Molecular Phylogenetics and Evolution 57 (2010) 485–494

Contents lists available at ScienceDirect

Molecular Phylogenetics and Evolution

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

Analyzing the relationship between sequence divergence and nodal support using Bayesian phylogenetic analyses

Robert Makowsky *, Christian L. Cox, Corey Roelke, Paul T. Chippindale

University of Texas at Arlington, Department of Biology, Box 19498, Arlington, TX 76019, USA

ARTICLE INFO

Article history:
Received 28 July 2009
Revised 6 May 2010
Accepted 11 May 2010
Available online 21 May 2010

Keywords:
Divergence
Phylogenetics
Bayesian
Sequence

ABSTRACT

Determining the appropriate gene for phylogeny reconstruction can be a difficult process. Rapidly evolving genes tend to resolve recent relationships, but suffer from alignment issues and increased homoplasy among distantly related species. Conversely, slowly evolving genes generally perform best for deeper relationships, but lack sufficient variation to resolve recent relationships. We determine the relationship between sequence divergence and Bayesian phylogenetic reconstruction ability using both natural and simulated datasets. The natural data are based on 28 well-supported relationships within the subphylum Vertebrata. Sequences of 12 genes were acquired and Bayesian analyses were used to determine phylogenetic support for correct relationships. Simulated datasets were designed to determine whether an optimal range of sequence divergence exists across extreme phylogenetic conditions. Across all genes we found that an optimal range of divergence for resolving the correct relationships does exist, although this level of divergence expectedly depends on the distance metric. Simulated datasets show that an optimal range of sequence divergence exists across diverse topologies and models of evolution. We determine that a simple to measure property of genetic sequences (genetic distance) is related to phylogenic reconstruction ability in Bayesian analyses. This information should be useful for selecting the most informative gene to resolve any relationships, especially those that are difficult to resolve, as well as minimizing both cost and confounding information during project design.

1. Introduction

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

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

Rate of molecular evolution within and among genes is a simple characteristic of a dataset that may affect phylogenetic performance. In the case of rapidly evolving sequences, alignment and determination of character homology may be difficult or impossible (Blouin et al., 1998; Lopez et al., 1999; Xia et al., 2003). For intraspe-

* Corresponding author. Fax: +1 817 272 2855.
E-mail address: makowsky@uab.edu (R. Makowsky).

1055-7903/$ - see front matter © 2010 Published by Elsevier Inc.
doi:10.1016/j.ympev.2010.05.009

cific analyses, many mitochondrial genes, as well as nuclear markers such as microsatellites and AFLPs, usually provide phylogenetic signal without saturation (Berendzen et al., 2003; Creer et al., 2004; Dawson, 2001; Downie, 2004; Koopman, 2005; Vekemans et al., 2002). For slowly evolving sequences, the number of variable sites (and therefore the number of informative sites) will be low and incomplete lineage sorting of ancestral polymorphisms may obscure relationships (Maddison, 1997; Maddison and Knowles, 2006; Takahashi et al., 2001). Deeper relationships require more slowly evolving genes (e.g. nuclear ribosomal genes) to recover the correct topology (Avise, 2000; Hare, 2001; Palumbi et al., 2001). However, the same genes may evolve at different absolute rates across lineages, so a taxonomic consideration is important. For example, cytochrome oxidase I may be the fastest evolving mitochondrial gene in some lineages, while NADH-II may be the fastest in others (Kumazawa et al., 2004; Mueller, 2006).

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

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

Here, we determine whether there exists an optimal range of sequence divergence with broad applicability across taxa and divergence times. Specifically, we determine if researchers should aim for a particular range of sequence divergence during phylogenetic analysis planning so as to maximize the probability of recovering the correct topology. Our goals are to (1) identify (if possible) a global range of sequence divergence that maximally recovers the correct topology and (2) determine whether different types of genes (mitochondrial or nuclear, protein encoding or ribosomal) exhibit specific ranges of divergence for optimal phylogenetic reconstruction. We use a well-corroborated phylogeny (that we treat as "known" for the purpose of analyses) and compare the trees recovered from 12 genes using Bayesian methods to the assumed true topology to determine at what levels of sequence divergence phylogenetic methods most often recover the correct topology. We also used simulations to test whether a relationship between sequence divergence and phylogenetic reconstruction ability exists across different topologies and models of evolution. Additionally, we test whether intrinsic properties of the trees (branch lengths and unequal rates of sequence evolution across taxa) affect node support.

2. Materials and methods

2.1. Natural datasets

For our model phylogeny we started with the "known" phylogeny presented in Russo et al. (1996) and added taxa based on sequence availability and strength of relationship support (Fig. 1). We followed the phylogenetic relationships presented in multiple independent analyses (citations below refer to support for each relationship) using multiple types of character sets. Within mammals, the whales in the Russo et al. tree were reduced to one taxonomic unit and five OTUs were added from the following lineages; Canidae (Node 23), Felidae (two taxa; Node 24), Marsupiala (Node 18), and Primata (Node 20) (Douady and Douzery, 2003; Hudelot et al., 2003; Lin et al., 2002; Liu et al., 2001; Murphy et al., 2001; Phillips and Penny, 2003; Prasad et al., 2008; Waddell and Shelley, 2003). We added Crocodilia (Node 17) and Squamata (three taxa; Node 15 and 16) to the lineage represented by chickens in the Russo et al. tree (Cao et al., 2000; Cotton and Page, 2002; Hedges and Poling, 1999). Sister to the Reptilia (Node 14) and Mammalia (Node 18) (i.e. Amniota (Node 13)) are Amphibia (Node 7), which were divided into Anura (three taxa; Node11) and Caudata (four taxa; Node 8). Within Caudata, two *Plethodon* salamanders (Node 10) are sister to *Eurycea* (Node 9) which collectively are sister to Ambystomatidae; within Anura, *Xenopus* (Node 11) is sister to the Bufonidae–Ranidae clade (Node 12) (Chippindale et al., 2004; Frost et al., 2006; Hugall et al., 2007; Min et al., 2005; Mueller et al., 2004). Collectively, Tetrapoda (Node 6) is sister to the Teleostei (Node 2), which are represented by five taxa; Cyprinidae, Salmonidae (two taxa; Node 5), and Tertraodontidae (two taxa; Node 4) (Cotton and Page, 2002; Mank et al., 2005; Miya et al., 2003). Chondrichthyes (*Mustelus manazo)* and Cephalochordata (*Brachiostoma japonicum*) were used as outgroups. Details on specific sequences can be found in Appendix A. While we acknowledge that the phylogeny is not known precisely, we think that the multiple lines of evidence cited above strongly support the phylogenetic relationships presented. A copy of the aligned, concatenated matrix is available through TreeBASE (http://purl.org/phylo/treebase/phylows/study/TB2:S10500).

