

# Bias in Phylogenetic Reconstruction of Vertebrate Rhodopsin Sequences

Belinda S. W. Chang<sup>1</sup> and Dana L. Campbell<sup>2</sup>

Department of Organismic and Evolutionary Biology, Harvard University

Two spurious nodes were found in phylogenetic analyses of vertebrate rhodopsin sequences in comparison with well-established vertebrate relationships. These spurious reconstructions were well supported in bootstrap analyses and occurred independently of the method of phylogenetic analysis used (parsimony, distance, or likelihood). Use of this data set of vertebrate rhodopsin sequences allowed us to exploit established vertebrate relationships, as well as the considerable amount known about the molecular evolution of this gene, in order to identify important factors contributing to the spurious reconstructions. Simulation studies using parametric bootstrapping indicate that it is unlikely that the spurious nodes in the parsimony analyses are due to long branches or other topological effects. Rather, they appear to be due to base compositional bias at third positions, codon bias, and convergent evolution at nucleotide positions encoding the hydrophobic residues isoleucine, leucine, and valine. LogDet distance methods, as well as maximum-likelihood methods which allow for nonstationary changes in base composition, reduce but do not entirely eliminate support for the spurious resolutions. Inclusion of five additional rhodopsin sequences in the phylogenetic analyses largely corrected one of the spurious reconstructions while leaving the other unaffected. The additional sequences not only were more proximal to the corrected node, but were also found to have intermediate levels of base composition and codon bias as compared with neighboring sequences on the tree. This study shows that the spurious reconstructions can be corrected either by excluding third positions, as well as those encoding the amino acids Ile, Val, and Leu (which may not be ideal, as these sites can contain useful phylogenetic signal for other parts of the tree), or by the addition of sequences that reduce problems associated with convergent evolution.

## Introduction

Phylogenetic analysis is a complex problem in inference. It is therefore not surprising that all existing phylogenetic methods are known to fail under some conditions and for a variety of reasons. In recent years, several issues have emerged as particularly thorny. Use of an oversimplified model of molecular evolution or strong violation of the assumptions of a model can result in convergence to an incorrect topology with greater certainty as sequence length increases (i.e., inconsistency). This type of problem is particularly relevant to parsimony analyses, especially in cases in which some branches are much longer than others, a problem which has been dubbed “long-branch attraction” (Felsenstein 1978). Phylogenetic methods based on explicit models of evolution, such as distance and maximum likelihood, tend to be less vulnerable to this type of problem, but even these are known to display inconsistency under conditions where their assumptions are strongly violated (Hillis, Huelsenbeck, and Cunningham 1994; Gaut, and Lewis 1995; Yang 1996; Huelsenbeck 1997; Sullivan and Swofford 1997; Huelsenbeck 1998). In addition, although taxon sampling has long been an important issue in phylogenetic analyses, it remains difficult to establish reasonable guidelines for sampling and to assess the effects that it may have on the accuracy of tree topologies (Hillis 1996, 1998; Poe 1998; Rannala et al. 1998).

Determining the particular conditions under which phylogenetic methods fail is critical to both understanding their limitations and developing new, improved models and algorithms better suited to the analysis of molecular data. For example, third positions have been thought to be problematic in many data sets due to the effects of base compositional bias (Saccone, Pesole, and Preparata 1989; Sidow and Wilson 1990; Sogin, Hinkle, and Lelpe 1993). This has led to the development of models that incorporate nonstationary changes in base composition for both distance and likelihood phylogenetic methods (Lockhart et al. 1994; Steel 1994; Galtier and Gouy 1995, 1998). However, in practice, it is often difficult to identify concrete examples of failure of phylogenetic methods in real data sets and to pinpoint the reasons for that failure. Another common feature of molecular data sets that may cause phylogenetic methods to fail is variation in codon bias across the tree, but examples of this in real data sets have yet to be isolated, and the challenges they pose for phylogenetic reconstruction have only just begun to be addressed (Goldman and Yang 1994; Muse 1996; Yang 1997).

Rhodopsin is an ideal genetic system for exploring issues in phylogenetic reconstruction, because it has been cloned from a variety of species, and much is known about its function and molecular evolution (Chang et al. 1995, 1996; Baylor 1996; Baylor and Burns 1998; Bowmaker 1998; Sakmar 1998; Townson et al. 1998). Rhodopsin is a single-copy nuclear gene encoding a seven-transmembrane G-protein-coupled receptor which forms the first step in the visual transduction cascade in the photoreceptors of the eye (Nathans 1992). In vertebrates, it is expressed at high levels in a single cell type, rod photoreceptor cells (Khorana 1992; Chang et al. 1996; Baylor and Burns 1998; Sakmar 1998). Rhodopsin has been found to exist in more than one copy only in rare instances, for example, in polyploid animals such as the carp, *Cyprinus carpio* (Lar-

<sup>1</sup> Present address: Department of Molecular Biology and Biochemistry, Rockefeller University.

<sup>2</sup> Present address: Department of Biology, University of Maryland, College Park.

Key words: molecular evolution, hydrophobic amino acids, base compositional bias, codon bias, parametric bootstrapping.

Address for correspondence and reprints: Belinda S. W. Chang, Rockefeller University, 1230 York Ave., Box 284, New York, New York 10021. E-mail: changb@rockvax.rockefeller.edu.

*Mol. Biol. Evol.* 17(8):1220–1231. 2000

© 2000 by the Society for Molecular Biology and Evolution. ISSN: 0737-4038

**Table 1**  
**Rhodopsin Sequences Used in this Study**

	Tree Label	Species Name	Accession No.	%GC <sub>3</sub>	Scaled $\chi^2$	ENC	
Fishes.....	Sandgoby	<i>Pomatoschistus minutus</i>	X62405	75	0.46	41	
	Zebrafish	<i>Brachydanio rerio</i>	L11014	89	0.82	33	
	Goldfish	<i>Carassius auratus</i>	L11863	74	0.45	42	
	Carp	<i>Cyprinus carpio</i>	S74449	77	0.48	41	
	Cavefish	<i>Astyanax mexicanus</i>	U12328	86	0.77	34	
	<b>MBerndti</b>	<b><i>Myripristis berndti</i></b>	<b>U57538</b>	<b>68</b>	<b>0.38</b>	<b>43</b>	
	<b>Anguilla</b>	<b><i>Anguilla anguilla</i></b>	<b>L78008</b>	<b>74</b>	<b>0.39</b>	<b>44</b>	
	<b>Conger</b>	<b><i>Conger conger</i></b>	<b>S82619</b>	<b>81</b>	<b>0.61</b>	<b>39</b>	
	Mammals.....	Cow	<i>Bos taurus</i>	K00502	89	0.74	35
Hamster		<i>Cricetulus griseus</i>	X61084	75	0.41	47	
Mouse		<i>Mus musculus</i>	M36695	79	0.49	42	
Dog		<i>Canis familiaris</i>	X71380	82	0.59	40	
Human		<i>Homo sapiens</i>	K02281	84	0.65	38	
Macaque		<i>Macaca fascicularis</i>	S76579	86	0.70	38	
Rabbit		<i>Oryctolagus cuniculus</i>	U21688	91	0.79	34	
Rat		<i>Rattus norvegicus</i>	U22180	80	0.57	38	
Lampreys.....		<b>Sea lamprey</b>	<b><i>Petromyzon marinus</i></b>	<b>U67127</b>	<b>87</b>	<b>0.71</b>	<b>36</b>
		Lamprey	<i>Lampetra japonica</i>	M63632	88	0.73	34
Amphibians.....	Rana	<i>Rana pipiens</i>	S49004	65	0.32	46	
	Xenopus	<i>Xenopus laevis</i>	L04692	59	0.26	49	
	Salamander	<i>Ambystoma tigrinum</i>	U36574	71	0.40	44	
Reptiles <sup>a</sup> .....	Alligator	<i>Alligator mississippiensis</i>	U23802	61	0.28	51	
	Chicken	<i>Gallus gallus</i>	M88178	84	0.59	40	
	Anolis	<i>Anolis carolinensis</i>	L31503	56	0.16	57	
Skate.....	<b>Skate</b>	<b><i>Raja erinacea</i></b>	<b>U81514</b>	<b>71</b>	<b>0.32</b>	<b>50</b>	

NOTE.—The five additional sequences in the “expanded data set” are shown in bold type. %GC<sub>3</sub> refers to GC content at third positions; scaled  $\chi^2$  and effective number of codons (ENC) are two measures of codon bias (Shields et al. 1988; Wright 1990).

