Syst. Biol. 47(1):61–76, 1998

Amphioxus Mitochondrial DNA, Chordate Phylogeny, and the Limits of Inference Based on Comparisons of Sequences

GAVIN J. P. NAYLOR[1] AND WESLEY M. BROWN[2]

[1]*Department of Zoology and Genetics, Iowa State University, Ames, Iowa 50011, USA; E-mail: gnaylor@iastate.edu*

[2]*Department of Biology, University of Michigan, Ann Arbor, Michigan 48109-1048, USA*

Abstract.—Analyses of both the nucleotide and amino acid sequences derived from all 13 mitochondrial protein-encoding genes (12,234 bp) of 19 metazoan species, including that of the lancelet *Branchiostoma floridae* ("amphioxus"), fail to yield the widely accepted phylogeny for chordates and, within chordates, for vertebrates. Given the breadth and the compelling nature of the data supporting that phylogeny, relationships supported by the mitochondrial sequence comparisons are almost certainly incorrect, despite their being supported by equally weighted parsimony, distance, and maximum-likelihood analyses. The incorrect groupings probably result in part from convergent base-compositional similarities among some of the taxa, similarities that are strong enough to overwhelm the historical signal. Comparisons among very distantly related taxa are likely to be particularly susceptible to such artifacts, because the historical signal is already greatly attenuated. Empirical results underscore the need for approaches to phylogenetic inference that go beyond simple site-by-site comparison of aligned sequences. This study and others indicate that, once a sequence sample of reasonable size has been obtained, accurate phylogenetic estimation may be better served by incorporating knowledge of molecular structures and processes into inference models and by seeking additional higher order characters embedded in those sequences, than by gathering ever larger sequence samples from the same organisms in the hope that the historical signal will eventually prevail. [Amphioxus; chordate phylogeny; homoplasy; mtDNA; molecular systematics; phylogenetic inference.]

The practice of inferring evolutionary trees from DNA sequences has flourished in recent years, its credibility bolstered by the observation that phylogenies of well-studied groups are usually supported by sequence data. When the sequence of a particular gene or other well-defined DNA segment yields an inference congruent with an accepted relationship for a particular group, there is a tendency to regard that segment as reliable for phylogenetic inference and to use it to determine phylogenies for taxa whose relationships are unknown (Graybeal, 1994; Cho et al., 1995). However, from the beginning of such studies it was recognized that any DNA segment can only be useful over a limited divergence range; outside that range the historical signal would be either too undeveloped or too attenuated to be reliable. Furthermore, with an increase in the number of such studies it also became apparent that the useful range varied among different taxa. Thus, there are instances in which sequence data provide accurate assessments for some relationships, and erroneous ones for others (Felsenstein, 1978; Hillis, 1991; Kim, 1996; Philippe et al., 1994). The latter occur whenever the embedded historical signal is overturned by a stronger, homoplasious signal among the DNA sequences.

Various methods are used for phylogenetic reconstruction, each implying a different model of evolutionary change and emphasizing different aspects of the observed character-state covariation among taxa. It is common practice to regard a phylogeny that is supported by several different methods as correct, and especially so when statistical methods for evaluating the strength of support [e.g., bootstrapping (Felsenstein, 1985) and decay indices (Bremer, 1988; Donoghue et al., 1992)] are compelling. This stems from a tacit assumption that an incorrect phylogeny, even if it is the best fit to the available data, will not receive significant statistical support when the result itself is evaluated. That assumption is incorrect. Statistical evaluations merely assess the strength of the signal used to order the data hierarchically (Swofford et al., 1996). Thus, if there is a hierarchical signal in the data that arises from a nonhistorical source and if that signal is sufficiently strong, it can overwhelm not only a weaker historical signal, but also a statistical evaluation of the result.

The ''total evidence'' approach to phylogenetic reconstruction (Eernisse and Kluge, 1993) is currently among the most widely applied. It ''uses character congruence to find the best fitting hypothesis for an unpartitioned set of synapomorphies, which is ideally all of the relevant available data'' (Eernisse and Kluge, 1993). In its purest implementation the approach weights all characters equally in order to dispense with any need to identify different classes of information. Proponents maintain that partitioning evidence into classes is artificial, ''because there is little reason to believe such categories are mind-independent categories with discoverable boundaries'' (Eernisse and Kluge, 1993). This somewhat narrow perspective has gained a following because it obviates a need to specify explicit (and often poorly known) processes about the way in which traits evolve. Advocates of the method espouse a view that homoplasy (character-state covariation among taxa due to influences other than shared history) will be randomly distributed with respect to taxa, and that hierarchically structured historical signal (character-state covariation among taxa due to shared history) will overshadow the homoplasy if enough data are collected (see Farris, 1983). In keeping with this view, it is assumed that any incorrect inferences will be due to stochastic error associated with an insufficiently large sample size of characters, and that they will disappear as more data are collected.

The sample size of sites required to ensure that the historical signal overturns the homoplasy depends to a large extent on the strength of the historical signal in a data set and on the grain size of the homoplasy. If the homoplasy is dispersed in a fine-grained fashion—that is, is distributed in small ''packets,'' each of which suggests a different nonhistorical association—it will likely appear randomly distributed at relatively small sample sizes. If homoplasy is dispersed in a coarse-grained fashion, so that several sites suggest the same nonhistorical grouping, then larger sample sizes (i.e., more sequence) will be required before patterns of homoplasy appear randomly distributed. In essence, the sample size of sites at which the randomness of homoplasy becomes apparent is dictated by the grain size of the homoplasy. Thus, even if homoplasy were randomly distributed among taxa at the level of the entire genome (an assumption that has not been empirically tested), it would appear to be highly nonrandomly distributed within a particular sample of sites if its grain size were coarse and the sample of sites insufficiently large. Note, in the current context, that the term ''grain size'' has no spatial connotation. When we refer to homoplasy as ''coarse-grained,'' we mean only that several sites within a fragment imply the same nonhistorical grouping; the misleading sites that collectively constitute a ''packet'' need not be spatially contiguous along the sequence.

The premise that homoplasy is randomly distributed or unstructured within data sets underlies the phylogenetically meaningful interpretations of bootstrapping, decay indices, and successive weighting (Farris, 1969). Groupings assessed as unreliable (i.e., those with little character support) are assumed to be due to chance, while those assessed as reliable are assumed to be so due to shared history. Unfortunately, if homoplasy is nonrandomly distributed or if it shows ''systematic error'' (Swofford et al., 1996), then analyses will not only yield erroneous phylogenetic inferences, but many of the tests designed to evaluate the reliability of their constituent nodes will lead to falsely confident assessments.

Given the appeal of the ''total evidence,'' equally weighted parsimony approach, it would be useful to evaluate the extent to which its required assumption for random distribution of homoplasy is actually met by molecular data sets. When phylogeny is

TABLE 1. Species used and their Genbank accession numbers.

