Molecular Phylogenetics and Evolution 84 (2015) 53–63

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

journal homepage: www.elsevier.com/locate/ympev

Building the avian tree of life using a large-scale, sparse supermatrix

J. Gordon Burleigh *, Rebecca T. Kimball, Edward L. Braun

Department of Biology, University of Florida, United States

ARTICLE INFO

Article history:
Received 9 July 2014
Revised 3 December 2014
Accepted 5 December 2014
Available online 27 December 2014

Keywords:
Supermatrix
Phylogeny
Birds

ABSTRACT

Birds are the most diverse tetrapod class, with about 10,000 extant species that represent a remarkable evolutionary radiation in which most taxa arose during a short period of time. There has been a tremendous increase in the amount of molecular data available from birds, and more than two-thirds of these species have some sequence data available. Here we assembled these available sequence data from birds to estimate a large-scale avian phylogeny. We performed an unconstrained maximum likelihood analysis of a sparse supermatrix comprising 22 nuclear loci and seven mitochondrial regions from 6714 species. We inferred a phylogeny with a backbone remarkably similar to that obtained by detailed analyses of multigene datasets, yet with the addition of thousands of more taxa. All orders were monophyletic with generally high support. While most families and genera were well supported, a number of them, especially within the oscine passerines, had little or no support. This likely reflects problems with the circumscription of these genera and families. Our results indicate that the amount of sequence data currently available is sufficient to produce a robust estimate of the avian tree of life using current methods of inference. The availability of a tree that is unconstrained by prior information, with branch lengths that have a direct connection to the underlying data, should be useful for comparative methods, taxonomic revisions, and prioritizing taxa that should be targeted for additional data collection.

1. Introduction

Understanding the tree of life has been a central challenge for evolutionary biologists since the time of Darwin. Our ability to infer the structure of that tree has been transformed by advances in DNA sequencing technology, phylogenetic algorithms, and high performance computing (reviewed by Delsuc et al., 2005; Whelan, 2008; McCormack et al., 2013a,b), which have enabled the generation of extensive phylogenies (e.g., Goloboff et al., 2009; Thomson and Shaffer, 2010; Pyron et al., 2013; Rabosky et al., 2013; Soltis et al., 2013). Such large, species-level phylogenetic hypotheses, ideally including measures of branch lengths and estimates of topological and branch length uncertainty, can be powerful tools for comparative analyses of species and character evolution (Harvey and Pagel, 1991; Ricklefs, 2007) and for examining patterns of biodiversity (Tingley and Dubey, 2012; Holt et al., 2013), ecological implications of global change (Edwards et al., 2007), the structure of communities and ecosystems (Emerson and Gillespie, 2008; Cavender-Bares et al., 2009; Vamosi et al., 2009), and conservation priorities (Forest et al., 2007; Diniz-Filho et al., 2013; Jetz et al., 2014). However, synthesizing heterogeneous phylogenetic data, and ultimately extending the taxonomic sampling of past phylogenies without sacrificing the accuracy or quality of the estimated phylogenetic trees, presents a major challenge for evolutionary biologists.

Large-scale phylogenies have been generated using several different approaches. Supertree approaches combine trees with partial taxon overlap (Sanderson et al., 1998; Bininda-Emonds et al., 2002; Bininda-Emonds, 2004) to generate large tree estimates (e.g., Liu et al., 2001; Bininda-Emonds et al., 2007). Since the trees are the input for supertree analyses, this allows the inclusion of trees based on many different types of data, including traditional taxonomies (e.g., Liu et al., 2001; Cardillo et al., 2006). Although most supertrees lack branch length data, several methods to incorporate branch lengths onto supertrees have been developed (Gittleman et al., 2004; Webb and Donoghue, 2004; Torices, 2010; Eastman et al., 2013). Alternatively, supermatrix approaches combine data from different sources into a single matrix for analyses. Often these data are molecular sequences (e.g., Driskell et al., 2004; Thomson and Shaffer, 2010), but sometimes they include both molecules and morphology (Gatesy et al., 2003; Goloboff et al., 2009). Given sufficient data, a supermatrix can provide credible estimates of phylogeny where the positions of all taxa and the branch lengths are directly linked to underlying data (Gatesy et al., 2004). Methods that combine features of supertree and

* Corresponding author at: Department of Biology, University of Florida, P.O. Box 118526, Gainesville, FL 32611, United States.
E-mail addresses: gburleigh@ufl.edu (J.G. Burleigh), rkimball@ufl.edu (R.T. Kimball), ebraun68@ufl.edu (E.L. Braun).

http://dx.doi.org/10.1016/j.ympev.2014.12.003
1055-7903/

supermatrix analyses, such as the mega-phylogeny approach, which generates character matrices by using taxonomic hierarchies to combine sequences (Smith et al., 2009), have also been proposed. Still other approaches incorporate species that are not represented by sequence data using traditional taxonomic information (Webb and Donoghue, 2004; Bininda-Emonds et al., 2007; Kuhn et al., 2011; Jetz et al., 2012; Smith et al., 2013a; Thomas et al., 2013).

