18S Ribosomal RNA and Tetrapod Phylogeny

XUHUA XIA,[1] ZHENG XIE,[1,2] AND KARL M. KJER[3]

[1]*Department of Biology, University of Ottawa, 150 Louis, P.O. Box 450, Station A, Ottawa, Ontario K1N 6N5, Canada;*
E-mail: xxia@uottawa.ca (X.X.)
[2]*Institute of Environmental Protection, Hunan University, Changsha, China*
[3]*Department of Entomology, Cook College, Rutgers University, Highland Park, New Jersey 08904, USA; E-mail: kjer@aesop.rutgers.edu*

Abstract.— Previous phylogenetic analyses of tetrapod 18S ribosomal RNA (rRNA) sequences support the grouping of birds with mammals, whereas other molecular data, and morphological and paleontological data favor the grouping of birds with crocodiles. The 18S rRNA gene has consequently been considered odd, serving as "definitive evidence of different genes providing significantly different estimates of phylogeny in higher organisms" (p. 156; Huelsenbeck et al., 1996, Trends Ecol. Evol. 11:152–158). Our research indicates that the previous discrepancy of phylogenetic results between the 18S rRNA gene and other genes is caused mainly by (1) the misalignment of the sequences, (2) the inappropriate use of the frequency parameters, and (3) poor sequence quality. When the sequences are aligned with the aide of the secondary structure of the 18S rRNA molecule and when the frequency parameters are estimated either from all sites or from the variable domains where substitutions have occurred, the 18S rRNA sequences no longer support the grouping of the avian species with the mammalian species. [alignment; 18S rRNA; RNA secondary structure; Indel; molecular phylogenetics; tetrapod phylogeny.]

One of the early controversies in the phylogenetic relationship among tetrapods is whether birds are more closely related to crocodilians (Romer, 1966; Carroll, 1988; Gauthier et al., 1988) or to mammals (Gardiner, 1982; Løvtrup, 1985). Hedges et al. (1990) collected a set of 18S ribosomal RNA (rRNA) sequences to evaluate the relationships among tetrapods and found that the bird–mammal grouping was much more strongly supported, with a bootstrap value of 88%, than the bird–crocodilian grouping. A subset of these 18S rRNA sequences was used subsequently in a statistical test, based on the minimum-evolution criterion, to evaluate relative support of these alternative phylogenetic hypotheses (Rzhetsky and Nei, 1992). The nine shortest trees, including the neighbor-joining tree, all grouped the avian and mammalian species together as a monophyletic taxon.

The bird–mammal grouping contradicts both the traditional classification and the results derived from a large amount of other molecular data (Hedges, 1994; Seutin et al., 1994; Caspers et al., 1996; Janke and Arnason, 1997; Zardoya and Meyer, 1998; Ausio et al., 1999) and morphological and paleontological data (Eernisse and Kluge, 1993). For this reason, the proposal of the bird–mammal grouping based on the 18S rRNA gene has received critical examination from many different perspectives to determine what bias could have been introduced in analyzing the 18S rRNA sequences among tetrapod species.

Four kinds of potential bias involving the 18S rRNA gene have been proposed. First, the genomes of homeotherms such as birds and mammals tend to be more GC rich than those of poikilotherms (Bernardi, 1993). This shift in nucleotide frequencies, i.e., the problem of nonstationarity in the substitution process, is generally not accommodated in either parsimony or maximum-likelihood phylogenetic methods. For this reason, the LogDet (Lockhart et al., 1994) distance, which is based on a substitution model that presumably should correct for the nucleotide frequency shift, was used to see whether the annoying bird–mammal grouping would

disappear (Huelsenbeck and Bull, 1996; Huelsenbeck et al., 1996). It did not.

Second, the substitution pattern is biased favoring $U \leftrightarrow C$ transitions in rRNA genes (Marshall, 1992). This bias is expected and is mainly caused by the fact that the nucleotide G can pair with either U or C in maintaining the secondary structure of the rRNA molecule. For example, in pairing with a G, a C can be replaced by a U with little effect on the secondary structure. This ease of substitution is partly the reason for the model (Tamura and Nei, 1993) that uses one parameter for the $T \leftrightarrow C$ transition and another for the $A \leftrightarrow G$ transition. However, the substitution pattern is not unique in the 18S rRNA gene and therefore cannot explain the difference in phylogenetic outcome between the 18S rRNA and the rest of rRNA molecules. The use of the most general distance such as the LogDet still does not break up the bird–mammal grouping when applied to the 18S rRNA sequences (Huelsenbeck and Bull, 1996; Huelsenbeck et al., 1996).

The combination of a higher GC content in homeotherms and the biased substitution pattern favoring $U \leftrightarrow C$ transitions may jointly increase the problems associated with long-branch attraction (Marshall, 1992; Huelsenbeck et al., 1996). If $U \leftrightarrow C$ transition is the predominant substitution type and if birds and mammals have experienced an increase in C, then convergent $U \rightarrow C$ transitions may occur independently in the lineages leading to birds and to mammals. However, Hedges and Maxon (1992) dismissed long-branch attraction because the 18S rRNA sequences do not seem to have experienced substitutional saturation.

A weighted parsimony method (Williams and Fitch, 1990; Fitch et al., 1995) produced equivocal results (Marshall, 1992). When the paleontological tree grouping birds with crocodilians was used as the starting tree, the method ended up supporting this starting tree. However, when the tree grouping birds with mammals was used as the starting tree, the method ended up supporting this new starting tree. The weighted parsimony method is

known to depend on the starting tree, and its relevance to the 18S rRNA sequences is not obvious (Hedges, 1992).

Third, sequence misalignment was suspected to have resulted in a biased phylogenetic estimate from the 18S rRNA sequences (Eernisse and Kluge, 1993). However, a realignment of the sequences that was not based on secondary structure (Eernisse and Kluge, 1993) generated results that are exactly the same as those of earlier studies (Hedges et al., 1990; Rzhetsky and Nei, 1992). As previously argued (Hedges et al., 1990), the sequence divergence between the amphibians and the amniotes is only 4.4%; thus, sequence alignment should not be a problem.

Fourth, negligence in accommodating variable substitution rates between the conserved and the variable domains in the 18S rRNA sequences might cause a problem (Van de Peer et al., 1993). Both the indel and nucleotide substitution events have occurred predominantly in the eight variable domains (Van de Peer et al., 1993). If some sequences have experienced a number of deletions at their variable domains and if the genetic distance between the two sequences is calculated by using all homologous sites between the two sequences, then the genetic distance involving the shortest sequence, i.e., the one with the shortest variable region, will be relatively underestimated (Van de Peer et al., 1993). Although this observation is insightful, it cannot explain the bird–mammal grouping in previous studies (Hedges et al., 1990; Rzhetsky and Nei, 1992) because in these studies all indel-containing sites were deleted before the phylogenetic analysis was performed so that the number of the homologous sites between any pair of sequences would be constant, i.e., all sequences have the same number of conserved and variable sites in these studies.

Huelsenbeck et al. (1996) made perhaps the most extensive and critical examination of the 18S rRNA sequences in relation to the tetrapod phylogeny. They tried almost all existing phylogenetic methods, such as distance methods with the LogDet distance and the maximum-likelihood method with several substitution models. However, the 18S rRNA sequences consistently produced the bird–mammal grouping, whereas other rRNA genes supported the bird–"reptile" grouping. This result led Huelsenbeck et al. (1996:156) to conclude that their analysis "offers definitive evidence of different genes providing significantly different estimates of phylogeny in higher organisms."

Is the 18S rRNA gene really so unique? This question is serious because the rRNA genes have been heralded as the universal yardstick in molecular phylogenetics (Olsen and Woese, 1993), and it would be truly frustrating if we often had "different genes providing significantly different estimates of phylogeny in higher organisms" (Huelsenbeck et al., 1996:156), especially when the universal yardstick is at fault.

In previous studies, including that of Eernisse and Kluge (1993), little attention was paid to two problems. The first problem is that of sequence alignment and the associated definition and treatment of alignment in ambiguous regions. The realignment of Eernisse and Kluge (1993) has not been published but appears to have been generated using a gap penalty slightly higher than that used by Hedges et al. (1990). No reference was made to the secondary structure of the rRNA molecule, which has been considered by some as essential for aligning rRNA sequences (Kjer, 1995; Notredame et al., 1997; Hickson et al., 2000).