<sup>a</sup> In this paper, we refer to the “reptiles” as including birds.

hammar and Risinger 1994). Most important for this study, phylogenetic relationships among vertebrates, for which rhodopsin sequences are available, have been well-characterized using fossil, morphological, and molecular data (Carroll 1997).

This study takes advantage of well-established vertebrate relationships to examine in detail molecular evolutionary forces which result in spurious reconstructions in a data set of vertebrate rhodopsin sequences. Once these factors have been identified, methods are explored to reduce their effects and eliminate the spurious reconstructions.

## Materials and Methods

Rhodopsin sequences were obtained from the GenBank database via NCBI's website (<http://www.ncbi.nlm.nih.gov/genbank/>). GenBank accession numbers for all the sequences used are given in table 1. Rhodopsin cDNA sequences were aligned using CLUSTAL W and modified by hand to allow only gaps between codons. This file was then translated to yield an equivalently aligned amino acid rhodopsin data set. Parsimony, distance, and maximum-likelihood phylogenetic analyses were performed using a beta-test version of PAUP\*, version 4 (Swofford 1999). Trees were rooted using the lamprey sequence as an outgroup. In addition, many of the analyses also included four paralogous rod-like cone opsin genes (GenBank accession numbers: gekko blue, M92035; chick green, M92038; goldfish green1, L11865; goldfish green2, L11866) as outgroup sequences in order to confirm the position of the root (Chang et al. 1995). The results of these analyses con-

firmed the position of the lamprey as the most basally diverging vertebrate rhodopsin.

In order to determine the best model for distance and likelihood analyses, likelihood scores were determined for five different models: JC (Jukes and Cantor 1969), K2P (Kimura 1980), F81 (Felsenstein 1981), HKY85 (Hasegawa, Kishino, and Yano 1985), and GTR (Yang 1994). Additionally, the effect of incorporating among-sites rate heterogeneity (using the  $\Gamma$ -distribution; Yang 1993) into each of the models was examined. Likelihood ratio tests were used to compare likelihood scores obtained for pairs of nested models to determine which model best fit our sequence data (Felsenstein 1991; Yang, Goldman, and Friday 1994).

In addition to equally weighted parsimony analyses, 2:1 transversion (Tv): transition (Ts) weighting was also used. Although other weighting schemes were explored, they produced less reliable trees (data not shown). In addition, this weighting scheme reflects the likelihood estimate of the Tv/Ts ratio (1.5). Distance bootstrap analyses were performed using the HKY85+ $\Gamma$ , HKY85, and K2P models and the neighbor-joining algorithm.

In order to assess phylogenetic signal in the data set, 10,000 random trees were generated in PAUP\* to calculate  $g_1$  statistics (Hillis and Huelsenbeck 1992). In addition, two measures of codon bias, scaled  $\chi^2$  and effective number of codons (ENC) (Shields et al. 1988; Wright 1990), were calculated to assess codon usage in each taxon. Measures of nucleotide and codon bias were calculated using the program MEA (generously provided by its author, E. Moriyama).

**Table 2**  
**Ranges of Pairwise Nucleotide Distances (HKY85-corrected) Among Major Vertebrate Groups**

	Lampreys	Skate	Fishes	Amphibians	Reptiles	Mammals
Lampreys (2) . . . . .	0.075	0.287–0.292	0.280–0.337	0.276–0.315	0.253–0.357	0.234–0.287
Skate (1) . . . . .		—	0.294–0.353	0.280–0.300	0.264–0.325	0.280–0.300
Fishes (8) . . . . .			0.062–0.264	0.263–0.347	0.242–0.381	0.235–0.343
Amphibians (3) . . . . .				0.160–0.200	0.203–0.279	0.223–0.232
Reptiles (3) . . . . .					0.184–0.258	0.176–0.322
Mammals (8) . . . . .						0.037–0.154

NOTE.—Number of sequences is indicated in parentheses beside each vertebrate group.

To test for long-branch attraction (Huelsenbeck 1998), 100 data sets were simulated by parametric bootstrapping using the program SIMINATOR (Huelsenbeck, Hillis, and Jones 1996) with parameters estimated from the original rhodopsin sequence data set. The simulated data sets were subsequently analyzed using equally weighted maximum parsimony in PAUP\*, version 4, with 100 replications of (nonparametric) bootstrapping, 10 random-addition replicates each.

## Results

### Phylogenetic Analyses

Phylogenetic analyses were performed on a data set of 20 vertebrate rhodopsin nucleotide sequences (table 1). Although this data set showed high levels of genetic variation (table 2) and generally performed well in reconstructing traditional relationships among vertebrates, phylogenetic analyses consistently show substantial bootstrap support for two groupings which contradict established vertebrate relationships: reptiles and amphibians form a clade (fig. 1B), instead of the more traditional reptiles and mammals (fig. 1A), and alligator and anolis form a clade (fig. 2B), instead of alligator and chicken (fig. 2A). In parsimony analysis with equal weights (table 3), bootstrap support was 86% for the reconstruction of amphibians as the sister group to reptiles (this node is hereinafter referred to as amph+rept) and 72% for the grouping of alligator + anolis as the sister lineage to the chicken (this node is hereinafter referred to as gator+anol). Support for these resolutions is robust to changes in the relative weightings of transversions and transitions: 91% for rept+amph and 65% for gator+anol with 2-to-1 Tv/Ts (table 3). Less than 5% bootstrap support was seen for more accepted resolutions of these nodes. This is in contrast to the robust

support for established relationships elsewhere in the tree (fig. 3). Note that this data set, like many other molecular data sets, does not recover the Glires clade (rodents + rabbits), but instead places the rodents basal to a clade containing artiodactyls and other mammals. On the other hand, most morphological data recover the Glires (de Jong 1998). The Glires controversy is beyond the scope of this paper and does not influence its major observations.

Another unusual aspect of this data set is that despite the substantial divergences among sequences (table 2), useful phylogenetic signal has been retained in third positions. Not only do third positions contribute the largest numbers of informative sites (out of 568 total informative sites, 315 were in third positions, 151 were in first positions, and 102 were in second positions), but they also contain enough signal that when analyzed alone (fig. 4), they recover a tree that is almost as well supported as the tree with all three codon positions included. In addition, analyses of the degree of skewness of a distribution of lengths of 10,000 randomly generated trees imply that some phylogenetic signal does reside in third positions (all positions:  $g_1 = -0.69$ ; third positions only:  $g_1 = -0.69$ ; first + second positions only:  $g_1 = -0.80$ ).

Although there is useful phylogenetic signal retained at third positions with respect to many nodes in the tree, this signal also appears to be underlying some of the support for the problematic reconstructions. Bootstrap support for these incongruent resolutions almost completely disappeared when third positions were excluded from parsimony analysis (<5% for rept+amph, 10% for gator+anol), an effect that is robust to changes in transversion-transition weighting (table 3). Furthermore, excluding third positions had the effect of increasing support for the more established chicken + alligator

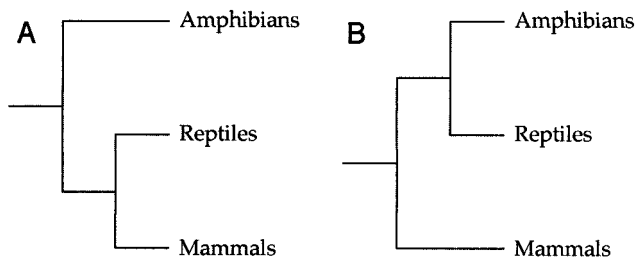


FIG. 1.—Alternate resolutions of the reptile-amphibian-mammal trichotomy. A, Expected resolution based on well-established vertebrate relationships. B, Spurious resolution supported by the rhodopsin data set.

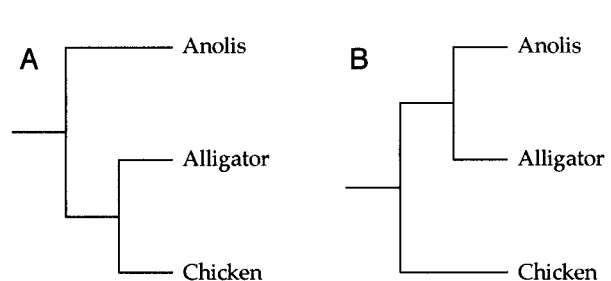


FIG. 2.—Alternate resolutions of the chicken-alligator-anolis trichotomy. A, Expected resolution based on well-established relationships. B, Spurious resolution supported by the rhodopsin data set.