Species name	Common name	GenBank accesion number
Mus musculus	Mouse	J01420
Rattus norvegicus	Rat	X14848
Bos taurus	Cow	J01394
Balaenopterus physalus	Fin-back whale	X61145
Balaenopterus musculus	Blue whale	X72204
Didelphis virginiana	Opposum	Z29573
Gallus gallus	Chicken	X52392
Xenopus laevis	Frog	M10217
		X01600
		X01601
		X02890
Cyprinus carpio	Carp	X61010
Oncorhynchus mykiss	Trout	L29771
Petromyzon marinus	Lamprey	U11880
Branchiostoma floridae	Lancelet	AF035164–AF035176
Paracentrotus lividus	Sea urchin 1	J04815
Strongylocentrotus purpuratus	Sea urchin 2	X12631
Drosophila yakuba	Fruit fly	X03240
Cepaea nemoralis	Snail	U23045
Anopheles gambiae	Mosquito	L20934
Ascaris suum	Nematode 1	X54253
Caenorhabditis elegans	Nematode 2	X54252

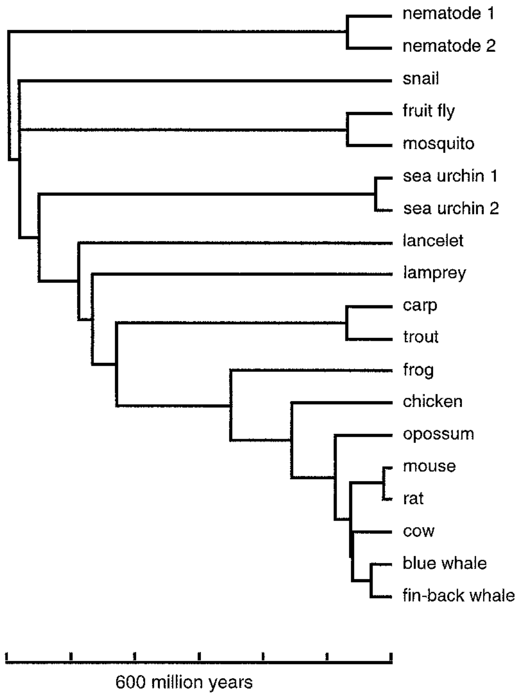


FIGURE 1. The expected pattern of phylogenetic relationships for 19 taxa. Branch lengths depicted are estimates from the fossil record (Benton, 1993). They reflect the earliest fossil occurrence assignable to the stem lineage of the extant form. In the case of the frog, the earliest fossil occurrence for the anuran stem group was used rather than the first fossil assignable to the amphibian grade.

known, nonrandom distribution of homoplasy can be inferred when the data set strongly supports an incorrect tree. The strength of departure from randomness can be assessed by evaluating the level of bootstrap support, or the decay index for the incorrect groups, or by subjecting the data to a Templeton (1983) test.

Although no phylogeny is known with certainty, a number are very well supported, perhaps the best known being that for echinoderms plus chordates (see Maisey, 1986, 1988; Gauthier et al., 1988, and references therein). Complete mitochondrial genomes have been sequenced for representatives of several vertebrate classes, two echinoderm classes, and a number of outgroups. We have recently sequenced a mitochondrial DNA (mtDNA) from the lancelet *Branchiostoma floridae* (''amphioxus''), a species of Cephalochordata, the immediate sister taxon to the Craniata. Thus, complete mtDNA sequences are now available from representatives of most key lineages in vertebrate evolution. Phylogenetic inferences derived from comparisons of these sequences can be contrasted with the accepted phylogeny, providing an opportunity to examine the distribution patterns of homoplasy in mtDNA sequences of this group.

MATERIALS AND METHODS

We assembled complete mitochondrial sequences for 19 taxa (Table 1) whose phylogenetic relationships are noncontroversial (Fig. 1). The protein encoding regions were aligned at the amino acid level using Clustal W (Thompson et al., 1994) and were checked for higher order structural concordance using the codon-coloring feature of Aligner (Eernisse, 1995). The resulting data set, consisting of 19 aligned 12,234-bp sequences, was subjected to a series of phylogenetic analyses using

TABLE 2. Inference errors resulting from bootstrap analysis of each gene individually and all genes combined.

Inferred grouping	ATP6	ATP8	CO1	CO2	CO3	CYTB	ND1	ND2	ND3	ND4	ND4L	ND5	ND6	All
All nucleotide substitutions—AGCT														
(frog, fish)		X												
(chicken, fish)	X		X								X		X	X
(frog, chicken, fish)	X		X		X							X	X	X
(frog (fish, amniotes)						X								
(opossum (frog, fish, amniotes))												X		
(lancelet (echinoderms, vertebrates))	X			X	X		X	X	X	X	X	X	X	X
(lancelet (flies, echinoderms, vertebrates))			X			X								
Transversions—RY														
(frog, fish)		X												
(chicken, frog)												X		
(frog, chicken, fish)			X									X		X
(lancelet (echinoderms, vertebrates))				X										
(lancelet (flies, echinoderms, vertebrates))			X					X						
Amino acids														
(frog, fish)	X	X								X		X	X	X
(chicken, fish)											X			
(chicken, frog, fish)										X			X	
(chicken, frog, fish, lamprey)	X													
(rodents (whales, cow, opossum))									X					
(lamprey (lancelet (echinoderms, vertebrates)))							X							
(lancelet (echinoderms, vertebrates))						X				X		X		X
(lancelet (flies, echinoderms, vertebrates))								X						

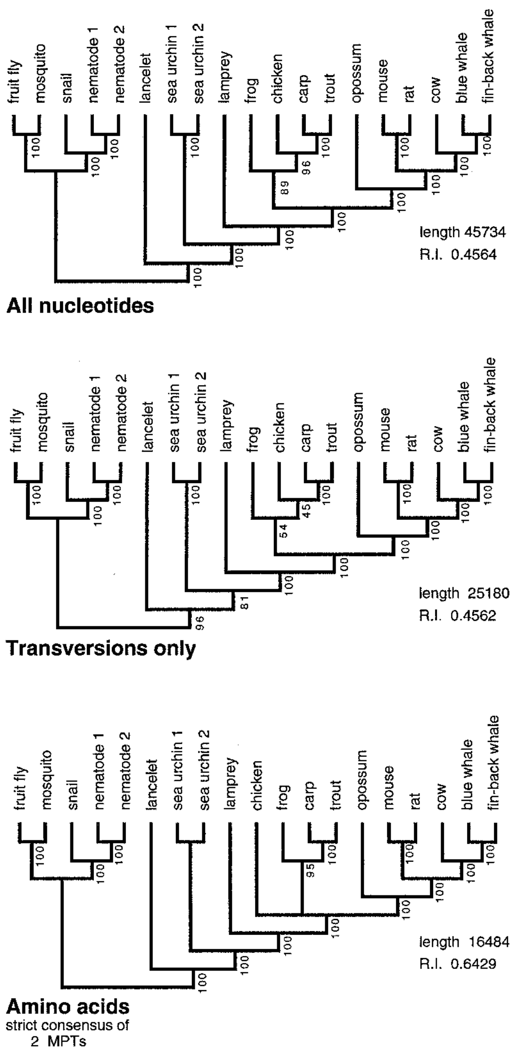
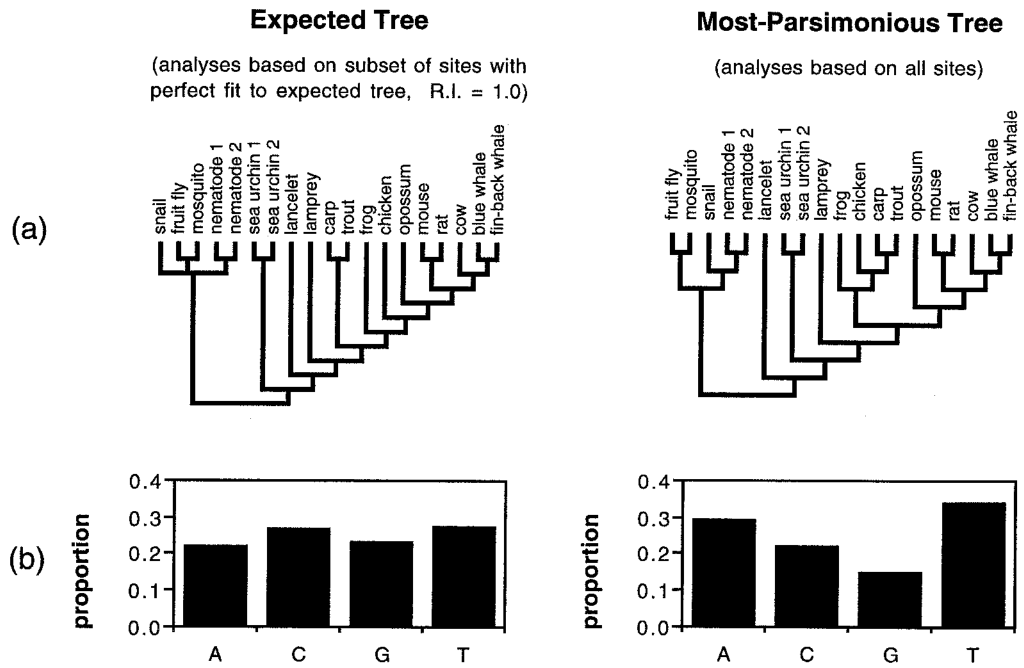


