J Mol Evol (1996) 42:183–193

JOURNAL OF MOLECULAR EVOLUTION

Evolutionary Relationships of Eukaryotic Kingdoms

Sudhir Kumar, Andrey Rzhetsky

Institute of Molecular Evolutionary Genetics and Department of Biology, 322 Mueller Laboratory, The Pennsylvania State University, University Park, PA 16802, USA

Received: 8 December 1994 / Accepted: 28 September 1995

Abstract. The evolutionary relationships of four eukaryotic kingdoms—Animalia, Plantae, Fungi, and Protista—remain unclear. In particular, statistical support for the closeness of animals to fungi rather than to plants is lacking, and a preferred branching order of these and other eukaryotic lineages is still controversial even though molecular sequences from diverse eukaryotic taxa have been analyzed. We report a statistical analysis of 214 sequences of nuclear small-subunit ribosomal RNA (srRNA) gene undertaken to clarify these evolutionary relationships. We have considered the variability of substitution rates and the nonindependence of nucleotide substitution across sites in the srRNA gene in testing alternative hypotheses regarding the branching patterns of eukaryote phylogeny. We find that the rates of evolution among sites in the srRNA sequences vary substantially and are approximately gamma distributed with size and shape parameter equal to 0.76. Our results suggest that (1) the animals and true fungi are indeed closer to each other than to any other ''crown'' group in the eukaryote tree, (2) red algae are the closest relatives of animals, true fungi, and green plants, and (3) the heterokonts and alveolates probably evolved prior to the divergence of red algae and animal–fungus–green–plant lineages. Furthermore, our analyses indicate that the branching order of the eukaryotic lineages that diverged prior to the evolution of alveolates may be generally difficult to resolve with the srRNA sequence data.

Key words: Small-subunit ribosomal RNA — Phylogeny — Animals — Fungi — Plants — Alveolates — Heterokonts — Stramenopiles

Introduction

An unequivocal phylogeny of eukaryotes has not emerged from morphological, biochemical, cytological, and physiological evidence because of the tremendous diversity of these data and scarcity of the fossil record (Corliss 1984; Margulis and Schwartz 1988; Sogin 1989; Hendriks et al. 1991; Hasegawa et al. 1992; Knoll 1992; Cavalier-Smith 1993; Wainright et al. 1993). Evolutionary models of deep eukaryote phylogeny based on qualitative comparisons of these characters remain largely unconfirmed because only a few common and usable characters are available that are known to be repeatedly acquired and independently lost in various eukaryotic lineages, making it difficult to draw reliable phylogenetic inference. Thus, an independent and reliable phylogeny of eukaryotes is needed to study the evolution of various molecular, cellular, and developmental characters. For this purpose, phylogenetic analyses of nucleotide sequences of small and large ribosomal RNA as well as amino acid sequences from a variety of protein-coding genes have been reported (Gouy and Li 1989; Hasegawa et al. 1992; Knoll 1992; Baldauf and Palmer 1993; Cavalier-Smith 1993; Wainright et al. 1993; Nikoh et al. 1994).

Of all the macromolecules available with which to study eukaryote evolution, the nuclear small-subunit ribosomal RNA (srRNA) gene has been used most extensively because of its ubiquity and slow rate of evolution

Correspondence to: S. Kumar

and because of the availability of data from many representative eukaryote taxa (Woese 1987; Sogin 1989; Patterson and Sogin 1992; Cavalier-Smith 1993; Van de Peer et al. 1993). Comparisons of srRNA sequences already have proven useful in establishing phylogenetic relationships of diverse prokaryotes and protists. However, the use of these data in elucidating the branching order of animals, plants, fungi, heterokonts, and alveolates has resulted in contradictory topologies (e.g., Gouy and Li 1989; Cavalier-Smith 1993; Van de Peer et al. 1993; Wainright et al. 1993; Smothers et al. 1994). In particular, the statistical resolution of higher-order relationships inferred in evolutionary studies using the srRNA gene has been poor even when only a small number of representative taxa have been used (e.g., Cavalier-Smith 1993; Wainright et al. 1993). The difficulty in unraveling a reliable branching order of eukaryote phyla in the ''crown'' of the eukaryote tree (Knoll 1992) has been attributed to the small number of nucleotide changes that separate these groups at the interphylum level. However, these analyses were based on oversimplified models of srRNA evolution. Here, we attempt to reevaluate the evolutionary relationships of eukaryotes by considering variation in functional constraints and nonindependence of evolution among sites of srRNA gene. In addition to reconstructing phylogenetic trees, we compare alternative hypotheses regarding branching patterns of eukaryote lineages to identify aspects of eukaryote evolution that can be inferred from srRNA data with high statistical confidence.

Materials and Methods

Nucleotide Sequences Used. A representative compilation of the aligned srRNA sequences was obtained from the Ribosomal Database Project (RDP) updated on August 1, 1993, and aligned with sequences kindly provided by Dr. M. Sogin (Larson et al. 1993; Wainright et al. 1993). A list of represented taxa is given in Table 1. The manual alignment was done using the multiple alignment editor program of the VOSTORG package (Zharkikh et al. 1991). Finally, we included only those sites that were used by Wainright et al. (1993) in the phylogenetic analysis (1,572 sites). We also obtained the amino acid sequences of translation elongation factor α, actin, and α- and β-tubulin genes that were used by Balduaf and Palmer (1993).

Patterns of Nucleotide Substitution. The process of nucleotide substitution in srRNA gene is known to be constrained by its stem-and-loop secondary structure. Large portions of the srRNA gene are nearly invariant and are distributed among moderately conserved and highly variable regions (Elwood et al. 1985; Noller et al. 1986). These highly conserved segments of srRNA facilitate the study of evolutionary relationships among distantly related taxa. The fact that stem-loop secondary structure needs to be maintained indicates that functional constraints vary across sites in this gene. This is expected to result in different rates of nucleotide substitution among different segments in the srRNA gene (see also Rzhetsky 1995).

To examine the extent of variation of substitution rates among sites, we computed the number of changes required at each site by the parsimonious reconstruction of ancestral states at each interior node in a neighbor-joining tree (Saitou and Nei 1987). The distribution of number of nucleotide changes obtained in this manner is given in Fig. 1B. The Poisson distribution predicted for a homogenous substitution rate across sequence sites does not fit the data, whereas a negative binomial distribution gives a much better fit (Fig. 1B) (Uzzell and Corbin 1971). This suggests that the substitution rates are approximately gamma distributed among sites, with size and shape parameter $a = 0.76$ (Golding 1983; Jin and Nei 1990). We found that the estimated value of a was quite robust to choice of sites as well as the phylogenetic tree used to infer the number of changes at each site, see also (Wakeley 1993).

Next, we estimated the average G + C composition of srRNA gene in major eukaryote phyla (Fig. 1A). Clearly, G + C content in *Giardia* (Metamonada) is significantly larger than 50%, but the average G + C contents for other eukaryotic groups used in this study were not very different from 50%. Therefore, we used Kimura's two-parameter model where substitution rates vary across sites according to a gamma distribution ($a = 0.76$) to correct for multiple substitutions at the same site (Kimura 1980; Jin and Nei 1990). Below we formulate assumptions used in the derivation of Jin and Nei's (1990) estimator of evolutionary distance and describe a simple extension of their formulas for consideration of nonindependence of evolution in some sites in the srRNA gene.