**Table 3**  
**Bootstrap Support (%) for Alternative Resolutions of the Incongruent Nodes in the Original Data Set**

	Reptiles + Mammals	Reptiles + Amphibians	Chicken + Alligator	Alligator + Anolis
Parsimony analyses				
All positions . . . . .	<5	86	<5	72
2:1 (Tv/Ts) <sup>a</sup> . . . . .	<5	91	9.5	65
Third positions only. . . . .	<5	68	<5	41
2:1 (Tv/Ts) <sup>b</sup> . . . . .	<5	80	<5	53
Including ILV codons. . . . .	<5	89	<5	59
Transversions only. . . . .	<5	48	<5	25
First + second positions only. . . . .	16	<5	72	10
2:1 (Tv/Ts). . . . .	10	<5	68	17
Excluding ILV codons. . . . .	44	<5	84	<5
Amino acid analyses <sup>c</sup> . . . . .	25	<5	<5	<5
Distance analyses				
K2P. . . . .	<5	78	16	43
HKY85. . . . .	<5	76	<5	44
HKY85+Γ. . . . .	<5	71	<5	47
LogDet <sup>d</sup> . . . . .	<5	56	38	38
Maximum-likelihood analyses				
All positions, HKY85+Γ <sup>e</sup> . . . . .	<5	79	<5	44

<sup>a</sup> See figure 3 for full phylogeny.  
<sup>b</sup> See figure 4.  
<sup>c</sup> See figure 5.  
<sup>d</sup> See figure 6.  
<sup>e</sup> See figure 7.

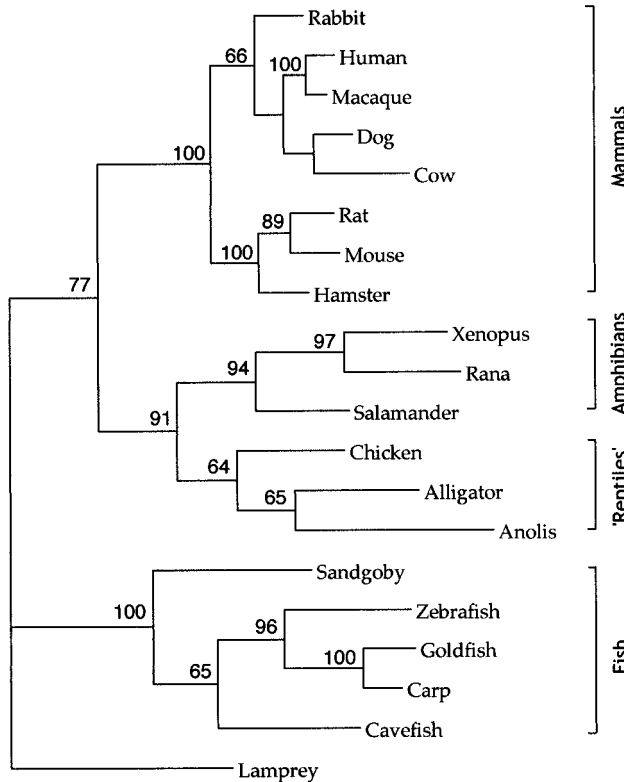


FIG. 3.—Maximum-parsimony bootstrap phylogeny of the 20 rhodopsin nucleotide sequences in the original data set. One hundred bootstrap replications of 20 vertebrate rhodopsin sequences were performed, with 2-to-1 weighting of transversions to transitions.

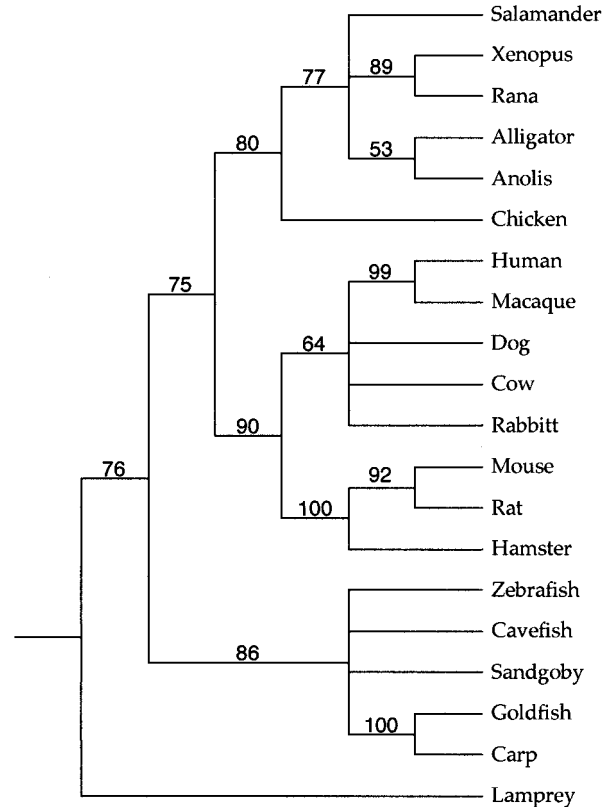


FIG. 4.—Maximum-parsimony bootstrap phylogeny of third positions only; 100 replications with 2-to-1 weighting of transversions to transitions.

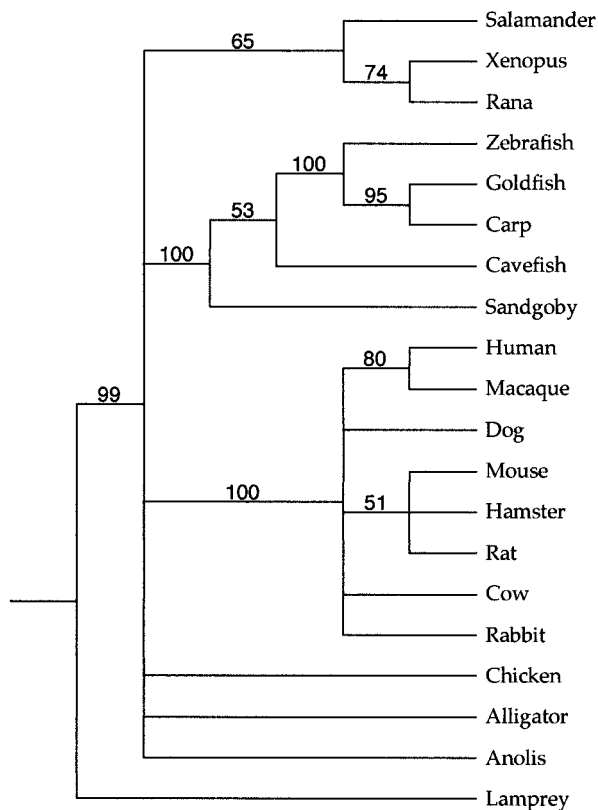


FIG. 5.—Maximum-parsimony bootstrap phylogeny of rhodopsin amino acid sequences. Equally weighted parsimony analysis with 100 bootstrap replications.

grouping (hereinafter chick+gator) to 72%, in contrast to the <5% bootstrap support shown when all positions were included in the analysis. Support for the reptile + mammal grouping (hereinafter rept+mamm) also increased, but not as much (16%), when third positions were excluded.

When analyzed alone, third positions showed substantial support for the spurious resolutions (68% for rept+amph, 41% for gator+anol) and no support for the well-corroborated relationships, an effect which was robust to changes in Tv/Ts weighting (table 3 and fig. 4). Analyses of the amino acid sequences, which should be free of the base compositional and codon bias effects particularly problematic for third-base positions and transitions, did not show any support for the incongruent relationships (table 3). However, the bootstrap phylogeny based on amino acids was rather poorly resolved in general (fig. 5).

Distance analyses which did not incorporate nonstationary changes in base composition did not fare much better than parsimony for this data set, and also tended to recover the problematic nodes with substantial bootstrap support (71% for rept+amph and 47% for gator+anol, HKY85+ $\Gamma$  model; table 3). These bootstrap values remained quite stable, even when the correction for rate heterogeneity was not included in the analysis or when models with fewer parameters were used (table 3).

