The Phylogenetic Relationship of Tetrapod, Coelacanth, and Lungfish Revealed by the Sequences of Forty-Four Nuclear Genes

Naoko Takezaki,*1 Felipe Figueroa,* Zofia Zaleska-Rutczynska,* Naoyuki Takahata,† and Jan Klein*2

*Max-Planck-Institut für Biologie, Abteilung Immunogenetik, Corrensstrasse Tübingen, Germany; and
†The Graduate University for Advanced Studies, Department of Biosystems Science, Kanagawa, Japan

The origin of tetrapods is a major outstanding issue in vertebrate phylogeny. Each of the three possible principal hypotheses (coelacanth, lungfish, or neither being the sister group of tetrapods) has found support in different sets of data. In an attempt to resolve the controversy, sequences of 44 nuclear genes encoding amino acid residues at 10,404 positions were obtained and analyzed. However, this large set of sequences did not support conclusively one of the three hypotheses. Apparently, the coelacanth, lungfish, and tetrapod lineages diverged within such a short time interval that at this level of analysis, their relationships appear to be an irresolvable trichotomy.

Introduction

In modern classification schemes, living vertebrates consist of two main groups, the jawless Agnatha (represented by hagfishes and lampreys) and the jawed Gnathostomata (Pough, Heiser, and McFarland 1989). Jawed vertebrates are divided again into two groups, the cartilaginous Chondrichthyces (represented by sharks, rays, and chimeras) and the bony Osteichthyes. The two main groups of bony vertebrates are the ray-finned fishes (Actinopterygii) and the lobe-limbed vertebrates (Sarcopterygii). The latter comprise three groups: coelacanths (Actinistia), lungfishes (Dipnoi), and four-limbed vertebrates, or Tetrapoda, encompassing amphibians, reptiles, birds, and mammals (Clack 2002). More commonly, however, the designation Sarcopterygii is applied to a group comprising the coelacanths and lungfishes—the lobe-finned fishes. The phylogenetic relationship among the vertebrate groups has been agreed upon for some time (Romer 1966; Carroll 1988), with two important exceptions. The first of these exceptions, the relationship of hagfishes and lampreys to each other and to the gnathostomes appears to have now been resolved with the help of concatenated sequences from a large number of nuclear genes. Takezaki et al. (2003) have provided convincing evidence that hagfishes and lampreys are related to each other more closely than either of them is to the jawed vertebrates. The second exception, the relationship among the three groups of extant lobe-limbed vertebrates remains unresolved (Clack 2002; Schultz and Trueb 1991). Taking the ray-finned fishes as an outgroup, the three groups can theoretically be related to one another in four different ways (fig. 1): (1) the lungfishes are more closely related to the tetrapods than to the coelacanths (Tree 1, figure 1A); (2) the coelacanths are most closely related to the tetrapods (Tree 2, figure 1B); (3) lungfishes and coelacanths are most closely related to each other in a sister group of Sarcopterygii (sensu stricto) (Tree 3, figure 1C); or (4) the three groups are equally related to each other in a trichotomous rather than the more common dichotomous relationship (Tree 4, figure 1D).

The situation becomes more complicated when extant species are taken into account (Panchen and Smithson 1987). A number of fossil forms have been described that document the transition from fishes to tetrapods (Carroll 1988; Benton 1993; Zhu and Yu 2002). They include forms closely related to lungfishes or coelacanths (both were described before their living relatives were discovered), forms more distantly related to the lungfishes (the Porolepiformes and their relatives), species more closely related to the tetrapods (the Osteolepiformes and their relatives), and a few other species not easily assignable to any of these main lineages (Carroll 1988; Schultz and Trueb 1991; Clack 2002).

Early attempts to formulate hypotheses of tetrapod origin were based on the traditional approach of inferring ancestor-descendent relationships from the sharing of characters judged subjectively as being most important phylogenetically (Romer 1966). After the introduction (Hennig 1966) and widespread acceptance (Rosen et al. 1981) of the method of cladistic analysis, efforts were made to define and use large sets of shared derived characters to draw cladograms relating the various vertebrate taxa to one another. Both approaches have led to a stalemate, however. Among the traditionalists, the view prevailed ultimately that coelacanths were the closest living relatives of tetrapods (Romer 1966). Similarly, although the cladistic approach has produced a prevailing view that lungfishes are the closest extant relatives of the tetrapods (Tree 1) (see Rosen et al. [1981], Gardiner [1984], Maisey [1986], Panchen and Smithson [1987], and Ahlberg [1991]), alternative views continue to draw staunch support. The coelacanth-tetrapod sister group relationship (Tree 2) is advocated by Fritzsch (1987), Long (1989), Young, Long, and Ritchie (1992), and Zhu and Schultz (1997) among others, whereas the coelacanth-lungfish sister group (Tree 3) is favored by Northcutt (1986), Chang (1991), and Forey, Gardiner, and Patterson (1991), for example. The reasons for the different conclusions reached by these authors seem to be the choice of taxa included in the analysis, the selection of characters and the interpretation of character polarity, and the use of

---

1 Present address: National Institute of Genetics, Yata 1111, Mishima, Shizuoka 411-8540, Japan.
2 Present address: Department of Biology, Pennsylvania State University, University Park, Pennsylvania.

Key words: phylogenetic analyses, tetrapod evolution, lobe-finned fish, coelacanth, lungfish.

E-mail: ntakezak@lab.nig.ac.jp.

doi:10.1093/molbev/msh150
Advance Access publication May 5, 2004

Molecular Biology and Evolution vol. 21 no. 8 © Society for Molecular Biology and Evolution 2004; all rights reserved.
African clawed toad (*Xenopus laevis*)

**Materials and Methods**

**Acquisition of Sequence Data**

Total RNA was isolated from the whole body of a zebrafish (*Danio rerio*) and the liver of a coelacanth (*Latimeria chalumnae*), African lungfishes (*Protopterus dolloi* and *P. aethiopicus*), domestic fowl (*Gallus gallus*), African clawed toad (*Xenopus laevis*), and catshark (*Scyliorhinus canicula*) and used for the cDNA synthesis with the help of the Smart cDNA library construction kit (Clontech). The primers were designed by making multiple alignments of available sequences for genes of interest and finding conserved regions. The cDNA was amplified by the polymerase chain reaction (PCR), and the amplification products were cloned and sequenced. PCR amplification was carried out in the PTC-200 Programmable Thermal Controller (Biozym) with the aid of the Advantage2PCR kit (Clontech) using 1 ml of the cDNA and 1 mM of each of the sense and antisense primers. Selected PCR products were isolated from low-melting-point agarose gels (Gibco BRL), purified with the GFX kit (Amersham Biosciences), and cloned into pcr 2.1 Topo vector with the aid of the TOPO TA kit (Invitrogen). Sequencing reactions were carried out on the Thermo sequenase Primer Cycle Sequencing Kit (Amersham Biosciences) and processed by the LI-COR Long ReadIR 4200 DNA Sequencer (MWG Biotech). Other sequences used in this study were obtained by a Blast search of the GenBank nonredundant protein database for the orthologs of the selected proteins. EST sequences of *G. gallus* and *X. laevis* were searched for in the GenBank and used in several cases. About 80% of coelacanth and lungfish sequences, 30% of bird and shark sequences, 20% of amphibian sequences, and one ray-finned fish sequence were obtained in this study. The list of accession numbers and species names of the sequences and multiple alignments of amino acid sequences used in this study are available as Supplementary Material online. Multiple alignments of amino acid sequences were produced by the ClustalW version 1.82 (Thompson, Higgins, and Gibson 1994) and checked by eye. Unalignable parts of the sequences and positions with indels were excluded from the analysis.