The reported sequence divergence of 4.4% between amphibians and amniotes (Hedges and Maxson, 1992) might have misled researchers into thinking that few changes have occurred during the evolution of 18S rRNA and that there is consequently little ambiguity in sequence alignment. In fact, many indel events have occurred, and it is extremely difficult to arrive at a definite alignment, even with the information provided by secondary structure. Hedges et al. (1990) excluded a segment of sequences because of the difficulty in aligning them. The reported 4.4% sequence divergence was obtained after deleting all indel-containing sites and is therefore not a reflection of sequence differences. With sequences of such low divergence, these hypervariable regions are important because the majority of characters may come from regions of ambiguous homology. In addition, even if 4.4% were an accurate estimate of pairwise divergence, Kjer (1995) cautioned against making general statements about divergence levels below which structural alignments would be unimportant because of the high variability of conservation among stems.

The 18S rRNA sequences in mammalian species are much longer than those in avian and "reptilian" species. The alignment program therefore has more room to slide the "reptilian" bases to match the mammalian sequences during sequence alignment. The similarity in nucleotide frequencies between birds and mammals (Bernardi, 1993) would increase the chance of spurious matching between the avian and mammalian sequences. For illustration, imagine four orthologous sequences, two having experienced no indel events and two having experienced many indel events (Fig. 1). The alignment of many sites, especially at sites 3 and 4 and 10 and 11,

```
               10        20        30        40
       ----|----|----|----|----|----|----|----|
Seq1   A-CC-----GC--------GA--CUUC--GA-CC-CG--G
Seq2   A-CC-----GU--------GA--CUUG--GA-CC-CG--G
Seq3   AACCCCGGUGUAGGGGGAAGAACCUUGAUGAACCUCGAUG
Seq4   AACCCCGGUGCAGGGGGAAGAACCUUCAUGAACCUCGAUG
```

FIGURE 1.   The problem of aligning short and long sequences.

is uncertain. Assuming that indel events are generally rare, we conclude that Seq1 and Seq2 are more similar to each other, as are Seq3 and Seq4. However, if we delete all indels, then the uncertainty in sequence alignment is forgotten, and all existing phylogenetic methods will generate the "best" tree, with Seq2 and Seq3 forming a monophyletic taxon. Previous studies that support the grouping of birds with mammals (Hedges et al., 1990; Rzhetsky and Nei, 1992; Huelsenbeck et al., 1996) happen to have aligned long mammalian sequences with short avian and "reptilian" sequences and then deleted all indels before phylogenetic analysis.

The information in the secondary structure of rRNA sequences has been recognized as helpful in guiding sequence alignment (Kjer, 1995; Hickson et al., 1996; Notredame et al., 1997; Buckley et al., 2000), and structure-based alignment has improved phylogenetic resolution in many studies (Dixon and Hillis, 1993; Kjer, 1995; Titus and Frost, 1996; Morrison and Ellis, 1997; Uchida et al., 1998; Mugridge et al., 1999; Cunningham et al., 2000; Gonzalez and Labarere, 2000; Hwang and Kim, 2000; Lydeard et al., 2000; Morin, 2000; Xia, 2000b). Structural alignments are performed manually and thus require the investigator to look at the data and make decisions, thus preventing some of the arbitrary statements about homology that are illustrated in Figure 1.

For the 18S rRNA sequences from the tetrapod species, even the inclusion of secondary structure information cannot guarantee unequivocal alignment of all homologous sites. Although many authors have proposed methods for handling indels (Swofford, 1993; Baldwin et al., 1995; Hibbett et al., 1995; Kjer, 1995; Crandall and Fitzpatrick, 1996; Kretzer et al., 1996; Manos, 1997; Flores-Villela et al., 2000; Kjer et al., 2001), ambiguously aligned sites have sometimes been handled in phylogenetics in two extreme and inappropriate ways, i.e., they are either totally discarded or totally included with the hope that the majority of sites have been aligned properly. Lutzoni et al. (2000) described a method for coding these ambiguously aligned regions, but their method has not yet been used with the 18S rRNA sequences for tetrapods.

The second problem shared by previous studies is the misuse of frequency parameters in the distance and maximum-likelihood methods. The vast majority of both substitution and indel events have occurred in just a few variable domains of the 18S rRNA sequences (Van de Peer et al., 1993). The variable domains have nucleotide frequencies different from those of the conserved domains in the 18S rRNA gene and the 28S rRNA gene (Zardoya and Meyer, 1996). In phylogenetic analyses involving the distance and maximum-likelihood methods, the frequency parameters most appropriate for the underlying substitution model must be used. The most appropriate estimate of the frequency parameters should be derived from the sites where substitution occurs, i.e., from the variable domains. However, in previous studies, many sites in the variable domains have been deleted after alignment because they contain indels. Consequently, the frequency parameters in those studies have been estimated mainly from the conserved domains

where nucleotide substitutions are rare. Such frequency parameters could be irrelevant to the underlying substitution models used.

In this study, we reexamined phylogenetic relationship among tetrapods by (1) using the 18S rRNA sequences aligned against the secondary structure and (2) estimating the frequency parameters from all sites (i.e., including indel-containing sites) or from variable sites only. In contrast to previous studies based on the 18S rRNA sequences, our reanalysis does not support the hypothesis that birds group with mammals.

MATERIALS AND METHODS

We used three sets of 18S rRNA sequences in this study. The first set of sequences was retrieved from the rRNA WWW server (Van de Peer et al., 2000; http://rrna.uia.ac.be/ssu/) and consists of 48 sequences, after excluding redundant sequences. The sequence files from the rRNA WWW server are plain-text files with a special distribution format to specify secondary structure information. The format uses square brackets to enclose a helix, parentheses to enclose a nonstandard base pair, and braces to enclose an internal loop. The computer software DAMBE (Xia, 2000a; Xia and Xie, 2001) can read the files and interpret the symbols properly. The alignment was refined by visual inspection against the secondary structure, and the final aligned sequences and the second and third sets of data are available at http://aix1.uottawa.ca/~xxia/research/data/XiaXieKjer.htm. The refinement of the alignment was required because, as with any huge electronic database of rRNA expanding at a rapid rate, some of the sequences we downloaded had been taxonomically aligned.

The second set of aligned 18S rRNA sequences was retrieved from the Ribosomal Database Project II (Maidak et al., 2000; ftp://ftp.cme.msu.edu/pub/RDP/SSU_rRNA/alignments/). This FTP site contains two relevant files (SSU_Euk.gb and SSU_Euk_rep.gb). All tetrapod sequences with >1200 resolved bases were included in this study, and the set consists of 15 aligned sequences. Most of these sequences have been used in previous studies to generate the best tree supporting the bird–mammal grouping (Hedges et al., 1990; Rzhetsky and Nei, 1992; Eernisse and Kluge, 1993; Huelsenbeck et al., 1996). We refined the alignment by visual inspection, and the final alignment is also available in the URL above.

The third sets of sequences were retrieved from GenBank, with four sequences not contained in the previous two sets: *Crocodylus niloticus* (crocodile), *Ornithorhynchus anatinus* (platypus), *Vombatus ursinus* (wombat), and *Didelphis virginiana* (opossum). The platypus and the two marsupial species help subdivide the branch leading to the placental mammal clade. The sequences are also aligned against the secondary structure, and the final alignment is also available at the URL above. Alignment-ambiguous regions were defined with reference to secondary structure (Kjer, 1997). We examined whether there could be something peculiar about the

sequences collected by Hedges et al. (1990) that could cause the avian and mammalian sequences to group together. We sequenced most of the 18S rRNA gene from the turtle *Trachemys scripta* and combined these new sequences with those of *Pseudemys scripta* from Hedges et al. (1990), forming a chimeric sequence. Other sequences from Hedges et al. (1990) were replaced with taxa collected by others, except for the lizard and snake sequences, which are the only representatives of their taxa available. Of the sequences that were retained from Hedges et al. (1990), we recorded as missing any specific site at which one nucleotide was recorded from all of the amniotes by Hedges et al. (1990) and another nucleotide was recorded from all other taxa by other researchers. These sequences were analyzed using parsimony methods, with a new method for coding alignment-ambiguous regions (Lutzoni et al., 2000), and analyzed using likelihood methods, with a general time-reversible model with a gamma correction for among-site rate variation and an estimate of invariant sites. Parameters were estimated from the parsimony tree with PAUP 4 (Swofford, 2000). The three sets of sequences are not mutually exclusive.