Given the variation in base composition in this data set, especially at third positions (see table 1), analyses

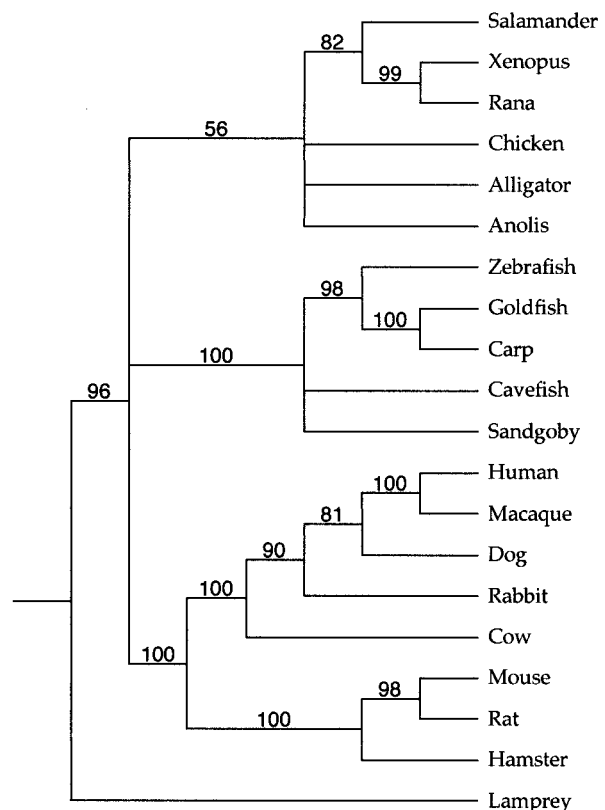


FIG. 6.—Neighbor-joining bootstrap phylogeny using LogDet distances (1,000 replications).

using LogDet/paralinear distance methods (Lake 1994; Lockhart et al. 1994; Steel 1994) were performed. These methods allow for nonstationary changes in base composition among sequences in a phylogeny and would be expected to perform better for data sets where this is a problem. Phylogenetic bootstrap analyses using LogDet distances did show reduced support for the problematic reconstructions (56% for rept+amph and 38% for gator+anol; table 3 and fig. 6). Moreover, for one of the problematic nodes, there was also slightly increased support for the correct reconstruction (38% for chick+gator; table 3).

Maximum-likelihood methods were also explored for this data set (fig. 7). Likelihood ratio tests were used to compare nested models of evolution in order to identify models that best fit our data set. These models were tested for a phylogeny of well-established vertebrate relationships (fig. 8A). Among the models tested, among-sites rate heterogeneity was the single most important parameter resulting in significantly better likelihood scores ( $\chi^2$  ranged from 2235.2 to 2412.8,  $P < 0.001$  for all comparisons; table 4). Among the models incorporating rate heterogeneity, GTR+ $\Gamma$  had significantly higher likelihood scores in pairwise comparisons with all other models ( $\chi^2 = 123.8$ –619.6,  $P < 0.001$  for all comparisons) except for the HKY85+ $\Gamma$  model ( $\chi^2 = 6$ ,  $P = 0.2$ ). The HKY85+ $\Gamma$  model, when compared with nested models with fewer parameters, had significantly better likelihood scores ( $\chi^2 = 117.8$ –306.8,  $P < 0.001$  for all comparisons). Since the GTR+ $\Gamma$  model was not

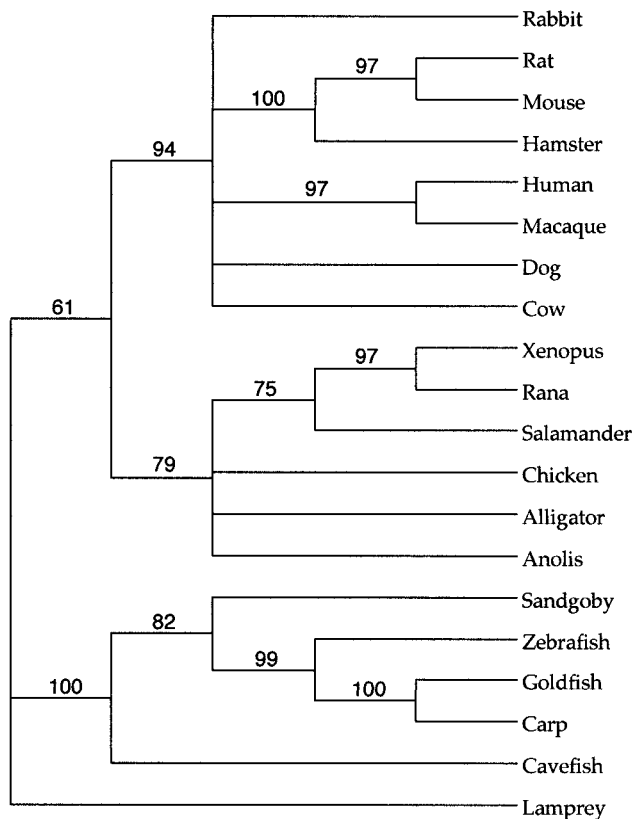


FIG. 7.—Maximum-likelihood bootstrap phylogeny under the HKY+ $\Gamma$  model (100 replications).

found to be significantly better than the HKY85+ $\Gamma$  model, the HKY85+ $\Gamma$  model was determined to be the best fit of those tested for this data set and was subsequently used in a full likelihood bootstrap analysis of the rhodopsin data set. However, maximum-likelihood phylogenetic methods under the HKY85+ $\Gamma$  model did not perform any better than distance or parsimony methods, showing substantial support for spurious resolutions

at both nodes (79% for rept+amph and 44% for gator+anol; fig. 7 and table 3).

Since base compositional effects appeared to be important in this data set, likelihood methods which allow for nonstationary GC content were also explored (Galtier and Gouy 1998). Due to the computational intensity of this method (which allows GC content to vary across all of the branches of the tree), which made a full maximum-likelihood bootstrap analysis prohibitive, likelihood scores for alternative resolutions of the reptile-mammal-amphibian node (see fig. 1) were determined instead. A topology representing a well-established tree of vertebrate relationships (fig. 8A) was found to have a lower log likelihood score ( $L = -9,566.12$ ) than a second topology with the alternate resolution ((amphibians, reptiles), mammals), represented in figure 1B ( $L = -9,544.30$ ). The higher likelihood score of the topology with the incongruent resolution at this node, as compared with established vertebrate relationships, indicates that even this model cannot fully account for the signal underlying the spurious reconstruction.

Finally, it has been suggested that hydrophobic amino acids may be less useful for phylogenetic reconstruction than other amino acids (Naylor and Brown 1997). To explore the effects of hydrophobic amino acids in the rhodopsin data set, nucleotide positions encoding the hydrophobic amino acids Ile, Leu, and Val were excluded in a parsimony analysis (189 nucleotide positions excluded, representing 63 amino acids). This analysis showed greatly reduced bootstrap support for the spurious resolutions (<5% for rept+amph and 33% for gator+anol, 2:1 Tv/Ts; table 3), indicating that positions encoding for these amino acids may underlie the spurious signal. If the spurious signal was due mainly to functional constraints on these hydrophobic amino acids, then excluding third positions should not affect the analysis. This was not the case, as the effect remained even when only third positions of the hydrophobic amino acids Ile, Leu, and Val were excluded (table 3).

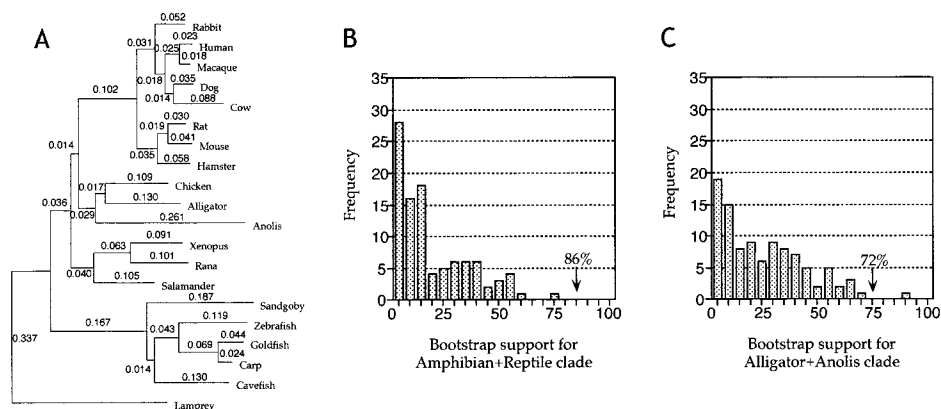


FIG. 8.—Results of simulation studies using parametric bootstrapping (see text for details concerning simulation conditions). Simulated data sets were analyzed using maximum parsimony with 100 nonparametric bootstrap replications. A, Tree topology used for simulations, with branch lengths indicated. B, Distribution of parsimony bootstrap support for the spurious reconstruction ((reptiles, amphibians), mammals) in the simulated data sets. C, Distribution of parsimony bootstrap support for the spurious reconstruction ((alligator, anolis), chicken) in the simulated data sets. These graphs represent the expected distribution of support for each of the spurious nodes under the model assumptions specified in the simulation experiment (see text). Arrows in histograms indicate the bootstrap values of the spurious nodes recovered in parsimony analysis of the real (non-simulated) rhodopsin data set. Histogram bins represent ranges of 5% bootstrap values.