2.2. Simulated datasets

All simulated datasets were created using Mesquite 2.6 (Maddison and Maddison, 2009). Parameter values for the simulated datasets were estimated from a total evidence analysis of the nat-

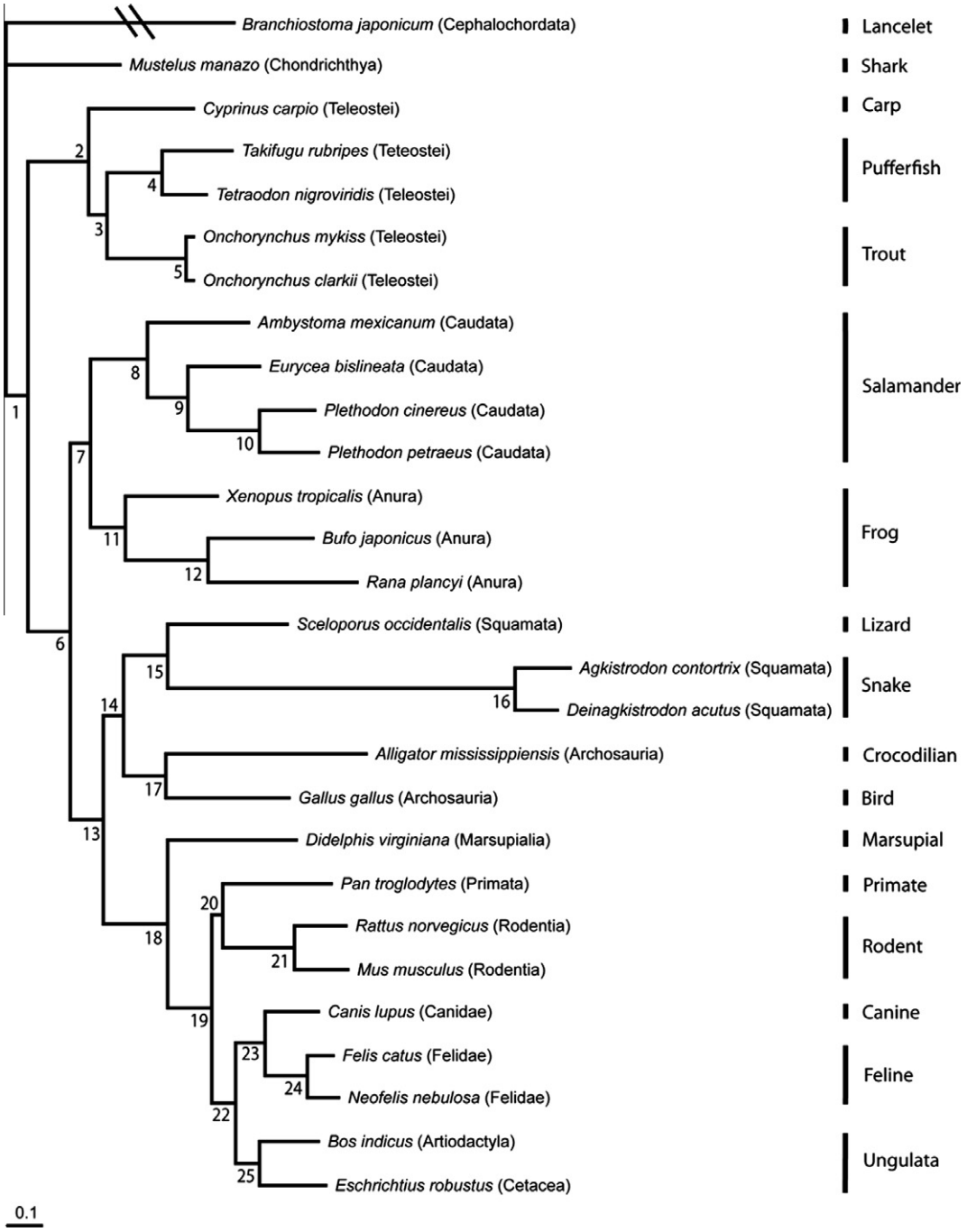

Fig. 1. The "known" phylogeny used for this study with branch lengths estimated from a Bayesian total evidence analysis. Lancelet and Shark are the outgroups. OTUs are labeled with both primary species and clade names corresponding to the text (see Appendix A and text for explanation). Numbers at nodes correspond to Table 3.

ural datasets. These include: nucleotide frequency (A = 0.3473, C = 0.2812, G = 0.1516, T = 0.2199), proportion of invariable sites (0.317), GTR rate matrix (A–C = 1.66, A–G = 3.51, A–T = 2.21, C–G = 0.67, C–T = 12.06), and gamma distribution of site rate variability (0.57). In addition, rate variability among codon positions was defined as the datasets average evolutionary rate multiplied by 0.49, 0.27, and 2.21 for first, second, and third positions, respectively. Datasets of varying evolutionary rate were evolved according to either the topology of the "known" tree or a specific variation. Variations include a topology with equal branch lengths, a "radiation" topology with terminal branch lengths ten times longer than all internal branch lengths that would represent a rapid radiation followed by stasis, and a topology in which a strict molecular clock was enforced (created using the randomly ultrametricized option in Mesquite 2.6).

To determine the effect of assuming an incorrectly parameterized model of evolution, 15 datasets of varying evolutionary rate were modeled to each of three evolutionary models that incorporate an increasing number of parameters. The first model incorporates only nucleotide frequencies and is equivalent to the Jukes Cantor (JC) model. The second model (GTR) includes nucleotide frequencies and a general time reversible rate matrix for nucleotide change. The third model (GTR + I + G) includes nucleotide frequencies, a general time reversible rate matrix for nucleotide change, the proportion of invariable sites and a gamma shaped distribution of site rate variability. To determine the effect of topology, 15

Table 1
Descriptions of how simulated datasets were modeled as well as results of statistical tests.

Simulation category name	Simulation model of evolution	Simulation topology	Number of significant K–S tests	Results of Mood's median test (df = 5)
JC	JC	Known	6	$\chi^2 = 119.66$, $P < 0.000$
GTR	GTR	Known	5	$\chi^2 = 63.46$, $P < 0.000$
GTR + I + G	GTR + I + G	Known	4	$\chi^2 = 36.39$, $P < 0.000$
Equal	GTR + I + G	Equal	4	$\chi^2 = 59.23$, $P < 0.000$
Radiation	GTR + I + G	Radiation	4	$\chi^2 = 28.80$, $P < 0.000$
Ultrametric	GTR + I + G	Ultrametric	2	$\chi^2 = 25.74$, $P < 0.000$

datasets of varying evolutionary rate were modeled upon the four topological variations described above using the GTR + I + G model. Overall, a total 90 simulated datasets were created and each one was analyzed separately. See Table 1 for a complete description of each simulated dataset.