**Sequence Collection**

The collection of sequences consists of mammalian (human), bird (fowl), amphibian (*Xenopus laevis*, a few other frogs, and a newt), coelacanth (mostly *Latimeria chalumnae*, with a few *L. menadoensis*), lungfish (mostly African lungfishes, with a few Australian and South American lungfishes), ray-finned fishes (different teleosts), and cartilaginous fishes (represented by shark sequences). The two coelacanth species (*L. chalumnae* from the Comoros Islands in the western Indian Ocean and *L. menadoensis* from Indonesia) diverged less than 10 MYA (Holder et al. 1999). The lungfish lineage is thought to be monophyletic, with African and South American lungfishes close to each other and the Australian lungfish most divergent (Carroll 1988). In mtDNA (Hedges, Hass, and Maxson 1993) and nuclear ribosomal RNA (Zardoya and Meyer 1996) studies, the three lungfish lineages form a monophyletic group. We found only a few of the Australian lungfish and either African or South American lungfish sequences in the database. In those cases, the sequences of the different lungfish lineages formed a monophyletic group tightly (data not shown).

About half of the genes obtained in our laboratory and from the database code for ribosomal proteins. On average, these relatively short genes evolve at about half the rate of nonribosomal proteins (table S1 in Supplementary Material online). They seem to contain different variants of the cladistic method. The use of molecular markers has not resolved the controversy, either. Mitochondrial DNA sequences of individual genes (Meyer and Wilson 1990; Meyer and Dolven 1992; Yokobori et al. 1994), gene fragments, or whole genomes (Hedges, Hass, and Maxson 1993; Cao et al. 1998; Zardoya et al. 1998) generally support a sister group relationship between lungfishes and tetrapods (Tree 1), although at least one gene (Yokobori et al. 1994) and the maximum-parsimony analysis of the whole genome (Zardoya et al. 1998) favor the coelacanth-lungfish sister group relationship (Tree 3). Various nuclear genes, on the other hand, supports Tree 1, 2, or 3, depending on the gene. Thus, Tree 1 is supported by DM20/PLP (Tohyama et al. 2000), RAG1, RAG2, POMC, and DM20/PLP (Venkatesh, Erdmann, and Brenner 2001) markers; Tree 2 is supported by the hemoglobin genes (Gorr, Kleinschmidt, and Fricker [1991]; but for criticism see, Forey [1991], Sharp and Lloyd [1991], Stock and Swofford [1991]. Meyer and Wilson [1991], Meyer [1995]) and the 18S ribosomal RNA gene (Stock et al. 1991); and Tree 3 is supported by the 28S ribosomal RNA gene (Zardoya and Meyer 1996).

In the present study, we attempted to resolve the phylogenetic relationships among coelacanths, lungfishes, and tetrapods by sequencing a large number of nuclear genes and by collecting sequences from the database. The data set included sequences of genes from mammals, birds, amphibians, coelacanths, lungfishes, ray-finned (teleost) fishes, and cartilaginous fishes.

**Materials and Methods**

Fig. 1.—The four possible phylogenetic relationships among tetrapod, coelacanth, and lungfish lineages.

A:B: Tetrapod

Lungfish
Coelacanth
Ray-finned fish

Tree 1

C: Tetrapod

Coelacanth
Lungfish
Ray-finned fish

Tree 2

D: Tetrapod

Coelacanth
Lungfish
Ray-finned fish

Tree 3

(Tree 4 (trichotomy))

Phylogeny of Tetrapod, Coelacanth, and Lungfish 1513
less phylogenetic information than nonribosomal proteins, but other than that, they do not show any other peculiarities, such as unusual selective constraints (see also Takezaki et al. [2003]).

We included genes encoding hemoglobin $\alpha$ and $\beta$ chains (Gorr, Kleinschmidt, and Fricke 1991) whose orthology has been questioned (Sharp and Lloyd 1991; Stock and Swofford 1991). At the time of the Gorr, Kleinschmidt, and Fricke (1991) study, only a small number of sequences were available for comparison. We examined the orthology of the sequences by comparing them with a large number of currently available gnathostome hemoglobin sequences (less than 300 of the $\alpha$ and $\beta$ chains each) and have come to the conclusion that the assumption of orthology of the sequences used in the present study is justified.

Orthology of Sequences

Orthology of sequences used in this study was examined by constructing neighbor-joining (NJ) trees (Saitou and Nei 1987). All the related sequences in the database were retrieved and the phylogenetic trees were constructed several times by gradually removing distantly related sequences and short sequences. In case of doubt, the locus was excluded in the final analysis. It is possible that the sequences used are not orthologous because of the unavailability of sequences in the database or loss of genes in some lineages. However, because almost complete sets of human and mouse sequences are available owing to the genome projects, we believe that the orthology of the sequences used can be justified for almost all the genes used in this study.

Data Sets Used in This Study

The sequences, together with those retrieved from databases, were assembled in three sets (table 1). A 41-gene set contained sequences of 41 loci from each of the following vertebrate groups: mammals, birds, amphibians, coelacanths, lungfishes, ray-finned fishes, and cartilaginous fishes. Each group was represented by one species. A 42-gene set contained sequences from all the above groups, except birds. A 44-gene set consisted of sequences from the same groups as in the 41-gene set, except birds and amphibians.

The issue of taxon sampling has been debated. Some studies indicated that adding more taxa to break long branches increases the probability of obtaining the correct tree topology (Poe and Swofford 1999) and that increasing the number of amino acid positions is a better way to obtain a correct tree topology (Poe and Swofford 1999; Rosenberg and Kumar 2001). This is the way taken in our study.