Figure 2. MPTs resulting from equally weighted analysis of the combined data set, for nucleotides (top), transversions only (center), and amino acids (bottom). Bootstrap support percentages are shown at each node. Tree length and RI are shown to the right of each topology. Note that two MPTs result for the amino acid analysis. The topology depicted is the strict consensus of the two MPTs.

PAUP*4.0 version 53 (written by David Swofford). Snail, fruit fly, mosquito, and two nematode species were used as a collective outgroup for all analyses.

We examined trees derived from equally weighted parsimony analysis for each of the 13 protein-encoding genes, both individually and in combination. Analyses were conducted at three levels: using all nucleotides; using transversions only; and using amino acid sequences. The degree of support for each node was evaluated using the bootstrap method of Felsenstein (1985). The nucleotide sequence data were also subjected to distance analyses using Jukes–Cantor (JC) (1969), Kimura two-parameter (K2P) (1980), Hasegawa–Kishino–Yano (HKY) (1985), and general time-reversible (GTR) (Lanave et al., 1984; Tavaré, 1986; Rodríguez et al., 1990) distances. Each of the four distances were used in conjunction with four different models of among-site rate variation (ASRV) (Sullivan et al., 1995, 1996; Yang, 1996): (a) no rate variation; (b) a proportion of sites assumed to be invariant, I, with the remainder having equal rates (Fitch and Margoliash, 1967); (c) rate variation following a discrete approximation to a gamma distribution, Γ (Yang, 1994); and (d) a proportion of sites assumed to be invariant, with the remainder following a discrete approximation to a gamma distribution, $I + \Gamma$ (Gu et al., 1995; Sullivan et al., submitted). In all, 16 (4×4) different models were investigated. Parameter values for each of the 16 models were obtained by fitting the expected tree to the data and optimizing values for that tree under maximum likelihood. Maximum-likelihood tests evaluating the fit of each of the models to the expected tree were carried out; the results are shown in the Appendix. Heuristic searches were conducted using maximum likelihood for the same 16 model conditions just described.

We fitted the expected tree to the data and measured the phylogenetic informativeness of each of the 12,234 sites for that tree using the retention index (RI; Archie, 1989; Farris, 1989). We measured base composition and its variation (deviation from stationarity) among the 19 taxa for the subset of sites with a perfect fit to the expected tree (those with RI = 1.0), and contrasted the values with those obtained for the entire population of sites. Principal-component analyses of nucleotide base composition and amino acid composition were plotted to provide a graphic repre-



(c)

Taxon	A	C	G	T
sea urchin 1	0.21	0.29	0.24	0.27
sea urchin 2	0.21	0.29	0.23	0.27
lancelet	0.20	0.27	0.24	0.29
lamprey	0.22	0.29	0.22	0.26
carp	0.21	0.30	0.23	0.25
trout	0.21	0.31	0.23	0.25
frog	0.22	0.30	0.22	0.25
chicken	0.22	0.32	0.23	0.24
opossum	0.23	0.30	0.22	0.25
blue whale	0.22	0.31	0.23	0.23
fin-back whale	0.22	0.31	0.23	0.23
cow	0.23	0.30	0.23	0.24
mouse	0.23	0.30	0.23	0.24
rat	0.23	0.30	0.23	0.24

Chi-square = 32.114801
(df=39), P = 0.7746122

Taxon	A	C	G	T
sea urchin 1	0.28	0.23	0.17	0.3
sea urchin 2	0.26	0.24	0.18	0.3
lancelet	0.25	0.17	0.21	0.3
lamprey	0.30	0.24	0.13	0.3
carp	0.30	0.28	0.15	0.2
trout	0.26	0.30	0.16	0.2
frog	0.31	0.24	0.13	0.3
chicken	0.28	0.34	0.13	0.2
opossum	0.33	0.22	0.12	0.3
blue whale	0.31	0.29	0.13	0.2
fin-back whale	0.31	0.28	0.13	0.2
cow	0.31	0.27	0.13	0.2
mouse	0.33	0.25	0.12	0.3
rat	0.32	0.27	0.12	0.2

Chi-square = 2617.295373
(df=39), P < 0.00000001

Figure 3. Comparison of base-compositional analysis of the 1207 sites with a perfect fit (RI = 1.0) to the expected tree and of the entire population of sites. (a) Expected tree (left); MPT yielded by equally weighted analysis of the entire data set (right). (b) Corresponding base-compositional profiles for each data set. Note the more balanced distribution of the four nucleotides in the subset of sites with a perfect fit (RI = 1.0) to the expected tree. (c) Deviation from stationarity among ingroup taxa was assessed using a chi-squared test. Base-compositional differences are significantly different from random expectation for the entire population of sites ($P < 0.0000001$), but are not significantly different for the subset of sites with a perfect fit to the expected tree. These tests are intended only as coarse heuristics and do not account for phylogenetic structure (Swofford, 1997).

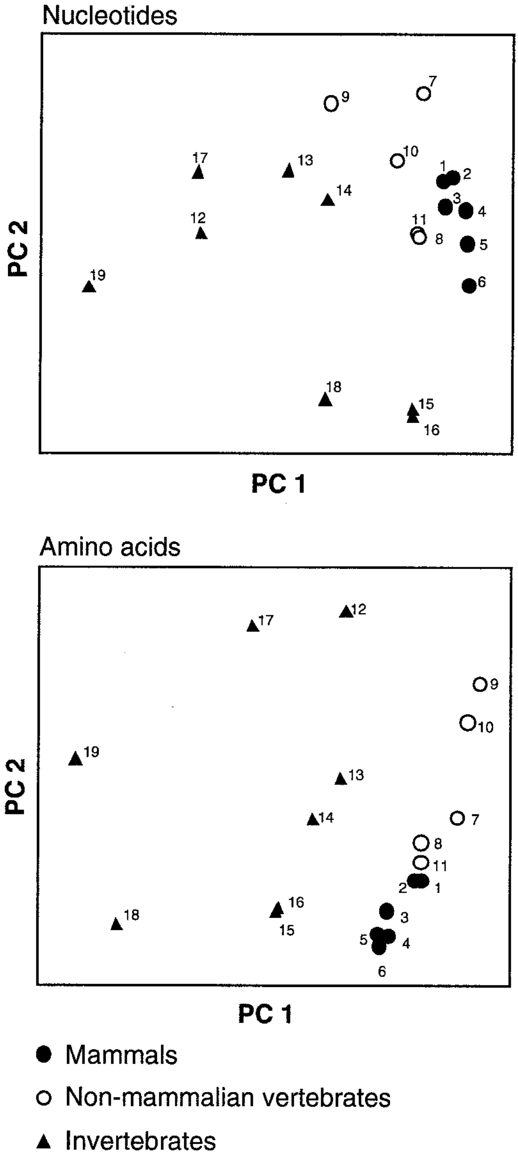


FIGURE 4. Compositional similarity among taxa for nucleotide bases and amino acid residues. Principal-component plots (PC1 vs. PC2) allow immediate identification of compositional similarity among the 19 taxa. The first two components account for 98% of the variation in nucleotide base composition and 67% of the variation in amino acid composition. All analyses are based on the correlation matrix. Solid circles represent mammals, open circles nonmammalian vertebrates, and solid triangles invertebrates. Numbers–species correspondences: 1 = fin-back whale, 2 = blue whale, 3 = cow, 4 = rat , 5 = mouse, 6 = possum, 7 = chicken, 8 = frog, 9 = trout, 10 = carp, 11 = lamprey, 12 = lancelet, 13 = sea urchin 1, 14 = sea urchin 2, 15 = mosquito, 16 = fruitfly,