2.3. Data collection

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

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

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

We ran a Bayesian phylogenetic analysis for each dataset (natural or simulated) using MrBayes 3.1.2 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) with the following parameters: nst = 6, rates = invgamma, ratepr = variable, statefreqpr = dirichlet (1, 1, 1, 1) and unlinked gamma shape parameters. For protein coding genes, each codon position was analyzed separately during analyses. We chose to use the GTR + I + G model for all genes because MrModeltest (Nylander, 2004) returned this model for 11 of 12 genes (using the Akaike Information Criterion) and model overparameterization should not negatively affect the analysis (Castoe et al., 2004; Lemmon and Moriarty, 2004). Each analysis included six chains, sampling every at 1000 generations, and was run for at least 7,500,000 generations (default parameters otherwise). Stationarity of the analysis was determined by examining the standard deviation of split frequencies (<0.01) and $-ln$ likelihood plots in AWTY (Nylander et al., 2008). Burnin calculations were conservative, between 2.5 and 5.0 million generations. To assess phylogenetic performance we used the posterior probabilities associated with each "correct" node (i.e. congruent with the "known" phylogeny) by examining the observed bipartitions in the 50 percent majority-rule consensus tree.

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

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

3. Results

3.1. Natural datasets

Pairwise divergences within genes ranged from 0.000 to 2.93 substitutions per site (based on the model of substitution, Table 2) and node standard deviations ranged from 0.0 to 0.152. The different correction measurement models yielded very similar patterns (the difference being scale of divergence axes), so only K2P corrected distances are presented in the following figures. Posterior probabilities for correct nodes ranged from 0.0% to 100% (Table 3).

The relationship between sequence divergence and posterior probability for all genes recovers an optimal range of divergence (Fig. 2). Specifically, sequences that were either too divergent or too similar recovered lower average posterior probabilities for correct nodes. Analyses using mitochondrial protein coding genes (there was little observable difference between the combined protein vs. divergence and the mitochondrial protein vs. divergence plots, so only one is reported) found an optimal K2P corrected sequence divergence of approximately 0.07 and recovered the correct topology with highly similar sequences. Analyses using nuclear protein and ribosomal genes showed an unexpected lack of any relationship that may be more an artifact of low sample size that the true pattern.

The standard deviation of node divergence was positively correlated with average K2P corrected pairwise divergence for ribo-

Table 2
Maximum pairwise divergence between taxa for different substitution models for the natural dataset.

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

Table 3
Recovered posterior probability of each node for each gene and the node's preceding branch length (Br len) in the natural dataset. Node numbers correspond to Fig. 1. Dashes (–) represent nodes that were not calculated due to incomplete taxon sampling.

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

some, protein, and combined data (Fig. 3), meaning that genes with high levels of divergence recovered higher levels of heterochrony. The posterior probability associated with a node was significantly related (Mood's median test; $\chi^2 = 16.47$, df = 3, $P < 0.001$) to the node's standard deviation for all datasets (Fig. 3), reaffirming the notion that high levels of heterochrony lead to reduced phylogenetic performance (Harrison and Larsson, 2008).

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

There was a significant positive relationship between branch length and posterior probability (Fig 4). Due to the long branch length of snakes we tested whether snakes were biasing the results, but found that, the significance of the relationship between branch and posterior probability also exists when snakes were excluded ($F = 6.05$, df = 24, $r^2 = 20.8\%$, $P = 0.022$ with snakes; $F = 11.97$, df = 23, $r^2 = 35.2\%$, $P = 0.002$ without snakes).

3.2. Simulated datasets

Simulated datasets recovered the same relationship between posterior probability and sequence divergence as the "known" phylogeny (Fig. 5), although the pattern is shifted in different directions. For example, analyses using the datasets produced by JC (A) and GTR (B) models of evolution recovered a broad range of divergences that had high corresponding posterior probabilities. Analyses of the GTR + I + G (C) and ultrametric tree (F) produced a pattern most similar to the one observed in the natural data. The equal branch length dataset analyses recovered the highest level of phylogenetic support across divergence levels, while analyses of the radiation dataset showed the lowest support levels. K–S tests and Mood's median tests (Table 1) found that the relationship is

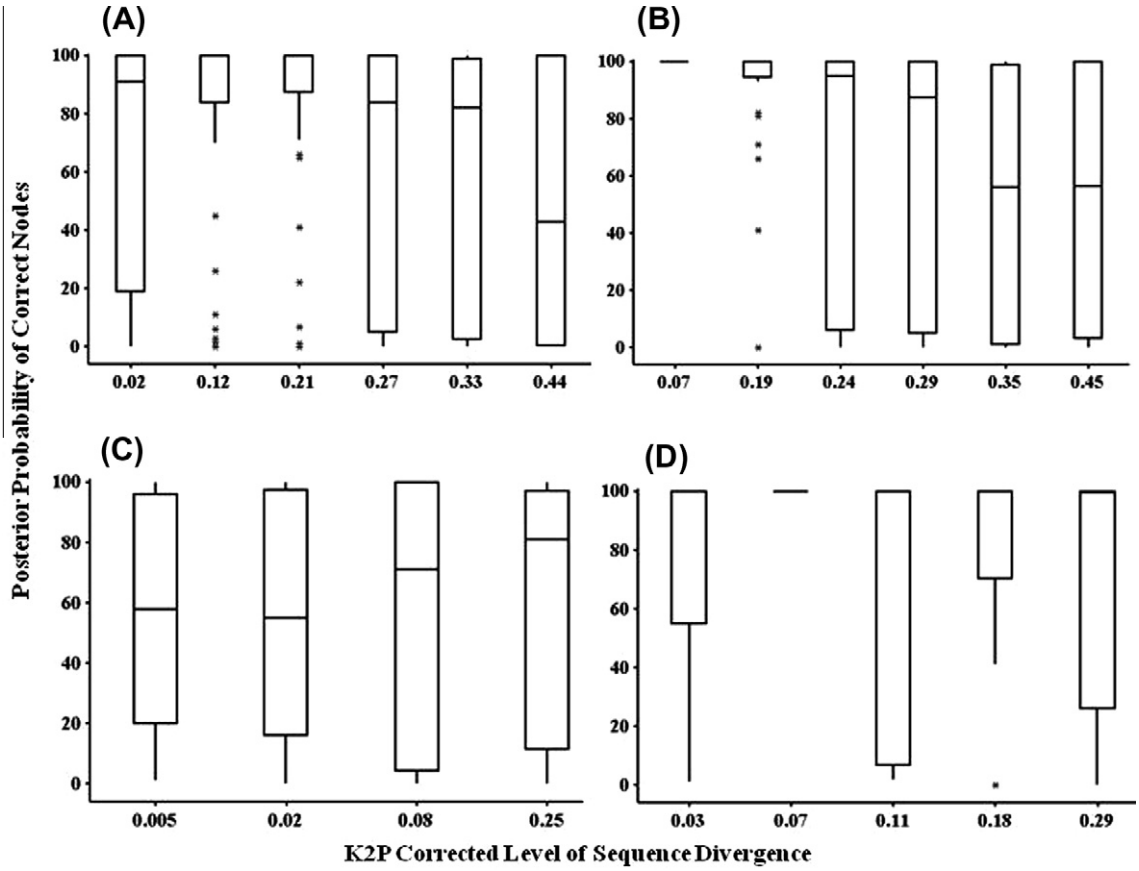

Fig. 2. Posterior probabilities of correct nodes at different levels of corrected sequence for the natural dataset. Data are presented using boxplots where the center is the median, box edges are the first and third quartiles, and stars are outliers. (A) All genes combined, (B) mitochondrial protein coding genes, (C) ribosomal genes, (D) nuclear protein coding genes.

