a new, large-scale, likelihood-based phylogeny for the colubroids, including 761 species sampled for up to five genes: cytochrome *b* (93% of 761 species sampled), ND4 (69%), ND2 (28%), c-mos (54%), and RAG-1 (13%), totaling up to 5814 bp per species. We also compare likelihood bootstrapping and a recently proposed ultra-fast measure of branch support (Shimodaira-Hasegawa-like [SHL] approximate likelihood ratio), and find that the SHL test shows strong support for several clades that were weakly-supported by bootstrapping in this or previous analyses (e.g., Dipsadinae, Lamprophiidae). We find that SHL values are positively related to branch lengths, but show stronger support for shorter branches than bootstrapping. Despite extensive missing data for many taxa (mean = 67% per species), neither bootstrap nor SHL support values for terminal species are related to their incompleteness, and that most highly incomplete taxa are placed in the expected families from previous taxonomy, typically with very strong support. The phylogeny indicates that the Neotropical colubrine genus *Scaphiodontophis* represents an unexpectedly ancient lineage within Colubridae. We present a revised higher-level classification of Colubroidea, which includes a new subfamily for *Scaphiodontophis* (Scaphiodontophiinae). Our study provides the most comprehensive phylogeny of Colubroidea to date, and suggests that SHL values may provide a useful complement to bootstrapping for estimating support on likelihood-based trees.

© 2010 Elsevier Inc. All rights reserved.

1. Introduction

Dense taxon sampling is extremely important for phylogenetic and evolutionary studies. For example, extensive taxon sampling may greatly increase phylogenetic accuracy under some conditions (e.g., Rannala et al., 1998; Zwickl and Hillis, 2002) and allow more accurate estimates of diversification rates (e.g., Heath et al., 2008; Cusimano and Renner, 2010). However, major challenges to inferring large-scale phylogenies remain. One is the expense and time

necessary to obtain tissue samples and comparable character sampling for hundreds of species and many genes. Another is the difficulty of estimating trees and support values using sophisticated model-based methods (e.g., maximum likelihood) on large-scale data matrices in a reasonable amount of time.

Some recent advances have made inferring large-scale phylogenies more tractable for many groups. One is the finding that large matrices with extensive missing data can yield well-supported trees that are largely congruent with traditional taxonomy (e.g., Driskell et al., 2004; Philippe et al., 2004; Wiens et al., 2005; Thomson and Shaffer, 2010). The supermatrix approach to phylogenetics involves gathering all or most available data and analyzing it simultaneously (de Queiroz and Gatesy, 2007). This easily permits new sequence data to be combined with existing information from databases such as GenBank (e.g., Sanderson et al., 2003) to yield densely sampled supermatrices. Second, recent computational innovations have greatly facilitated estimating large-scale,

* Corresponding author. Present address: Department of Biological Sciences, The George Washington University, 2023 G St. NW, Washington, DC 20052, United States.

E-mail addresses: rpyron@colubroid.org (R.A. Pyron), frank.burbrink@csi.cuny.edu (F.T. Burbrink), grcolli@unb.br (G.R. Colli), anietomontesdeoca@me.com (A.N.M. de Oca), vitt@ou.edu (L.J. Vitt), kate@life.bio.sunysb.edu (C.A. Kuczynski), wiensj@life.bio.sunysb.edu (J.J. Wiens).

likelihood-based phylogenies (e.g., Stamatakis et al., 2008; Guindon et al., 2010; Price et al., 2010). Third, fast new methods for assessing clade confidence for likelihood trees have been developed that provide an alternative to traditional, time-intensive methods such as non-parametric bootstrapping (e.g., Anisimova and Gascuel, 2006; Guindon et al., 2010).

For many species-rich groups, supermatrix strategies (e.g., Wiens et al., 2005; de Queiroz and Gatesy, 2007; Thomson and Shaffer, 2010) offer a useful approach for inferring large-scale phylogenies when different amounts of character data are available across taxa. Other potential methods for large-scale phylogenetic inference include supertree construction (e.g., Bininda-Emonds, 2004) and mega-phylogeny approaches (e.g., Smith et al., 2009). The supertree method involves grafting trees inferred from different datasets into a single phylogeny, but it suffers from the need for *a priori* assumptions about which species belong to which taxa, limiting the potential for new discoveries about the phylogeny. The mega-phylogeny approach is similar to the supermatrix strategy, but uses an automated pipeline to identify gene regions and homologous sequence clusters of interest. We prefer the supermatrix strategy, as it most directly incorporates the largest amount of sequence data into the phylogenetic analysis, without assuming placement of species within groups *a priori*.

Here, we produce a new, large-scale phylogeny for the superfamily Colubroidea, the advanced snakes (*sensu* Lawson et al., 2005). Colubroids are among the most diverse groups of extant terrestrial vertebrates (>2500 species; Lawson et al., 2005) despite their relatively recent origin in the Cenozoic (Burbrink and Pyron, 2008; Vidal et al., 2009). They occur on every continent except Antarctica (Vitt and Caldwell, 2009) and include many common and familiar groups (e.g., racers, garter, rat, king, and milk snakes), and all known dangerously venomous snake species, such as elapids (cobras, sea snakes, and mambas) and viperids (e.g., rattlesnakes, adders, and vipers). These venomous colubroids are responsible for ∼20,000–94,000 human fatalities every year (Kasturiratne et al., 2008). Given their diversity and broad distribution, colubroids have been the focus of many phylogeny-based studies in historical biogeography (e.g., Keogh, 1998; Pinou et al., 2004; Alfaro et al., 2008; Pyron and Burbrink, 2009a; Daza et al., 2010) and evolutionary biology (e.g., Fry and Wüster, 2004; Lynch, 2009; Pyron and Burbrink, 2009b,c; Burbrink and Pyron, 2010). However, despite the great biological and medical significance of this group, no study has offered a comprehensive assessment of the higher-level phylogeny of Colubroidea. For example, none has included representatives of all currently recognized subfamilies in a single analysis.