←

17 = snail, 18 = nematode 1, 19 = nematode 2. Note that in both plots the two sea urchin taxa (13 and 14) are closer to the vertebrate taxa than is the lancelet (12).

sentation of overall compositional similarities among taxa. Maximum-likelihood tests evaluating the fit of the 16 models to the expected tree were carried out for the subset of sites with an RI of 1.0 and are contrasted with similar analyses using all 12,234 sites (see Appendix). Kishino–Hasegawa (1989) tests contrasting the topology of the expected tree with that of the most parsimonious tree (MPT) yielded by the complete nucleotide data set are shown for different models in the Appendix.

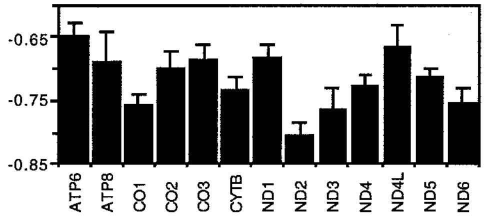
Each site in the alignment was classified according to gene, codon position, amino acid (the modal amino acid across taxa in the alignment), chemical property, charge, and relative hydrophobicity of the modal amino acid for that site in the alignment. An analysis of variance assessing the effect of each of these six factors on phylogenetic informativeness (RI for the expected tree) was then carried out.

RESULTS

In the equally weighted parsimony analyses, none of the genes, either individually or in combination, yielded the expected tree, nor were any of the fully resolved trees resulting from bootstrap resampling of the individual genes consistent with the expected tree. There was considerable consistency among genes in the pattern of inferred errors (Table 2). When all substitutions were analyzed, 10 of the 13 genes indicated *Branchiostoma* to be the sister taxon to a (vertebrate + echinoderm) clade, 4 genes (ATP6, CO1, ND4l, and ND6) indicated chicken to be the sister taxon to fishes, and 5 genes (ATP6, CO1, CO3, ND5, and ND6) indicated a monophyletic (frog, fish, chicken) clade. These repeated error patterns imply that homoplasy is highly nonrandomly distributed. The bootstrap consensus trees from the transversion analyses were less resolved and had fewer conflicts with the accepted tree; however,

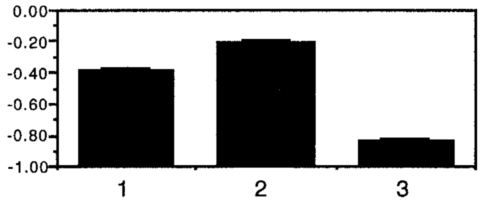
Gene

Sum of Squares	=	13.05
F Ratio	=	3.96
DF	=	12
Prob>F	=	0.0005



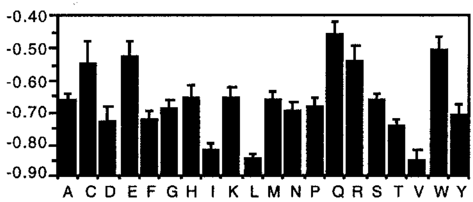
Codon position

Sum of Squares	=	555.02
F Ratio	=	1352.73
DF	=	2
Prob>F	=	0.0005



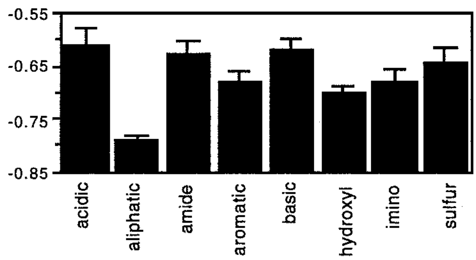
Amino acid

Sum of Squares	=	72.76
F Ratio	=	14.34
DF	=	19
Prob>F	=	0.0005



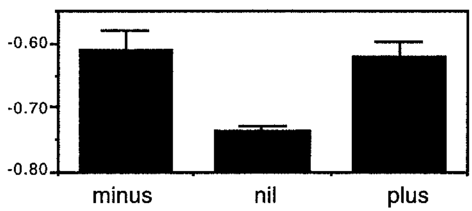
Chemical property

Sum of Squares	=	34.59
F Ratio	=	18.21
DF	=	7
Prob>F	=	0.0005



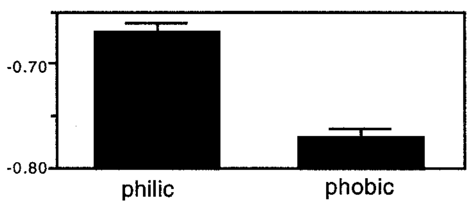
Charge

Sum of Squares	=	9.44
F Ratio	=	17.20
DF	=	2
Prob>F	=	0.0005



Hydrophobicity

Sum of Squares	=	20.22
F Ratio	=	74.05
DF	=	1
Prob>F	=	0.0005



CO1, CO2, and ND2 indicated *Branchiostoma* to be outside a (vertebrate + echinoderm) clade, and CO1 and ND5 indicated a (frog, fish, chicken) clade. There was less consistency in the pattern of errors with the amino acid sequences; nevertheless, some of the same inference errors seen in the nucleotide analyses resurfaced. When all genes were combined, the MPT and the corresponding bootstrap consensus differed from the expected tree at all three (nucleotide, transversion, and amino acid) levels of analysis, and the incorrect groupings often had high levels of bootstrap support (see Fig. 2).

Distance and maximum-likelihood analyses of the complete nucleotide data set failed to yield the expected topology for any one of the 16 model/ASRV combinations tested. These analyses, like the parsimony analysis (Fig. 2), all placed *Branchiostoma* outside echinoderms and the frog, fish, and chicken in a clade of their own. Although no single analysis yielded the expected tree, the more parameter-rich models (HKY and GTR with ASRV) yielded trees that were not significantly different from the expected tree when subjected to Kishino–Hasegawa (1989) tests (see Appendix). This suggests an improved fit between model and the data for the parameter-rich models.

That the entire protein-encoding portion of the mtDNA, a total of 12,234 sites, yields an inference that is both incorrect and supported by high bootstrap values in an equally weighted parsimony analysis is sobering. The fact that distance and maximum-likelihood analyses of the data under a variety of models (in which rate matrix parameters were optimized by first fitting the expected tree to the data set) also fail to yield the expected tree supports our original supposition that homoplasy is nonrandomly distributed within this large sample of sites. It is possible that most of the structured or misleading homoplasy is concentrated within a few genes. Indeed, we found that when we subjected a combined data set comprising amino acid sequences from ND1, ND4, CO1, CO2, CO3, and CYTB (2,302 amino acid sites) to an equally weighted parsimony bootstrap analysis, the expected tree resulted with 100% bootstrap support for all but three nodes. However, the utility of this finding is questionable, since the genes yielding correct results might vary among data sets and thus not be determinable a priori.