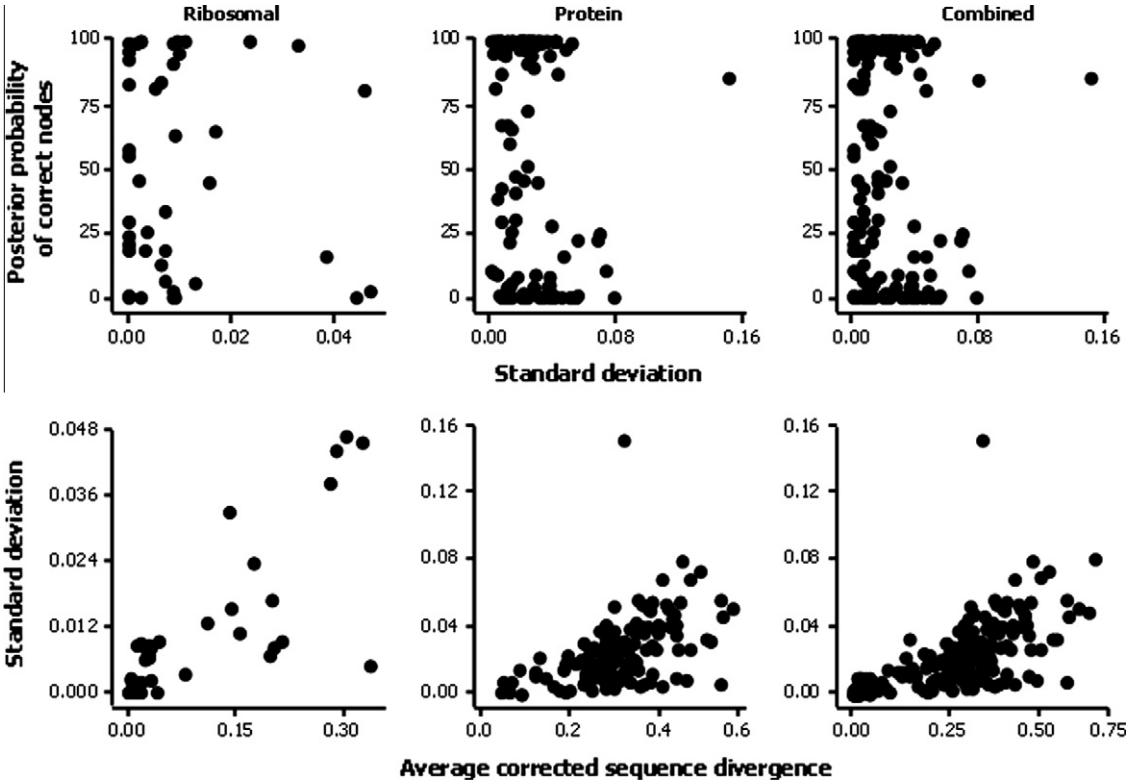

Fig. 3. Relationship between posterior probability, standard deviation of corrected sequence divergence, and mean divergence of each node for ribosomal genes, protein encoding genes, and all genes combined for the natural dataset.

Table 4
K–S pairwise comparison results (*P*-values) between divergence groups for both the mitochondrial protein dataset (top right, bolded) and complete dataset (bottom left). Notice that the K2P corrected sequence divergence for the groups is different for the two datasets. The Bonferroni corrected *P*-value is 0.003 and statistically significant comparisons are marked with an ∗.

	0.07	0.19	0.24	0.29	0.35	0.45
0.02	–	**0.832**	**0.001**∗	**0.002**∗	**0.000**∗	**0.001**∗
0.12	0.041	–	**0.006**	**0.006**	**0.000**∗	**0.000**∗
0.21	0.040	0.607	–	**0.961**	**0.640**	**0.640**
0.27	0.301	0.002∗	0.040	–	**0.640**	**0.999**
0.33	0.196	0.000∗	0.005∗	0.781	–	**0.832**
0.44	0.195	0.000∗	0.000∗	0.780	0.440	–

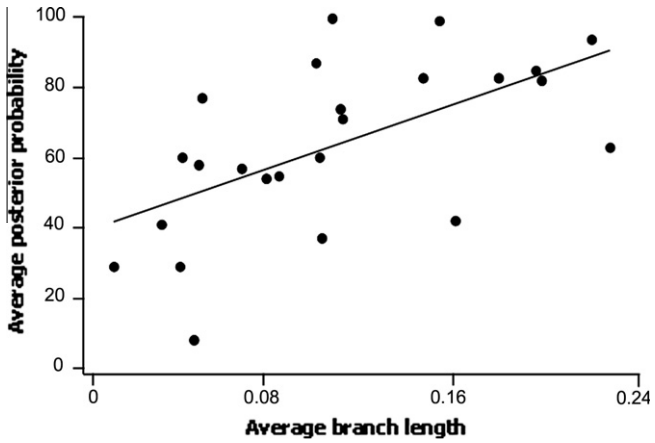

Fig. 4. Relationship between branch length and posterior probability for the natural dataset. Snake has been removed due to the exceptionally long branch length, but this does not affect the significance of the relationship.

significant for each simulation category. This suggests that an optimal level of divergence exists regardless of the assumed model of evolution or topology, although the optimal ranges of genetic divergence as well as median posterior probabilities are different across the six simulations.

4. Discussion

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

Combining mitochondrial, ribosomal and nuclear genes, we found an optimal divergence range of approximately 0.12–0.21 K2P corrected (0.09–0.18 uncorrected p, 0.14–0.26 T–N gamma corrected) substitutions per site for the natural dataset. Interestingly, posterior probabilities for correct nodes declined more precipitously with greater divergence from the optimal range than with less divergence. We also analyzed the data by gene category; ribosome, nuclear protein, and mitochondrial protein. Protein coding genes, especially mitochondrial ones, recover high support for correct relationships even when divergence levels are very low (0.05) and work best at K2P corrected divergences under 0.20 (0.19 for uncorrected p, 0.28 for T–N gamma corrected). This is in sharp contrast with ribosomal genes, which recover similar support values for correct nodes at all divergence levels tested (0.005–0.25 K2P). A dataset that spanned more evolutionary time and incorporated more taxa will be necessary to better understand the relationship between sequence divergence and nodal support for ribosomal genes. For nuclear protein genes, a strategy that focuses on organisms with complete, annotated genomes and well-resolved phylogenetic relationships will be necessary.