Birds represent an excellent group for the generation of a large, taxon-rich phylogeny. There has been extensive research on avian ecology, behavior, conservation, development, genetics, and physiology (e.g., Bonnet et al., 2002), and over three quarters of extant species are represented by at least one publication in the Zoological Record database (Ducatez and Lefebvre, 2014). These data, when put in a comparative framework, have the potential to yield new insights into many biological processes. While relationships among avian families and orders have been controversial, recent studies are converging on a set of basal relationships, both for birds as a whole (Ericson et al., 2006; Hackett et al., 2008; McCormack et al., 2013a,b; Kimball et al., 2013) and for the most diverse avian order (Passeriformes; Barker et al., 2002; Barker et al., 2004; Ericson et al., 2014). Thus, the current state of avian phylogeny is ready for the generation of more taxon-rich phylogenies. Indeed, comprehensive taxon-rich phylogenies of birds are beginning to be generated. Jetz et al. (2012) used a modified supermatrix approach that was able to place taxa not represented by sequence data. Since Jetz et al. (2012) used relatively few loci in the supermatrix, some nodes were constrained to match the Ericson et al. (2006) or Hackett et al. (2008) trees, and the tree was constructed using a "staged" approach. Two large-scale avian supertrees have also been published recently (Holt et al., 2013; Davis and Page, 2014); however, neither tree has meaningful branch lengths, and the use of trees without branch lengths has been criticized (Kreft and Jetz, 2013).

Similar to other taxonomic groups, the data available for birds are heterogeneous, creating several challenges for the assembly of a supermatrix. Some taxa are sampled for many loci, while other taxa are only represented by a single locus; some loci are sampled for many species, but no locus is sequenced for all sampled species. This uneven distribution of data can create analytical difficulties (Lemmon et al., 2009; Sanderson et al., 2011; Simmons and Goloboff, 2013). There are now sequence data for over 70% of all avian species, most of which are sampled for one of two mitochondrial regions, which may ameliorate problems associated with lack of sequence overlap among taxa. However, it is unclear whether a supermatrix generated from these data can be used to estimate a robust phylogeny congruent with expected nodes. Therefore, we assembled the available molecular data from birds to estimate a taxon-rich avian phylogeny that reflects the tremendous increases in data and advances in computational phylogenetics.

2. Methods

2.1. Data assembly

To assemble a supermatrix with the greatest possible taxonomic and gene sampling while remaining computationally feasible for a maximum likelihood (ML) phylogenetic analysis, we downloaded all core nucleotide sequences in GenBank (http://www.ncbi.nlm.nih.gov; Benson et al., 2009) from birds (class Aves) that were available by June, 2011. We sought to maximize the amount of data in the character matrix while minimizing the overall size of the matrix, a balance achieved by including genes with sequences from many taxa and much taxonomic overlap with the other genes. Additionally, we needed to align genes across all birds and focus on genes that were phylogenetically informative across divergent avian taxa. Thus, we chose to focus on 22 nuclear loci with broad sampling to form the backbone phylogenetic hypothesis of birds and seven mitochondrial gene regions for which sequences were available from the most bird species (Table 1).

For each gene, we identified available sequences in GenBank using BLAST searches (Altshul et al., 1990). We did this to directly identify all available homologous sequences rather than relying on the highly variable (since they are user-defined) GenBank gene name annotations for individual sequences. For the nuclear loci from Hackett et al. (2008; Table 1), we used the original sequences from Hackett et al. (2008) as a query for a BLASTN search against all other core nucleotide sequences from Aves in GenBank (Altshul et al., 1990). We defined as homologs all sequences that had a maximum *E*-value of 10^{-10} (1.0E−10) and $\geqslant$50% overlap of both the target and query sequences. We excluded one of the loci (BDNF) used by Hackett et al. (2008) because it exhibits base compositional convergence that results in a strong misleading signal (Harshman et al., 2008; Kimball et al., 2013; Smith et al., 2013b). To identify sets of homologous sequences from the remaining nuclear and mitochondrial loci, we clustered sequences less than 10,000 bp in length based on results from an all-by-all pairwise BLAST analysis. The all-by-all BLASTN search was conducted using blastall (Altshul et al., 1990) with the default parameters. A Perl script was used to identify the largest single-linkage clusters of sequences, in which all sequences in a cluster form a connected component, with edges liking sequences (nodes) representing a significant BLAST hit, and each sequence has a significant BLAST hit against at least one other sequence in the cluster.

Table 1
Loci used in the supermatrix, the number of species with sequences (based on the GenBank taxonomy), and the length of the gene alignment after pruning columns with less than 4 nucleotides. An asterisk next to the locus name indicates that the locus was also used in the Hackett et al. (2008) matrix.

Locus	NCBI species	Length (bp)
Mitochondrial		
12S	1113	861
COI	2159	1580
COII	1030	1077
CytB	4898	1145
ND2	4232	1037
ND3	1111	625
ND4	279	903
Nuclear		
ALDOB*	184	2429
CLTC*	169	2082
CMOS	477	608
CRYAA*	151	1379
EEF2*	144	2059
EGR1UTR*	292	533
EGR1exon*	369	1226
FGBint4*	178	759
FGBint5*	1113	765
FGBint67	1343	2107
GH1*	206	1588
HMGN2*	104	1949
IRF2*	163	633
MB*	1528	867
MUSK*	317	704
MYC*	280	1369
NGF*	169	750
NTF3*	170	728
ODC1	1106	961
PCBD1*	162	1390
RAG1	1435	2930
RAG2	595	1155
RHO*	492	2103
TGFB2*	731	741
TPM1*	226	568