**Table 4**  
**Likelihood Ratio Tests Comparing Nested Models**

EVOLUTION- ARY MODELS	LN $L^a$		LIKELIHOOD RATIO TESTS <sup>b</sup>		
	no $\Gamma$	+ $\Gamma$	With $\Gamma$ vs.	vs. HKY85+ $\Gamma$	vs. GTR+ $\Gamma$
			Without $\Gamma$		
JC . . . . .	-13,742.6	-12,598.4	2,288.4* (1)	306.8* (4)	619.6* (8)
F81 . . . . .	-13,704.8	-12,544.2	2,321.2* (1)	252.6* (1)	511.2* (5)
K2P . . . . .	-13,518.1	-12,350.5	2,335.2* (1)	117.8* (3)	123.8* (7)
HKY85 . . . . .	-13,498.0	-12,291.6	2,412.8* (1)	—	6.0 (4)
GTR . . . . .	-13,406.2	-12,288.6	2,235.2* (1)	—	—

<sup>a</sup> Likelihood scores for a tree of established vertebrate relationships (fig. 8A).

<sup>b</sup> Likelihood ratio test,  $\chi^2 = 2\Delta\ln L$ , with degrees of freedom indicated in parentheses.

\* Significant (assuming a  $\chi^2$  distribution);  $P < 0.001$ .

### Statistical Tests Comparing Trees

Several statistical tests were performed using the rhodopsin nucleotide data set in order to determine if phylogenies with and without the two spurious reconstructions were significantly different. The Templeton (1983) test and the “winning sites” test (Prager and Wilson 1988) compare trees under the parsimony criteria, whereas the Kishino-Hasegawa test (Kishino and Hasegawa 1989) was formulated to compare trees under either likelihood or parsimony. Tests under the parsimony criteria are shown in table 5, and tests under the likelihood criteria are shown table 6. Each of the two spurious reconstructions (rept+amph, gator+anol) was tested separately in pairwise tests of trees with and without each spurious reconstruction. These tests confirmed the results of the phylogenetic bootstrap analysis, pinpointing third positions and nucleotides encoding Ile, Leu, and Val as the sites supporting the spurious reconstructions. Although neither spurious reconstruction (rept+amph, gator+anol) was significantly better with all nucleotide sites included, when only third positions and sites encoding Ile, Leu, or Val were considered, trees with the spurious reconstructions became significantly better than those without. This was true under both parsimony (table 5) and likelihood (table 6). Conversely, when only first and second positions, excluding those sites encoding Ile, Leu, or Val, were considered, the tree without spurious reconstructions was found to be better than either one of the trees with the spurious reconstructions. This result was significant under parsimony, but not under likelihood (tables 5 and 6).

### Simulation Studies Using Parametric Bootstrapping

Long-branch attraction has been identified as a potential reason for problematic groupings in several studies (Huelsenbeck, Hillis, and Jones 1996; Huelsenbeck 1997, 1998). One of the criteria for identifying long-branch attraction as a potential problem in phylogenetic analyses of a particular data set, according to Huelsenbeck (1997), is to show that branches of the topology are indeed long enough to attract using simulation studies. In order to test for long-branch attraction as a contributing factor to the spurious reconstructions of the rhodopsin data set, simulations were performed using parametric bootstrapping techniques. Parameters and

branch lengths for the simulations were estimated using maximum likelihood on an established tree of vertebrate relationships (Carroll 1997; figure 8A), under the HKY+ $\Gamma$  model (maximum-likelihood-estimated parameters:  $K = 3.12$ ,  $\alpha = 0.33$ , frequency of A = 0.19, frequency of C = 0.35, frequency of G = 0.24, frequency of T = 0.23; branch lengths are given in fig. 8A). Although a few of the resolutions of taxa present in this tree do remain somewhat controversial (e.g., the placement of the rabbits as basal to artiodactyls instead of with rodents), these are unlikely to affect the simulations with respect to the nodes in question.

Results from parsimony bootstrap analysis of the 100 simulated data sets are graphed in figure 8B and C, representing the expected null distribution of parsimony bootstrap values for each spurious reconstruction (rept+amph, gator+anol). Note that support for these spurious clades was being examined under conditions where the data were simulated from topologies reflecting the more established relationships (rept+mamm, gator+chick). The median level of bootstrap support for the incorrect rept+amph clade was 10.5% and that for the gator+anol clade was 19% in the simulated data sets. In the real rhodopsin data set, bootstrap support for both spurious resolutions was significantly higher than expected from the null distribution of simulated data sets generated by parametric bootstrapping (86% for rept+amph and 72% for gator+anol;  $P < 0.05$  in both cases). This indicates that the level of support seen for the problematic reconstructions is higher than would be expected given the conditions of the simulations, and therefore unlikely to be due to long-branch attraction.

### Base Composition and Codon Bias Measures

Since the results of the phylogenetic analyses and statistical tests comparing phylogenies implied that third positions, as well as transitions, underlie the bootstrap support of the spurious reconstructions, base composition and codon bias measures were examined for evidence of convergent evolution. First- and second-position nucleotide compositions were fairly homogeneous across all sequences. However, at third positions, reptile and amphibian rhodopsins tended to have lower %GC than other sequences (table 1). This pattern of convergent evolution may confound phylogenetic analyses and

**Table 5**  
**P Values for Statistical Tests of Parsimony Length Differences Between Trees With and Without Spurious Reconstructions**

	SHORTER TREE IN PAIRWISE COMPARISONS	ORIGINAL DATA SET			EXPANDED DATA SET		
		K-H Test	Templ. Test	Win. Sites Test	K-H Test	Templ. Test	Win. Sites Test
All positions . . . . .	Tree 2	0.0771	0.0771	0.1116	0.8349	0.8348	1.0000
	Tree 3	0.3660	0.3657	0.4510	0.3766	0.3763	0.4610
2:1 (Tv/Ts) . . . . .	Tree 2 <sup>a</sup>	0.3322	0.3484	0.3487	0.7632	0.7441	1.0000
	Tree 3	0.6082	0.6903	0.3496	0.6735	0.7311	0.4962
Third positions only . . . . .	Tree 2	0.0122*	0.0124*	0.0192*	0.3465	0.3458	0.4807
	Tree 3	0.0337*	0.0339*	0.0518	0.0394*	0.0396*	0.0592
2:1 (Tv/Ts) . . . . .	Tree 2	0.0741	0.0707	0.1763	0.7242	0.7250	0.8450
	Tree 3	0.0627	0.0692	0.0500*	0.0706	0.0730	0.0896
First + second positions only . . . . .	Tree 1	0.1799	0.1797	0.3750	0.1799	0.1797	0.3750
	Tree 1	0.0833	0.0833	0.1460	0.0833	0.0833	0.1460
2:1 (Tv/Ts) . . . . .	Tree 1	0.1969	0.1634	0.6875	0.2485	0.2342	0.6875
	Tree 1	0.0679	0.0696	0.1460	0.0679	0.0696	0.1460
Third positions including ILV codons . . . . .	Tree 2	0.0080*	0.0082*	0.0125*	0.3716	0.3711	0.5034
	Tree 3	0.0235*	0.0236*	0.0367*	0.0278*	0.0280*	0.0425*
2:1 (Tv/Ts) . . . . .	Tree 2	0.0567	0.0545	0.1336	0.7320	0.7327	0.8506
	Tree 3	0.0488*	0.0543	0.0369*	0.0547	0.0569	0.0673
First + second positions, excluding ILV codons . . . . .	Tree 1	0.0454*	0.0455*	0.1250	0.0833	0.0833	0.2500
	Tree 1	0.0347*	0.0348*	0.0654	0.0347*	0.0348*	0.0654
2:1 (Tv/Ts) . . . . .	Tree 1	0.1089	0.0977	0.3750	0.2062	0.1936	0.6250
	Tree 1	0.0410*	0.0493*	0.0654	0.0410*	0.0493*	0.0654

NOTE.—Pairwise tests are between trees with (tree 2) and without (tree 1) the problematic reptile + amphibian grouping, followed by another test with (tree 3) and without (tree 1) the problematic alligator + anolis grouping. K-H = Kishino-Hasegawa test (Kishino and Hasegawa 1989); Templ. = Templeton (1983) test; Win. Sites = winning sites test (Prager and Wilson 1988).

<sup>a</sup> For this comparison, tree 1 is shorter than tree 2 in the expanded data set.