We used simulated datasets to determine the generality of the observed relationship between posterior probability and sequence divergence seen in the natural dataset. Specifically, we examined phylogenetic performance under differing degrees of model over-parameterization and variations in topology. Neither model over-parameterization nor topology was found to affect the overall pattern. When the JC, GTR, and GTR + I + G simulations are compared, the major difference is in the level of sequence divergence associated with optimal phylogenetic reconstruction between GTR + I + G and the other two models. This is mostly likely due to the incorporation of a specific number of invariable sites in the analysis, which causes some sites to evolve very quickly and become highly saturated at low levels of sequence divergence. This saturation results in an underestimated level of sequence divergence for the GTR + I + G datasets. For datasets where the model of evolution was held constant (GTR + I + G) and the topology was varied, the same relationship between sequence divergence and phylogenetic reconstruction ability was observed, although the pattern is shifted.

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

Divergence optima for phylogenetic reconstruction occur for a variety of reasons. Besides the reasons already discussed in the introduction, we found that as the standard deviation of the pairwise sequence divergences at a node increases (i.e. molecular clock violations or heterochrony), the average posterior probability decreases (Fig. 3). Previous researchers have documented that strong deviations from a molecular clock reduce the effectiveness of most phylogenetic reconstruction methods (e.g. Felsenstein, 1983, 2004; Rzhetsky and Sitnikova, 1996), so this may be another reason why analyses using highly divergent sequences recover low posterior probabilities for correct nodes.

Beyond the determination of divergence optima, we also observed two notable patterns involving the relationships among posterior probability, topology, and sequence divergence. First, in

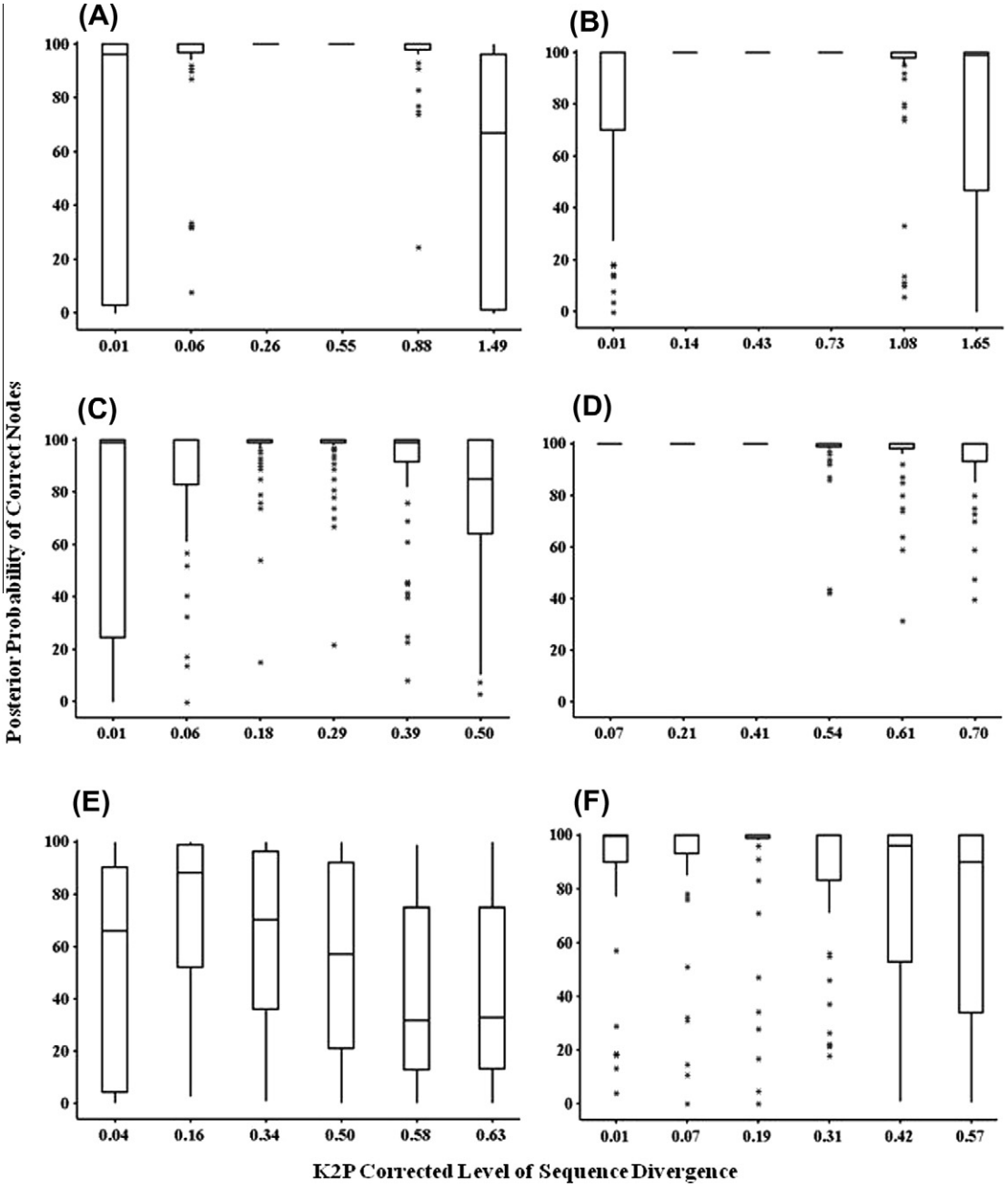

Fig. 5. Relationship between posterior probability of correct nodes and corrected sequence divergence for simulated datasets. Dataset categories are (A) JC, (B) GTR, (C) GTR + I + G, (D) Equal, (E) Radiation, and (F) Ultrametric.

the natural dataset, no single gene analyses recovered all of the correct relationships. This is not surprising given the length of evolutionary time (c. 500 million years) that our known phylogeny encompasses and the correspondingly large differences in average corrected divergence associated with each node. This is similar to other findings that single gene trees have a very low probability of fully recovering the true relationships (Cao et al., 1994; Rokas et al., 2003). The problem is further confounded when gene trees do not represent species trees, probably reflected in this study as the low recovered posterior probabilities for "correct" nodes when divergences are optimal (Fig. 2).

Second, we found that several nodes consistently exhibited low posterior probability support values across genes, while others consistently exhibited high support values across genes (Table 3). Interestingly, analyses using some mitochondrial genes recovered strong support for nodes ca. 300 million years old. The poorly supported nodes (<90% PP) were generally "deeper" ones, but not always (e.g. Archonta, Theria, Tertraodontidae). One likely cause is the branch length associated with each node (e.g. Rokas and Carroll, 2006; Wiens et al., 2008). Snakes were both included and excluded because of an unusually long branch length, most likely due to their unusually rapid mitochondrial evolution (Castoe et al., 2008; Jiang et al., 2007). We found a significant positive correlation between estimated branch length and posterior probability ($P = 0.022$ with snakes; $P = 0.002$ without snakes) in the natural dataset.