Phylogenetic Analysis

Phylogenetic trees were obtained by using the maximum-parsimony (MP) method in PAUP* version 4.0b10 (Swofford 2002), the NJ method, and the maximum-likelihood (ML) method in PAML version 3.13 software (Yang 1999). A branch-and-bound search was conducted in the MP method. Poisson correction distances with or without the gamma correction (Nei and Kumar 2000) and Dayhoff distance in the PHYLIP package (Felsenstein 1995) were used in the NJ method. The gamma shape parameter was estimated by the ML method. In the ML method, Poisson, JTT (Jones, Taylor, and Thornton 1992), and Dayhoff (Dayhoff, Schwartz, and Orcutt 1978) models were used with or without assuming rate variation across positions that follow the gamma distribution. The best fit to the data for most of the genes and for the concatenated sequences was obtained by using the JTT model. For the rest of the genes, the best fit was given by the Dayhoff model (see details in Supplementary Material online). Because the results obtained by using the different substitution models and different distance measures were essentially the same, only the results obtained from the JTT model are shown for the ML method and those obtained from the Poisson correction distance are shown for the NJ method without the assumption of rate variation across positions that follow the gamma distribution. The likelihood values were computed for all the possible tree topologies (three for four taxa, 15 for five taxa, 105 for six taxa, and 945 for seven taxa). In the ML method, bootstrap probability (BP) for the amino acid sequence data was obtained by the RELL method (Kishino, Miyata, and Hasegawa 1990) with 10,000 replications. The BPs for internal branches were computed by summing up the BPs for the tree topologies that contain the branch (Takezaki et al. 2003).

The Kishino-Hasegawa test (Kishino and Hasegawa 1989) was conducted for the ML and the MP trees. The results of the test were consistent with the results the bootstrap test, and none of them were significant. Therefore, only the result of the bootstrap test is shown in the Results section.

Phylogenetic trees obtained by using the Bayesian method in MrBayes version 3.0 software (Ronquist and Huelsenbeck 2003) had generally the same topologies as

<table>
<thead>
<tr>
<th>Gene Set</th>
<th>Vertebrate Lineages</th>
<th>Method (BP)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>MP</td>
<td>NJ</td>
</tr>
<tr>
<td>44</td>
<td>MCLR</td>
<td>3 (46)</td>
</tr>
<tr>
<td></td>
<td>MCLRS</td>
<td>3 (82)</td>
</tr>
<tr>
<td>42</td>
<td>MCLR</td>
<td>1 (51)</td>
</tr>
<tr>
<td></td>
<td>MACLRS</td>
<td>3 (87)</td>
</tr>
<tr>
<td></td>
<td>MCLRS</td>
<td>3 (35)</td>
</tr>
<tr>
<td></td>
<td>MCLR</td>
<td>3 (62)</td>
</tr>
<tr>
<td>41</td>
<td>MBACLRS</td>
<td>1 (56)</td>
</tr>
<tr>
<td></td>
<td>MBACLRS</td>
<td>3 (86)</td>
</tr>
<tr>
<td></td>
<td>MCLRS</td>
<td>1 (35)</td>
</tr>
<tr>
<td></td>
<td>MCLR</td>
<td>3 (58)</td>
</tr>
</tbody>
</table>

NOTE.—Topologies of tree types 1 through 3 are shown in figure 1. A = amphibian; B = bird; C = coelacanth; L = lungfish; M = mammal; R = ray-finned fish; S = cartilaginous fish (shark); BP = bootstrap probability; MP = maximum parsimony; NJ = neighbor-joining; ML = maximum-likelihood.
those obtained by the ML method but with much higher posterior probabilities than the bootstrap probabilities. However, a number of studies based on computer simulation (Alfaro, Zoller, and Lutzoni 2003; Douady et al. 2003b) and the actual data (Murphy et al. 2001; Whittingham et al. 2002; Douady et al. 2003a) have shown that Bayesian posterior probabilities tend to be higher than the bootstrap probabilities of the ML method even for false nodes (Douady et al. 2003b). In some cases, the conflicting branching patterns were supported by high posterior probabilities (Buckley et al. 2002; Douady et al. 2003c). Furthermore, a computer simulation study has shown that a significantly high posterior probability was given to the trees drawn by using concatenated sequences generated under conflicting tree topologies (Suzuki, Glazko, and Nei 2002). Therefore, Bayesian method was not used in the final analysis.

Analysis of the Pattern of Amino Acid Replacements in Sequence Comparisons

Useful phylogenetic information is revealed by the comparison of character states (here amino acid residues at a given position) of sequences of different lineages. For four different protein sequences, 15 different amino acid residue configurations are possible at any given position. If we designate the residues A, B, C, and D, the configurations are AAAA (all four sequences have the same residue at a given position), BAAA (the first sequence has a different residue than the remaining three), AABB (the first two sequences share a residue that is different from the residue shared by the other two sequences), and so on, up to ABCD (each of the four sequences has a different residue at a given position of the sequences).

In the case of four sequences, there are three phylogenetically informative configurations (PICs) (AABB, ABAB, and ABBA) for the MP method. Assuming that the four sequences are tetrapod (mammal), coelacanth, lungfish, and ray-finned fish, in this order, the conditions under which Tree 3 becomes the MP tree are, first, n(ABBA) > n(AABB) (condition 1), and second, n(ABBA) > n(ABAB) (condition 2), where n(.) indicates the number of positions that have the configuration in parentheses. The conditions under which Tree 3 becomes an NJ tree are, third, d_{13} + d_{24} > d_{14} + d_{23} (condition 3), and fourth, d_{12} + d_{34} > d_{14} + d_{23} (condition 4), where d_{ij} stands for a distance of sequences i and j (four-point condition [Li and Graur 1991, page 110]). Condition 3 corresponds to the situation in which the sum of branch lengths of Tree 3 becomes smaller than that of Tree 1. Condition 4 corresponds to the situation in which the sum of branch lengths of Tree 3 becomes smaller than that of Tree 2. Using p-distance (proportion of differences) as a distance measure, conditions 3 and 4 in which Tree 3 becomes an NJ tree can be written by using the numbers of the configurations as follows:

\[
2 \times n(ABBA) + n(ABCA) + n(BAAC) > 2 \times n(ABAB) + n(ABAC) + n(BCAA)
\]

for condition 3 and

\[
2 \times n(ABBA) + n(ABCA) + n(BAAC) > 2 \times n(ABAB) + n(ABAC) + n(BCAA)
\]

for condition 4.

To determine whether conditions 1 through 4 are supported significantly by the data, we conducted a multinomial test. Conditions 1 and 2 can be written as

\[
d1 = n(ABBA) - n(AABB) > 0 \quad (1) \quad \text{and} \quad d2 = n(ABBA) - n(ABAB) > 0 \quad (2).
\]

The variance of \(d1\) can be calculated as

\[
V(d1) = V[n(ABBA)] + V[n(AABB)] - 2 Cov[n(ABBA), n(AABB)],
\]

where \(V[n(ABBA)] = n \cdot pABBA(1 - pABBA), n\) is the total number of amino acid positions examined, and \(pABBA = n(ABBA)/n. V[n(AABB)]\) can be computed similarly. \(Cov[n(ABBA), n(AABB)] = -n \cdot pABBA\) where \(pABBA = n(ABBA)/n\).

