Molecular Phylogenetics and Evolution 69 (2013) 1021–1032

Contents lists available at SciVerse ScienceDirect

Molecular Phylogenetics and Evolution

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

Identifying localized biases in large datasets: A case study using the avian tree of life

Rebecca T. Kimball [a,*], Ning Wang [a,b,1], Victoria Heimer-McGinn [a,2], Carly Ferguson [a,3], Edward L. Braun [a]

[a] Department of Biology, University of Florida, Gainesville, FL 32611, United States
[b] MOE Key Laboratory for Biodiversity Sciences and Ecological Engineering, College of Life Sciences, Beijing Normal University, Beijing 100875, China

ARTICLE INFO

Article history:
Received 28 September 2012
Revised 12 May 2013
Accepted 29 May 2013
Available online 20 June 2013

Keywords:
Phylogenomics
Incongruence
Localized biases
Gene tree discordance

ABSTRACT

Large-scale multi-locus studies have become common in molecular phylogenetics, with new studies continually adding to previous datasets in an effort to fully resolve the tree of life. Total evidence analyses that combine existing data with newly collected data are expected to increase the power of phylogenetic analyses to resolve difficult relationships. However, they might be subject to localized biases, with one or a few loci having a strong and potentially misleading influence upon the results. To examine this possibility we combined a newly collected 31-locus dataset that includes representatives of all major avian lineages with a published dataset of 19 loci that has a comparable number of sites (Hackett et al., 2008. Science 320, 1763–1768). This allowed us to explore the advantages of conducting total evidence analyses, and to determine whether it was also important to analyze new datasets independent of published ones. The total evidence analysis yielded results very similar to the published results, with only slightly increased support at a few nodes. However, analyzing the 31- and 19-locus datasets separately highlighted several differences. Two clades received strong support in the published dataset and total evidence analysis, but the support appeared to reflect bias at a single locus (β-fibrinogen [FGB]). The signal in FGB that supported these relationships was sufficient to result in their recovery with bootstrap support, even when combined with 49 loci lacking that signal. FGB did not appear to have a substantial impact upon the results of species tree methods, but another locus (brain-derived neurotrophic factor [BDNF]) did have an impact upon those analyses. These results demonstrated that localized biases can influence large-scale phylogenetic analyses but they also indicated that considering independent evidence and exploring multiple analytical approaches could reveal them.

© 2013 Elsevier Inc. All rights reserved.

1. Introduction

Improvements in technology have made it possible to collect large, multi-locus datasets for phylogenetic studies. These multi-locus datasets may be extracted from whole genomes (Rokas et al., 2003; Wildman et al., 2007), reflect large-scale *de novo* data collection (Dunn et al., 2008; Hackett et al., 2008), the combination of data from multiple studies into a total evidence analysis (Kimball et al., 2011), or some combination of those approaches. These large datasets have led to increased resolution and support for many nodes in the Tree of Life. However, even with large datasets that appear likely to have the power to robustly identify phylogenetic relationships, conflicts among large-scale datasets have been identified (e.g., compare Dunn et al., 2008; Philippe et al., 2009; Schierwater et al., 2009). Although analyses that can identify some sources of conflict have been proposed (e.g., Philippe et al., 2011), it seems clear that unexpected clades recovered in phylogenetic analyses of large datasets, even those with high support values, should be considered hypotheses that should be subjected to additional tests.

Several phenomena can lead to high support in analyses of large molecular datasets. The simplest, and probably most common, is that the support reflects evolutionary history. However, both systematic and localized biases can result in incorrect estimates of phylogeny, sometimes with high levels of support. There has been substantial attention paid to systematic biases, such as long-branch attraction (Felsenstein, 1978) and convergence in base composition (e.g., Jeffroy et al., 2006; Phillips et al., 2004), which can result in strong non-historical signal, though the use of better-fitting models and noise reduction methods can sometimes address these biases (e.g., Braun and Kimball, 2002; Pratt et al., 2009).

* Corresponding author. Address: Department of Biology, P.O. Box 118525, University of Florida, Gainesville, FL 32611, United States. Fax: +1 352 392 3704.
E-mail address: rkimball@ufl.edu (R.T. Kimball).
[1] Present address: MOE Key Laboratory for Tropical Plant and Animal Ecology, College of Life Sciences, Hainan Normal University, Haikou 571158, China.
[2] Present address: Department of Biochemistry, University College Cork, Cork, Ireland.
[3] Present address: Department of Chemistry & Biochemistry, University of California Los Angeles, Los Angeles, CA, United States.

1055-7903/$ - see front matter © 2013 Elsevier Inc. All rights reserved.
http://dx.doi.org/10.1016/j.ympev.2013.05.029

Likewise, there are specific cases where the majority of gene trees differ from the species tree (Degnan and Rosenberg, 2006). Additional challenges for phylogenetic estimation include alignment errors (Lake, 1991; Liu et al., 2010) and the incorrect identification of orthologs (Philippe et al., 2011). Finally, there are also examples where unexpected phylogenetic signal appears to be limited to individual genes or specific subsets of the genome (Katsu et al., 2009; Rokas et al., 2003). However, it is unclear how often, if at all, these localized biases result in misleading conclusions when large-scale datasets are analyzed, but the possibility that they can be problematic needs to be explored.