The Jin and Nei (1990) model is based on the assumption that the nucleotide substitution in each site follows a continuous-time Markov process where instantaneous rates are equal for all transitional substitutions and for all transversional substitutions in accordance with Kimura's (1980) two-parameter model. Furthermore, the ratio of instantaneous transition and transversion rates is assumed to be identical for all sequence sites, but the absolute substitution rates are allowed to vary randomly among nucleotide sites following a gamma distribution (Jin and Nei 1990). That is, the actual number of nucleotide substitutions, k, separating two present-day sequences at a particular nucleotide site follows a Poisson distribution with parameter (xd)

$$P(k = i \mid x) = \frac{(xd)^i e^{-xd}}{i!} \tag{1}$$

where d is a positive constant and is same for all the sites of the two sequences compared (average number of substitutions per site; see later), and x is a random variable following a gamma distribution with density

$$f(x) = \begin{cases} \dfrac{a^a}{\Gamma(a)} x^{a-1} e^{-ax}, & x \geqslant 0 \\ 0, & x < 0 \end{cases} \tag{2}$$

Here, a determines the shape of the gamma distribution. The expected value and variance of x are equal to 1 and $1/a$, respectively. Now, it is straightforward to show that the unconditional distribution of k is a negative binomial distribution and it is the same for every site of the sequences compared.

$$P(k = i) = \frac{\Gamma(a + i)}{\Gamma(a) i!} \left(\frac{a}{a + d}\right)^a \left(\frac{d}{a + d}\right)^i \tag{3}$$

Similarly, the unconditional joint distribution of the actual number of transitional (k_s) and transversional (k_v) substitutions at a site can be shown to be equal to

$$P(k_s = i, k_v = j) =$$

$$\frac{\Gamma(a + i + j)}{\Gamma(a) i! j!} \left(\frac{a}{a + d}\right)^a \left[\frac{dR}{(a + d)(1 + R)}\right]^i \left[\frac{d}{(a + d)(1 + R)}\right]^j \tag{4}$$

Table 1. Eukaryote and prokaryote genera used in this study (number of individual species examined shown in square brackets)

Animals [35]
Triploblasts [29]: *Acyrthosiphon, Aedes, Alligator, Ambystoma, Argulus, Artemia, Bufo, Caenorhabditis, Callinectes, Drosophila, Echinorhinus, Eurypelma, Fundulus, Herdmania, Heterodon, Homo, Hyla, Mus, Opisthorchis, Oryctolagus, Placopecten, Porocephalus, Rattus, Schistosoma, Sebastolobus, Squalus, Strongyloides, Tenebrio, Xenopus*
Cnidaria [2]: *Anemonia, Tripedalia*; Placozoa [1]: *Trichoplax*; Porifera [2]: *Microciona, Scypha;* Ctenophore [1]: *Mnemiopsis*
Fungi [45]
Ascosphaera, Aspergillus, Athelia, Aureobasidium, Blastocladiella, Blastomyces, Byssochlamys, Candida [8], Chaetomium, Chytridium, Coccidioides, Debaryomyces, Endogone, Eremascus, Gigaspora, Glomus, Hansenula, Kluyveromyces, Leucosporidium, Leucostoma, Monascus, Mucor, Neocallimastix, Neurospora, Ophiostom [2], Penniclium, Pneumocystis, Podospora, Saccharomyces, Schizosaccharomyces, Spizellomyces, Talaromyces, Thanatephorus, Thermoascus, Torulaspora, Torulopsis, Yarrowia
Green Plants [44]
Land Plants [19]: *Arabidopsis, Buckleya, Buxus, Cornus [2], Dendrophthora, Euonymus, Fragaria, Glycine, Hedera, Hydrocotyle, Lycopersicon, Nyssa, Oryza, Phoradendron, Schoepfia, Sinapis, Zamia, Zea*
Green Algae [25]: *Ankistrodesmus, Characium [3], Chlamydomonas, Chlorella [8], Chlorococcopsis, Dunaliella, Friedmannia, Hydrodictyon, Nanochlorum, Neochloris, Parietochloris, Pediastrum, Prototheca [2], Scenedesmus, Volvox*
Red Algae [5]
Gracilaria [3], Gracilariopsis, Porphyra
Heterokonts [19]
Oomycetes [3]: *Achlya, Lagenidium, Phytophthora*
Golden-Brown Algae [6]: *Chromulina, Hibberdia, Mallomonas [2], Ochromonas, Synura*
Diatoms [6]: *Bacillaria, Cylindrotheca, Nitzschia, Rhizosolenia, Sceletonema, Stephanopyxisi*
Yellow-Green Algae [1]: *Tribonema*
Brown Algae [2]: *Costaria, Fucus*
Eustimatophycea [1]: *Nannochloropsis*
Other Chromists [3]
Haptophyta [1]: *Emiliania*
Crytomonads [2]: *Pyrenomonas, Crytomonas*
Alveolates [34]
Ciliates [21]: *Blepharisma, Colpidium, Colpoda, Euplotes, Glaucoma, Onychodromus, Opisthonecta, Oxytricha [2], Paramecium, Stylonychia, Tetrahymena [10]*
Dinoflagellates [4]: *Crypthecodinium, Prorocentrum, Symbiodinium [2]*
Apicomplexans [9]: *Babesia, Sarcocystis, Theileria, Plasmodium [6]*
Heterolobosea [4]
Naegleria, Paratetramitus, Tetramitus, Vahlkampfia
Kinetoplastids [11]
Bodo, Crithidia, Endotrypanum, Leishmania [5], Leptomonas, Trypanosoma [2]
Parabasalia [1]
Tritrichomonas
Microsporidia [1]
Vairimorpha
Metamonada [2]
Giardia [2]
Choanoflagellates [2]
Acanthocoepsis, Diaphanoeca
Other Eukaryotes [4]
Acanthamoeba; *Dictyostelium*; *Euglena, Entamoeba*
Bacteria [4]
Anacystis, Escherichia, Neisseria, Planctomyces

where $d = E(k_s) + E(k_v)$ and $R = E(k_s)/E(k_v)$. Next, the probabilities of having transitional and transversional differences between two sequences turn out to be identical for all sites in this model and are given by

$$P = \frac{1}{2}\left\{\frac{1}{2} - \left[\frac{a}{a + d(1+2R)/(1+R)}\right]^a + \frac{1}{2}\left[\frac{a}{a + 2d/(1+R)}\right]^a\right\} \quad (5)$$

and

$$Q = \frac{1}{2}\left\{1 - \left[\frac{a}{a + 2d/(1+R)}\right]^a\right\}, \quad (6)$$

respectively.

In this case, the expected number of nucleotide substitutions per site is given by

$$d = \frac{a}{2}\left[(1-2P-Q)^{-1/a} + \frac{1}{2}(1-2Q)^{-1/a} - \frac{3}{2}\right] \quad (7)$$

where P is the expected proportion of transitional differences and Q is the expected proportion of transversional differences between a pair of sequences (Jin and Nei 1990), and is usually estimated by substituting parameters in the right-hand side of (7) by their estimates.