Strong heterogeneity in substitution rate is expected to exist in the 18S rRNA sequences, and we used the method of Gu and Zhang (1997) implemented in DAMBE (Xia, 2000a; Xia and Xie, 2001) to estimate the alpha parameter of the gamma distribution. We also estimated the proportion of invariant sites using a modification of Gu and Zhang's method as follows. The estimated alpha value is used with the new version of DNAML (Felsenstein, 1993) and for correcting distance estimation based on the TN93 model (Tamura and Nei, 1993).

The genomes of homeotherms such as birds and mammals tend to be more GC rich than those of poikilotherms (Bernardi, 1993), and the 18S rRNA sequences from the avian and mammalian species are also more GC rich than those of other species. If the high GC content in avian and mammalian sequences has been gained independently in the two lineages, then we should use a substitution model that would accommodate nonstationarity in the substitution process. At present, only the underlying model for the LogDet (Lockhart et al., 1994) and the paralinear (Lake, 1994) distances accommodate nonstationarity, and thus much of our phylogenetic analysis was limited to distance-based methods.

The presence of a large proportion of invariant sites may bias phylogenetic estimation (Lockhart et al., 1996). It is important to distinguish between the invariant sites and those sites with no observed substitution, the former being a subset of the latter. The proportion of sites with no observed substitution (designated p) is made of two components: the proportion of sites expected to have experienced no substitution (p_1) under certain substitution model and the truly invariant sites, i.e., those where a change will have a very deleterious effect and will be strongly selected against (p_2). To estimate p_2, we allowed the p_2 value to fluctuate between 0 and p (and the p_1 value consequently fluctuated from p to 0) and fit the observed substitution data to a negative binomial dis-

tribution. The resulting p_2 value that produced the best fit to the substitution data was used as the proportion of invariant sites. This method has been implemented in DAMBE (Xia, 2000a; Xia and Xie, 2001). A similar approach was used by Lockhart et al. (1996).

Unless specified otherwise, the nucleotide frequencies were estimated using all sites, including the indel-containing sites. This approach differs from that used in previous studies (Hedges et al., 1990; Rzhetsky and Nei, 1992; Eernisse and Kluge, 1993; Huelsenbeck et al., 1996), in which all indel-containing sites were deleted before the phylogenetic analysis was performed.

RESULTS AND DISCUSSION

The mammalian and avian sequences are consistently more GC rich than the sequences from poikilotherms (Table 1), a finding with two implications. First, avian 18S rRNA sequences are much shorter than mammalian sequences, and alignment of the short avian sequences against the long mammalian sequences allows a great

TABLE 1.   Frequencies of A+T and G+C for the fish, amphibian, "reptilian," mammalian, and avian species.

Taxon	Scientific name	Accesion no.	Frequencies	
			A+T	C+G
Fish	*Latimeria chalumnae*	L11288	0.4766	0.5234
Amphibian	*Xenopus laevis*	X02995	0.4619	0.5380
	Xenopus laevis	X04025	0.4622	0.5378
	Ranodon sibiricus	AJ279506	0.4788	0.5212
Crocodilian	*Alligator mississippiensis*	AF173605	0.4634	0.5367
Tuatara	*Sphenodon punctatus*	AF115860	0.4602	0.5398
Mammal	*Mus musculus*	X00686	0.4398	0.5602
	Mus musculus	X82564	0.4396	0.5605
	Rattus norvegicus	K01593	0.4424	0.5576
	Rattus norvegicus	M11188	0.4422	0.5577
	Rattus norvegicus	V01270	0.4429	0.5571
	Oryctolagus cuniculus	X06778	0.4465	0.5534
	Homo sapiens	K03432	0.4393	0.5607
	Homo sapiens	M10098	0.4395	0.5605
	Homo sapiens	U13369	0.4388	0.5612
	Homo sapiens	X03205	0.4388	0.5612
Bird	*Anas platyrhynchos*	AF173614	0.4467	0.5533
	Dromaius novaehollandiae	AF173610	0.4537	0.5463
	Tockus nasutus	AF173626	0.4410	0.5590
	Chordeiles acutipennis	AF173622	0.4398	0.5602
	Charadrius semipalmatus	AF173638	0.4410	0.5591
	Larus glaucoides	AF173637	0.4381	0.5619
	Urocolius macrourus	AF173617	0.4386	0.5613
	Columba livia	AF173630	0.4404	0.5596
	Coracias caudata	AF173625	0.4404	0.5596
	Cuculus pallidus	AF173628	0.4410	0.5590
	Galbula pastazae	AF173624	0.4410	0.5590
	Ortalis guttata	AF173613	0.4399	0.5602
	Coturnix pectoralis	AF173611	0.4502	0.5498
	Gallus gallus	AF173612	0.4525	0.5475
	Grus canadensis	AF173632	0.4410	0.5590
	Gallirex porphyreolophus	AF173618	0.4393	0.5608
	Picoides pubescens	AF173615	0.4410	0.5590
	Tyrannus tyrannus	AF173616	0.4398	0.5602
	Ciconia nigra	AF173636	0.4398	0.5602
	Apus affinus	AF173619	0.4392	0.5607
	Trogon collaris	AF173623	0.4404	0.5596
	Turnix sylvatica	AF173631	0.4404	0.5596
	Upupa epops	AF173627	0.4404	0.5596
	Apteryx australis	AF173609	0.4537	0.5463

deal of freedom in sliding the avian bases to match the mammalian bases. This length mismatch and the similarity in nucleotide frequencies between avian and mammalian sequences lead to "optimal" alignment (i.e., the best alignment score); therefore, it is necessary to use a nucleotide substitution model that accommodates the inherent nonstationary substitution process. At present, only the paralinear distance (Lake, 1994) and the LogDet distance (Lockhart et al., 1994) methods are appropriate for the phylogenetic analysis of these sequences.

There is a subtle difference between the paralinear and the LogDet distances. To highlight the difference, we reproduced the distance between two nucleotide sequences (1 and 2):

$$d_{12} = -\frac{1}{4} \ln \left(\frac{\det J_{12}}{\sqrt{\prod_{i=1}^{4} p_{1i} \prod_{i=1}^{4} p_{2i}}} \right), \tag{1}$$

where J_{12} is the observed substitution matrix, p_1 and p_2 are nucleotide frequencies for sequences 1 and 2, respectively, and $\det J_{12}$ means the determinant of J_{12}. In the formulation of the paralinear distance, J_{12} are numbers and p_1 and p_2 are reconstituted from J_{12} (Lake, 1994). Consequently, p_1 and p_2 are based on aligned sites only, i.e., sites with no indels. However, this approach causes a new problem for analyzing the 18S rRNA sequences. Both substitution and indel events have occurred almost exclusively in just a few variable domains of the 18S rRNA sequences (Van de Peer et al., 1993). The variable domains have nucleotide frequencies different from those of the conserved domains in the 18S rRNA gene and in the 28S rRNA gene (Zardoya and Meyer, 1996). In phylogenetic analyses involving distance and maximum-likelihood methods, frequency parameters most appropriate for the underlying substitution model must be used. The most appropriate estimate of the frequency parameters should be from the sites where substitution occurs, i.e., from the variable domains. However, variable domains in the 18S rRNA sequences are poorly represented in the aligned sites because of the presence of many indels in these domains. Thus, p_1 and p_2 in Equation 1 are mainly based on invariable domains and consequently are not appropriate for phylogenetic reconstruction. PAUP 4 (Swofford, 2000) uses this original formulation for calculating the pairwise Lake/LogDet distances.

Two modifications can be made to alleviate the problem of using inappropriately estimated frequency parameters. The first is to use polymorphic sites only in phylogenetic reconstruction. This would produce proper estimates of p_1 and p_2 but has the disadvantage of generating extraordinarily large distances. An alternative is to use to use the LogDet distance (Lockhart et al., 1994), which defines J_{12} as a substitution matrix in proportions summing up to 1 and p_1 and p_2 as vectors of proportions summing up to 1. This permits the computation of empirical frequencies from all sites, including sites containing indels. Both DAMBE and the DNADIST program in PHYLIP (Felsenstein, 1993) use all sites in computing p_1 and p_2. This approach allows sites in the variable domains of the 18S rRNA sequences to be better represented in computing nucleotide frequencies and is the approach that we have taken in analyzing the 18S rRNA sequences.