By using the Z statistic, we find that conditions 1 and 2 are not significantly supported by the 44-gene set (\(Z = d1/\sqrt{V(d1)} = 0.99\) and \(Z = d2/\sqrt{V(d2)} = 0.73\)).

Conditions 3 and 4 can be written as

\[
D1 = 2 \times n(ABBA) + n(ABCA) + n(BAAC) - 2 \times n(ABAB) - n(ABAC) - n(BCAA) > 0 \quad \text{and}
\]

\[
D2 = 2 \times n(ABBA) + n(ABCA) + n(BAAC) - 2 \times n(ABAB) - n(ABAC) - n(BCAA) > 0,
\]

respectively.

The significance of \(D1\) and \(D2\) can be tested similarly as the significance of \(d1\) and \(d2\). The Z statistics for \(D1\) and \(D2\) were 2.10 and 1.77, respectively; only the former is significant at the 5\% level.

Computer Simulation Procedure

The observed pattern of amino acid replacements in the quartets of sequences was analyzed and compared with that obtained by the analysis of computer-generated sequences in which different models of sequence evolution were used. Specifically, computer simulation was utilized to obtain the expected frequencies of the different amino acid configurations and the probabilities of different tree topologies, as well as bootstrap probabilities. The JTT model was used to generate the amino acid sequences with or without the assumption of rate variation across positions that followed the gamma distribution (variable or uniform rate, respectively). To obtain the expected numbers of the configurations, 10,000 replications were carried out in the case of the uniform rate and 1,000 replications were carried out in the case of the variable rate, and the average numbers of the configurations and their standard errors were computed. In the preliminary study, 100 replications were carried out to obtain the probabilities of the different tree topologies and bootstrap probabilities. PAUP* and PAML software were used to draw the MP and ML trees in each replication. The branch lengths and the gamma
parameters estimated by the ML method were used to generate the sequences.

Simulations to obtain the expected numbers of configurations were conducted by using two different schemes, A and B, each according to two different protocols (1 and 2 in the case of scheme A and 3 and 4 in the case of scheme B). The assumption underlying scheme A was that the phylogenies of all the genes have the Tree 3 or Tree 4 topologies. In protocol 1, the branch lengths estimated for the concatenated sequences of the 41-gene or 44-gene sets were used to generate artificial sequences having the length of the concatenated sequences. In protocol 2, the branch lengths estimated for each gene were used to generate the artificial sequences that have the lengths of these genes, and the sequences were concatenated afterwards. Protocol 2 was used to take into account interlocus variation in rate and in-branch length ratio.

The assumption underlying scheme B was that tree topologies vary from gene to gene. In protocol 3, real sequences at each locus were first used to obtain ML trees, and the sequences were then divided into three groups corresponding to Trees 1, 2, and 3. In each group, the sequences were concatenated, the concatenated sequences were used to obtain an ML tree, and the branch lengths of the tree were determined. The topologies of the three trees and the branch lengths of the trees were then used to generate artificial sequences that correspond to each of the three groups. Finally, the artificial sequences from the three groups were concatenated, and the concatenated sequences were analyzed for amino acid configurations. In protocol 4, the loci were not grouped, but instead, an ML tree was obtained for each locus, the topology and branch lengths were used to generate corresponding artificial sequences, and the sequences of different loci were then concatenated. In both schemes, the numbers of configurations obtained in the two different protocols were similar, which indicates that rate variation among the genes is not important. However, because numbers of configurations generated by assuming uniform or varying rate are different, rate variation across amino acid positions is important. For this reason, only the results obtained by using protocols 1 and 3 under the assumption of the variable rate across positions are shown.

A computer simulation was also conducted in which it was assumed that a proportion \( p \) of the amino acid positions evolved according to the two-state model, in which substitutions occur only between two kinds of amino acids (where \( p \) ranged from 5% to 10%). The rest of the positions evolved according to the JTT model, with the assumption of variable rate across positions. The two-state positions were assumed to evolve at a uniform rate that was \( k \)-times higher than the rate for positions that evolve according to the JTT model, with variable rate across positions (where \( k \) ranged from 5 to 10 and the rate variability was defined by the gamma distribution). The branch lengths estimated from Tree 4 and the JTT model, with the assumption of variable rate across positions for the 41-gene set, were used to generate the portion of sequences that evolve according to the JTT model, with variable rate. Branch lengths \( k \) times longer than these were used to generate the sequences that evolve according to the two-state model.

### Results

**Phylogenetic Relationships Among Gnathostome Lineages**

Phylogenetic trees were obtained by MP, ML, and NJ methods from concatenated protein sequences translated from the 44-gene, 42-gene, and 41-gene sets. From each set, trees were also obtained after the exclusion of some of the taxa: cartilaginous fishes from all three sets, amphibians from the 42-gene and 41-gene sets, and birds from the 41-gene set (table 1). Disregarding the coelacanths and the lungfishes, the phylogenetic relationships were as expected; cartilaginous fishes diverged first, followed by ray-finned fishes, amphibians, birds, and mammals (fig. 2). On taking coelacanths and lungfishes into account, ambiguity entered the analysis in that the different tree topologies that were obtained depended on the tree-drawing method and the taxa included in the analysis. Of the 30 variants of the analysis (three different methods times 10 different
combinations of taxa), 24 yielded Tree 3, six yielded Tree 1, and none yielded Tree 2 or Tree 4 (table 1).

Thus, seemingly, the analysis excluded the coelacanth as a contender for the position of the tetrapod’s closest relative and favored a sister-group relationship between coelacanths and lungfishes. Of the three algorithms used, the NJ method consistently yielded Tree 3, whereas the MP and ML methods yielded either Tree 3 or Tree 1, depending on the combination of taxa tested. However, after the exclusion of cartilaginous fishes from the analysis, the MP and ML methods also yielded Tree 3 consistently (table 1), albeit with low bootstrap probability (BP ≥ 58%) in most of the cases. In general, the bootstrap support for Tree 3 was high only when the NJ method was used (BP = 88% to 99%). It was low to intermediate for trees obtained by the MP and ML methods (58% to 89%).

Both the low bootstrap probability and the apparent effect of taxon selection on the tree topology cast doubt on Tree 3 as representing the true relationship among tetrapods, coelacanths, and lungfishes. Because of this concern, the sequence data were subjected to further analysis.