To account for the nonindependence of sites in the stem regions of the srRNA secondary structure, we note that the observed values $\hat{P}$ and

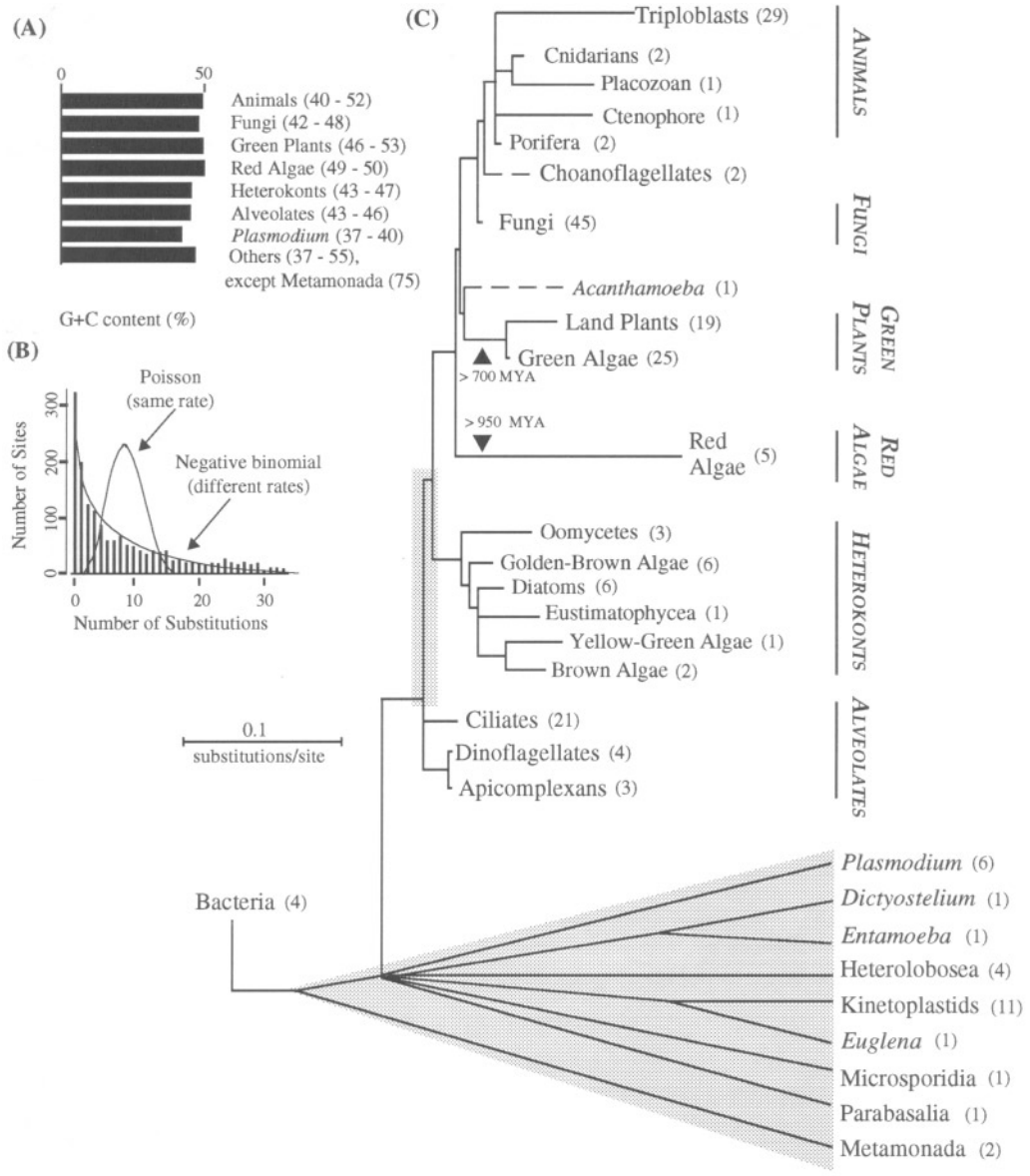

Fig. 1. **A** Average G + C contents (%) of srRNA gene in major eukaryote groups. *Numbers in parentheses* indicate the minimum and the maximum G + C contents in each group. **B** The distribution of the number of nucleotide substitutions (mean = 7.45, variance = 80.01) across sites (see text for details). **C** A preferred branching order of the major eukaryotic groups (Table 1). The *rectangular shaded area* indicates relationships that could not be established with high statistical confidence because of highly divergent outgroup taxa. *Dashed lines and shaded areas* indicate statistically unresolved branching patterns. In the ''crown'' of the tree, the lengths of the terminal branches are *truncated* at the common ancestor of each monophyletic group (number of taxa in *parentheses*). For early diverging eukaryotic lineages branch lengths are not drawn to the scale because the branching order is unresolved. The dates of divergence indicated are based on the fossil record (Butterfield et al. 1988; Butterfield et al. 1990; Knoll 1992).

$\hat{Q}$ are consistent and unbiased estimates of P and Q, respectively, even though the sites are not independent. Therefore, the estimator obtained from (7) by substituting P and Q with their estimates is consistent, but can be biased when the number of sites sampled, n, is small (Tajima 1993). The magnitude of this bias is determined by the values of variances, covariances, and higher-order moments of $\hat{P}$ and $\hat{Q}$. Although this bias can be reduced, it is usually small when sequences are long and can be safely neglected in most applications (Tajima 1993; Rzhetsky and Nei 1994b). Therefore, eqn. (7) can be used for distance estimation even if some of the sites evolve in a nonindependent fashion.

The variances and covariances of $\hat{P}$ and $\hat{Q}$ are also used to compute variance of $\hat{d}$. The variance of $\hat{d}$ can be computed using the delta technique in statistics

$$V(d) \cong \left(\frac{\partial d}{\partial P}\right)^2 V(\hat{P}) + \left(\frac{\partial d}{\partial Q}\right)^2 V(\hat{Q}) + 2\left(\frac{\partial d}{\partial P}\right)\left(\frac{\partial d}{\partial Q}\right) Cov(\hat{P},\hat{Q}) \qquad (8)$$

Since the variance of the estimate of a is small compared to the variances of $\hat{P}$ and $\hat{Q}$ (as stated above, the same estimate of a was obtained for various subsets of the alignment of sites and different tree topologies), it is not considered in computation of $V(\hat{d})$.

To estimate the increase in $V(\hat{d})$ due to nonindependence of sites, we assume that some sites are completely dependent (compensatory changes), which is equivalent to the assumption that some of the sites are sampled twice. This assumption is used to simplify the computation of variance and will result in overestimate of variances because the

complementary sites in the stem regions are not always completely dependent. Therefore, the confidence probability values obtained are likely to be conservative. Since the nucleotide substitutions in complementary sites of the srRNA stem region are assumed to be completely positively correlated, if a transitional substitution occurs in one strand of the stem, a compensatory change in the complementary strand will also be a transitional substitution. Similarly, a transversional substitution in one strand would result in a transversional change in the complementary strand. In the present model,

$$\hat{Q} = \frac{1}{n}\sum_{i=1}^{n} I_{Qi}, \quad E(\hat{Q}) = Q \tag{9}$$

$$\hat{P} = \frac{1}{n}\sum_{i=1}^{n} I_{Pi}, \quad E(\hat{P}) = P \tag{10}$$

$I_{Pi} = \begin{cases} 1 & \text{if the two sequences have a transition difference in the} \\ 0 & i\text{-th site, and otherwise; and} \end{cases}$

$I_{Qi} = \begin{cases} 1 & \text{if the two sequences have a transversion difference in the} \\ 0 & i\text{-th site, and otherwise} \end{cases}$

If n_S sites out of n are pairwise correlated, then

$$V(\hat{P}) = \frac{1}{n^2} E\left[\left(\sum_{i=1}^{n} I_{Pi}\right)^2\right] - \frac{1}{n^2} E^2\left(\sum_{i=1}^{n} I_{Pi}\right)$$

Since,

$$E(I_{Pi}I_{Pj}) = \begin{cases} E(I_{Pi})E(I_{Pj}) & = P^2 \text{ for unpaired sites, and } i \neq j \\ E(I_{Pi}) & = P \text{ for paired sites, or } i = j \end{cases}$$

and

$$E\left(\sum_{i=1}^{n} I_{Pi}\right) = nP$$

We obtain

$$V(\hat{P}) = \frac{n^2 - n_S - n}{n^2} P^2 + \frac{n + n_S}{n^2} P - P^2 = \frac{n + n_S}{n} \cdot \frac{P(1-P)}{n} \tag{11}$$

That is, the variance of $\hat{P}$ is $(n + n_S)/n$ times larger than the same values when the sites are independent. A similar argument applies for an $(n + n_S)/n$ times increase in the variance of $\hat{Q}$. It can be easily shown that the covariance of $\hat{P}$ and $\hat{Q}$ increases by the same factor:

$$Cov(\hat{P},\hat{Q}) = \frac{n^2 - n_S - n}{n^2} PQ + \frac{n + n_S}{n^2} \cdot 0 - PQ = \frac{n + n_S}{n} \cdot \frac{-PQ}{n} \tag{12}$$

Therefore, the variance of $\hat{d}$ computed with delta method for nonindependent sites should be $(n + n_s)/n$ times larger than the variance of $\hat{d}$ computed for independent sites. In case of the srRNA sequence of *Saccharomyces cerevisiae*, about 900 out of the 1,572 sites used participate in the hairpin stem regions (Rubtsov et al. 1980; Georgiev et al. 1981). If this number is assumed for all the rRNA sequences in this study, the variance of the distance estimates will increase by about 1.58 times. As noted before, the confidence probability values obtained using these variances are likely to be conservative.