Distance-based phylogenetic reconstruction demands both the unbiased estimation of the distance matrix and an efficient and accurate method that uses the input distance matrix to search for the best tree based on a biologically meaningful optimization criterion. The latter component has been much advanced in recent years, with the development of new methods implemented in Weighbor (Bruno et al., 2000), BIONJ (Gascuel, 1997), and FastME (Desper and Gascuel, 2002). In particular, FastME represents one of the first successful implementations of the global minimum evolution (ME) criterion in phylogenetic analysis. Previous implementations of the ME criterion, such as METREE (Rzhetsky and Nei, 1994) and FITCH in the PHYLIP package (Felsenstein, 1993), use the ordinary least-square method for evaluating branch lengths and do not handle the resulting negative branch lengths in a meaningful way. FastME is fast and achieves high topological accuracy by the combination of a very efficient branch-swapping algorithm and a fast tree-evaluating method equivalent to the weighted least-square method.

The tree produced by FastME with default options and with the LogDet distance for the 48 sequences in the first set (Fig. 2) revealed a group of odd-looking sequences: *Ambystoma mexicanum* (salamander; GenBank M59384), *Nesomantis thomasseti* (salamander; M59396), *Bufo valliceps* (frog; M59386), *Turdus migratorius* (bird; M59402), *Pseudemys scripta* (turtle; M59398), *Heterodon platyrhinos* (snake; M59392), and *Alligator mississippiensis* (M59383). These amphibian, "reptilian," and avian species have relatively long branches, do not cluster with their taxonomic sister taxa, and form a cluster among themselves. These sequences are all from the first study (Hedges et al., 1990) in which the avian and mammalian species formed a monophyletic group.

A close examination of these sequences shows that all have many unresolved sites, which suggests that the neighboring resolved sites in the sequences might also be unreliable. The long branches associated with these sequences may not mean that they all have extraordinarily rapid evolutionary rates but rather are more likely to be the result of sequencing errors. The grouping of these heterogeneous sequences together to the exclusion of their respective sister taxa cannot be satisfactorily explained without invoking sequencing errors. A site-by-site examination of the data confirms this explanation. Examination of this odd group of sequences suggests that the grouping of the avian and mammalian species by previous studies based on this group of sequences (Hedges et al., 1990; Rzhetsky and Nei, 1992; Eernisse and Kluge, 1993; Huelsenbeck and Bull, 1996) is at least partially attributable to sequencing error.

In subsequent analyses of the first set of sequences, we excluded these seven sequences and one of the two
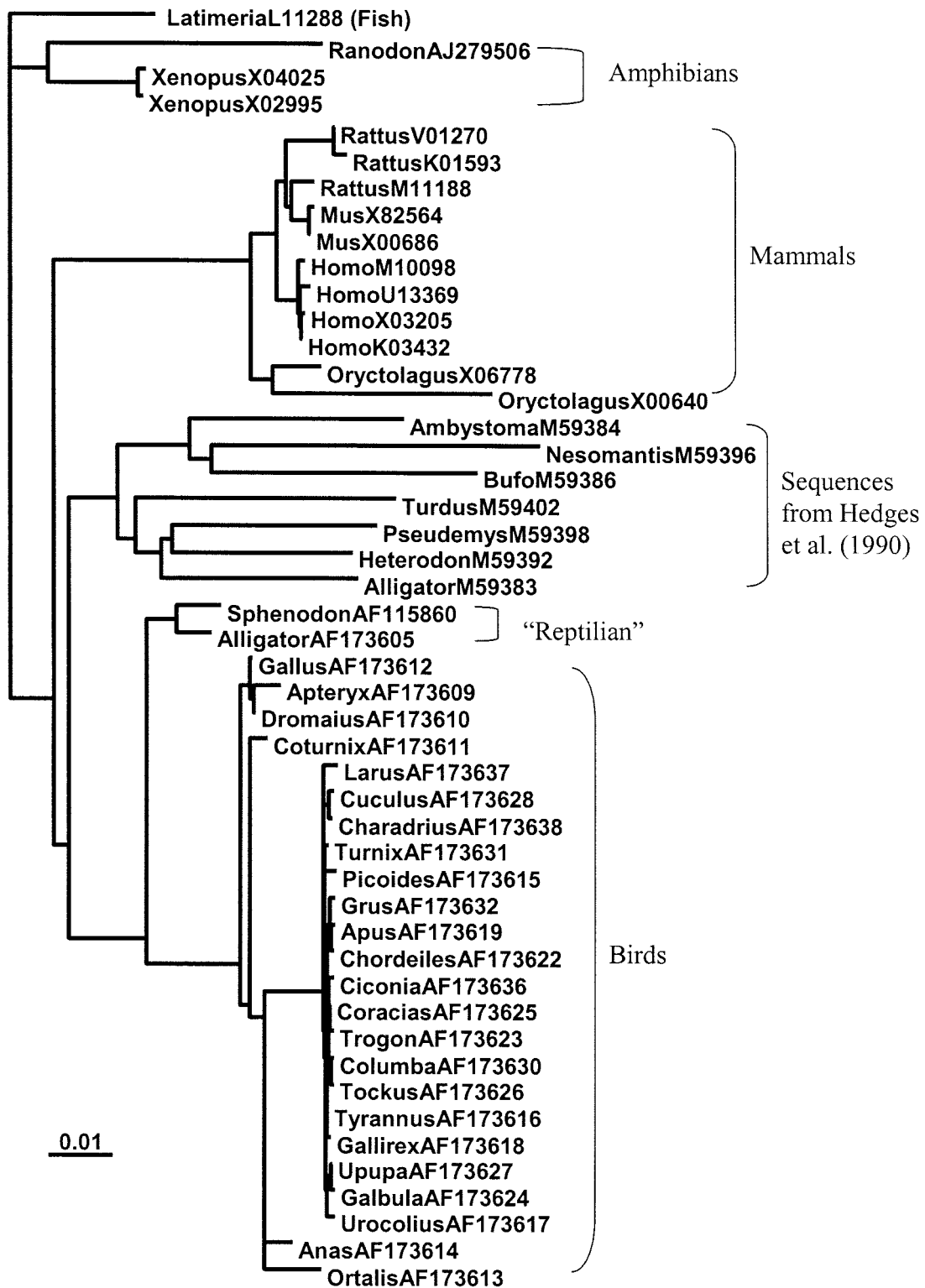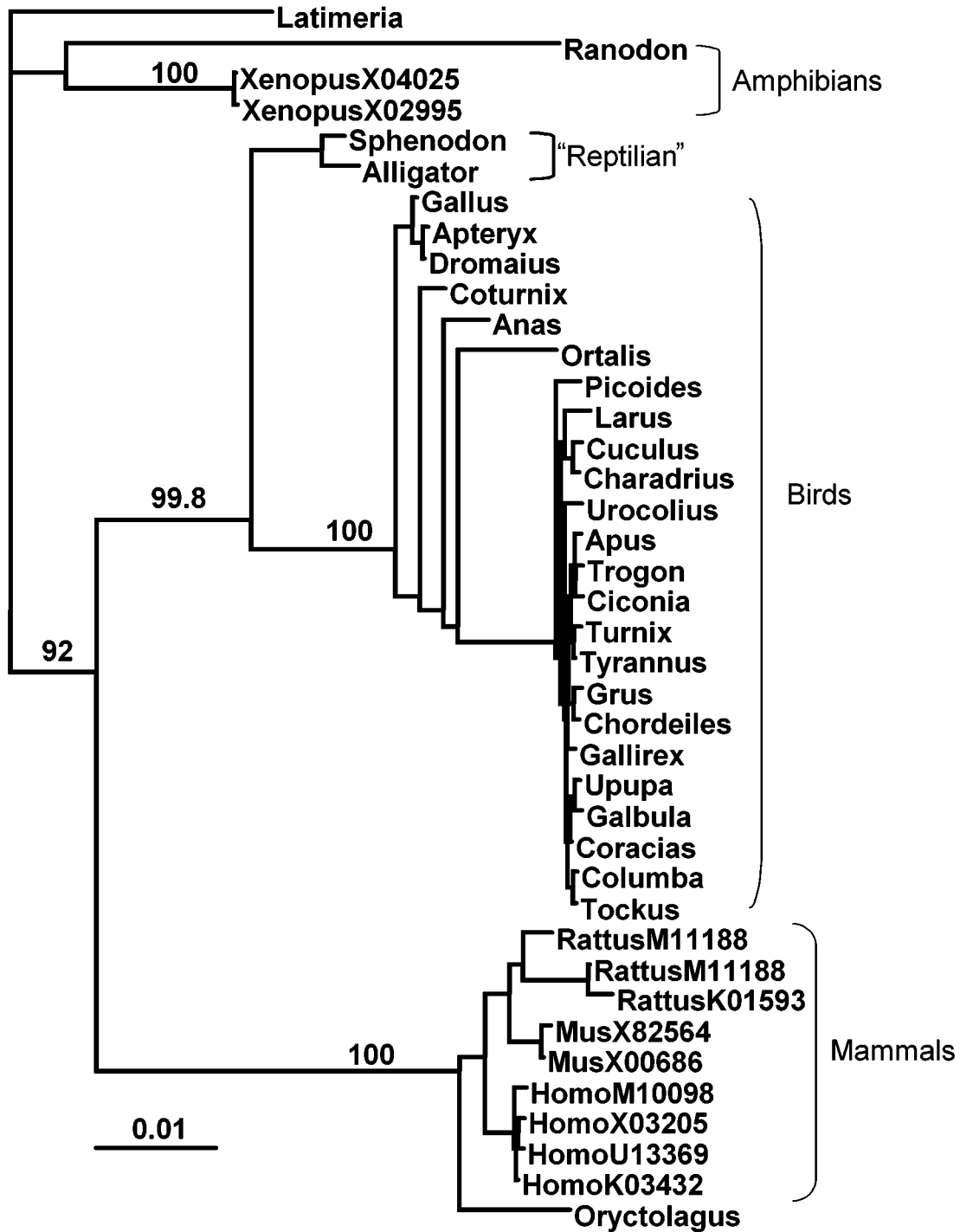
FIGURE 2. Phylogenetic tree obtained from the FastME method with LogDet distances. All sites were included in counting nucleotide frequencies for computing the LogDet distance.