To examine the effect of concatenation on the results of the analysis, trees were obtained for each of the 44 loci separately. Using the NJ method, 12, 14, and 18 loci produced Trees 1, 2, and 3, respectively. For the ML method, the numbers were 12, 16, and 15, respectively (excluding one locus for which the likelihood values for Trees 1, 2, and 3 were indistinguishable). For the MP method, the numbers were 11, 11, and 14, respectively (excluding eight loci for which two or more equally parsimonious trees were obtained [see Supplementary Material online]). Similar results were also obtained with the 42-gene and 41-gene data sets. These observations reinforce the concern about the interpretation of the concatenated sequence analysis.

Analysis of Amino Acid Replacement Patterns

To find out why the NJ method identified unambiguously and with high bootstrap probability Tree 3 as representing the phylogenetic interrelationships among tetrapods, coelacanths, lungfishes, and ray-finned fishes, whereas the MP and ML method equivocated, with low bootstrap probability between Trees 3 and 1, a position-by-position analysis of proteins specified by the 44 loci in each of the four taxa was carried out. The specific purpose of the analysis was to ascertain how the observed pattern of amino acid replacements at the individual positions deviated from the expected pattern and how this deviation influenced the phylogenetic reconstruction by the NJ, MP, and ML methods. The theory underlying the pattern analysis is described in the Materials and Methods. Taking into account only whether or not amino acid residues at a position are of the same kind, each position in a collection of four sequences can assume 15 different configurations: all four sequences can share the same residue (generally, AAAA); three sequences can share a residue while the fourth has a different residue (configurations ABBB, BABB, BBAB, and BBBA); two sequences share one residue while the other two share a different residue (configurations AABB, ABAB, and ABBA); two sequences have the same residue while each of the other two has a different residue (configurations AABC, ABAC, ABCA, BACA, BACA, and BCAA); or all four sequences have a different residue at a given position (configuration ABCD). In all these configurations, the letters refer to tetrapod, coelacanth, lungfish, and ray-finned sequences, in this order.

Pattern analysis reveals the reasons for the difference in the phylogenetic messages conveyed by the MP and NJ methods. Because of the different principles on which these two methods are based, the MP method identifies only three of the 15 configurations as being PICs. They are configuration numbers 6 (AABB), 7 (ABAB), and 8 (ABBA), which give Trees 2, 1, and 3, respectively (table 2). Hence, for a sequence set to yield Tree 3 unambiguously, it has to have significantly more configurations 8 than configurations 6 or 7. The real sequence set contains 92, 96, and 106 configuration numbers 6, 7, and 8, respectively (table 2). A multinomial test shows the differences in the number of configurations not to be significant statistically for the MP method to choose Tree 3 over Trees 1 or 2 unambiguously.

The situation with the application of the ML method is more complex because, here, it matters not only whether the sequences share an amino acid residue of the same kind at a given position but also which residues are actually present in the sequences at this position. Because, however, the method gives similar results as the MP method, the same positions as in the MP method are apparently responsible for the observed ambiguity.

In contrast to the MP method, the NJ method takes into account not only configurations 6 through 8 but also configurations 9 through 14; the remaining six configurations are phylogenetically noninformative. This method is, in essence, a search for a tree topology that gives the smallest sum of branch lengths. In the case of four sequences, Tree 3 becomes a minimum tree under two conditions: (1) The sum (S3) of the number of configuration 8 and half the numbers of configurations 11 and 12 exceeds significantly the sum (S2) of the number of configuration 6 and half the numbers of configurations 9 and 14. (2) S3 exceeds significantly the sum (S1) of the number of configuration 7 and half the numbers of configuration 10 and 13. As pointed out above, the excess of configuration 8 over the configurations 6 and 7 in the observed data is not significant statistically. For configurations 9 through 14, the situation is different. Here the total number of observed configurations 11 and 12 is 228 as compared with 183 configurations 9 and 14 and to 186 configurations 10 and 13, and at least the difference between S3 and S2 is statistically significant by the multinomial test (see Materials and Methods). It is apparently this excess that is responsible for the high bootstrap support of Tree 3 when the NJ method is applied to the sequence data. However, computer simulations show that such a difference in bootstrap support by different methods can happen by chance when the phylogeny of all the genes is multifurcating or when it varies from gene to gene.

Note that p-distance was assumed to be used for the conditions for the NJ method to choose different tree
...topologies based on the numbers of configurations mentioned above. When gamma correction or Poisson correction distance is used, the conditions are more complicated than those for $p$-distance. However, the tree topologies and the bootstrap values given by the NJ method using different distances were essentially the same. Differences in the numbers of configurations that are responsible for the NJ tree when $p$-distance is used must be affecting the bootstrap value of the NJ trees that use the other distances.

Analysis of Sequences Generated by Computer Simulation

A tree of four sequences consists of a single internal branch that connects two pairs (clusters) of external branches. In the case of an ambiguous relationship among the four sequences, one might expect the internal branch to be very short, which, thus, makes the external branches emerge nearly from the same node. Yet, in the case of Tree 3, based on the observed sequences, the internal branch is quite long, regardless of the method used to draw the tree; its length ranges from 10% to 20% of the external branch lengths. Furthermore, computer simulations indicate that the observed long length of the internal branch should lead to a clear resolution of the tree topology based on the real sequences, which, as we have seen, it does not. Finally, the high, nearly equal numbers of configurations 6 through 8 assign a long internal branch to all three theoretically possible bifurcating four-sequence trees, and this is highly unusual. Because of these unexpected findings, computer simulations were conducted to search for a model that could account for them.

The starting point of the simulations was the JTT model, which was used to assign specific amino acid residues to individual positions of artificially generated sequences that corresponded in lengths and other parameters to the real sequences of the 44-gene set. The artificial sequences were then allowed to evolve under conditions specified by the chosen tree topology (either Tree 3 or Tree 4) and by the branch lengths of trees based on the real sequences. Amino acid replacements were introduced under the assumption of a varying substitution rate (i.e., the probability of a replacement at a given position being specified by the gamma distribution). Three evolutionary models were tested. In the first and second model, the interrelationships of the four sequences encoded in each locus of the 44-gene set was assumed to have either Tree 3 or Tree 4 topology. In the third model, the tree topology was allowed to vary from locus to locus.

In the second model, there are two possibilities: (1) all the genes evolved according to the same tree topology in which the internal branch is short enough to be regarded as multifurcation, or (2) the tree topology varies from gene to gene because the differential fixation of polymorphism in the ancestral population of tetrapods, coelacanths, and lungfishes. The second possibility can happen if the divergence of the three lineages had occurred in a very short time (< 1 to 2 Myr). However, trees with such short internal branches can be regarded essentially as multifurcating trees.