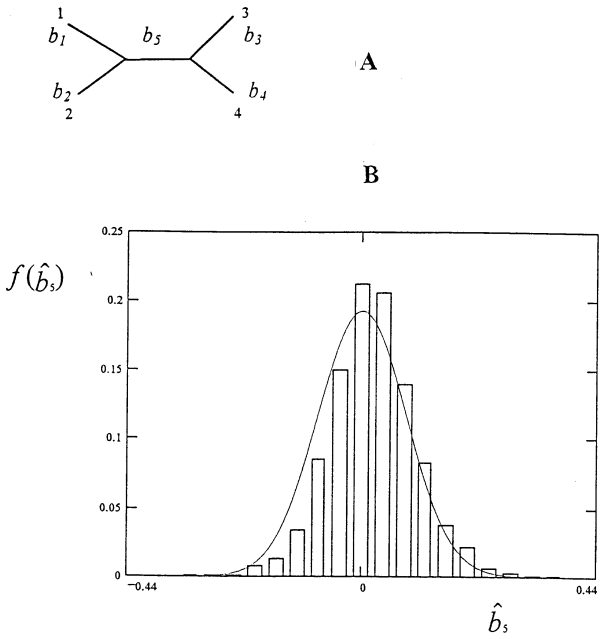

Fig. 2. **A** An unrooted model tree used to compute the distribution of the estimate of interior branch, $\hat{b}_5$. **B** The distribution of $\hat{b}_5$ (histogram) as obtained from 30,000 simulation replications where Kimura's two-parameter model (transition/transversion = 10) with gamma-distributed substitution rates ($a = 0.76$) was used. A normal distribution with the same mean and variance is shown in *solid line*. The fit is expected to improve as the number of sequences increases. For this simulation we used $b_1 = b_2 = b_3 = b_4 = 0.3$ and $b_5 = 0.001$ (T.L. Sitnikova, personal communication).

Phylogenetic Analysis. We first searched for the minimum evolution tree for the compiled data set of 210 eukaryotic and 4 bacterial sequences that were 1,572 nucleotides long (Rzhetsky and Nei 1993, 1994a). The evolutionary distances were estimated using equation (7) with $a = 0.76$ and the pairwise-deletion option (Jin and Nei 1990; Kumar et al. 1993). In this tree, animals, land plants, green algae, true fungi, heterokonts, and alveolates (except *Plasmodium*) formed monophyletic groups (Fig. 1C). The monophyletic relationships were robust to values of a from 0.7 to 0.8. However, if the rate variation across sites is ignored (*i.e.*, $a = \infty$), animals ceased to be monophyletic; see also Fig. 1 of Cavalier-Smith (1993).

We used the four-cluster analysis (Rzhetsky et al. 1995) to evaluate the statistical confidence of the branching order of major eukaryotic groups as implemented in the PHYLTEST program (Kumar 1995). In the four-cluster analysis, all three possible topologies (X, Y, and Z) relating four monophyletic groups A, B, C, and D (containing n_A, n_B, n_C, and n_D sequences, respectively) are compared directly. If S_X, S_Y, and S_Z are sums of branch lengths for trees X, Y, and Z, respectively (corresponding to cluster arrangements [[A,B][C,D]], [[A,C][B,D]], and [[A,D][B,C]], respectively), and if $S_X - S_Y < 0$ and $S_X - S_Z < 0$, then X is most likely to be the correct tree according to the minimum evolution principle. For any two trees (say, I and II), the difference $S_I - S_{II}$ and its variance, $V(S_I - S_{II})$, are computed from the estimates of evolutionary distances, d_{ij}'s, and their variances and covariances. The statistical significance of the difference $S_I - S_{II}$ is then obtained by a normal deviate test since the distribution of $S_I - S_{II}$ is approximately normal (Fig. 2). To account for the nonindependence of sites involved in hairpin stem regions in the srRNA secondary structure, the variance of the difference $S_I - S_{II}$ was increased by 1.58 times, as explained above.

For protein-coding genes, we conducted phylogenetic analyses using the neighbor-joining method in the MEGA package (Kumar et al. 1994). For estimating pairwise distances, Poisson correction and

gamma distances were employed. The phylogenetic trees obtained using these two distance measures were almost identical. Furthermore, these trees were quite similar to those obtained by Baldauf and Palmer (1993) with parsimony analysis. In addition to the bootstrap test, we conducted four-cluster analysis to evaluate the reliability of the higher-order branchings in the phylogenetic trees reconstructed.

Results and Discussion

The phylogenetic framework presented in Fig. 1C shows a preferred branching order of major eukaryote lineages as inferred from the srRNA gene. In this tree, animals with three germ layers (triploblasts) are monophyletic (Lake 1990), but a clear-cut evolutionary dichotomy between diploblasts and triploblasts was not observed. Land plants and green algae formed the green-plants lineage. As expected, the true fungi appeared as a natural clade consisting of Ascomycetes, Basidiomycetes, Chytridiomycetes, and Zygomycetes; and, pseudofungi (oomycetes) were only remotely related to them (Cavalier-Smith 1987b). Several photosynthetic and heterotrophic protist taxa, characterized by the presence of tripartite tubular hairs on the anterior flagellum, comprised a monophyletic group of heterokonts (Stramenopiles) (Cavalier-Smith 1987a; Patterson 1989; Anderson 1991; Barr 1992; Bhattacharya et al. 1992). Furthermore, groups of ciliates, apicomplexans (except *Plasmodium*), and dinoflagellates also were monophyletic. These monophyletic groups have been independently recognized on the basis of other molecular and morphological data, and, thus, we assumed their monophyly for the purpose of the statistical analysis (Corliss 1984; Margulis and Schwartz 1988; Sogin 1989; Lake 1990; Patterson and Sogin 1992; Cavalier-Smith 1993).

The bootstrap method provides one way of assessing the reliability of branching patterns in a phylogenetic tree (Felsenstein 1985; Cavalier-Smith 1993; Wainright et al. 1993). However, in the computation of the statistical confidence for eukaryotic relationships when using srRNA sequences, we needed to take into account (1) the nonindependence of evolutionary changes in stem regions of the srRNA gene and (2) monophyly of various eukaryotic clades, in addition to the heterogeneity of evolutionary rates among sites. Moreover, we analyzed a large data set that is not limited by the narrow representation of major eukaryotic groups (Table 1). For these reasons, we used computationally efficient four-cluster analysis, which allowed us to incorporate such biological information in the comparisons of alternative phylogenetic hypotheses (Rzhetsky et al. 1995). In this analysis, the estimation of actual evolutionary relationships among the groups considered is not affected by the phylogenetic relationships within each group if these groups are indeed monophyletic (Rzhetsky et al. 1995).