Oryctolagus cuniculus sequences (rabbit; X00640). This sequence was obtained many years ago (Connaughton et al., 1984), and its suspiciously long branch (Fig. 2) suggests that it is unreliable. The phylogenetic tree for the remaining 40 sequences, based on the FastME method with the LogDet distances, clustered the avian and "reptilian" species to the exclusion of mammalian species (Fig. 3a). The bootstrap values from 500 resamples leave little ambiguity in such a grouping (Fig. 3a). The combination of the DNADIST (producing a matrix of LogDet distance matrix) and NEIGHBOR programs in PHYLIP (Felsenstein, 1993) also groups the avian and the

(a)

FIGURE 3. Phylogenetic tree obtained from the FastME method (a) and the Fitch–Margoliash method (b) with the LogDet distances. Sequences of poor quality have been removed. The numbers are bootstrap values. (*Continued on next page*)
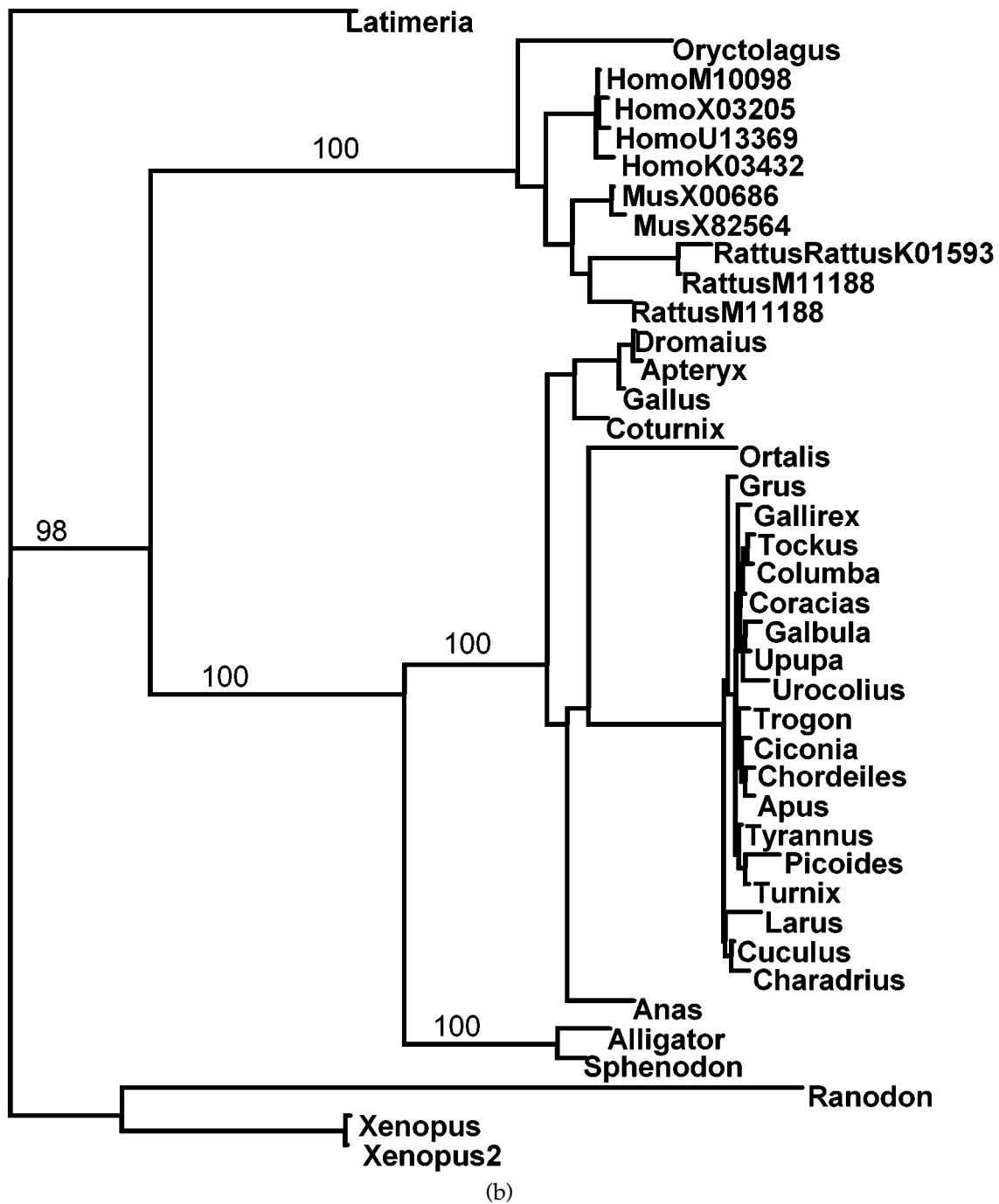
(b)

FIGURE 3. Continued.

"reptilian" species together to the exclusion of mammalian species. The reconstructed tree from Weighbor (Bruno et al., 2000) is similar to that of the neighbor-joining (NJ) method, and both trees share the annoying outcome of grouping one of the three rat sequences with the mouse sequences.

The phylogenetic tree based on the Fitch-Margoliash (FM) method (Fitch and Margoliash, 1967), implemented in the FITCH program of the PHYLIP package (Felsenstein, 1993) and in DAMBE (Xia, 2000a; Xia and Xie, 2001), is similar to the FastME tree in that avian and "reptilian" species are clustered together with high bootstrapping values (Fig. 3b). Although the FM method has a global optimization criterion whereas the NJ method achieves only local optimization, this advantage of the FM method over the NJ method is typically lost in practical computation. The FM method is slow, and current implementations of the method, such as those in PHYLIP and DAMBE, adopt a greedy algorithm by starting the tree reconstruction with three operational taxonomic

units (OTUs) and then add new OTUs to the growing tree sequentially. Thus, the so-called global optimization is only applied to successive local trees, and it is misleading to call this approach global optimization. In contrast, FastME explores the tree space much more thoroughly than does the FM method implemented in FITCH and DAMBE.

There is a small difference in the implementation of the FM method between DAMBE and PHYLIP. DAMBE starts with three OTUs that have the greatest average distance from the other OTUs and then adds other OTUs sequentially to the tree. Once all the trees have been added, the first three OTUs are then taken off and replanted. This process should produce a tree that is better than that produced by the FITCH program using its default mode but may not be as good as that produced by the FITCH program when all optimization switches are turned on.

The second set of 18S rRNA sequences is mostly composed of sequences used originally to produce the bird–mammal grouping with high bootstrap values (Hedges et al., 1990; Rzhetsky and Nei, 1992; Eernisse and Kluge, 1993; Huelsenbeck et al., 1996). However, when the sequences are aligned according to secondary structure and all sites are used for counting nucleotide frequencies in computing LogDet distances, the avian and "reptilian" sequences form a monophyletic group with unambiguous bootstrapping support (Fig. 4). Phylogenetic reconstruction with the FM method produced identical topology and almost identical bootstrapping values. Thus, the bird–mammal grouping is still not recovered with proper

phylogenetic methods even when the sequence quality is low.