In the third model, the tree topology varies from gene to gene. However, in contrast to the above case in which the effect of ancestral polymorphism is considered, all the three possible trees have relatively long internal branches. Such a case is possible in the following scenario: A large-scale gene duplication occurred in the ancestor of tetrapod (T), coelacanths (C), lungfish (L). In the duplicated loci (D1 and D2), there were two sets of sequences for the three lineages. One of the duplicated genes had been lost independently in each lineage. The remaining sequences of the three lineages are various combination of the sequences.
The Observed and Expected Numbers of the Amino Acid Configurations at Positions in Protein Sequences of the Indicated Four Vertebrate Lineages and Specified by Loci of the 41-Gene Set

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Observed</th>
<th>Expected</th>
<th>Observed</th>
<th>Expected</th>
<th>Observed</th>
<th>Expected</th>
<th>Observed</th>
<th>Expected</th>
<th>Observed</th>
<th>Expected</th>
</tr>
</thead>
<tbody>
<tr>
<td>6 MBCR</td>
<td>7175</td>
<td>7218.4 ± 36.0</td>
<td>7134</td>
<td>7174.8 ± 36.9</td>
<td>6998</td>
<td>7040.5 ± 37.8</td>
<td>6954</td>
<td>6994.0 ± 39.6</td>
<td>7024</td>
<td>7070.9 ± 36.4</td>
</tr>
<tr>
<td>6 MBLR</td>
<td>174</td>
<td>195.0 ± 13.4</td>
<td>177</td>
<td>195.0 ± 13.4</td>
<td>244</td>
<td>251.0 ± 14.3</td>
<td>233</td>
<td>247.5 ± 15.0</td>
<td>218</td>
<td>229.8 ± 14.1</td>
</tr>
<tr>
<td>6 MACR</td>
<td>149</td>
<td>172.2 ± 12.6</td>
<td>136</td>
<td>172.2 ± 12.6</td>
<td>326</td>
<td>353.5 ± 17.4</td>
<td>316</td>
<td>342.9 ± 17.1</td>
<td>325</td>
<td>348.6 ± 16.4</td>
</tr>
<tr>
<td>6 MALR</td>
<td>360</td>
<td>347.2 ± 17.5</td>
<td>401</td>
<td>347.2 ± 17.5</td>
<td>299</td>
<td>305.5 ± 15.7</td>
<td>343</td>
<td>346.1 ± 17.7</td>
<td>307</td>
<td>314.0 ± 16.1</td>
</tr>
<tr>
<td>6 BACR</td>
<td>565</td>
<td>570.1 ± 21.2</td>
<td>568</td>
<td>570.1 ± 21.2</td>
<td>486</td>
<td>509.1 ± 20.5</td>
<td>476</td>
<td>507.1 ± 20.1</td>
<td>492</td>
<td>507.2 ± 20.5</td>
</tr>
<tr>
<td>6 BALR</td>
<td>214</td>
<td>197.7 ± 13.0</td>
<td>201</td>
<td>197.7 ± 13.0</td>
<td>135</td>
<td>114.9 ± 9.6</td>
<td>139</td>
<td>115.2 ± 10.6</td>
<td>134</td>
<td>114.8 ± 9.9</td>
</tr>
<tr>
<td>7 AAAA</td>
<td>48</td>
<td>16.6 ± 4.1</td>
<td>46</td>
<td>16.6 ± 4.1</td>
<td>81</td>
<td>25.2 ± 5.1</td>
<td>85</td>
<td>25.4 ± 5.0</td>
<td>70</td>
<td>24.3 ± 4.8</td>
</tr>
<tr>
<td>7 ABAB</td>
<td>219</td>
<td>213.1 ± 14.5</td>
<td>219</td>
<td>213.1 ± 14.5</td>
<td>136</td>
<td>143.3 ± 11.8</td>
<td>142</td>
<td>156.3 ± 12.2</td>
<td>150</td>
<td>148.4 ± 12.4</td>
</tr>
<tr>
<td>7 ABCA</td>
<td>72</td>
<td>63.1 ± 7.8</td>
<td>66</td>
<td>63.1 ± 7.8</td>
<td>118</td>
<td>116.1 ± 10.4</td>
<td>119</td>
<td>111.2 ± 10.5</td>
<td>123</td>
<td>116.7 ± 10.5</td>
</tr>
<tr>
<td>7 ACAB</td>
<td>45</td>
<td>40.9 ± 6.6</td>
<td>46</td>
<td>40.9 ± 6.6</td>
<td>78</td>
<td>72.1 ± 8.6</td>
<td>84</td>
<td>77.9 ± 8.9</td>
<td>85</td>
<td>79.7 ± 8.9</td>
</tr>
<tr>
<td>7 BACA</td>
<td>72</td>
<td>68.4 ± 8.1</td>
<td>83</td>
<td>68.4 ± 8.1</td>
<td>93</td>
<td>84.1 ± 9.4</td>
<td>94</td>
<td>83.0 ± 9.1</td>
<td>84</td>
<td>79.4 ± 8.5</td>
</tr>
<tr>
<td>7 ABAC</td>
<td>45</td>
<td>43.2 ± 6.7</td>
<td>44</td>
<td>43.2 ± 6.7</td>
<td>53</td>
<td>54.7 ± 7.6</td>
<td>60</td>
<td>59.9 ± 7.7</td>
<td>56</td>
<td>53.1 ± 7.0</td>
</tr>
<tr>
<td>14 BCAC</td>
<td>68</td>
<td>61.2 ± 7.9</td>
<td>66</td>
<td>61.2 ± 7.9</td>
<td>77</td>
<td>91.7 ± 9.5</td>
<td>72</td>
<td>89.1 ± 9.2</td>
<td>79</td>
<td>86.2 ± 9.0</td>
</tr>
<tr>
<td>15 ABCD</td>
<td>104</td>
<td>108.3 ± 10.6</td>
<td>101</td>
<td>108.3 ± 10.6</td>
<td>148</td>
<td>152.0 ± 12.1</td>
<td>151</td>
<td>157.8 ± 12.4</td>
<td>140</td>
<td>149.3 ± 11.8</td>
</tr>
</tbody>
</table>

Note.—The number of shared amino acid positions was 9337. M = mammal (human); B = bird; A = amphibian; C = coelacanth; L = lungfish; R = ray-finned fish; Obs = observed; Exp = expected. The expected values are shown only for the varying rate case. The gamma parameter was 0.35 in all the cases.
Table 4
The Numbers of Observed and Expected Amino Acid Configurations in Proteins Encoded by Loci of the 41-Gene Set, Assuming the Two-State Model of Protein Evolution