The data set analyzed in this paper included most of the major eukaryotic groups (Table 1) and, therefore, a large number of four-cluster analyses were conducted. Only results for which the confidence probabilities (CP value) were higher than 90% were generally used in drawing conclusions. Since the number of all possible four-cluster analyses was enormous for the present data set, we reduced the number of possible combinations by first statistically examining the sister-group relationships of the animals and fungi and that of the land plants and green algae. (This approach was taken in further analysis also.) If these relationships are statistically supported, we reduce the number of groups by two by forming two clades consisting of two groups each. Because the four-cluster analysis is an unrooted analysis, we established the sister-group relationships of the lineages in the crown of the eukaryotic tree in two ways: (1) by conducting a series of nested four-cluster analyses and (2) by using early diverging eukaryotic lineages as the fourth group (referred to as the rooted analysis). In the latter approach, we found that the comparisons involving Bacteria as the fourth group generally produced CP values much lower than the critical cutoff value, and whenever the results of four-cluster analyses were statistically significant, the internal branch length was found to be significantly positive in two alternative trees. This is probably because the srRNA genes are under different evolutionary constraints in prokaryotes and because Bacteria are the most distantly related outgroup to the crown taxa. Therefore, taxa that appeared to have evolved early in the eukaryote tree in Fig. 1C were mainly used in the rooted four-cluster analyses.

Results of the nested and rooted four-cluster analyses in Table 2A indicate that the land plants and green algae together constitute an independent lineage. Somewhat lower statistical confidence values for this relationship are observed in the rooted four-cluster analyses because the presumed outgroup taxa show large genetic divergences from the ''crown'' groups. Results of the nested and rooted four-cluster analyses reported in Table 2B and C show that the animals and true fungi are closer to each other than either is to red algae, heterokonts, or alveolates. However, rooted four-cluster analyses for obtaining a preferred branching order of animal, true fungi, and green-plant lineages did not resolve the animal, true fungi, and green-plant trichotomy with high statistical confidence (Table 2C). As mentioned above, this is probably because the eukaryotic lineages that fall outside the crown of the tree show much larger distances than those among the ''crown'' groups (see Fig. 1C). Therefore, if we assume that heterokonts (or alveolates) constitute an outgroup lineage, the animals and true fungi are closer to each other than either is to green plants (Table 2B). The closer relationship of animals to true fungi (than to green plants) is also reflected in similarities in characteristics other than those defined by polynucleotides, including the utilization and synthesis of chitin, storage of food in the form of glycogen (not starch), synthesis of hemes

Table 2. Comparison of alternative branching orders in the ''crown'' of the eukaryote tree[a]

Taxa[b]	Best tree	Alternative trees			
		I	CP_I	II	CP_{II}
A) Relationship of land plants (L) and green algae (G)[c]					
Animals (A) and fungi (F)	([L,G],[A,F])	([L,A],[G,F])	99%	([L,F],[G,A])	99%
Animals (A) and red algae (R)	([L,G],[A,R])	([L,A],[G,R])	99%	([L,R],[G,A])	99%
Fungi (F) and red algae (R)	([L,G],[F,R])	([L,F],[G,R])	99%	([L,R],[G,F])	97%
Red algae (R) and heterokonts (H)	([L,G],[R,H])	([L,R],[G,H])	99%	([L,H],[G,R])	99%
Animals (A) and **kinetoplastids** (K)	([L,G],[A,K])	([L,A],[G,K])	86%	([L,K],[G,A])	99%
Fungi (F) and **kinetoplastids** (K)	([L,G],[F,K])	([L,F],[G,K])	86%	([L,K],[G,F])	99%
Red algae (R) and **kinetoplastids** (K)	([L,G],[R,K])	([L,R],[G,K])	76%	([L,K],[G,R])	91%
Heterokonts (H) and **kinetoplastids** (K)	([L,G],[H,K])	([L,H],[G,K])	99%	([L,K],[G,H])	99%
Dinoflagellates (D) and **kinetoplastids** (K)	([L,G],[D,K])	([L,D],[G,K])	99%	([L,K],[G,D])	99%
B) Relationship of animals (A) and fungi (F) with other ''crown'' groups					
Green plants (P) and red algae (R)	([A,F],[P,R])	([A,P],[F,R])	79%	([A,R],[F,P])	55%
Green plants (P) and heterokonts (H)	([A,F],[P,H])	([A,P],[F,H])	94%	([A,H],[F,P])	98%
Green plants (P) and dinoflagellates (D)	([A,F],[P,D])	([A,P],[F,D])	94%	([A,D],[F,P])	99%
Green plants (P) and apicomplexans (I)	([A,F],[P,I])	([A,P],[F,I])	86%	([A,I],[F,P])	93%
Red algae (R) and heterokonts (H)	([A,F],[R,H])	([A,R],[F,H])	86%	([A,I],[F,R])	97%
Red algae (R) and dinoflagellates (D)	([A,F],[R,D])	([A,R],[F,D])	84%	([A,D],[F,R])	91%
Heterokonts (H) and dinoflagellates (D)	([A,F],[H,D])	([A,H],[F,D])	99%	([A,D],[F,H])	99%
Heterokonts (H) and apicomplexans (I)	([A,F],[H,I])	([A,H],[F,I])	99%	([A,I],[F,H])	99%
Heterokonts (H) and ciliates (C)	([A,F],[H,C])	([A,H],[F,C])	96%	([A,C],[F,H])	99%
C) Relationship of animals (A) and fungi (F) with other crown groups in rooted comparisons					
Green plants (P) and **Parabasalia** (S)	([A,F],[P,S])	([A,P],[F,S])	91%	([A,S],[F,P])	48%
Red algae (R) and **Euglena** (E)	([A,F],[R,E])	([A,R],[F,E])	88%	([A,E],[F,R])	89%
Red algae (R) and **Parabasalia** (S)	([A,F],[R,S])	([A,R],[F,S])	98%	([A,S],[F,R])	97%
Heterokonts (H) and **kinetoplastids** (K)	([A,F],[H,K])	([A,H],[F,K])	95%	([A,K],[F,H])	95%
Heterokonts (H) and **Euglena** (E)	([A,F],[H,E])	([A,H],[F,E])	98%	([A,E],[F,H])	95%
Heterokonts (H) and **Parabasalia** (S)	([A,F],[H,S])	([A,H],[F,S])	99%	([A,S],[F,H])	96%
Dinoflagellates (D) and **Euglena** (E)	([A,F],[D,E])	([A,D],[F,E])	93%	([A,E],[F,D])	92%
Apicomplexans (I) and **kinetoplastids** (K)	([A,F],[I,K])	([A,I],[F,K])	92%	([A,K],[F,I])	92%
Apicomplexans (I) and **Euglena** (E)	([A,F],[I,E])	([A,I],[F,E])	99%	([A,E],[F,I])	99%
Apicomplexans (I) and **Parabasalia** (S)	([A,F],[I,S])	([A,I],[F,S])	95%	([A,S],[F,I])	94%
D) Relationship of animals–fungi clade (N) and green plants (P)					
Red algae (R) and heterokonts (H)	([N,P],[R,H])	([N,R],[P,H])	22%	([N,H],[P,R])	49%
Heterokonts (H) and dinoflagellates (D)	([N,P],[H,D])	([N,H],[P,D])	99%	([N,D],[P,H])	99%
Heterokonts (H) and apicomplexans (I)	([N,P],[H,I])	([N,H],[P,I])	99%	([N,I],[P,H])	99%
Red algae (R) and **Euglena** (E)	([N,P],[R,E])	([N,R],[P,E])	86%	([N,E],[P,R])	85%
Red algae (R) and **Parabasalia** (S)	([N,P],[R,S])	([N,R],[P,S])	92%	([N,S],[P,R])	93%
Heterokonts (H) and **kinetoplastids** (K)	([N,P],[H,K])	([N,H],[P,K])	99%	([N,K],[P,H])	90%
Heterokonts (H) and **Euglena** (E)	([N,P],[H,E])	([N,H],[P,E])	97%	([N,E],[P,H])	96%
Heterokonts (H) and **Parabasalia** (S)	([N,P],[H,S])	([N,H],[P,S])	93%	([N,S],[P,H])	96%
Dinoflagellates (D) and **Euglena** (E)	([N,P],[D,E])	([N,D],[P,E])	90%	([N,E],[P,D])	89%
Apicomplexa (I) and **kinetoplastid** (K)	([N,P],[I,K])	([N,I],[P,K])	99%	([N,K],[P,I])	98%
E) Relationship of animals–fungi–green plants (Z) and red algae (R)					
Heterokonts (H) and dinoflagellates (D)	([Z,P],[H,D])	([Z,H],[P,D])	91%	([Z,D],[P,H])	97%
Heterokonts (H) and apicomplexans (I)	([Z,P],[H,I])	([Z,H],[P,I])	90%	([Z,I],[P,H])	95%
F) Relationship of alveolates: dinoflagellates (D), apicomplexans (I), and ciliates (C)					
Animals (A)	([D,I],[C,A])	([D,C],[I,A])	96%	([D,A],[I,C])	96%
Fungi (F)	([D,I],[C,F])	([D,C],[I,F])	99%	([D,F],[I,C])	99%
Green plants (P)	([D,I],[C,P])	([D,C],[I,P])	96%	([D,P],[I,C])	91%
Heterokonts (H)	([D,I],[C,H])	([D,C],[I,H])	98%	([D,H],[I,C])	95%