Since analyses using large datasets are expected to reduce the variance of the estimated phylogeny there has been limited concern regarding the specific gene regions collected for various studies. Moreover, it has been common to combine data collected as part of previous studies into these larger datasets (e.g., Gatesy et al., 2002; Kimball and Braun, 2008; Pratt et al., 2009; Shen et al., 2012; Thomson and Shaffer, 2010). This practice intrinsically results in datasets with overlapping genes and it could be problematic if one or more genes included in these analyses exhibit strong localized biases. It has been suggested that when enough loci are sampled, any misleading phylogenetic signal localized to specific loci should not affect the conclusions of phylogenetic analyses (Rokas et al., 2003). Indeed, the Rokas et al. (2003) phylogenomic analyses revealed that analyses using most collections of 20 or more genes supported the same phylogeny, despite the existence of substantial, often well-supported, incongruence among estimates of phylogeny based upon individual genes (suggesting some localized biases were likely present). Nonetheless it remains important to examine this more broadly to determine whether localized biases are generally unimportant in large-scale datasets and only systematic biases need to be considered.

The avian tree of life represents an interesting test case for this type of analysis. The topology of the avian tree has been particularly difficult to elucidate due to a rapid radiation at the base of the largest group of birds, Neoaves (which represents over 95% of all avian species; Sibley and Ahlquist, 1990). In fact, Neoaves has been suggested to represent a hard polytomy (Poe and Chubb, 2004), though two large-scale analyses (Ericson et al., 2006; Hackett et al., 2008) have identified supraordinal clades that appear strongly supported. Moreover, there was no incongruence between these two studies for well-supported nodes, though analyses of the larger dataset from Hackett et al. (2008) resulted in more nodes with support than Ericson et al. (2006). However, those two studies used some of the same loci, and thus could be affected by similar localized biases.

Two of the novel and strongly supported relationships in Hackett et al. (2008) have been re-evaluated using datasets that had no overlapping loci (Smith et al., 2013; Wang et al., 2012) and transposable element (TE) insertions (Haddrath and Baker, 2012; Suh et al., 2011). Both Wang et al. (2012) and Smith et al. (2013) searched for misleading phylogenetic signal, and uncovered no evidence that either localized or global biases affected the nodes in question. The conclusions from both of these studies were congruent with Hackett et al. (2008) for the specific relationships being examined (the limited taxon sampling in those studies prevented many additional relationships from being compared). McCormack et al. (2013) used a large number of ultraconserved elements (UCEs) from up to 32 species in Neoaves, providing the ability to test some additional clades identified by Hackett et al. (2008). The strongly supported groups in McCormack et al. (2013) largely corroborated the conclusions of Hackett et al. (2008), with a single conflict in one of two analyses (cf. McCormack et al., 2013, Fig. 2A *versus* B). However, the more limited taxon sampling of these subsequent studies make it difficult to determine whether localized biases can influence the conclusions of large-scale datasets like that used by Hackett et al. (2008).

Here we extended the Hackett et al. (2008) data matrix (hereafter the 19-locus dataset) by adding data from 31 loci, providing a total of 50 loci for analysis. To allow examination of all the higher-level relationships proposed by Hackett et al. (2008), we chose a sample of 77 taxa representing all major avian clades that were selected to break up long branches and to largely overlap with the taxa in Hackett et al. (2008). The additional 31 loci were focused on non-coding regions and resulted in a dataset that was similar in size to Hackett et al. (2008). We concatenated the two datasets into a 50-locus dataset and analyzed this using ML and partitioned ML methods with two different alignment approaches. After conducting the total molecular evidence analysis, we explored whether separate analyses of the 31-locus and 19-locus datasets supported similar clades and exhibited similar levels of bootstrap support relative to each other and to the combined 50-locus dataset. We also searched for localized biases with the potential to drive incongruence by comparing results from the 31- and 19-locus datasets. Finally, we estimated the species tree using individual gene trees. Although our approaches provided corroboration for many of the relationships found by Ericson et al. (2006) and Hackett et al. (2008), it also highlighted a localized bias that affected a small number of relationships that were supported by both studies. These results indicated that exploring independent evidence and multiple analytical strategies may provide useful information that is complementary to total evidence analyses.

2. Methods

2.1. Data collection

To generate the 31-locus data matrix, we added sequences to the data that were collected by Braun et al. (2011), Kimball et al. (2009), Smith et al. (2013), and Wang et al. (2012) (loci are listed in Supplementary material Table S1). None of the loci included in the 31-locus dataset were included in Hackett et al. (2008), so this dataset was independent of that study (as well as Ericson et al., 2006). The loci were non-coding regions, primarily introns (with the short segments of coding exon that flanked introns trimmed prior to analyses) but also including two untranslated regions (UTRs). The 50 loci were located on 17 chromosomes in the chicken genome; loci on the same chromosome are separated (e.g., Kimball et al., 2009) and thus unlikely to be linked. Since there appears to be strong conservation of chromosome structure in birds (Griffin et al., 2007), separation in the chicken genome suggests there should also be little or no linkage in other taxa.

The taxa used included all those from Smith et al. (2013), Wang et al. (2012), and the "moderate effort" taxon sample of Kimball et al. (2009), plus additional taxa to subdivide long branches and target the inclusion of at least two species in all major clades when possible (Supplementary material Table S2). Most species were included in Hackett et al. (2008); however, we added several additional species, including the zebra finch (*Taeniopygia guttata*) where data was taken from the draft genome (Warren et al., 2010), the kea (*Nestor notabilis*; added in Wang et al. (2012)), Darwin's rhea (*Pterocnemia pennata*; included in Harshman et al. (2008) and Smith et al. (2013)) and the black-legged seriema (*Chunga burmeisteri*; to provide a second taxon in Cariamidae, a family placed in an unexpected position by Hackett et al. (2008)). For these taxa we downloaded zebra finch data for the Hackett et al. (2008) loci and we amplified and sequenced some loci that were used in Hackett et al. (2008) for the other species added, allowing us to include these taxa in both datasets.

Download English Version:

https://daneshyari.com/en/article/5919731

Download Persian Version:

https://daneshyari.com/article/5919731

Daneshyari.com