<table>
<thead>
<tr>
<th>Configuration</th>
<th>Observed</th>
<th>Expected $p = 0.05$</th>
<th>Expected $p = 0.10$</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAAA</td>
<td>7006</td>
<td>6788.2 ± 34.8</td>
<td>6534.4 ± 37.3</td>
</tr>
<tr>
<td>ABBB</td>
<td>299</td>
<td>357.6 ± 16.8</td>
<td>395.3 ± 18.4</td>
</tr>
<tr>
<td>BABB</td>
<td>264</td>
<td>329.1 ± 16.7</td>
<td>368.3 ± 17.8</td>
</tr>
<tr>
<td>BBAB</td>
<td>318</td>
<td>383.0 ± 17.7</td>
<td>421.7 ± 19.0</td>
</tr>
<tr>
<td>BBBA</td>
<td>489</td>
<td>562.4 ± 21.9</td>
<td>604.9 ± 22.5</td>
</tr>
<tr>
<td>AABB</td>
<td>87</td>
<td>754.4 ± 8.6</td>
<td>121.7 ± 10.3</td>
</tr>
<tr>
<td>ABAB</td>
<td>89</td>
<td>72.4 ± 8.0</td>
<td>118.2 ± 10.3</td>
</tr>
<tr>
<td>ABBB</td>
<td>96</td>
<td>73.3 ± 8.3</td>
<td>119.5 ± 10.7</td>
</tr>
<tr>
<td>AABC</td>
<td>109</td>
<td>118.1 ± 10.4</td>
<td>111.4 ± 10.4</td>
</tr>
<tr>
<td>ABAC</td>
<td>102</td>
<td>101.3 ± 9.7</td>
<td>95.1 ± 9.6</td>
</tr>
<tr>
<td>ABCA</td>
<td>82</td>
<td>70.1 ± 8.5</td>
<td>65.9 ± 7.8</td>
</tr>
<tr>
<td>BAAC</td>
<td>119</td>
<td>109.9 ± 10.3</td>
<td>104.3 ± 9.9</td>
</tr>
<tr>
<td>BACA</td>
<td>68</td>
<td>75.0 ± 8.5</td>
<td>71.0 ± 8.2</td>
</tr>
<tr>
<td>BCAC</td>
<td>58</td>
<td>66.2 ± 7.9</td>
<td>62.6 ± 8.1</td>
</tr>
<tr>
<td>ABCD</td>
<td>151</td>
<td>155.2 ± 11.9</td>
<td>146.7 ± 12.1</td>
</tr>
</tbody>
</table>

Note.—Obs = observed. $k$ is the rate of amino acid replacements based on the two-state model. The rate for the rest of the sequences based on the JTT model is 1. $p$ is proportion of amino acid positions evolving according to the two-state model. The 41-gene data set was used. The four vertebrate lineages used were mammal, coelacanth, lungfish, and ray-finned fish.

This finding indicates that during the evolution of the tetrapod lineages, more multiple hits occurred than expected under the JTT model with a variable substitution rate. Also, both the observed and the expected numbers of configuration 6 (AABB), which supports the clustering of two tetrapod lineages (mammal and bird, mammal and amphibian, or bird and amphibian), are clearly higher than the numbers of configurations 7 and 8. This finding indicates that where the phylogeny is known, the distribution of PICs observed in our analysis is in agreement with it. Furthermore, the magnitude of this difference is largest for the MBCR and MBLR and smallest for the MACR, MALR, BACR, and BALR combinations (where M, B, C, L, and R stand for mammal, bird, coelacanth, lungfish, and ray-finned fish, respectively). This finding again indicates that the data set reconstructs known phylogenies correctly and also that in the cases where the phylogeny is known, the difference between the observed and expected numbers of PICs is most likely caused by an excess of multiple hits. Finally, the numbers of PICs are very similar between combinations that include the lungfish and those that include the coelacanth. This finding indicates that the coelacanth and lungfish diverged from the tetrapod lineage at nearly the same time.

Simulation Under Mixed Substitution Rate Conditions

The JTT model with a variable substitution rate, which assumes a higher probability of multiple hits at certain sites than at others, does not explain fully the observed data. The possibility was, therefore, tested that a portion $p$ (= 0.05 or 0.1) of the sites underwent multiple hits with higher probability than assumed by the variable rate model. In this version of the simulation, 90% to 95% of the positions were allowed to evolve according to the JTT model with a variable substitution rate, whereas the remaining positions evolved according to the two-state model with $k$ (= 5 or 10) times higher rate. The simulation showed that when $p = 0.05$ and $k = 10$, the expected numbers of PICs match the observed values well (see table 4).

Distribution of Multiple Hits Among PICs

Assuming that multiple hits affect the three types of PICs (configurations 6, 7, and 8) equally, the phylogenetic relationship among the tetrapod, lungfish, and coelacanth lineages can be interpreted as representing a multifurcation (a trichotomy). An alternative possibility, however, is that multiple hits affect the three PIC types unequally, either because of differences in the number of nucleotide changes in codons involved in the A→B amino acid replacements or because of differences in the properties of the amino acid residues that distinguish the three types. In either case, the effect might mimic that caused by an increase in the number of multiple hits. To test these two possibilities, two additional types of analysis were carried out. In the first type, codons specifying amino acid residues in the real sequences were examined one by one and classified according to the minimum number of nucleotide changes required to effect the A→B replacements. The examination revealed that more than 90% of the replacements in the PICs can be explained by single-nucleotide substitutions in the codons involved. This result is in accordance with the expectation under the JTT model. Importantly, however, no significant differences were found among the configurations 6, 7, and 8 in the observed frequencies of single-nucleotide substitutions. To search for a possible effect caused by differences in amino acid properties, average Miyata-Yasunaga distances (Miyata, Miyazawa, and Yasunaga 1979), which are a measure of amino acid property difference based on polarity and volume, were computed from the real sequences under the assumption of either Tree 3 or Tree 4 phylogeny (see table 5). Significant differences
The observed and expected numbers of PICs divided by the minimum number of nucleotide changes required to change amino acid A to B or vice versa, and the average Miyata-Yasunaga distance (dMY).

<table>
<thead>
<tr>
<th>Combination of vertebrate lineages</th>
<th>PIC 1</th>
<th>PIC 2</th>
<th>PIC 3</th>
<th>dMY 1</th>
<th>dMY 2</th>
<th>dMY 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>MCLR ABBB</td>
<td>81</td>
<td>6</td>
<td>0</td>
<td>0.97</td>
<td>0.12</td>
<td>0.07</td>
</tr>
<tr>
<td>MCLR ABAB</td>
<td>82</td>
<td>7</td>
<td>0</td>
<td>0.97</td>
<td>0.12</td>
<td>0.07</td>
</tr>
<tr>
<td>MCLR ABBA</td>
<td>89</td>
<td>5</td>
<td>0</td>
<td>0.97</td>
<td>0.12</td>
<td>0.07</td>
</tr>
<tr>
<td>ACGR ABBB</td>
<td>72</td>
<td>6</td>
<td>0</td>
<td>0.97</td>
<td>0.12</td>
<td>0.07</td>
</tr>
<tr>
<td>ACGR ABAB</td>
<td>73</td>
<td>7</td>
<td>0</td>
<td>0.97</td>
<td>0.12</td>
<td>0.07</td>
</tr>
<tr>
<td>ACGR ABBA</td>
<td>96</td>
<td>4</td>
<td>0</td>
<td>0.97</td>
<td>0.12</td>
<td>0.07</td>
</tr>
</tbody>
</table>

The table shows the observed and expected numbers of PICs for different combinations of vertebrate lineages. The expected numbers are calculated for each PIC, and the average Miyata-Yasunaga distance (dMY) is provided for each combination.