[a] Statistical confidence was obtained by adjusting for extreme site dependence in stem regions and by considering variation in functional constraints in srRNA gene. Confidence probabilities (CP) are expressed as $100 \cdot (1 - P)$%, where P is the probability value obtained in a two-tailed normal deviate test. Most of the four-cluster combinations were examined, but results for which CP $\geq$ 90% are presented, except where low statistical resolution is intended to be shown. CP < 90% are underlined

[b] In rooted analyses the outgroup is shown in bold

[c] Many other statistically significant results using Euglena, Metamonada, and Parabasalia as outgroups are not shown

through δ-aminolevulinic acid (not glutamate) pathway, production of sterol using lanosterol (not cycloartenol), and the use of the codon UGA to specify the amino acid tryptophan in mitochondria (not a stop codon) (Corliss 1984; Cavalier-Smith 1987b; Ragan 1989; Sleigh 1989). In fact, virtually no characters exist that clearly unite green plants with true fungi or animals.

Both choanoflagellate species in our study occupied a

Table 3. Evolutionary relationships of diploblastic and triploblastic animals[a]

Taxa	Best tree	Alternative trees			
		I	CP_I	II	CP_{II}
A) Relationships within diploblastic lineages; land plants (P) used as outgroups					
Porifera (R), Cnidaria (C), Placozoa (L)	([P,R],[C,L])	([P,C],[R,L])	99%	([P,L],[R,C])	99%
Porifera (R), Cnidaria (C), Ctenophore (E)	([P,R],[C,E])	([P,C],[R,E])	41%	([P,E],[R,C])	93%
Porifera (R), Ctenophore (E), Placozoa (L)	([P,R],[E,L])	([P,E],[R,L])	98%	([P,L],[R,C])	94%
Ctenophore (E), Cnidaria (C), Placozoa (L)	([P,E],[C,L])	([P,C],[E,L])	0%	([P,L],[E,C])	88%
B) Relationships of different diploblastic lineages with triploblasts (T), using land plants (P) as an outgroup					
Cnidaria (C), Placozoa (L)	([P,T],[C,L])	([P,C],[T,L])	95%	([P,L],[T,C])	96%
Cnidaria (C), and Ctenophore (E)	([P,T],[C,E])	([P,C],[T,E])	92%	([P,E],[T,C])	85%
Cnidaria (C) and Porifera (R)	([P,T],[C,R])	([P,C],[T,R])	16%	([P,R],[T,C])	93%
Placozoa (L) and Ctenophore (E)	([P,T],[L,E])	([P,L],[T,E])	98%	([P,E],[T,L])	97%
Placozoa (L) and Porifera (R)	([P,T],[L,R])	([P,L],[T,R])	29%	([P,R],[T,L])	3%
Ctenophore (E) and Porifera (R)	([P,T],[E,R])	([P,E],[T,R])	97%	([P,R],[T,E])	97%

[a] See note to Table 2

position between the true fungi and animals clades (Wainright et al. 1993). However, the sister-taxa relationship of choanoflagellate with animals was not statistically established (results not shown). Choanoflagellates share a number of morphological features with sponges (see Wainright et al. 1993) as well as with the primitive fungal group Chytridiomycetes (Cavalier-Smith 1993). Just as do animals and true fungi, choanoflagellates have mitochondria with flattened cristae and lack plastids. In our tree (Fig. 1C), a short interior branch that is not significantly different from zero separates choanoflagellates and true fungi. This indicates that animals, fungi, and choanoflagellates may have diverged from a common ancestor, probably a choanomonad protist, in a relatively short period of time (Cavalier-Smith 1987b; Cavalier-Smith 1993). Within animals, our analyses were inconclusive regarding the evolutionary dichotomy of triploblasts and diploblasts as well as concerning the evolutionary relationships of the four diploblastic lineages (Table 3). Within true fungi, Chytridiomycetes appeared to be of most ancient origin; however, this result was not statistically supported in all the four-cluster analyses (results not shown). Furthermore, results presented in Table 2D suggest that the animal–fungus and green-plant lineages are closer to each other than either is to red algae, heterokonts, or alveolates.

Some previous phylogenetic analyses have suggested that heterokonts shared a most recent common ancestor with animals, true fungi, and green plants (Sogin 1989; Bhattacharya et al. 1992; Cavalier-Smith 1993). In Fig. 1C, however, the red algae lineage appears to be closest to the animal, fungus, and green-plant lineages. In the four-cluster analysis, the red algae and animal–fungus–green-plant lineages are indeed closer to each other than either is to heterokonts or alveolates (Table 2E). The apparent difference in the placement of red algae lineage between our study and others probably occurs because the evolutionary divergences between the red algae and other higher eukaryotes were severely underestimated in previous studies in which the rate heterogeneity among sites was not considered. However, results of rooted four-cluster analysis were inconclusive regarding the order in which red algae, heterokonts, and alveolates branched off. In the srRNA tree, the red algae clade is separated from animal–fungus and green-plant clades by a rather short branch, and, thus, these results need to be verified with additional data from other genes. In any case, animals, fungi, green plants, and red algae are similar in the presence of plate-like cristae in mitochondria (not tubular as in most heterokonts) and the use of α1–4 glucan for energy storage (not β1–3 glucan as in heterokonts). Furthermore, both red algae and green plants have plastids located in cytosol, not in the rough endoplasmic reticulum and enclosed by a periplast membrane, as in photosynthetic heterokonts.