Collective Properties of Sites with a Perfect Fit to the Expected Tree

When the sequence data were fitted to the topology of the expected tree, we identified 1,207 phylogenetically informative sites with a perfect fit to that tree (i.e., 1,207 sites with an RI of 1.0). Base composition for this subset of sites was less skewed and showed no significant deviation from stationarity for the ingroup taxa, in marked contrast to the situation observed for the entire population of sites (Fig. 3). Moreover, in principal-component plots of nucleotide base composition and amino acid composition (Fig. 4), the vertebrate taxa have profiles that are clearly more similar to those of the two echinoderms than to that of *Branchiostoma*. These results are consistent with the prediction, made on the basis of simulations (Saccone et al., 1989, 1990, 1993; Steel et al., 1993; Lockhart et al., 1994; Steel, 1994; Pesole et al., 1995), that base-compositional deviations from stationarity can result in hierarchically structured homoplasy and, consequently, lead to incorrect phylogenetic inference.

We emphasize, however, that the base-

←

FIGURE 5. Tests of the association between functional characteristics and the phylogenetic informativeness of a site when the combined data set is fitted to the expected tree. The degree of informativeness was assessed using RI. Analysis of variance indicates that all six factors (gene, codon position, amino acid, chemical property, charge, and hydrophobicity) have highly significant effects on RI (log-transformed). The relative effects of the different levels of each factor are plotted against log RI (ordinate) as response sample means. Bars correspond to one standard error.

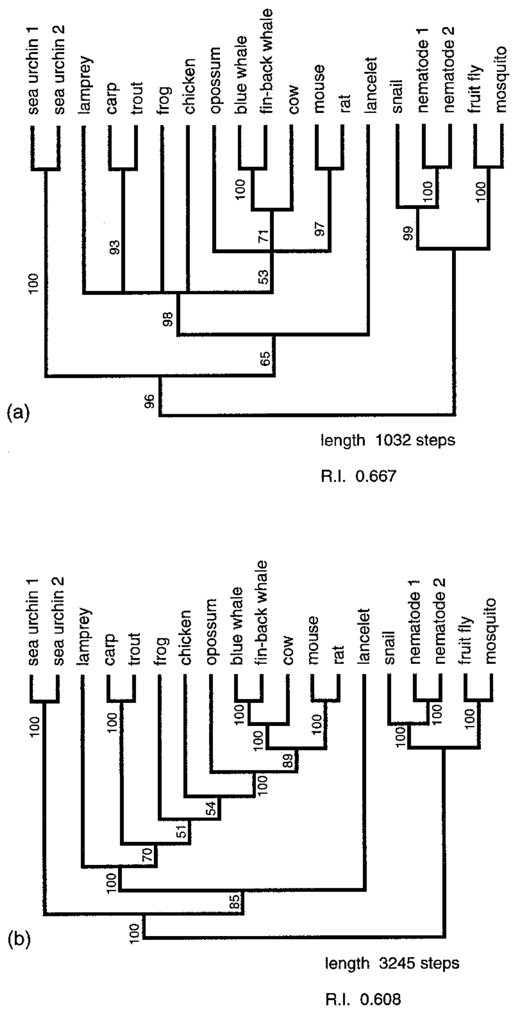


FIGURE 6. MPTs based on functional subsets of amino acids. Bootstrap support percentages are shown at each node. (a) Strict consensus of two MPTs resulting from the analysis of first and second codon positions for sites whose modal amino acid was proline or cysteine. (b) Single MPT resulting from analysis of first and second codon positions for sites whose modal amino acid was proline, cysteine, methionine, glutamine, and asparagine (the imino, sulfur, and amide side-chain groups, respectively).

compositional differences do not completely account for the inference errors in this data set. The base-composition plots for the subset of sites with a perfect fit to the expected tree, although markedly different from those for the entire data set, also show the vertebrate taxa to have profiles that are more similar to those of the two echinoderms than to that of *Branchiostoma*. There is thus no simple additive correspondence between base-compositional bias and the inferred phylogeny. Furthermore, erroneous inferences are not ameliorated by LogDet neighbor-joining analysis, a procedure demonstrated through simulation to retrieve correct phylogenies in the face of nonstationary base compositions when sites are independent (Steel et al., 1993; Lockhart et al., 1994; Steel, 1994). This is the case even when a proportion of sites are assumed to be invariant to accommodate bias due to among-site rate heterogeneity (Waddell, 1995; Swofford et al., 1996).

Relationship Between Function and Phylogenetic Informativeness

Analysis of variance indicated that all six factors tested (gene, codon position, amino acid, chemical properties, charge, and relative hydrophobicity) have highly significant effects ($P < 0.0005$) on the phylogenetic informativeness (RI) of a site (Fig. 5). This result reflects the importance of these properties to molecular structure and function. Significant interaction terms were found among some of the properties. For example, first positions had markedly higher RIs for hydrophilic than for hydrophobic sites, an association not seen at second or third positions. An interaction was also seen between gene and codon position ($P < 0.005$). Effect tests for this interaction revealed that third position sites had significantly higher RIs ($P < 0.05$) in ATP8 and ND4L than in other genes, suggesting that third codon position constraints may differ among genes.

Based on these analyses, we were able to identify classes of sites that yielded the expected tree when subjected to parsimony analysis. The greatest overall support resulted from an analysis of first and second codon positions of sites modally coding for proline, cysteine, methionine, glutamine, and asparagine. Parsimony analysis of the first two sites of all codons in positions modally coding for proline and cysteine yielded an incompletely resolved bootstrap consensus tree that was compatible with the expected vertebrate tree and had 65% bootstrap support for a monophyletic Chordata (cephalochordates + vertebrates; Fig. 6a). When the first two sites of all codons in positions modally coding for methionine, glutamine, and asparagine were added to this analysis the expected tree was obtained in fully resolved form, with strengthened (85%) bootstrap support for a monophyletic Chordata (Fig. 6b). Although there is an undeniable element of circularity involved in using the expected tree to determine sites that are informative, it is interesting and probably significant that those we identified are associated with conservative molecular motifs that are frequently important for protein structure and function. By contrast, analysis shows sites modally coding for the rapidly evolving hydrophobic amino acids leucine, isoleucine, and valine (Fig. 5) to have especially poor fits to the expected tree. Although poor fits are generally thought to be associated with saturated sites that have lost their signal, our analysis suggests something more problematic for phylogenetic inference: These sites have not only lost their historical signal, but contain a nonrandom signal that is misleading. Interestingly, a Templeton test indicates that the MPT (Fig. 2) is significantly ($P < 0.0001$) different from the expected tree (Fig. 1) when all 12,234 sites are included in the analysis, but not significantly different ($P = 0.94$) when isoleucine, leucine, valine, and third position sites are excluded. Similar results are seen with Kishino–Hasegawa (1989) tests. Details are presented in the Appendix.

It is possible that further work may show some of the patterns identified here to be more widespread. At present, however, we regard them as specific to this study and, at best, applicable only to studies using sequences from these same genes among metazoan taxa over a comparable range of divergence. Had we analyzed this same set of taxa using sequences from a different set of genes (e.g., genes for monomeric enzymes of the cytosol), different classes of informative sites might have been obtained, and a comparison of these same genes from more recently diverged taxa would almost certainly yield a different suite of informative sites. We also acknowledge that a denser sampling of echinoderm and chordate taxa for the same set of genes would likely change (and possibly improve) the phylogenetic estimate based on the entire data set (Lecointre et al., 1993; Hillis, 1996; Kim, 1996).