One limitation of this work is that the only phylogenetic reconstruction method we tested was Bayesian analysis, using MrBayes software. Other phylogenetic methods, such as parsimony and maximum likelihood, are commonly used and should be tested

for optimal sequence divergence. We speculate that maximum likelihood will yield results similar to those of this current study, while parsimony will probably have a lower optimal sequence divergence (because parsimony does not take into account complex models of molecular evolution). Unfortunately, these methods do not have nodal support values that are equivalent to posterior probabilities. Regardless of the equivalence (or lack thereof) between posterior probabilities and bootstrap proportions, several studies have demonstrated a correlation between the two values (Cummings et al., 2003; Erixon et al., 2003), at least under some conditions, so we predict that the overall results would be similar.

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

In conclusion, we have demonstrated an optimal divergence for sequences of approximately 0.12–0.21 K2P corrected pairwise distance yield the highest support for correct nodes. Divergences as low as 0.025 and as high as 0.30 also recovered high support for correct relationships, but divergences over 0.30 show a sharp decline in support for correct nodes. However, we cannot determine if different types of gene (protein or ribosomal) as well as where the gene is encoded (nuclear or mitochondrial) may be important factors to take into account. Our natural dataset allows us to draw conclusions for mitochondrial protein encoding genes, but we are reluctant to draw conclusions for nuclear protein or ribosomal genes. For mitochondrial protein encoding genes, it appears that lower sequence divergences are associated with higher support values, with a large drop in posterior probability at K2P corrected divergences over 0.20 (0.19 for uncorrected p, 0.28 for T–N gamma corrected). Simulated datasets exhibited the same relationship between sequence divergence and phylogenetic reconstruction ability regardless of topology or model of evolution adequacy. This information has the most utility for relationships that are difficult to resolve, but can also be used for project design to ensure that sequence data are as informative as possible. Future work on combining genes (i.e. supertrees) that evolve at different rates so that nodes are weighted towards more optimally divergent levels could also greatly help evolutionary biologists get the most correct information from their data while minimizing confounding information.

Acknowledgments

The authors would like to thank the many people who helped with the development of this manuscript; Jesse Meik, Brian Fontenot, Walter Schargel, Esther Betran, Charles Tracy, John Pace, Clement Gilbert, Jeff Streicher, Phi Sigma Biolunch participants, and the anonymous reviewers.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.ympev.2010.05.009.

References

Aguileta, G., Marthey, S., Chiapello, H., Lebrun, M.H., Rodolphe, F., Fournier, E., Gendrault-Jacquemard, A., Giraud, T., 2008. Assessing the performance of single-copy genes for recovering robust phylogenies. Syst. Biol. 57, 613–627.
Avise, J., 2000. Phylogeography. Harvard University Press, Cambridge, Massachusetts.
Avise, J.C., Saunders, N.C., 1984. Hybridization and introgression among species of sunfish (*Lepomis*): analysis by mitochondrial DNA and allozyme analysis. Genetics 108, 237–255.
Berendzen, P., Simons, A., Wood, R., 2003. Phylogeography of the northern hogsucker, *Hypentelium nigricans* (Teleostei: Cypriniformes): genetic evidence for the evidence of the ancient Teays River. J. Biogeogr. 30, 1139–1152.
Blouin, M.S., Yowell, C.A., Courtney, C.H., Dame, J.B., 1998. Substitution bias, rapid saturation, and the use of mtDNA for nematode systematics. Mol. Biol. Evol. 15, 1719–1727.
Blouin, C., Butt, D., Roger, A., 2004. Impact of taxon sampling on the estimation of rates of evolution at sites. Mol. Biol. Evol. 22, 784–791.
Bruce, R., 2005. Did *Desmognathus* salamanders reinvent the larval stage? Herpetol. Rev. 36, 107–112.
Bull, J.J., Cunningham, C.W., Molineux, I.J., Badgett, M.R., Hillis, D.M., 1993. Experimental molecular evolution of bacteriophage T7. Evolution 47, 993–1007.
Cao, Y., Adachi, J., Janke, A., Pääbo, S., Hasegawa, M., 1994. Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: instability of a tree based on a single gene. J. Mol. Evol. 39, 519–527.
Cao, Y., Sorenson, M.D., Kumazawa, Y., Mindell, D.P., Hasegawa, M., 2000. Phylogenetic position of turtles among amniotes: evidence from mitochondrial and nuclear genes. Gene 259, 139–148.
Castoe, T.A., Doan, T.M., Parkinson, C.L., 2004. Data partitions and complex models in Bayesian analysis: the phylogeny of gymnophthalmid lizards. Syst. Biol. 53, 448–469.
Castoe, T.A., Jiang, Z.J., Gu, W., Wang, Z.O., Pollock, D.D., 2008. Adaptive evolution and functional redesign of core metabolic proteins in snakes. PLoS One 3, e2201.
Chippindale, P., Bonett, R., Baldwin, A., Wiens, J., 2004. Phylogenetic evidence for a major reversal of life-history evolution in plethodontid salamanders. Evolution 58, 2809–2822.
Collins, T.M., Fedrigo, O., Naylor, G.J.P., 2005. Choosing the best genes for the job: the case for stationary genes in genome-scale phylogenetics. Syst. Biol. 54, 493–500.
Cotton, J.A., Page, R.D.M., 2002. Going nuclear: gene family evolution and vertebrate phylogeny reconciled. Proc. R. Soc. Lond., Ser. B: Biol. Sci. 269, 1555–1561.
Creer, S., Thorpe, R.S., Malhotra, A., Chou, W.H., Stenson, A.G., 2004. The utility of AFLPs for supporting mitochondrial DNA phylogeographical analyses in the Taiwanese bamboo viper, *Trimeresurus stejnegeri*. J. Evol. Biol. 17, 100–107.
Cummings, M.P., Handley, S.A., Myers, D.S., Reed, D.L., Rokas, A., Winka, K., 2003. Comparing bootstrap and posterior probability values in the four-taxon case. Syst. Biol. 52, 477–487.
Dawson, M., 2001. Phylogeography in coastal marine animals: a solution from California. J. Biogeogr. 28, 723–736.
Douady, C.J., Douzery, E.J.P., 2003. Molecular estimation of eulipotyphlan divergence times and the evolution of "Insectivora". Mol. Phylogenet. Evol. 28, 285–296.
Downie, D., 2004. Phylogeography in a galling insect, grape phylloxera, *Daktulosphaira vitifoliae* (Phylloxeridae) in the fragmented habitat of the Southwest USA. J. Biogeogr. 31, 1759–1768.
Erixon, P., Svennblad, B., Britton, T., Oxelman, B., 2003. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. Syst. Biol. 52, 665–673.
Felsenstein, J., 1983. Parsimony in systematics: biological and statistical issues. Ann. Rev. Ecol. Syst. 14, 313–333.
Felsenstein, J., 2004. Inferring Phylogenies. Sinauer Associates, Sunderland, Massachusetts.
Frost, D., Grant, T., Faivovich, J., Bain, R., Haas, A., Haddad, C., de Sa, R., Channing, A., Wilkinson, M., Donnellan, S., Raxworthy, C., Campbell, J., Blotto, B., Moler, P., Drewes, R., Nussbaum, R., Lynch, J., Green, D., Wheeler, W., 2006. The amphibian tree of life. Bull. Am. Museum Nat. Hist. 297, 1–370.
Graur, D., Li, W.-H., 2000. Fundamentals of Molecular Evolution. Sinauer Associates, Sunderland, Massachusetts.
Graybeal, A., 1993. The phylogenetic utility of cytochrome *b*: lessons from bufonid frogs. Mol. Phylogenet. Evol. 2, 256–269.
Hare, M.P., 2001. Prospects for nuclear gene phylogeography. Trends Ecol. Evol. 16, 700–706.
Harrison, L.B., Larsson, H.C.E., 2008. Estimating evolution of temporal sequence changes: a practical approach to inferring ancestral developmental sequences and sequence heterochrony. Syst. Biol. 57, 378–387.
Heath, T.A., Hedtke, S.M., Hillis, D.M., 2008. Taxon sampling and the accuracy of phylogenetic analysis. J. Syst. Evol. 46, 239–257.