In the four-cluster analysis, dinoflagellates and apicomplexans share a most recent common ancestor (Table 2F) to the exclusion of ciliates. It appears that these two major alveolate lineages have diverged in a relatively short period of time (Gajadhar et al. 1991; Wolters 1991). Furthermore, ciliates and the group of dinoflagellates and apicomplexans were significantly closer to each other than to animals, true fungi, green plants, red algae, or heterokonts in the nested four-cluster analyses (results not shown). However, attempts to identify monophyly of alveolates by rooted four-cluster analysis produced results that were not statistically significant, except in two cases: (1) Ciliates appear to be closer to true fungi than to the group of dinoflagellates and apicomplexans when using Euglena as an outgroup (CP = 91%), and (2) ciliates and the group of dinoflagellates and apicomplexans cluster together when red algae and Parabasalia are the other two groups in the four-cluster analysis (CP = 97%). Monophyly of alveolates (including *Plasmodium*) has been shown with high bootstrap *P* values previously (e.g., 96% in Cavalier-Smith 1993). However, oversimplified models of nucleotide substitution without the consideration of variability in substitution rates among sites

Table 4. Comparison of alternative branching orders in early diverging eukaryotic lineages[a]

Taxa	Best tree	Alternative trees			
		I	CP_I	II	CP_{II}
A) Relationship of kinetoplastids (K) and *Euglena* (E), using Bacteria (B) as an outgroup					
Microsporidia (M)	([K,E],[M,B])	([K,M],[E,B])	96%	([K,B],[M,E])	96%
Parabasalia (S)	([K,E],[S,B])	([K,S],[E,B])	99%	([K,B],[S,E])	95%
B) Relationship of *Dictyostelium* (D) and Entamoeba (E), using Bacteria (B) as an outgroup					
Kintoplastids + *Euglena* (K)	([D,E],[K,B])	([D,K],[E,B])	97%	([D,B],[E,K])	97%
Parabasalia (P)	([D,E],[P,B])	([D,P],[E,B])	99%	([D,B],[E,P])	99%
Microsporidia (I)	([D,E],[I,B])	([D,I],[E,B])	97%	([D,B],[E,I])	99%
Metamonada (M)	([D,E],[M,B])	([D,M],[E,B])	99%	([D,B],[E,M])	99%
C) Earliest-diverging eukaryotic lineage: Metamonada (M), using Bacteria (B) as a root					
Microsporidia (I) and Heterolobosea (L)	([M,B],[I,L])	([M,I],[B,L])	99%	([M,L],[B,I])	68%
Microsporidia (I) and Parabasalia (S)	([M,B],[I,S])	([M,I],[B,S])	96%	([M,S],[B,I])	50%
Microsporidia (I) and Entamoeba (E)	([M,B],[I,E])	([M,I],[B,E])	96%	([M,E],[B,I])	96%
Heterolobosea (L) and Parabasalia (S)	([M,B],[L,S])	([M,L],[B,S])	73%	([M,S],[B,L])	97%
Heterolobosea (L) and Euglenozoa (K)	([M,B],[L,K])	([M,L],[B,K])	98%	([M,K],[B,L])	99%
Heterolobosea (L) and Entamoeba (E)	([M,B],[L,W])	([M,L],[B,E])	99%	([M,E],[B,L])	99%
Euglenozoa (K) and Entamoeba (E)	([M,B],[K,T])	([M,K],[B,T])	98%	([M,E],[B,K])	97%

[a] See note to Table 2. Results shown in C were obtained by using only 1,254 sites where at least one of the bacterial sequences contained at least one unambiguous nucleotide. CP values obtained were somewhat smaller than those produced when all 1,572 sites were used. Euglenozoa is the group of kinetoplastids and *Euglena*

was used in such studies. The bootstrap *P* value is directly influenced by the variance of the distance estimator used. Because the Jukes-Cantor distance with *uniform* rates among sites has much smaller variance than the model used in the present study and because the non-independence of evolution in stem sites was ignored in previous studies, the confidence in the groupings observed may be overstated in those studies. Severe underestimation of evolutionary distances when using simple models may also be responsible for differences in the placement of *Plasmodium* (an apicomplexan) in our srRNA tree and in other studies. In four-cluster analyses rooted by using Bacteria, the topology in which *Plasmodium* is closer to alveolates is almost always better than the alternatives when only protistan lineages that fall outside the crown of the eukaryote tree are considered. However, the statistical resolution becomes very low when the crown groups are included in these rooted four-cluster analyses.

In the srRNA tree, the heterokonts and alveolates (Gajadhar et al. 1991) evolved prior to the divergence of red algae and the animal–fungus–green-plant lineages. But, the order in which the heterokonts and alveolates branched off in the eukaryote tree remains uncertain because of the high genetic divergences between the early diverging protistan lineages and the ''crown'' groups.

In addition to heterokonts, our data set contained two other members of kingdom Chromista: haptophyta and cryptomonadea. In four-cluster analyses, these two groups appeared to be closer to heterokonts than to animals, true fungi, green plants, red algae, or alveolates. However, the statistical significance for these relationships was less than the critical 90% value and the results from the four-cluster analysis were inconclusive regarding the monophyly of all members of Chromista.

Cellular slime mold and other protist lineages seem to have diverged prior to the diversification of alveolates. In the srRNA tree, kinetoplastids formed an independent lineage with euglena (Kivic and Walne 1984) (Table 4A) and *Dictyostelium* appear to be closer to Entamoeba than to other protists (Table 4B). Regarding the position of *Dictyostelium*, it has been suggested following an analysis of multiple protein data sets that the animals–fungi clade is closer to *Dictyostelium* than to plants (Kuma et al. 1995). However, our analysis strongly supports a closer association of the group of animals and true fungi to green plants than to *Dictyostelium* (CP = 99%) when kinetoplastids, Parabasalia, or Microsporidia are used for rooting the four-cluster analysis. In fact, none of the protein sequence data sets in Kuma et al.'s (1995) study appear to support a unique tree which is significantly better than the alternatives at the 1 standard error significance level, indicating a lack of phylogenetic resolution in those data.

In order to determine the earliest-diverging eukaryotic lineage we conducted four-cluster analyses using Bacteria as outgroup and found that metamonads were likely to be the first protistan lineage to diverge (Table 4C; see also Cavalier-Smith 1993). However, statistically insignificant results from the four-cluster analyses performed to find a preferred branching order of Microsporidia, Heterolobosea, Parabasalia, Euglenozoa (kinetoplastids and *Euglena*), and the *Dictyostelium* and *Entamoeba* group indicated that their relative branching order may be difficult to elucidate from srRNA sequences (Fig. 1C).

The phylogenetic relationships in the star-like crown (Knoll 1992) of the eukaryote tree are becoming clearer with the use of molecular sequence data. Here, we have shown that the branching order (alveolates, heterokonts,

(red algae, green plants, (animals, fungi))) as deduced from the analysis of srRNA sequences is preferred. But sometimes a gene phylogeny may not reflect the actual species tree (Pamilo and Nei 1988). The closer relationship of animals to true fungi than to green plants has been generally supported in analyses of amino acid sequences of various protein-coding genes (Baldauf and Palmer 1993; Nikoh et al. 1994). Our four-cluster analyses with amino acid sequence data from translation elongation factor α, actin, and α- and β-tubulin genes also supported a closer association of animal with fungi than with plants (CP > 90%); however, see Gouy and Li (1989) and Sidow and Thomas (1994). But the phylogenetic trees obtained from these data did not place animals–fungi and green plants as sister taxa, as suggested by the srRNA gene. In fact, the higher-order relationships of the eukaryotes as obtained from these protein-coding genes were incongruent and statistically unresolved in the bootstrap test as well as the four-cluster analyses. In these phylogenies, well-defined groups such as the group of ciliates were not monophyletic (see also Fig. 1 in Baldauf and Palmer 1993). In addition, for these and other protein-coding genes, sequence data from only a limited number of eukaryotic lineages are available. In the future, complete sequences of nuclear genes from a wide variety of eukaryote lineages will allow more comprehensive phylogenetic analyses. Until then, evolutionary inferences from the srRNA gene will continue to provide some valuable insights in eukaryote phylogeny for use in molecular evolutionary and systematic studies.