CONCLUSIONS

The assumption that historical signal will prevail if enough sites are sampled is widely held among evolutionary and systematic biologists. It is explicitly championed by the ''total evidence parsimony'' school and is often implicit in the work of those who embrace evolutionary models (Churchill et al., 1992; Huelsenbeck and Hillis, 1993). For example, Cummings et al. (1995) attempted to determine a sequence-sampling strategy that would approximate inferences yielded by entire mtDNAs, believing that the inferences yielded by the entire sequence would be more ''reliable'' than would any particular subsample. Russo et al. (1996), in evaluating the performance of different phylogenetic inference methods, stated: ''The most important factor in constructing reliable phylogenetic trees seems to be the number of amino acids or nucleotides used.'' Results presented in the current study demonstrate that there are circumstances in which this is simply not the case. Despite a very large sample—12,234 protein-coding sites, the maximum obtainable from metazoan mtDNA—an erroneous yet robust topology resulted—a topology contradicted by a wealth of other data. Clearly, the models underlying inference methods, whether implicit as is the case for parsimony or explicit as is the case for distance and maximum-likelihood models, are not accommodating the processes that have shaped the data. In the present data set, several methods actually converge on an incorrect topology as more sequence is added. These results are consistent with predictions based on simulations by Huelsenbeck and Hillis (1993). More data are better than fewer data only when the inference model accommodates, in an unbiased way, the evolutionary forces that have shaped character-state distributions. Any disparities (biases) that exist between a model (implied or explicit) and the evolutionary process will be magnified with increasing amounts of data.

This study provides an empirical demonstration that further sequencing does not automatically lead to an improved phylogenetic estimate. Once sequences from a few genes have been obtained, we believe that time and effort would be better spent investigating how knowledge of the structures and functions of those sequences and the products they encode can be integrated and incorporated into phylogenetic inference methods, rather than by adding more sequence data. In stating this, it is not our intent to discourage sequencing efforts, but to emphasize that it is useful to incorporate knowledge about what a sequence does as well as about what it is into the inference models we use. Evolutionary biologists rarely analyze information contained in sequence data beyond an aggregate pooling of information derived from individual nucleotide sites, even though such information is available for many of the sequences that are routinely used for phylogenetic inference. The structural and functional attributes of a particular gene product persist and can often be followed long after the historical signal in the underlying individual sequence elements has been lost. It is becoming increasingly possible to empirically assess character-state change probabilities for sites associated with such structural and functional attributes. Once these have been estimated for a particular gene, they can be incorporated into methods of inference in much the same way as has been done with estimates of relative rates of transitions and transversions. Comparisons that make use of such information may ultimately provide the key to resolving phylogenetic questions, such as those involving relationships among deeply diverged groups, that are unresolvable by analysis of the individual sequence elements themselves.

ACKNOWLEDGMENTS

We are grateful to Stan Blum, Susan Brown, Tim Collins, Elizabeth Knurek, Fred Kraus, Christian Pazmandi, Chris Simon, Una Smith, Jack Sullivan, and Dave Swofford for critical comments. This work was supported by National Science Foundation grant DEB-9220640 to W.M.B. and by a Sloan Postdoctoral Fellowship to G.J.P.N.

REFERENCES

ARCHIE, J. W. 1989. Homoplasy excess ratios: New indices for measuring levels of homoplasy in phylogenetic systematics and a critique of the consistency index. Syst. Zool. 38:253–269.

BENTON, M. J. 1993. The fossil record 2. Chapman and Hall, London.

BREMER, K. 1988. The limits of amino acid sequence data in angiosperm phylogenetic reconstruction. Evolution 42:795–803.

CHO, S., A. MITCHELL, J. C. REGIER, C. MITTER, R. W. POOLE, T. P. FRIEDLANDER, AND S. ZHAO. 1995. A highly conserved nuclear gene for low-level phylogenetics: Elongation factor-1α recovers morphology-based tree for heliothine moths. Mol. Biol. Evol. 12: 650–656.

CHURCHILL, G. A., A. VON HAESSLER, AND W. C. NAVIDI. 1992. Sample size for a phylogenetic inference. Mol. Biol. Evol. 9:753–769.

CUMMINGS, M. P., S. P. OTTO, AND J. WAKELEY. 1995. Sampling properties of DNA sequence data in phylogenetic analysis.Mol. Biol. Evol. 12:814–822.

DONOGHUE, M. J., R. G. OLMSTEAD, J. F. SMITH, AND J. D. PALMER. 1992. Phylogenetic relationships of dipscales based on *rbc*L sequences. Ann. Missouri Bot. Garden 79:333–345.

EERNISSE, D. J., AND A. KLUGE. 1993. Taxonomic congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules and morphology. Mol. Biol. Evol. 10:1170–1195.

EERNISSE, D. J. 1995. DNA Stacks: HyperCard software utilities for molecular systematists, version 1.1. Published electronically. Available at ftp://ftp.biology.indiana.edu.

FARRIS, J. S. 1969. A successive approximation approach to character weighting. Syst. Zool. 18:374-385.

FARRIS, J. S. 1983. The logical basis of phylogenetic analysis. Pages 7–36 *in* Advances in cladistics, Volume II (N. I. Platnick.and V. A. Funk, eds.). Columbia Press, New York.

FARRIS, J. S. 1989. The retention index and the rescaled consistency index. Cladistics 5:417–419.

FELSENSTEIN, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool 27:401–416.

FELSENSTEIN, J. 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39:783–791.

FITCH, W. M., AND E. MARGOLIASH. 1967. A method for estimating the number of invariant amino acid positions in a gene using cytochrome c as a model case. Biochem. Genet. 1:65–71.

GAUTHIER J., A. G. KLUGE, AND T. ROWE. 1988. Amniote phylogeny and the importance of fossils. Cladistics 4:105–209.

GOLDMAN, N. 1993. Statistical tests of models of DNA substitution. J. Mol. Evol. 36:182–198.

GRAYBEAL, A. 1994. Evaluating the phylogenetic utility of genes: A search for genes informative about deep divergences among vertebrates. Syst. Biol. 43: 174–193.

GU, X., Y.-X. FU, AND W.-H. LI. 1995. Maximum likelihood estimation of the heterogeneity of substitution rate among nucleotide sites. Mol. Biol. Evol. 12: 546–557.

HASEGAWA, M., H. KISHINO, AND T. YANO. 1985. Dating of the human–ape splitting by a molecular clock of mitochondrial DNA. J. Mol. Evol. 21:160–174.

HILLIS, D. M. 1991. Discriminating between phylogenetic signal and random noise in DNA sequences. Pages 278–294 *in* Phylogenetic analysis of DNA sequences (M. M. Miyamoto and J. Cracraft, eds). Oxford Univ. Press, New York.

HILLIS, D. M. 1996. Inferring complex phylogenies. Nature 383:130–131.

HUELSENBECK, J. P., AND D. M. HILLIS. 1993. Success of phylogenetic methods in the four-taxon case. Syst. Biol. 42:247–264.

JUKES, T. H., AND C. R. CANTOR. 1969. Evolution of protein molecules. Pages 21–132 *in* Mammalian protein metabolism (H. N. Munro, ed.). Academic Press, New York.

KIM, J. 1996. General inconsistency conditions for maximum parsimony: Effects of branch lengths and increasing numbers of taxa. Syst. Biol. 45:363–374.

KIMURA, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. J. Mol. Evol. 16:111–120.

KISHINO, H., AND M. HASEGAWA. 1989. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order of the Hominoidea. J. Mol. Evol. 29:170–179.