\* Significant;  $P < 0.05$ .

result in the spurious grouping, as shown by mapping the GC content on the phylogeny (fig. 9). Furthermore, amphibian and reptile rhodopsins are less biased in their codon usage, as shown by scaled  $\chi^2$  and ENC codon bias measures, than are the rhodopsins of other vertebrate groups (table 1). Not only are there convergences in the overall degree of codon bias, but there are also convergences in the usage frequencies of specific codons that reflect the spurious groupings. This convergent pattern was evident when the codon usage frequencies were mapped on a tree. For example, convergences in the frequency of GGC, one of four codons coding for glycine, are shown mapped on the tree in figure 9.

The results of the phylogenetic analyses and statistical tests comparing alternative phylogenies also implicated positions encoding hydrophobic residues Ile, Leu, and Val as contributing to the high bootstrap support of the spurious reconstructions. In order to further explore this effect, base composition was examined at these sites for evidence of convergent evolution. Third positions in general had already been shown to be convergent in this data set (see above, fig. 9); therefore, for these amino acids, attention was focused on first and second positions. Second positions did not vary, as the amino acids Ile, Leu, and Val are all encoded by the same nucleotide, T. However, at first positions, at these sites, it was found

**Table 6**  
**Kishino-Hasegawa Tests Using Maximum Likelihood Under the HKY+ $\Gamma$  Model**

	TREE WITH HIGHER LNL	ORIGINAL DATA SET			EXPANDED DATA SET		
		$\Delta\ln L$	SD	$P$	$\Delta\ln L$	SD	$P$
All positions . . . . .	Tree 2	12.80	8.14	0.1162	9.73	8.56	0.2556
	Tree 3	1.20	5.40	0.8243	0.63	5.34	0.9054
Third positions only . . . . .	Tree 2	7.84	7.52	0.2982	8.09	5.51	0.1434
	Tree 3	0.78	3.44	0.8210	1.02	3.27	0.7558
First + second positions only . . . . .	Tree 1	1.09	1.72	0.5259	4.39	3.77	0.2447
	Tree 1	8.43	5.13	0.1006	7.72	4.97	0.1209
Third positions, including ILV codons . . . . .	Tree 2	15.46	7.77	0.0472*	11.99	6.69	0.0735
	Tree 3	3.25	3.59	0.3659	2.53	3.39	0.4549
First + second positions, excluding ILV codons . . . . .	Tree 1	1.60	2.12	0.4516	2.57	2.85	0.3684
	Tree 1	10.12	5.71	0.0771	9.78	5.59	0.0807

NOTE.—Pairwise tests are between trees with (tree 2) and without (tree 1) the problematic reptile + amphibian grouping, followed by another test with (tree 3) and without (tree 1) the problematic alligator + anolis grouping.

\* Significant;  $P < 0.05$ .



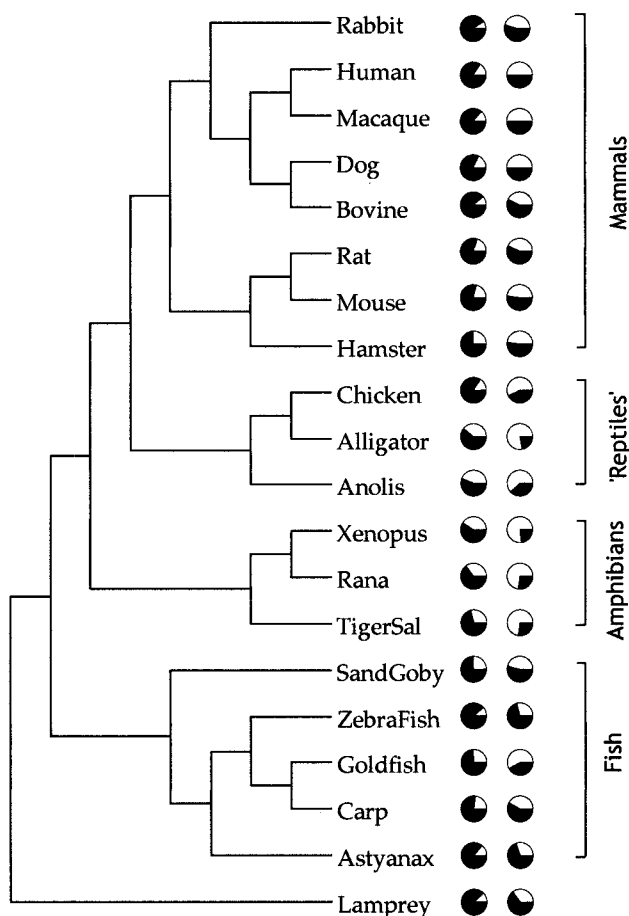


FIG. 9.—Phylogeny of expected vertebrate relationships showing convergences in %GC at third positions, and an example of convergence in codon bias of a sample codon of glycine. %GC is represented by the black portion of the pies in the first column, and the proportion of the total codons of Gly represented by the codon GGC is represented by the black portion of the pies in the second column.

that reptile and amphibian rhodopsins tended to have more A's (32.01%) than all other sequences (28.86%). This effect was not as marked in first positions that encoded amino acids other than Ile, Leu, and Val (27.28% for amph+rept, 26.43% for all others). This pattern of convergent evolution resulting in increased numbers of A's in first positions also results in an increased proportion of Ile's in reptile and amphibian rhodopsins relative to the total numbers of Ile, Leu, and Val present (28.1% for amph+rept, 26.5% for all others).

#### Effect of Increased Sampling

If the spurious reconstructions seen in this data set were due to convergent evolution, perhaps better sampling across the tree could ameliorate this effect. Rhodopsin sequences from five basally diverging taxa that were recent additions to GenBank were added to the data set: sea lamprey, *Conger* eel, *Anguilla* eel, skate, and *Myripristis berndti*, a holocentrid marine fish (table 1). It is important to note that not only do these sequences represent basal species poorly sampled in the original data set, but several of them also display values

of base composition at third positions and/or codon bias quite different from their closest neighbors on the tree, and are thus more likely to "break up" convergent effects.

*Myripristis berndti* and *Anguilla* rhodopsins have only 67.81% and 73.65% GC content at third positions, as compared with other fish rhodopsins, which average 80.16% (table 1). Similarly, skate rhodopsin has much lower %GC at third positions (70.70%) than the nearest basal lineage, lamprey rhodopsin (87.57%). The two measures of codon bias, scaled  $\chi^2$  and ENC, also showed the *M. berndti*, skate, and *Conger* rhodopsins to be atypically low in codon bias compared with neighboring fish and lamprey sequences (table 1).

For this expanded data set, equally weighted parsimony analysis of all positions showed reduced bootstrap support for the spurious rept+amph clade (48%) as compared with the original data set (86% without the additional sequences) and increased support for the correct rept+mamm clade, which rose from <5% in the original data set (table 3) to 25% in the expanded data set (table 7). Unlike analyses of the original data set, in which there was virtually no difference between equal weights versus 2-to-1 Tv/Ts weighting schemes, analysis of the expanded data set was highly sensitive to differences in weighting, particularly in the resolution of the reptile-mammal-amphibian node. When Tv/Ts weighting was used, bootstrap support for the correct rept+mamm clade jumped from 25% (equal weights) to 70% (2:1 Tv/Ts; fig. 10). In contrast, bootstrap support for the spurious gator+anol clade remains substantial in the analysis of the expanded data set (73%), and the high degree of sensitivity to differences in Tv/Ts weighting was not seen here (table 7).

In both cases, bootstrap support for the spurious resolutions disappeared entirely when third positions were excluded from parsimony analysis, regardless of Tv/Ts weighting (table 7). These results are similar to those of the analysis of the original data set (table 3). However, in contrast to the original data set, when third positions were excluded in the expanded data set, bootstrap support for the more established resolutions was increased (44% for rept+mamm and 78% for chick+gator, equal weights). When analyzed alone, third positions showed substantial support for the spurious resolutions and no support for the well-corroborated resolutions of these nodes, regardless of Tv/Ts weighting (table 7).

The patterns of bootstrap support in distance analyses of the expanded data set (table 7) remained very similar to those of the original data set (table 3), with very little difference in support between the models used, showing neither decreased support for spurious resolutions nor increased support for correct resolutions. Maximum-likelihood reconstructions under HKY85+ $\Gamma$  in the expanded data set also showed results similar to those found for the original data set and did not show reduced support for the spurious nodes nor heightened support for the correct nodes in the expanded data set (table 7).