Previous studies with roughly the same set of sequences grouped avian and mammalian species together with the LogDet distances (Huelsenbeck and Bull, 1996; Huelsenbeck et al., 1996). There are several possibilities for the discrepancy. First, the sequences used in previous studies may have been aligned differently. Second, the previous studies may have included the LogDet distances specified in equation 1 of Lockhart et al. (1994) instead of those of equation 3 of Lockhart et al. (1994). The latter is identical to our Equation 1 and the paralinear distance (Lake, 1994) in form, but that defined in equation 1 of Lockhart et al. (1994) is different and equals $-\ln(\det J_{12})$. Third, the previous studies may have also included the LogDet distances as defined in our Equation 1 but may have done the calculations based on sites containing no indels. This would imply that p_1 and p_2 in our Equation 1 were dominated by sites in the conserved domains of the 18S rRNA sequences and consequently may not be appropriate in characterizing a substitution pattern involving substitutions mostly in sites of the variable domains.

The inclusion of the indel-containing sites in our calculation of the LogDet distances also suffers from a possible bias. Both the indel events and the nucleotide substitution events occurred mostly in the variable domains of rRNA sequences (Van de Peer et al., 1993). If some sequences have experienced a number of deletions at their variable domains and if the genetic distance between the two sequences is calculated using all sites
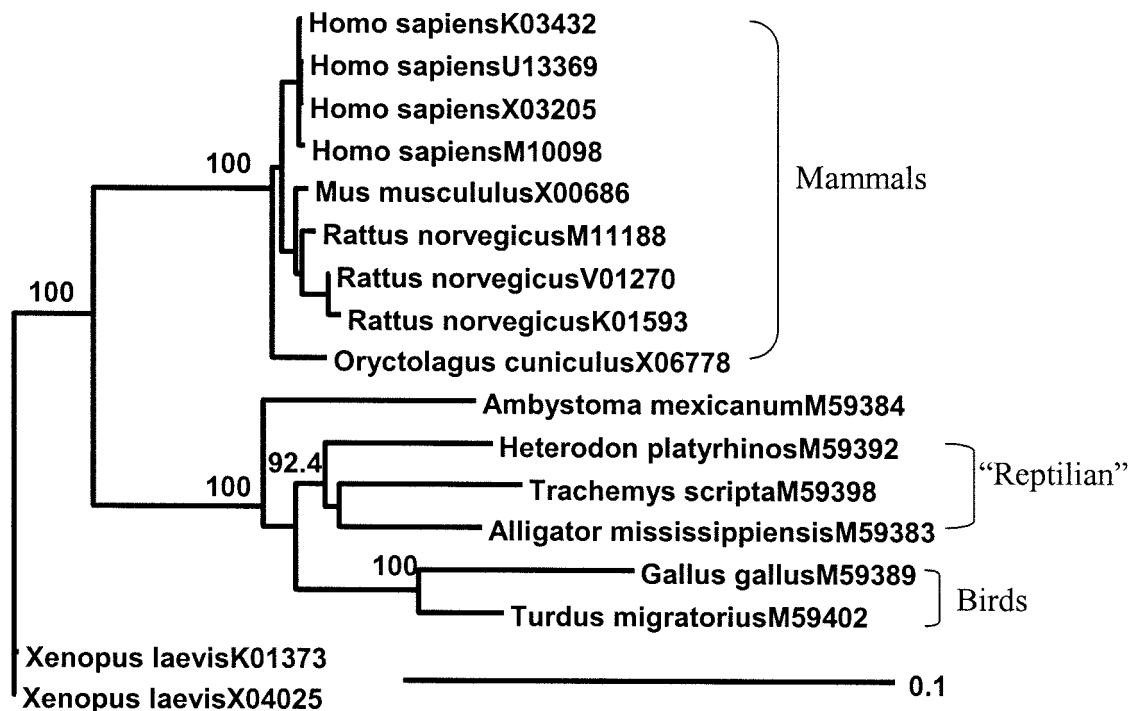

FIGURE 4. Phylogenetic tree obtained from the FastME method with LogDet distances and the second set of sequences used previously to support the bird–mammal grouping. The numbers are bootstrap values. The Fitch–Margoliash method produces the same topology and almost identical bootstrap values.
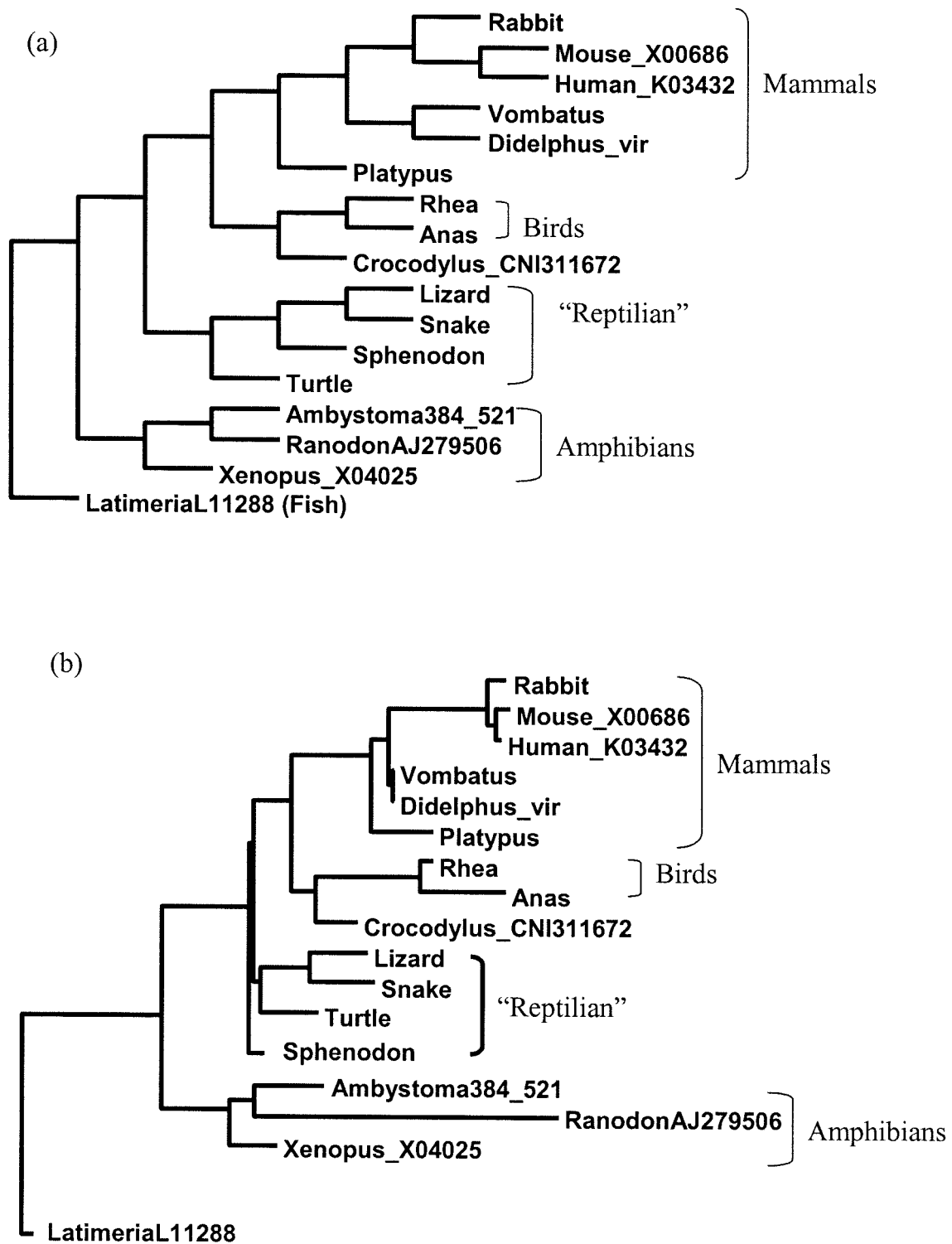
FIGURE 5. Maximum parsimony (μP) (a) and maximum likelihood (b) trees based on the third set of sequences, including the turtle and the more primitive mammalian species. The branch lengths of the MP tree are not estimated and are set to the same length for display.

between the two sequences (as we have done), then the genetic distance involving the shortest sequence, i.e., the one with the shortest variable region, will be relatively underestimated (Van de Peer et al., 1993). The avian and "reptilian" 18S rRNA sequences are shorter than those of mammalian species. If avian and "reptilian" sequences share a number of independent deletions of homologous variable domains, then our calculation of the LogDet distances would tend to underestimate the genetic distances involving the avian and the "reptilian" sequences. This problem is shared by results from both the first and the second data sets.