**Discussion**

To summarize, in the phylogenetic analysis of the 44 (42, 41) nuclear loci, either separately or as concatenated sequences, only the NJ method appears to resolve the relationship of tetrapod, lungfish, and coelacanth lineages unambiguously in that it favors consistently and with high bootstrap probability a sister-group relationship between lungfishes and coelacanthus (Tree 3, table 1). The MP and ML methods also favor this relationship in most cases but mostly with low bootstrap probability, whereas in other cases, depending on the taxon included in the analysis, they favor lungfish as the closest relative of tetrapods (Tree 1). Position-by-position analysis of the sequences, combined with computer simulation (table 2), reveals the unambiguous support of the Tree 3 phylogeny by the NJ method to be an artifact. Analysis of positions with phylogenetically informative configurations reveals them to support nearly equally Trees 1, 2, and 3. In other words, it supports the interpretation that the likely mode of divergence of the tetrapod, lungfish, and coelacanth lineages was close to multifurcation (i.e., Tree 4). There was an excess of three PICs that support each of the three possible trees topologies for coelacanth, lungfish, and tetrapod (Trees 1, 2, and 3) in the data, compared with those expected under the JTT model of amino acid substitution. The search for an evolutionary model that could explain the observed results by comparing them with the results of computer simulations under different input assumptions suggests that a small fraction of the amino acid positions (nucleotide sites) has suffered multiple hits in excess of those expected under the JTT model (tables 3 and 4). As the hits affect nearly equally the three types of positions favoring the three different bifurcating trees (table 5), for stochastic reasons sequences encoded in genes at different loci favor different trees and the overall effect is an unresolved phylogeny.

Another possible explanation for the excess of the PICs is that the variable tree topologies for the different genes are caused by paralogous relationships between loci that result from large-scale gene duplications and differential gene loss in coelacanth, lungfish, and tetrapod. According to this explanation, the duplicated loci separated before the divergence of the three lineages, and the longer internal branches resulted from an excessive number of the PICs in the paralogous sequences. However, the multiple-hit hypothesis is the most likely explanation because the three PICs appeared equally frequently in each locus analysis (data not shown). Furthermore, the excess of PICs was observed also for different combination of lineages ([mammal, bird, coelacanth or lungfish, ray-finned fish], [mammal, amphibian, coelacanth or lungfish, ray-finned fish]) for which the phylogenetic relationships are established (table 3).
Although each of the loci included in the 44-gene set was carefully screened for orthology among the sequences derived from the different species, the possibility of unrecognized paralogy at a few of the loci cannot be excluded. However, the contribution of such paralogous sequences could not have been large enough to influence the phylogenetic reconstruction because the same data set reconstructed the known tetrapod phylogeny correctly.

However, Vogl et al. (2003) found that for 48 plastid genes collected for several lineages including algae and plant, many of the gene trees were incongruent. They suggested a possibility of paralogy that resulted from ancient duplication of loci and subsequent loss as a cause of incongruence in gene trees.

Differential fixation of ancestral polymorphism is a widespread phenomenon, but at most loci, it influences phylogenies in which divergences occur within a time interval of less than 1 to 2 Myr (Klein et al. 1998). To estimate the length of the interval of transition from fish to tetrapod, we used computer simulation to determine the minimum lengths of the internal branch necessary to differentiate Tree 3 phylogeny from multifurcating divergence. By increasing the internal branch length from 0 (multifurcation) to a point at which the number of configuration 8 differed significantly from the number of configurations 6 and 7, we could determine the minimum length to be 0.005. Because the external branches leading to mammals, birds, or amphibians are approximately 0.1 long, which corresponds to approximately 400 Myr of divergence time (Carroll 1988), the length of the fish-tetrapod transition interval is estimated to have been less than 20 Myr. This value is reasonably close to that estimated from paleontological evidence. Fossils identified unambiguously as tetrapods and lobe-finned ancestors of tetrapods have been dated to early Frasnian and late Famennian stages of the Devonian, respectively, separated by a time interval of 15 Myr (Carroll 1997). This interval is far too long for ancestral polymorphisms to span it (Klein et al. 1998).

Recently, many studies have used multiple-gene sequences to resolve phylogenetic relationships (e.g., Murphy et al. 2001; Madsen et al. 2001; Wolf, Rogozin, and Koonin 2004). However, resolution of the phylogenetic relationship seems to depend on factors such as the shortness of the internal branches, the lengths of external branches leading to the taxa (Saitou and Nei 1986), and evolutionary rate variation in different lineages (Hedges 2002). In one study (Rokas et al. 2003), although phylogeny of yeast species was resolved by use of more than 100 genes, similar results were obtained by use of about 20 genes. (However, reanalysis of their data [Phillips, Desulc, and Penny 2004] indicated that a different tree topology was supported by 100% bootstrap probability when different nucleotide substitution models and a different [minimum-evolution] method were used.) By contrast, in the study of amoeba species (Bapteste et al. 2002), the position of Stramenopiles (a group of eukaryotes) and the relationships among Conosa (amoeba and slime mold), Opisthokonta (fungi and animal) and plant could not be resolved by use of more than 100 genes. At the order of magnitude of sequence data collection feasible to us, the trichotomy has become irresolvable. However, for the reasons mentioned above, the question whether the trichotomy can be resolved by increasing the sequence collection by higher orders of magnitude remains open. By computer simulation similar to the one described earlier, but changing not only the internal branch length (0.001, 0.002, 0.003, . . .) but also the number of loci (the number of amino acid positions) (100, 200, 300, . . .), we estimate that if the fish-tetrapod transition interval was 10 to 20 Myr long and the phylogeny was dichotomous, more than 200 loci would have to be analyzed to resolve it. With the shortening of the interval, the number of loci needing to be analyzed increases correspondingly. Here, whole genome sequences of a lungfish and a coelancanth might provide the answer to whether the trichotomy is resolvable.

Acknowledgments

We thank Dr. Herbert Tichy for help in acquisition of tissue sample, Ryszard Lorenz for technical assistance, and Jane Kraushaar for editorial assistance.

Literature Cited


Thompson, J. D., D. G. Higgins, and T. J. Gibson. 1994. CLUSTAL W: improving the sensitivity of progressive


William Martin, Associate Editor

Accepted April 14, 2004