Statistical comparisons of trees with and without the spurious reconstructions (rept+amph, gator+anol)

**Table 7**  
**Bootstrap Support (%) for the Incongruent Nodes in an Expanded Data Set with Five Additional Sequences**

	Reptiles + Mammals	Reptiles + Amphibians	Chickens + Alligators	Alligators + Anolis
Parsimony analyses				
All positions . . . . .	25	48	5.2	73
2:1 (Tv/Ts) <sup>a</sup> . . . . .	70	26	8.2	68
Third positions only . . . . .	<5	51	<5	59
2:1 (Tv/Ts) . . . . .	<5	69	<5	64
Transversions only . . . . .	33	22	16	10
First + second positions only . . . . .	44	<5	78	<5
2:1 (Tv/Ts) . . . . .	62	<5	77	<5
Distance analyses				
HKY85 . . . . .	<5	75	30	27
LogDet . . . . .	<5	82	40	19
Maximum-likelihood analyses				
All positions, HKY85+Γ . . . . .	<5	64	41	36

<sup>a</sup> See figure 10 for full phylogeny.

were consistent with the phylogenetic bootstrap analyses. A tree with the gator+anol clade was still better than one without this spurious reconstruction when only third positions and sites encoding Ile, Leu, and Val were considered. This result was significant under the parsimony criterion (table 5) and not quite significant under the likelihood criterion ( $P = 0.07$ ; table 6). However, the sites which clearly supported the spurious gator+anol reconstruction in both the original and extended data sets and also supported the spurious rept+amph reconstruction in the original data set were no longer

capable of distinguishing between a tree with the spurious rept+amph reconstruction and one without in the extended data set (tables 5 and 6). This result is again consistent with the phylogenetic bootstrap analyses, which suggest that the additional sequences aid in breaking up convergences among the sequences, but only for the spurious rept+amph reconstruction, which is more proximal to the additional sequences, leaving the spurious gator+anol reconstruction largely unaffected.

## Discussion

Our results indicate that the two problematic reconstructions in the original rhodopsin data set were probably not the result of topological effects such as long-branch attraction. This is demonstrated by the persistence of these spurious nodes when maximum-likelihood methods were used and by the fact that the bootstrap support for these spurious nodes was well outside of the distribution of support obtained for each node from simulated data sets generated by parametric bootstrapping. Rather, these spurious reconstructions were most likely due to convergences in base compositional bias at third positions, in codon bias, and in positions encoding for the hydrophobic amino acids Ile, Val, and Leu, which tend to group unrelated sequences. This represents a strong violation of phylogenetic model assumptions of stationary base composition and codon frequencies across the tree, which would cause methods not directly addressing these problems to fail under these conditions.

Base compositional bias at third positions has often been found to be problematic for phylogenetic reconstruction, and several methods have been developed in an attempt to address this problem (Lockhart et al. 1994; Galtier and Gouy 1995, 1998). Although these methods did reduce support for the spurious reconstructions in the rhodopsin data set, they were not completely effective in eliminating the problematic nodes, and it seems clear that base compositional bias is not the only reason for the spurious nodes. In fact, simulation studies on a

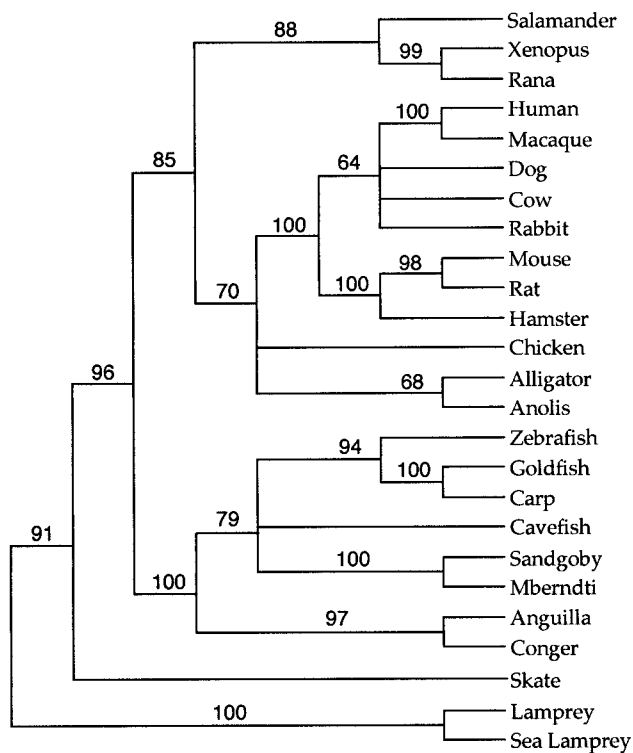


FIG. 10.—Maximum-parsimony analysis of the expanded data set with five additional sequences (100 bootstrap replications with 2-to-1 weighting of transversions to transitions).

data set of bat sequences have shown that levels of base compositional bias must be extremely high (>90% AT) in order to show any evidence of spurious reconstructions (Van Den Bussche et al. 1998). Although fairly high, levels of base compositional bias are not so extreme in the rhodopsin data set.

In addition to base compositional bias, convergent effects in codon bias and in positions encoding hydrophobic amino acids also appear to be supporting the spurious reconstructions in the rhodopsin data set. Other phylogenetic studies that have also found problematic reconstructions have attributed these to various problems such as not incorporating rate heterogeneity across sites into the phylogenetic model (Takezaki and Gojobori 1999), which is clearly not the case here. However, there is growing evidence that convergent or parallel evolution at the level of nucleotides (or amino acids) is a common feature of many molecular data sets and may pose a significant challenge in attempting to reconstruct unbiased phylogenies (Naylor and Brown 1997, 1998; Cao et al. 1998; Foster and Hickey 1999; Lee 1999). In particular, nucleotide sites encoding the hydrophobic amino acids Ile, Leu, and Val have been shown in other studies to display lower retention indices than other sites (Naylor and Brown 1997), and the analyses of the rhodopsin data set presented here provide more evidence of the importance of this effect. The reasons for it still remain unclear but may be related to relaxed constraints on hydrophobic amino acids contained within transmembrane domains.

There are several ways to address these problems of bias in base composition, codon frequencies, and sites encoding hydrophobic amino acids. All of these positions could be excluded from a parsimony phylogenetic analysis. This method can be effective in principle, but in fact may not be ideal, as these positions often contain useful phylogenetic signal in addition to the spurious signal, and excluding them can result in loss of resolution in the phylogenetic reconstructions (e.g., see Campbell, Brower, and Pierce 2000). Another way of addressing this problem would be to develop more complex models of evolution which incorporate these assumptions about base composition, codon bias, and amino acid composition. However, this may require the addition of many more parameters to the model, which may become problematic.

In addition to advances in phylogenetic methodology, this problem may be effectively addressed, albeit indirectly, via better sampling of species. Note that here "better sampling" means the addition of sequences not only proximal to problematic nodes, but also intermediate in base composition and codon bias. In other words, it is not only important when considering sampling issues to "break up" long branches that can lead to the failure of methods such as parsimony, but even more important to "break up" convergences in base composition and codon bias that can cause all types of phylogenetic methods, not just parsimony, to fail. In fact, it should be noted that of all the phylogenetic methods used here, only weighted parsimony methods are able to recover the correct topology once appropriately

sampled sequences are included in the analysis, and thus these methods outperform both distance and maximum-likelihood methods in this regard. This may reflect greater sensitivity of maximum-likelihood and distance methods to incorrect assumptions in the underlying models (with respect to nonstationary nucleotide and codon bias and hydrophobic sites) in comparison with parsimony methods, which sometimes may prove more robust to violations of these assumptions despite the fact that maximum-likelihood methods are known to be consistent over a larger set of conditions than are parsimony methods (Hillis, Huelsenbeck, and Cunningham 1994; Huelsenbeck 1997; Sullivan and Swofford 1997).

## Acknowledgments

We thank Z. Yang, R. Honeycutt, and two anonymous reviewers for many helpful comments on the manuscript, and N. Pierce and M. Donoghue for discussion and advice. B.S.W.C. is an NSF/Alfred P. Sloan Fellow in Molecular Evolution.