For the third set of data with more primitive mammalian lineages, the phylogenetic result supports the avian–crocodilian grouping (Fig. 5) in both parsimony and likelihood analyses. This set of sequences was aligned independently from the other two sets of sequences, and none of these three independently aligned sequences support the bird–mammal grouping. This leaves little doubt that the 18S rRNA gene is not as odd as previous studies have suggested (Hedges et al., 1990; Rzhetsky and Nei, 1992; Eernisse and Kluge, 1993; Huelsenbeck et al., 1996). In particular, this last set of data does not have the potential bias outlined in the previous paragraph involving our calculation of the LogDet distances.

Although it appears premature to conclude that the 18S rRNA sequences supply "definitive evidence of different genes providing significantly different estimates of phylogeny in higher organisms" (Huelsenbeck et al., 1996:156), it is important to properly choose the substitution model and phylogenetic methods. When we delete all indel-containing sites in the first and the second data sets so that the nucleotide frequencies are dominated by the invariable sites, then all major distance methods (NJ, FM, Weighbor, FastME) with any of the genetic distances (including LogDet and paralinear distances) group the avian and mammalian sequences as a monophyletic group, just as shown in previous studies. PAUP 4 (Swofford, 2000) ignores all indel-containing sites in calculating the pairwise Lake/LogDet distances, and the distance-based tree-making methods implemented in PAUP will always group the mammalian and avian species together to the exclusion of the "reptilian" species.

Similarly, when we apply to the first and the second data sets any existing maximum likelihood, maximum parsimony, or any distance-based method that does not accommodate the nonstationary nature of the substitution process involved, we again have avian sequences strongly grouped with the mammalian sequences to the exclusion of "reptilian" species (data not shown).

The sequences exhibit strong heterogeneity in substitution rate over sites, with estimated alpha values of 0.1643 and 0.1432 for the first and the second data sets, respectively. We have used the maximum-likelihood method with gamma-distributed rates by using the new version of DNAML and BASEML, and the resulting trees always grouped the birds and mammals together. When we do not use the LogDet distances but instead use

the genetic distance based on the three-parameter TN93 model (Tamura and Nei, 1993) or any other substitution model, the resulting trees also group the avian and mammalian species together, regardless of whether the distance is corrected with the estimated alpha value or not. This result highlights the importance of accommodating nonstationarity in the substitution process.

Structurally aligned 18S rRNA sequences from major tetrapod taxa produce topologies similar to those based on other genes, morphological characters, and paleontological evidence. The rRNA sequences must be aligned using the secondary structure as a template, and frequency parameters appropriate for the underlying substitution model must be used. Secondary structure information should also be used to determine the boundaries between aligned and alignment-ambiguous regions (Kjer, 1997) so that these regions can be objectively examined according to the coding method (Lutzoni et al., 2000). This study highlights the problem of applying a battery of computer programs to the data without first checking the quality of the data and emphasizes the importance of becoming intimately familiar with the data. Many of these conclusions could not have been made without looking at the data.

REFERENCES

AUSIO, J., J. T. SOLEY, W. BURGER, J. D. LEWIS, D. BARREDA, AND K. M. CHENG. 1999. The histidine-rich protamine from ostrich and tinamou sperm. A link between reptile and bird protamines. Biochemistry (Moscow) 38:180–184.

BALDWIN, B. G., M. J. SANDERSON, J. M. PORTER, M. F. WOJCIECHOWSKI, C. C. CAMPBELL, AND M. J. DONOGHUE. 1995. The ITS region of nuclear ribosomal DNA: A valuable source of evidence on angiosperm phylogeny. Ann. Mo. Bot. Gard. 82:257–277.

BERNARDI, G. 1993. The vertebrate genome: Isochores and evolution. Mol. Biol. Evol. 10:186–204.

BRUNO, W. J., N. D. SOCCI, AND A. L. HALPERN. 2000. Weighted neighbor joining: A likelihood-based approach to distance-based phylogeny reconstruction. Mol. Biol. Evol. 17:189–197.

BUCKLEY, T. R., C. SIMON, P. K. FLOOK, AND B. MISOF. 2000. Secondary structure and conserved motifs of the frequently sequenced domains IV and V of the insect mitochondrial large subunit rRNA gene. Insect Mol. Biol. 9:565–580.

CARROLL, R. L. 1988. Vertebrate paleontology and evolution. W. H. Freeman, New York.

CASPERS, G. J., G. J. REINDERS, J. A. M. LEUNISSEN, J. WATTEL, AND W. W. DEJONG. 1996. Protein sequences indicate that turtles branched off from the amniote tree after mammals. J. Mol. Evol. 42:580–586.

CONNAUGHTON, J. F., A. RAIRKAR, R. E. LOCKARD, AND A. KUMAR. 1984. Primary structure of rabbit 18S ribosomal RNA determined by direct RNA sequence analysis. Nucleic Acids Res. 12:4731–4745.

CRANDALL, K. A., AND J. J. F. FITZPATRICK. 1996. Crayfish molecular systematics: Using a combination of procedures to estimate phylogeny. Syst. Biol. 45:1–26.

CUNNINGHAM, C. O., H. ALIESKY, AND C. M. COLLINS. 2000. Sequence and secondary structure variation in the *Gyrodactylus* (Platyhelminthes: Monogenea) ribosomal RNA gene array. J. Parasitol. 86:567–576.

DESPER, R., AND O. GASCUEL. 2002. Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. J. Comput. Biol. 9:687–705.

DIXON, M. T., AND D. M. HILLIS. 1993. Ribosomal RNA secondary structure: Compensatory mutations and implications for phylogenetic analysis. Mol. Biol. Evol. 10:256–267.

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

FELSENSTEIN, J. 1993. PHYLIP 3.5 (phylogeny inference package), version 3.5. Department of Genetics, Univ. Washington, Seattle.

FITCH, D. H. A., B. BUGAJGAWEDA, AND S. W. EMMONS. 1995. 18S ribosomal-RNA gene phylogeny for some Rhabditidae related to *Caenorhabditis*. Mol. Biol. Evol. 12:346–358.

FITCH, W. M., AND E. MARGOLIASH. 1967. Construction of phylogenetic trees. Science 155:279–284.

FLORES-VILLELA, O., K. M. KJER, M. BENABIB, AND J. W. SITES. 2000. Multiple data sets, congruence and hypothesis testing for the phylogeny of basal groups of the lizard genus *Sceloporus* (Squamata, Phrynosomatidae). Syst. Biol. 49:713–739.

GARDINER, B. G. 1982. Tetrapod classification. Zool. J. Linn. Soc. 74:207–232.

GASCUEL, O. 1997. BIONJ: An improved version of the NJ algorithm based on a simple model of sequence data. Mol. Biol. Evol. 14:685–695.

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

GONZALEZ, P., AND J. LABARERE. 2000. Phylogenetic relationships of *Pleurotus* species according to the sequence and secondary structure of the mitochondrial small-subunit rRNA V4, V6 and V9 domains. Microbiology 146:209–221.

GU, X., AND J. ZHANG. 1997. A simple method for estimating the parameter of substitution rate variation among sites. Mol. Biol. Evol. 14:1106–1113.

HEDGES, S. B. 1992. The number of replications needed for accurate estimation of bootstrap *P*-value in phylogenetic studies. Mol. Biol. Evol. 9:366–369.

HEDGES, S. B. 1994. Molecular evidence for the origin of birds. Proc. Natl. Acad. Sci. USA 91:2621–2624.

HEDGES, S. B., AND L. R. MAXSON. 1992. 18S-ribosomal-RNA sequences and amniote phylogeny—Reply to Marshall. Mol. Biol. Evol. 9:374–377.

HEDGES, S. B., K. D. MOBERG, AND L. R. MAXSON. 1990. Tetrapod phylogeny inferred from 18S and 28S ribosomal RNA sequences and a review of the evidence for amniote relationships. Mol. Biol. Evol. 7:607–633.

HIBBETT, D. S., Y. FUKUMASA-NAKAI, A. TSUNEDA, AND M. J. DONOGHUE. 1995. Phyogenetic diversity in shiitake inferred from nuclear ribosomal DNA. Mycologia 87:618–638.