Several recent authors have addressed relationships within Colubroidea using DNA sequence data (e.g., Lawson et al., 2005; Burbrink and Pyron, 2008; Wiens et al., 2008; Kelly et al., 2009; Vidal et al., 2009; Zaher et al., 2009), typically sampling either many genes for relatively few taxa (e.g., seven genes for 24 species in Vidal et al., 2007; 20 genes for 29 species in Wiens et al., 2008) or many taxa for fewer genes (e.g., three genes for 131 species in Zaher et al., 2009). Major changes to colubroid taxonomy have been proposed based on these studies (e.g., Lawson et al., 2005; Burbrink et al., 2007; Vidal et al., 2007; Zaher et al., 2009). Yet, relatively few species and genera were included in these phylogenies, leaving the classification of many genera in question. These gaps in taxon sampling may hide radical differences between traditional taxonomies and molecular phylogenies. For example, the genus *Oxyrhabdium* was traditionally thought to belong to Xenodermatidae (Vitt and Caldwell, 2009), but molecular phylogenetic analyses showed it to be nested within Lamprophiidae (see Lawson et al., 2005; Kelly et al., 2009; Zaher et al., 2009).

Many of these recent molecular phylogenies agree regarding some relationships, such as monophyly of Homalopsidae and Viperidae. However, substantive disagreements remain regarding many parts of the phylogeny. One is the monophyly of the predominantly African assemblage "Lamprophiidae" (Kelly et al., 2009). Some studies have supported the monophyly of this group (Vidal et al., 2007; Burbrink and Pyron, 2008; Wiens et al., 2008; Zaher et al., 2009), whereas others have found it to be paraphyletic with respect to Elapidae (Kelly et al., 2009). Another is the placement of the family Homalopsidae as the sister taxon either to Elapidae + Lamprophiidae (Burbrink and Pyron, 2008), or to Elapidae + Lamprophiidae + Colubridae (Lawson et al., 2005; Vidal et al., 2007; Wiens et al., 2008). Yet another is the placement of the colubrid subfamily Dipsadinae as the sister taxon either to Natricinae (Vidal et al., 2008; Kelly et al., 2009; Zaher et al., 2009), Colubrinae (Vidal et al., 2007), or Colubrinae + Natricinae (Lawson et al., 2005; Wiens et al., 2008), with other colubrid subfamilies, such as Calamariinae and Pseudoxenodontinae (if sampled), often found interdigitated among these clades (Lawson et al., 2005; Zaher et al., 2009). These issues are important for numerous reasons, including understanding the relationships among medically significant taxa, and the interpretation of historical biogeographic scenarios. All of these questions are best addressed through a large-scale phylogenetic analysis of Colubroidea, using as many taxa as possible to resolve relationships within the group.

Here, we address colubroid relationships and classification using a supermatrix approach that combines data for two nuclear genes, three mitochondrial genes, and 761 colubroid species in 299 genera, totaling 70% of the 426 known genera and 29% of the 2654 identified species (The Reptile Database: Uetz, 2009; http://www.reptile-database.org/). In previous studies, dozens of researchers have generated sequences for hundreds of colubroid species, with five genes being commonly used (mitochondrial cytochrome *b*, ND2, and ND4; and nuclear c-mos and RAG-1). We also present new sequence data for 41 additional species (38 genera) from the two most species-rich colubroid subfamilies (Dipsadinae and Colubrinae), most of which have never been included in a molecular phylogenetic analysis. We combine these new sequences with existing data from previous studies to produce the largest analysis of Colubroidea to date, containing nearly six times as many species as any previous estimate, and including all known families and subfamilies in the same analysis for the first time.

We also compare two methods for estimating clade support for large-scale, likelihood-based phylogenies. Specifically, we compare support values from traditional, non-parametric bootstrapping (BS hereafter; Felsenstein, 1985, 2004) and the non-parametric Shimodaira-Hasegawa-like approximation of the likelihood-ratio test statistic (SHL hereafter; Anisimova and Gascuel, 2006; Guindon et al., 2010). Using SHL support values may be desirable, especially for large trees, as calculating them can be several orders of magnitude faster than assessing traditional BS support (Anisimova and Gascuel, 2006) and both measures seem to give similar values (Guindon et al., 2010). Although the thorough study by Guindon et al. (2010) addressed many aspects of the relative performance of these methods using empirical and simulated datasets, some key questions remain. Here, we assess the relationship between these values and branch lengths and missing data (in terminal taxa), questions that were not addressed in previous studies. We hypothesize that SHL values will be positively related to branch lengths (as shown for likelihood BS values; Wiens et al., 2008), but will show higher values than BS values on shorter branches (suggested but not explicitly tested by Guindon et al., 2010). We also hypothesize that there will be little or no relationship between SHL support values for clades and the proportion of missing data in the terminal taxa in those clades, nor for likelihood BS values (as shown for parsimony BS values and Bayesian posterior probabilities; Wiens et al., 2005). A key assumption of the supermatrix approach is that extensive missing data in terminal taxa need not prevent them from being placed in the tree with strong support.