Hedges, S.B., Poling, L.L., 1999. A molecular phylogeny of reptiles. Science 283, 998–1001.
Hillis, D.M., 1999. SINEs of the perfect character. Proc. Natl. Acad. Sci. USA 96, 9979–9981.
Hillis, D.M., Huelsenbeck, J.P., 1994. To tree the truth: biological and numerical simulations of phylogeny. In: Fambrough, D.M. (Ed.), Molecular Evolution of Physiological Processes. Rockefeller University Press, pp. 55–67.
Hillis, D.M., Moritz, C., Mable, B.K. (Eds.), 1996. Molecular Systematics. Sinauer Associates, Sunderland, Massachusetts.
Hudelot, C., Gowri-Shankar, V., Jow, H., Rattray, M., Higgs, P.G., 2003. RNA-based phylogenetic methods: application to mammalian mitochondrial RNA sequences. Mol. Phylogenet. Evol. 28, 241–252.
Huelsenbeck, J., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogeny. Bioinformatics 17, 754–755.
Hugall, A.F., Foster, R., Lee, M.S.Y., 2007. Calibration choice, rate smoothing, and the pattern of tetrapod diversification according to the long nuclear gene RAG-1. Syst. Biol. 56, 543–563.
Jermiin, L.S., Poladian, L., Charleston, M.A., 2005. Is the "Big Bang" in animal evolution real? Science 310, 1910–1911.
Jiang, Z., Castoe, T., Austin, C., Burbrink, F., Herron, M., McGuire, J., Parkinson, C., Pollock, D., 2007. Comparative mitochondrial genomics of snakes: extraordinary substitution rate dynamics and functionality of the duplicate control region. BMC Evol. Biol. 7, 123.
Ketmaier, V., Giusti, F., Caccone, A., 2006. Molecular phylogeny and historical biogeography of the land snail genus *Solatopupa* (Pulmonata) in the peri-Tyrrhenian area. Mol. Phylogenet. Evol. 39, 439–451.
Koopman, W.I.M., 2005. Phylogenetic signal in AFLP data sets. Syst. Biol. 54, 197–217.
Kumazawa, Y., Azuma, Y., Nishida, M., 2004. Tempo of mitochondrial gene evolution: Can mitochondrial DNA be used to date old divergences? Endocytobiosis Cell Res. 15, 136–142.
Lemmon, A., Moriarty, E., 2004. The importance of proper model assumption in Bayesian phylogenetics. Syst. Biol. 53, 265–277.
Lin, Y.-H., McLenachan, P.A., Gore, A.R., Phillips, M.J., Ota, R., Hendy, M.D., Penny, D., 2002. Four new mitochondrial genomes and the increased stability of evolutionary trees of mammals from improved taxon sampling. Mol. Biol. Evol. 19, 2060–2070.
Linder, P., Hardy, C., Rutschmann, F., 2005. Taxon sampling effects in molecular clock dating: an example from the African Restionaceae. Mol. Phylogenet. Evol. 35, 569–582.
Liu, F.-G.R., Miyamoto, M.M., Freire, N.P., Ong, P.Q., Tennant, M.R., Young, T.S., Gugel, K.F., 2001. Molecular and morphological supertrees for eutherian (placental) mammals. Science 291, 1786–1789.
Lopez, P., Forterre, P., Philippe, H., 1999. The root of the tree of life in the light of the covarion model. J. Mol. Evol. 49, 496–508.
Maddison, W.P., 1997. Gene trees in species trees. Syst. Biol. 46, 523–536.
Maddison, W.P., Knowles, L.L., 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55, 21–30.
Maddison, W.P., Maddison, D.R., 2009. Mesquite: a modular system for evolutionary analyses. Version 2.6. Available from: <http://mesquiteproject.org>.
Mank, J.E., Promislow, D.E.L., Avise, J.C., 2005. Phylogenetic perspectives in the evolution of parental care in ray-finned fishes. Evolution 59, 1570–1578.
Meyer, A., 1994. Shortcomings of the cytochrome *b* gene as a molecular marker. Trends Ecol. Evol. 9, 278–280.
Min, M., Yang, S., Bonett, R., Vieites, D., Brandon, R., Wake, D., 2005. Discovery of the first Asian plethodontid salamander. Nature 435, 87–90.
Miya, M., Takeshima, H., Endo, H., Ishiguro, N.B., Inoue, J.G., Mukai, T., Satoh, T.P., Yamaguchi, M., Kawaguchi, A., Mabuchi, K., Shirai, S.M., Nishida, M., 2003. Major patterns of higher teleostean phylogenies: a new perspective based on 100 complete mitochondrial DNA sequences. Mol. Phylogenet. Evol. 26, 121–138.
Morgan, J.A.T., Blair, D., 1998. Relative merits of nuclear ribosomal internal transcribed spacers and mitochondrial CO1 and ND1 genes for distinguishing among *Echinostoma* species (Trematoda). Parasitology 116, 289–297.
Mueller, R.L., 2006. Evolutionary rates, divergence dates, and the performance of mitochondrial genes in Bayesian phylogenetic analysis. Syst. Biol. 55, 289–300.
Mueller, R., Macey, J., Jaekel, M., Wake, D., Boore, J., 2004. Morphological homoplasy, life history evolution, and historical biogeography of plethodontid salamanders inferred from complete mitochondrial genomes. Proc. Natl. Acad. Sci. USA 101, 13820–13825.
Murata, S., Takasaki, N., Saitoh, M., Okada, N., 1993. Determination of the phylogenetic relationships among Pacific salmonids by using short interspersed elements (SINEs) as temporal landmarks of evolution. Proc. Natl. Acad. Sci. USA 90, 6995–6999.
Murphy, W.J., Eizirik, E., O'Brien, S.J., Madsen, O., Scally, M., Douady, C.J., Teeling, E., Ryder, O.A., Stanhope, M.J., de Jong, W.W., Springer, M.S., 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294, 2348–2351.
Nylander, J., 2004. MrModeltest (Program distributed by the author). Evolutionary Biology Centre, Uppsala University.
Nylander, J.A., Ronquist, F., Huelsenbeck, J.P., Nieves-Aldrey, J.L., 2004. Bayesian phylogenetic analysis of combined data. Syst. Biol. 53, 47–67.