HICKSON, R. E., C. SIMON, A. COOPER, G. S. SPICER, J. SULLIVAN, AND D. PENNY. 1996. Conserved sequence motifs, alignment, and secondary structure for the third domain of animal 12S rRNA. Mol. Biol. Evol. 13:150–169.

HICKSON, R. E., C. SIMON, AND S. W. PERREY. 2000. The performance of several multiple-sequence alignment programs in relation to secondary-structure features for an rRNA sequence. Mol. Biol. Evol. 17:530–539.

HUELSENBECK, J. P., AND J. J. BULL. 1996. A likelihood ratio test to detect conflicting phylogenetic signal. Syst. Biol. 45:92–98.

HUELSENBECK, J. P., J. J. BULL, AND C. W. CUNNINGHAM. 1996. Combining data in phylogenetic analysis. Trends Ecol. Evol. 11:152–158.

HWANG, S. K., AND J. G. KIM. 2000. Secondary structural and phylogenetic implications of nuclear large subunit ribosomal RNA in the ectomycorrhizal fungus *Tricholoma matsutake*. Curr. Microbiol. 40:250–256.

JANKE, A., AND U. ARNASON. 1997. The complete mitochondrial genome of *Alligator mississippiensis* and the separation between recent archosauria (birds and crocodiles). Mol. Biol. Evol. 14:1266–1272.

KJER, K. M. 1995. Use of ribosomal-RNA secondary structure in phylogenetic studies to identify homologous positions—an example of alignment and data presentation from the frogs. Mol. Phylogenet. Evol. 4:314–330.

KJER, K. M. 1997. Conserved primary and secondary structural motifs of amphibian 12S rRNA, domain III. J. Herpetol. 31:599–604.

KJER, K. M., R. J. BLAHNIK, AND R. HOLZENTHAL. 2001. Phylogeny of Trichoptera (Caddisflies): Characterization of signal and noise within multiple datasets. Syst. Biol. 50:781–816.

KRETZER, A., Y. LI, T. SZARO, AND T. D. BRUNS. 1996. Internal transcribed spacer sequences from 38 recognized species of *Suillus* sensu lato: Phylogenetic and taxonomic implications. Mycologia 88:776–785.

LAKE, J. A. 1994. Reconstructing evolutionary trees from DNA and protein sequences: Paralinear distances. Proc. Natl. Acad. Sci. USA 91:1455–1459.

LOCKHART, P. J., A. W. LARKUM, M. STEEL, P. J. WADDELL, AND D. PENNY. 1996. Evolution of chlorophyll and bacteriochlorophyll: The problem of invariant sites in sequence analysis. Proc. Natl. Acad. Sci. USA 93:1930–1934.

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.

LØVTRUP, S. 1985. On the classification of the taxon Tetrapoda. Syst. Zool. 34:463–470.

LUTZONI, F., P. WAGENER, V. REEV, AND S. ZOLLER. 2000. Integrating ambiguously aligned regions of DNA sequence in phylogenetic analyses without violating positional homology. Syst. Biol. 49:628–651.

LYDEARD, C., W. E. HOLZNAGEL, M. N. SCHNARE, AND R. R. GUTELL. 2000. Phylogenetic analysis of molluscan mitochondrial LSU rDNA sequences and secondary structures. Mol. Phylogenet. Evol. 15:83–102.

MAIDAK, B. L., J. R. COLE, T. G. LILBURN, C. T. PARKER, JR., P. R. SAXMAN, J. M. STREDWICK, G. M. GARRITY, B. LI, G. J. OLSEN, S. PRAMANIK, T. M. SCHMIDT, AND J. M. TIEDJE. 2000. The RDP (Ribosomal Database Project) continues. Nucleic Acids Res. 28:173–174.

MANOS, P. S. 1997. Systematics of *Nothofagus* (Nothofagaceae) based on rDNA spacer sequences (ITS): Taxonomic congruence with morphology and plastid sequences. Am. J. Bot. 84:1137–1155.

MARSHALL, C. R. 1992. Substitution bias, weighted parsimony, and amniote phylogeny as inferred from 18S-ribosomal-RNA sequences. Mol. Biol. Evol. 9:370–373.

MORIN, L. 2000. Long branch attraction effects and the status of "basal eukaryotes": Phylogeny and structural analysis of the ribosomal RNA gene cluster of the free-living diplomonad *Trepomonas agilis*. J. Eukaryot. Microbiol. 47:167–177.

MORRISON, D. A., AND J. T. ELLIS. 1997. Effects of nucleotide sequence alignment on phylogeny estimation: A case study of 18S rDNAs of Apicomplexa. Mol. Biol. Evol. 14:428–441.

MUGRIDGE, N. B., D. A. MORRISON, A. M. JOHNSON, K. LUTON, J. P. DUBEY, J. VOTYPKA, AND A. M. TENTER. 1999. Phylogenetic relationships of the genus *Frenkelia*: A review of its history and new knowledge gained from comparison of large subunit ribosomal ribonucleic acid gene sequences. Int. J. Parasitol. 29:957–972.

NOTREDAME, C., E. A. O'BRIEN, AND D. G. HIGGINS. 1997. RAGA: RNA sequence alignment by genetic algorithm. Nucleic Acids Res. 25:4570–4580.

OLSEN, G. J., AND C. R. WOESE. 1993. Ribosomal RNA: A key to phylogeny. Fed. Am. Soc. Exp. Biol. J. 7:113–123.

ROMER, A. S. 1966. Vertebrate paleontology. Univ. Chicago Press, Chicago.

RZHETSKY, A., AND M. NEI. 1992. A simple method for estimating and testing minimum-evolution trees. Mol. Biol. Evol. 9:945–967.

RZHETSKY, A., AND M. NEI. 1994. METREE: A program package for inferring and testing minimum-evolution trees. Comput. Appl. Biosci. 10:409–412.

SEUTIN, G., B. F. LANG, D. P. MINDELL, AND R. MORAIS. 1994. Evolution of the WANCY region in amniote mitochondrial-DNA. Mol. Biol. Evol. 11:329–340.

SWOFFORD, D. L. 1993. PAUP: Phylogenetic analysis using parsimony. Illinois Natural History Survey, Champaign.

SWOFFORD, D. L. 2000. PAUP: Phylogenetic analysis using parsimony* (*and other methods), version 4. Sinauer, Sunderland, Massachusetts.

TAMURA, K., AND M. NEI. 1993. Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees. Mol. Biol. Evol. 10:512–526.

TITUS, T. A., AND D. R. FROST. 1996. Molecular homology assessment and phylogeny in the lizard family Opluridae (Squamata: Iguania). Mol. Phylogenet. Evol. 6:49–62.

UCHIDA, H., K. KITAE, K. I. TOMIZAWA, AND A. YOKOTA. 1998. Comparison of the nucleotide sequence and secondary structure of the 5.8S ribosomal RNA gene of *Chlamydomonas tetragama* with those of green algae. DNA Seq. 8:403–408.

VAN DE PEER, Y., P. DE RIJK, J. WUYTS, T. WINKELMANS, AND R. DE WACHTER. 2000. The European small subunit ribosomal RNA database. Nucleic Acids Res. 28:175–176.

VAN DE PEER, Y., J. M. NEEFS, P. DE RIJK, AND R. DE WACHTER. 1993. Reconstructing evolution from eukaryotic small-ribosomal-subunit RNA sequences: Calibration of the molecular clock. J. Mol. Evol. 37:221–232.

WILLIAMS, P. L., AND W. M. FITCH. 1990. Phylogeny determination using dynamically weighted parsimony method. Methods Enzymol. 183:615–626.

XIA, X. 2000a. Data analysis in molecular biology and evolution. Kluwer, Boston.

XIA, X. 2000b. Phylogenetic relationship among horseshoe crab species: The effect of substitution models on phylogenetic analyses. Syst. Biol. 49:87–100.

XIA, X., AND Z. XIE. 2001. DAMBE: Software package for data analysis in molecular biology and evolution. J. Hered. 92:371–373.

ZARDOYA, R., AND A. MEYER. 1996. Phylogenetic performance of mitochondrial protein-coding genes in resolving relationships among vertebrates. Mol. Biol. Evol. 13:933–942.

ZARDOYA, R., AND A. MEYER. 1998. Complete mitochondrial genome suggests diapsid affinities of turtles. Proc. Natl. Acad. Sci. USA 95:14226–14231.