LANAVE, C., G. PREPARATA, C. SACCONE, AND G. SERIO. 1984. A new method for calculating evolutionary substitution rates. J. Mol. Evol. 20:86–93.

LECOINTRE, G., H. PHILIPPE, H. L VAN LÊ, AND H. LE GUYADER. 1993. Species sampling has a major impact on phylogenetic inference. Mol. Phyl. Evol. 2: 205–224.

LOCKHART, P. J., M. A. STEEL, M. D. HENDY, AND D. PENNY. 1994. Recovering evolutionary trees under a more realistic model of sequence evolution. Mol. Biol. Evol. 11:605–612.

MAISEY, J. G. 1986. Heads and tails: A chordate phylogeny. Cladistics 2:201–256.

MAISEY, J. G. 1988. Phylogeny of early vertebrate skeletal induction and ossification patterns. Evolutionary biology, Volume 22 (M. Hecht, B. Wallace, and G. T. Prance, eds.). Plenum, New York.

PESOLE, G., G. DELLISANTI, G. PREPARATA, AND C. SACCONE. 1995. The importance of base composition in the correct assessment of genetic distance. J. Mol. Evol. 41:1124–1127.

PHILIPPE H. A. CHENUIL, AND A. ADOUTTE. 1994. Can the Cambrian explosion be inferred through molecular phylogeny? Development (suppl.):15–25.

RODRÍGUEZ, F., J. L. OLIVER, A. MARÍN, AND J. R. MEDINA. 1990. The general stochastic model of nucleotide substitution. J. Theor. Biol. 142:485–501.

RUSSO, C. A. M, N. TAKEZAKI, AND M. NEI. 1996. Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. Mol. Biol. Evol. 13:525–536.

SACCONE, C., G. PESOLE, AND G. PREPARATA. 1989. DNA microenvironments and the molecular clock. J. Mol. Evol. 29:407–411.

SACCONE, C., C. LANAVE, G. PESOLE, AND G. PREPARATA. 1990. Influence of base composition on quantitative estimates of gene evolution. Methods Enzymol. 183:570–583.

SACCONE, C., C. LANAVE, AND G. PESOLE. 1993. Time and biosequences. J. Mol. Evol. 37:154–159.

STEEL, M. A., P. J. LOCKHART, AND D. PENNY. 1993. Confidence in evolutionary trees from biological sequence data. Nature 364:440–442.

STEEL, M. A. 1994. Recovering a tree from the leaf colorations it generates under a Markov model. Appl. Math. Lett. 7:19–23.

SULLIVAN, J., K. E. HOLSINGER, AND C. SIMON. 1995. Among-site rate variation and phylogenetic analysis of 12S rRNA data in sigmodontine rodents. Mol. Biol. Evol. 12:988–1001.

SULLIVAN, J., K. E. HOLSINGER, AND C. SIMON. 1996. The effect of topology on estimates of among site rate variation. J. Mol. Evol. 42:308–312.

SULLIVAN, J., D. L. SWOFFORD, AND G. J. P. NAYLOR. Uncertainty in estimating parameters of mixed-distribution models of rate heterogeneity. (submitted to Syst. Biol.)

SWOFFORD, D. L., G. J. OLSEN, P. J. WADDELL, AND D. M. HILLIS. 1996. Phylogenetic inference. Pages 407–514 *in* Molecular systematics, 2nd edition (D. M. Hillis, C. Moritz, and B. K. Mable, eds.). Sinauer Associates, Sunderland, Massachusetts.

TAVARÉ, S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lec. Math. Life Sci. 17:57–86.

TEMPLETON, A. R. 1983. Convergent evolution and non-parametric inferences from restriction fragment and DNA sequence data. Pages 151–179 *in* Statistical analysis of DNA sequence data (B. Weir, ed.). Marcel Dekker, New York.

THOMPSON, J. D., D. G. HIGGINS, AND T. J. GIBSON. 1994. CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position specific gap penalties and weight matrix choice. Nucleic Acids Res. 22: 4673–4680.

WADDELL, P. J. 1995. Statistical methods of phylogenetic analysis, including Hadamard conjugations, LogDet transforms and maximum likelihood. Ph.D. Dissertation, Massey Univ., New Zealand.

YANG, Z. 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: Approximate methods. J. Mol. Evol. 39: 306–314.

YANG, Z. 1996. Among site rate variation and its impact on phylogenetic analyses. TREE 11:367–372.

Received 13 March 1997; accepted 31 July 1997
Associate Editor: C. Simon

APPENDIX
MODELS OF SUBSTITUTION

Four different substitution models of increasing complexity were evaluated. The simplest, the Jukes–Cantor (1969) model, assumes both an even base composition and an equal probability of change for all six transformation types. The Kimura (1980) two-parameter model assumes equal base frequencies but allows a transition:transversion ratio to be specified. The Hasegawa–Kishino–Yano (1985) model allows for an uneven base composition and a transition:transversion ratio. The general time-reversible model (Lanave et al., 1984; Tavaré, 1986; Rodríguez et al., 1990) allows for an uneven base composition and separate probabilities of change for each of the six possible transformation types. None of the four models accommodates deviation from stationarity in either base composition or substitution dynamics.

Four among-site rate-heterogeneity models were evaluated for each of the following substitution models: (a) equal rates; (b) a proportion of sites assumed to be invariant among taxa, the remainder assumed to evolve at equal rates (I; Fitch and Margoliash, 1967); (c) rates assumed to follow a discrete approximation of the gamma distribution (Γ; Yang, 1994); (d) a proportion of sites assumed to be invariant, the remainder to follow a discrete approximation of the gamma distribution (I + Γ; Gu et al., 1995). Thus, 16 (4 x 4) substitution/among-site and rate-variation combinations were evaluated.

Swofford et al. (1996) have pointed out the tradeoffs between the consistency provided by a model's complexity and its sensitivity to random error. In general, it is desirable to use the simplest effective model to explain observations, that is, to choose a model that has enough parameters to explain the data satisfactorily, but not so many that statistical power is compromised.

TABLE 3. Likelihood-ratio test values for different substitution models.

	df	−log likelihood	X^2
JC	10	190,018.657	30,277.3712
JC + I	9	184,834.253	19,908.5647
JC + Γ	9	181,487.388	13,214.8342
JC + I + Γ	8	181,340.654	12,921.3662
K2P	9	189,228.248	28,696.5546
K2P + I	8	183,988.963	18,217.9833
K2P + Γ	8	180,223.396	10,686.8493
K2P + I + Γ	7	180,109.353	10,458.7644
HKY85	6	186,023.715	22,287.4874
HKY85 + I	5	180,474.223	11,188.5048
HKY85 + Γ	5	175,238.975	718.0086
HKY85 + I + Γ	4	175,160.793	561.64454
GTR	2	184,936.447	20,112.9514
GTR + I	1	179,573.355	9,386.76752
GTR + Γ	1	174,980.197	200.45272
GTR + I + Γ	0	174,879.971	0

Which Model Best Explains the Data?

In order to identify the most appropriate substitution model the –ln likelihood scores for the expected tree were compared for the 16 different models. A likelihood-ratio test statistic was computed and contrasted with a chi-squared approximation of the null distribution (Goldman, 1993). Results (Table 3) indicate that all models fit the data significantly worse ($P < 0.01$) than does the parameter-rich GTR + I + Γ model. This should not be interpreted to mean that the most parameter-rich model will find the expected tree (in fact it does not), but rather that simpler models will fare even more poorly.