Nylander, J.A., Wilgenbusch, J.C., Warren, D.L., Swofford, D.L., 2008. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics 24, 581–583.
Palumbi, S.R., Cipriano, F., Hare, M.P., 2001. Predicting nuclear gene coalescence from mitochondrial data: the three-times rule. Evolution 55, 5.
Phillips, M.J., Penny, D., 2003. The root of the mammalian tree inferred from whole mitochondrial genomes. Mol. Phylogenet. Evol. 28, 171–185.
Pollock, D.D., Zwickl, D.J., McGuire, J.A., Hillis, D.M., 2002. Increased taxon sampling is advantageous for phylogenetic inference. Syst. Biol. 51, 664–671.
Prasad, A.B., Allard, M.W., Program, N.C.S., Green, E.D., 2008. Confirming the phylogeny of mammals by use of large comparative sequence datasets. Mol. Biol. Evol. 25, 1795–1808.
Pratt, R.C., Gibb, G.C., Morgan-Richards, M., Phillips, M.J., Hendy, M.D., Penny, D., 2009. Toward resolving deep neoaves phylogeny: data, signal enhancement, and priors. Mol. Biol. Evol. 26, 313–326.
Rannala, B., Huelsenbeck, J., Yang, Z., Nielsen, R., 1998. Taxon sampling and the accuracy of large phylogenies. Syst. Biol. 47, 702–710.
Ranwez, V., Delsuc, F., Ranwez, S., Belkhir, K., Tilak, M.-K., Douzery, E., 2007. OrthoMaM: a database of orthologous genomic markers for placental mammal phylogenetics. BMC Evol. Biol. 7, 241.
Richard, M., Thorpe, R.S., 2001. Can microsatellites be used to infer phylogenies? Evidence from population affinities of the Western Canary Island Lizard (*Gallotia galloti*). Mol. Phylogenet. Evol. 20, 351–360.
Ripplinger, J., Sullivan, J., 2008. Does choice in model selection affect maximum likelihood analysis? Syst. Biol. 57, 76–85.
Rodiguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Lang, B.F., Philippe, H., 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst. Biol. 56, 389–399.
Rokas, A., Carroll, S., 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Mol. Biol. Evol. 22, 1337–1344.
Rokas, A., Carroll, S.B., 2006. Bushes in the tree of life. PLoS Biol. 4, e352.
Rokas, A., Williams, B.L., King, N., Carroll, S.B., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425, 798–804.
Ronquist, F., Huelsenbeck, J., 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572–1574.
Ros, V.I.D., Breeuwer, J.A.J., 2007. Spider mite (Acari: Tetranychidae) mitochondrial COI phylogeny reviewed: host plant relationships, phylogeography, reproductive parasites and barcoding. Exp. Appl. Acarol. 42, 239–262.
Rosenberg, M.S., Kumar, S., 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. Proc. Natl. Acad. Sci. USA 98, 10751–10756.
Russo, C.A., Takezaki, N., Nei, M., 1996. Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. Mol. Biol. Evol. 13, 525–536.
Rzhetsky, A., Sitnikova, T., 1996. When is it safe to use an oversimplified substitution model in tree-making? Mol. Biol. Evol. 13, 1255–1265.
Seo, T.-K., Kishino, H., 2008. Synonymous substitutions substantially improve evolutionary inference from highly diverged proteins. Syst. Biol. 57, 367–377.
Sullivan, J., Lavoue, S., Arnegard, M., Hopkins, C., 2004. AFLPs resolve phylogeny and reveal mitochondrial introgression within a species flock of African Electric fish (Mormyroidea: Teleostei). Evolution 58, 825–841.
Takahashi, K., Terai, Y., Nishida, M., Okada, N., 2001. Phylogenetic relationships and ancient incomplete lineage sorting among cichlid fishes in Lake Tanganyika as revealed by analysis of the insertion of retroposons. Mol. Biol. Evol. 18, 2057–2066.
Tamura, K., Dudley, J., Nei, M., Kumar, S., 2007. MEGA4: molecular evolutionary genetics analysis (MEGA) software version 4.0. Mol. Biol. Evol. 24, 1596–1599.
Townsend, J.P., 2007. Profiling phylogenetic informativeness. Syst. Biol. 56, 222–231.
Vekemans, X., Beauwens, T., Lemaire, M., Roldan-Ruiz, I., 2002. Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size. Mol. Ecol. 11, 139–151.
Vos, P., Hogers, R., Bleeker, M., Reijans, M., Lee, T.v.d., Hornes, M., Frijters, A., Pot, J., Peleman, J., Kuiper, M., 1995. AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. 23, 4407–4414.
Waddell, P.J., Shelley, S., 2003. Evaluating placental inter-ordinal phylogenies with novel sequences including RAG1, [gamma]-fibrinogen, ND6, and mt-tRNA, plus MCMC-driven nucleotide, amino acid, and codon models. Mol. Phylogenet. Evol. 28, 197–224.
Weisrock, D., Harmon, L., Larson, A., 2005. Resolving deep phylogenetic relationships in salamanders: analyses of mitochondrial and nuclear genomic data. Syst. Biol. 54, 758–777.
Wiens, J.J., Bonett, R.M., Chippindale, P.T., 2005. Ontogeny discombobulates phylogeny: paedomorphosis and higher-level salamander relationships. Syst. Biol. 54, 91–110.
Wiens, J.J., Kuczynski, C.A., Smith, S.A., Mulcahy, D.G., Sites, J.W., Townsend, T.M., Reeder, T.W., 2008. Branch lengths, support, and congruence: testing the phylogenomic approach with 20 nuclear loci in snakes. Syst. Biol. 57, 420–431.
Xia, X., Xie, Z., Salemi, M., Chen, L., Wang, Y., 2003. An index of substitution saturation and its application. Mol. Phylogenet. Evol. 26, 1–7.
Yang, Z., 1998. On the best evolutionary rate for phylogenetic analysis. Syst. Biol. 47, 125–133.
Zwickl, D.J., Hillis, D.M., 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51, 588–598.