## LITERATURE CITED

- BAYLOR, D. 1996. How photons start vision. *Proc. Natl. Acad. Sci. USA* **93**:560–565.
- BAYLOR, D. A., and M. E. BURNS. 1998. Control of rhodopsin activity in vision. *Eye* **12**:521–525.
- BOWMAKER, J. 1998. Evolution of colour vision in vertebrates. *Eye* **12**:541–547.
- CAMPBELL, D. L., A. V. Z. BROWER, and N. E. PIERCE. 2000. Molecular evolution of the *Wingless* gene and its implications for the phylogenetic placement of the butterfly family Riodinidae (Lepidoptera: Papilionoidea). *Mol. Biol. Evol.* **17**:684–696.
- CAO, Y., A. JANKE, P. J. WADDELL, M. WESTERMAN, O. TAKENAKA, S. MURATA, N. OKADA, S. PAABO, and M. HASEGAWA. 1998. Conflict among individual mitochondrial proteins in resolving the phylogeny of eutherian orders. *J. Mol. Evol.* **47**:307–322.
- CARROLL, R. L. 1997. Patterns and processes of vertebrate evolution. Cambridge University Press, Cambridge, England.
- CHANG, B. S. W., D. AYERS, W. C. SMITH, and N. E. PIERCE. 1996. Cloning of the gene encoding honeybee long-wavelength rhodopsin: a new class of insect visual pigments. *Gene* **173**:215–219.
- CHANG, B. S. W., K. S. CRANDALL, J. P. CARULLI, and D. L. HARTL. 1995. Opsin phylogeny and evolution: a model for blue shifts in wavelength regulation. *Mol. Phylogenet. Evol.* **4**:31–43.
- DE JONG, W. W. 1998. Molecules remodel the mammalian tree. *Trends Ecol. Evol.* **13**:270–275.
- FELSENSTEIN, J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst. Zool.* **27**:401–410.
- . 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**:368–376.
- . 1991. PHYLIP: phylogeny inference package. Version 3.4. University of Washington, Seattle.
- FOSTER, P. G., and D. A. HICKEY. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *J. Mol. Evol.* **48**:284–290.
- GALTIER, N., and M. GOUY. 1995. Inferring phylogenies from sequences of unequal base compositions. *Proc. Natl. Acad. Sci. USA* **92**:11317–11321.

- . 1998. Inferring pattern and process: maximum-likelihood implementation of a nonhomogeneous model of DNA sequence evolution for phylogenetic analysis. *Mol. Biol. Evol.* **15**:871–879.
- GAUT, B. S., and P. O. LEWIS. 1995. Success of maximum likelihood phylogeny inference in the four-taxon case. *Mol. Biol. Evol.* **12**:152–162.
- GOLDMAN, N., and Z. YANG. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* **11**:725–736.
- HASEGAWA, M., H. KISHINO, and T. YANO. 1985. Dating the human-ape splitting by a molecular clock of mitochondrial DNA. *J. Mol. Evol.* **22**:672–677.
- HILLIS, D. M. 1996. Inferring complex phylogenies. *Nature* **383**:130–131.
- . 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. *Syst. Biol.* **47**:3–8.
- HILLIS, D. M., and J. P. HUELSENBECK. 1992. Signal, noise, and reliability in molecular phylogenetic analyses. *J. Hered.* **83**:189–195.
- HILLIS, D. M., J. P. HUELSENBECK, and C. W. CUNNINGHAM. 1994. Application and accuracy of molecular phylogenies. *Science* **164**:671–677.
- HUELSENBECK, J. P. 1997. Is the Felsenstein zone a fly trap? *Syst. Biol.* **46**:69–74.
- . 1998. Systematic bias in phylogenetic analysis: is the Strepsiptera problem solved? *Syst. Biol.* **47**:519–537.
- HUELSENBECK, J. P., D. M. HILLIS, and R. JONES. 1996. Parametric bootstrapping in molecular phylogenetics: applications and performance. Pp. 19–45 in J. D. FERRARIS and S. R. PALUMBI, eds. *Molecular zoology*. Wiley and Sons, New York.
- JUKES, T. H., and C. R. CANTOR. 1969. Evolution of protein molecules. Pp. 21–132 in H. N. MUNRO, ed. *Mammalian protein metabolism*. Academic Press, New York.
- KHORANA, H. G. 1992. Rhodopsin, photoreceptor of the rod cell. *J. Biol. Chem.* **267**:1–4.
- 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 in Hominoidea. *J. Mol. Evol.* **29**:170–179.
- LAKE, J. A. 1994. Reconstructing evolutionary trees from DNA and protein sequences: paralogous distances. *Proc. Natl. Acad. Sci. USA* **91**:1455–1459.
- LARHAMMAR, D., and C. RISINGER. 1994. Molecular genetic aspects of tetraploidy in the common carp, *Cyprinus carpio*. *Mol. Phylogenet. Evol.* **1**:59–68.
- LEE, M. S. Y. 1999. Molecular phylogenies become functional. *Trends Ecol. Evol.* **14**:177–178.
- 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.
- MUSE, S. V. 1996. Estimating synonymous and nonsynonymous substitution rates. *Mol. Biol. Evol.* **13**:105–114.
- NATHANS, J. 1992. Rhodopsin: structure, function, and genetics. *Biochemistry* **31**:4923–4931.
- NAYLOR, G. J. P., and W. M. BROWN. 1997. Structural biology and phylogenetic estimation. *Nature* **388**:527–528.
- . 1998. Amphioxus mitochondrial DNA, chordate phylogeny, and the limits of inference based on comparisons of sequences. *Syst. Biol.* **47**:61–76.
- POE, S. 1998. The effect of taxonomic sampling on accuracy of phylogeny estimation: test case of a known phylogeny. *Mol. Biol. Evol.* **15**:1086–1090.
- PRAGER, E. M., and A. C. WILSON. 1988. Ancient origin of lactalbumin from lysozyme: analysis of DNA and amino acid sequences. *J. Mol. Evol.* **27**:326–335.
- RANNALA, B., J. P. HUELSENBECK, Z. YANG, and R. NIELSEN. 1998. Taxon sampling and the accuracy of large phylogenies. *Syst. Biol.* **47**:702–710.
- SACCONE, C., G. PESOLE, and G. PREPARATA. 1989. DNA microenvironments and the molecular clock. *J. Mol. Evol.* **29**:407–411.
- SAKMAR, T. P. 1998. Rhodopsin: a prototypical G protein-coupled receptor. *Prog. Nucleic Acid Res. Mol. Biol.* **59**:1–34.
- SHIELDS, D. C., P. M. SHARP, D. G. HIGGINS, and F. WRIGHT. 1988. “Silent” sites in *Drosophila* genes are not neutral: evidence of selection among synonymous codons. *Mol. Biol. Evol.* **5**:704–716.
- SIDOW, A., and A. C. WILSON. 1990. Compositional statistics: an improvement of evolutionary parsimony and its deep branches in the tree of life. *J. Mol. Evol.* **31**:51–68.
- SOGIN, M. L., G. HINKLE, and D. D. LELPE. 1993. Universal tree of life. *Nature* **362**:795.
- STEEL, M. 1994. Recovering a tree from the Markov leaf colourations it generates under a Markov model. *Appl. Math. Lett.* **7**:19–23.
- SULLIVAN, J., and D. L. SWOFFORD. 1997. Are guinea pigs rodents? The importance of adequate models in molecular phylogenetics. *J. Mamm. Evol.* **4**:77–86.
- SWOFFORD, D. L. 1999. PAUP\*, phylogenetic analysis using parsimony (\*and other methods). Version 4.0. Sinauer, Sunderland, Mass.
- TAKEZAKI, N., and T. GOJOBORI. 1999. Correct and incorrect vertebrate phylogenies obtained by the entire mitochondrial DNA sequences. *Mol. Biol. Evol.* **16**:590–601.
- TEMPLETON, A. R. 1983. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the humans and apes. *Evolution* **37**:221–244.
- TOWNSON, S. M., B. S. W. CHANG, E. SALCEDO, L. CHADWELL, N. E. PIERCE, and S. G. BRITT. 1998. Isolation and physiological characterization of the genes encoding the blue and ultraviolet sensitive opsins of the honeybee, *Apis mellifera*. *J. Neurosci.* **18**:2412–2422.
- VAN DEN BUSSCHE, R. A., R. J. BAKER, J. P. HUELSENBECK, and D. M. HILLIS. 1998. Base compositional bias and phylogenetic analyses: a test of the “flying DNA” hypothesis. *Mol. Phylogenet. Evol.* **10**:408–416.
- WRIGHT, F. 1990. The ‘effective number of codons’ used in a gene. *Gene* **87**:23–29.
- YANG, Z. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. *Mol. Biol. Evol.* **10**:1396–1401.
- . 1994. Estimating the pattern of nucleotide substitution. *J. Mol. Evol.* **39**:105–111.
- . 1996. Phylogenetic analysis using parsimony and likelihood methods. *J. Mol. Evol.* **42**:294–307.
- . 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput. Appl. Biosci.* **13**:555–556.
- YANG, Z., N. GOLDMAN, and A. FRIDAY. 1994. Comparison of models for nucleotide substitution used in maximum-likelihood phylogenetic estimation. *Mol. Biol. Evol.* **11**:316–324.

RODNEY HONEYCUTT, reviewing editor

Accepted April 14, 2000