A similar test was carried out for the subset of 1,207 sites maximally informative for the expected tree under parsimony (those with RI = 1.0). The results were comparable to those obtained for the entire data set, insofar as all models fitted the data significantly less well than did the parameter-rich GTR + I + Γ model. However, X^2 values were much lower for the maximally informative subset of data, with values ranging from 17.3 to 40.3 (cf. 200.5–30,277.4). This is probably because the maximally informative sites can be more easily reconciled to the expected tree with simple models. (These sites collectively exhibit a more even base composition and less deviation from stationarity and thus do not require extra parameters to accommodate base-compositional unevenness and among-site rate variation.)

Under Which Substitution Models Is the Expected Tree Significantly Different From the Most-Parsimonious Tree?

Kishino–Hasegawa (1989) tests were conducted to contrast the likelihood score for the MPT resulting from equally weighted parsimony (Fig. 2) with that for the expected tree (Fig. 1) under the 16 models for all 12,234 sites. In no case (Table 4) did the expected tree fit the data significantly better. For the simpler models (JC and K2P), the MPT had a significantly better score than the expected tree. However, as more parameter-rich models, accommodating base composition (HKY85, GTR) and among-site rate heterogeneity (I + Γ) were used, the differences in the –ln likelihood scores diminished in significance. Results show that both rate heterogeneity and base composition must be incorporated before the MPT and the expected tree are no longer significantly different.

Kishino–Hasegawa tests were also carried out for a 5,566-bp subset of the data from which third codon positions and sites modally coding for isoleucine, leucine, and valine were excluded (Table 5). Under all models, the expected tree had a better score than the MPT topology from Figure 2. However the difference in score did not become significant until both gamma-distributed rate heterogeneity and base composition were incorporated. We note that although the expected tree has a more likely score than the topology de-

TABLE 4. Likelihood scores and P values from Kishino-Hasegawa (1989) tests for the analysis of the entire data set. The most likely score is underlined. 1 = Expected tree; 2 = most parsimonious tree.

	JC	K2P	HKY85	GTR
Equal rates	1 = 190,018.6565 2 = 189,706.5707 $P < 0.0001$	1 = 189,228.2482 2 = 188,966.6047 $P < 0.0001$	1 = 186,023.7146 2 = 185,874.1456 $P = 0.0004$	1 = 184,936.447 2 = 184,809.982 $P = 0.0020$
I	1 = 184,834.253 2 = 184,611.3826 $P < 0.0001$	1 = 183,988.9626 2 = 183,811.3409 $P < 0.0001$	1 = 180,474.2233 2 = 180,403.5445 $P = 0.31$	1 = 179,573.3547 2 = 179,525.1419 $P = 0.1265$
Γ	1 = 181,487.388 2 = 181,34.7965 $P < 0.0001$	1 = 180,223.3956 2 = 180,146.4203 $P = 0.012$	1 = 175,238.975 2 = 175,252.093 $P = 0.6165$	1 = 174,980.1973 2 = 175,003.3906 $P = 0.3796$
I + Γ	1 = 181,340.654 2 = 181,204.7859 $P < 0.0001$	1 = 180,109.353 2 = 180,030.772 $P = 0.0089$	1 = 175,160.793 2 = 175,175.196 $P = 0.5739$	1 = 174,879.9709 2 = 174,903.3324 $P = 0.3642$

TABLE 5. Likelihood scores and P values from Kishino-Hasegawa (1989) tests for this subset of the data (third position, isoleucine, leucine, and valine sites excluded). The most likely score is underlined. 1 = Expected tree; 2 = most parsimonious tree.

	JC	K2P	HKY85	GTR
Equal rates	1 = 59,463.524 2 = 59,485.268 $P = 0.5123$	1 = 59,425.251 2 = 59,451.074 $P = 0.4367$	1 = 59,340.893 2 = 59,375.084 $P = 0.2987$	1 = 59,114.897 2 = 59,143.418 $P = 0.3870$
I	1 = 57,696.94 2 = 57,711.336 $P = 0.5716$	1 = 57,650.961 2 = 57,669.307 $P = 0.4715$	1 = 57,553.331 2 = 57,579.049 $P = 0.3064$	1 = 57,314.134 2 = 57,335.5039 $P = 0.3998$
Γ	1 = 56,455.274 2 = 56,483.425 $P = 0.1645$	1 = 56,394.168 2 = 56,425.445 $P = 0.1213$	1 = 56,235.43 2 = 56,271.685 $P = 0.0686$	1 = 56,105.484 2 = 56,142.104 $P = 0.0706$
I + Γ	1 = 56,454.932 2 = 56,482.845 $P = 0.1668$	1 = 56,393.699 2 = 56,424.717 $P = 0.1231$	1 = 56,235.09 2 = 56,271.145 $P = 0.0693$	1 = 56,104.476 2 = 56,140.645 $P = 0.0728$

TABLE 6. Likelihood scores and P values from Kishino-Hasegawa (1989) tests for the subset of the data using the 1,207 sites that are maximally informative for the expected tree under parsimony. The most likely score is underlined. 1 = Expected tree; 2 = most parsimonious tree.

	JC	K2P	HKY85	GTR
Equal rates	1 = 9,221.3168 2 = 9,275.9519 $P = 0.0001$	1 = 9,023.3549 2 = 9,068.3934 $P = 0.0003$	1 = 9,015.3612 2 = 9,059.9317 $P = 0.0003$	1 = 9,006.7110 2 = 9,052.3787 $P = 0.0003$
I	1 = 9,026.8563 2 = 9,071.4046 $P = 0.0003$	1 = 9,023.3549 2 = 9,068.3934 $P = 0.0003$	1 = 9,015.3612 2 = 9,059.9317 $P = 0.0003$	1 = 9,006.7110 2 = 9,052.3787 $P = 0.0003$
Γ	1 = 9,026.8563 2 = 9,071.4046 $P = 0.0003$	1 = 9,023.3549 2 = 9,068.3934 $P = 0.0003$	1 = 9,015.3612 2 = 9,059.9317 $P = 0.0003$	1 = 9,006.7110 2 = 9,052.3787 $P = 0.0003$
I + Γ	1 = 9,045.1688 2 = 9,081.1777 $P = 0.0015$	1 = 9,040.2982 2 = 9,076.6480 $P = 0.0014$	1 = 9,030.3767 2 = 9,066.1219 $P = 0.0015$	1 = 9,023.4482 2 = 9,060.3618 $P = 0.0013$

TABLE 7. Results of the Templeton tests comparing the expected tree with the most parsimonious tree. The shorter of two trees is underlined.

Data subset	Length of expected tree	Length of Fig. 2 topology (MPT)	*P* value
Complete data (12,234 bp)	46,058	<u>45,734</u>	<0.0001
No I, L, V, or 3rd positions (5,566 bp)	12,464	<u>12,462</u>	0.9433
RI = 1.0 sites only (1,207 bp)	<u>1,945</u>	1,996	<0.0001

picted in Figure 2 for this subset of the data, other tree topologies (that are neither the expected tree nor the tree shown in Fig. 2) have still better scores. Indeed, the MPT for this particular subset of sites is different from that resulting from an analysis of all 12,234 sites.

Kishino–Hasegawa tests were carried out for a second subset of the data: those 1,207 sites maximally informative for the expected tree under parsimony (i.e., those with RI = 1.0). In this case the expected tree (Table 6) has a significantly better score than does the MPT (Fig. 2), in all 16 cases. Inclusion of extra parameters to accommodate among-site rate heterogeneity and base-compositional differences has no effect on the level of significance between the two trees tested, because sites with a perfect fit do not show appreciable among-site rate variation or uneven base composition.

The results of the Templeton tests parallel those seen for the Kishino–Hasegawa (1989) tests (Table 7).