Acknowledgments. We thank M.L. Sogin and S. Baldauf for providing some of the sequence data used in the analysis, Drs. C.A. Hass, M.L. Sogin, T. Cavalier-Smith, S. Muse, S.B. Hedges, and M. Nei for helpful suggestions, and N. Takezaki and T.L. Sitnikova for assistance. We also thank two anonymous reviewers whose comments led to significant improvements in the contents and the presentation of results. This research was supported with grants from the NSF and NIH to M. Nei.

References

Anderson RA (1991) The cytoskeleton of chromophyte algae. Protoplasma 164:143–159

Baldauf SL, Palmer JD (1993) Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. Proc Natl Acad Sci USA 90:11558–11562

Barr DJS (1992) Evolution and kingdoms of organisms from the perspective of a mycologist. Mycologia 84:1–11

Bhattacharya D, Medlin L, Wainright PO, Ariztia EV, Claude B, Stickel SK, Sogin ML (1992) Algae containing chlorophyles $a + c$ are paraphyletic: molecular evolutionary analysis of the chromophyta. Evolution 46:1801–1817

Butterfield NJ, Knoll AH, Swett K (1988) Exceptional preservation of fossils in an upper proterozoic shale. Nature 334:424–427

Butterfield NJ, Knoll AH, Swett K (1990) A Bangiophyte red alga from the proterozoic of arctic Canada. Science 250:104–107

Cavalier-Smith T (1987a) The origin of eukaryote and archaebacterial cells. Ann NY Acad Sci 503:17–54

Cavalier-Smith T (1987b) The origin of fungi and pseudofungi. In: Rayner ADM, Brasier CM, Moore D (eds) Evolutionary biology of the Fungi. Cambridge University Press, Cambridge, pp 339–353

Cavalier-Smith T (1993) Kingdom protozoa and its 18 phyla. Microbiol Rev 57:953–994

Corliss JO (1984) The kingdom Protista and its 45 phyla. Biosystems 17:87–126

Elwood HJ, Olsen GJ, Sogin ML (1985) The small subunit ribosomal RNA gene sequences from the hypotricous ciliates *Oxytricha nova* and *Stylonchia pustulata.* Mol Biol Evol 2:399–410

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791

Gajadhar AA, Marquardt WC, Hall R, Gunderson J, Ariztia-Carmona EV, Sogin ML (1991) Ribosomal RNA sequences of *Sarcocystis muris, Theileria annulata,* and *Crypthecodinium cohnii* reveal evolutionary relationships among apicomplexans, dinoflagellates, and ciliates. Mol Biochem Parasitol 45:147–154

Georgiev OI, Nikolaev N, Hadjiolov AA, Skryabin KG, Zakharyev VM, Bayev AA (1981) Secondary structure of the yeast ribosomal RNA genes. 4. Complete sequence of the 25S rRNA gene from *Saccharomyces cerevisiae.* Nucleic Acids Res 9:6953–6958

Golding GB (1983) Estimates of DNA and protein sequence divergence: an examination of some assumptions. Mol Biol Evol 1:125–142

Gouy M, Li W-H (1989) Molecular phylogeny of the kingdoms animalia, plantae, and fungi. Mol Biol Evol 6:109–122

Hasegawa M, Hashimoto T, Adachi J (1992) Origin and evolution of eukaryotes as inferred from protein sequence data. In: Hartman H, Matsuno K, (eds) The origin and evolution of the cell. World Scientific, New Jersey, pp 107–130

Hendriks L, De Baere R, Van de Peer Y, Neefs J, Goris A, De Wachter R (1991) The evolutionary position of the rhodophyte *Prophyra umbilicalis* and the basidiomycete *Leucosporidium scottii* among other eukaryotes as deduced from compute sequences of small ribosomal subunit RNA. J Mol Evol 32:167–177

Jin L, Nei M (1990) Limitations of the evolutionary parsimony method of phylogenetic analysis. Mol Biol Evol 7:82–102

Kimura M (1980) A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol 16:111–120

Kivic PA, Walne PL (1984) An evaluation of a possible phylogenetic relationship between the Euglenophyta and Kinetoplastida. Orig Life 13:269–288

Knoll AH (1992) The early evolution of eukaryotes: a geological perspective. Science 256:622–627

Kuma K, Nikoh N, Iwabe N, Miyata T (1995) Phylogenetic position of *Dictyostelium* inferred from multiple protein data sets. J Mol Evol 41:238–246

Kumar S (1995) PHYLTEST: phylogeny hypothesis testing software. The Pennsylvania State University, University Park

Kumar S, Tamura K, Nei M (1993) Manual for MEGA: molecular evolutionary genetic analysis software. Pennsylvania State University, University Park

Kumar S, Tamura K, Nei M (1994) MEGA: molecular evolutionary genetics analysis software for microcomputers. Comput Appl Biosci 10:189–191

Lake JA (1990) Origin of the metazoa. Proc Natl Acad Sci USA 87: 763–766

Larson N, Olsen GJ, Maidak BL, McCaughey MJ, Overbeek R, Macke TJ, Marsh TL, Woese CR (1993) The ribosomal database project. Nucleic Acids Res 21:3021–3023

Margulis L, Schwartz KV (1988) Five kingdoms. WH Freeman and Company, New York

Nikoh N, Hayase N, Iwabe N, Kuma K, Miyata T (1994) Phylogenetic relationships of the kingdoms Animalia, Plantae, and Fungi as inferred from 23 different protein species. Mol Biol Evol 11:762–768

Noller HF, Asire M, Barta A, Douthwaite S, Goldstein T, Gutell RR, Moazed D, Normanly J, Prince JB, Stern S, Triman K, Turner S, Von Stolk B, Wheaton V, Weiser B, Woese CR (1986) Studies on

the structure and function of ribosomal RNA. In: Hardesty B, Kramer G (eds) Structure, function, and genetics of ribosomes. Springer Verlag, Berlin, pp 141–163

Pamilo P, Nei M (1988) Relationships between gene trees and species trees. Mol Biol Evol 5:568–583

Patterson DJ (1989) Stramenopiles: chromophytes from a protistan perspective. In: Green JC, Leadbeater BSC, Diver WL (eds) The chromophyte algae: problems and perspectives. Clarendon Press, Oxford, pp 357–379

Patterson DJ, Sogin ML (1992) Eukaryote origins and protistan diversity. In: Hartman H, Matsuno K (eds) The origin and evolution of the cell. World Scientific, New Jersey, pp 13–46

Ragan MA (1989) Biochemical pathways and the phylogeny of the eukaryotes. In: Fernholm B, Bremer K, Jornvall H (eds) The hierarchy of life. Elsevier Science Publisher, pp 145–160

Rubtsov PM, Musakhanov MM, Zakharyev VM, Karayev AA, Skryabin KG, Bayev AA (1980) The structure of the yeast ribosomal NA genes. I. The complete nucleotide sequence of the 18S ribosomal RNA gene from *Saccharomyces cerevisiae.* Nucleic Acids Res 8:5779–5794

Rzhetsky A (1995) Estimating substitution rates in ribosomal RNA genes. Genetics 141:771–783

Rzhetsky A, Kumar S, Nei M (1995) Four-cluster analysis: a simple method to test phylogenetic hypotheses. Mol Biol Evol 12:163–167

Rzhetsky A, Nei M (1993) Theoretical foundation of the minimum-evolution method of phylogenetic inference. Mol Biol Evol 10: 1073–1095

Rzhetsky A, Nei M (1994a) METREE: a program package for inferring and testing minimum-evolution trees. Comput Appl Biosci 10:409–412

Rzhetsky A, Nei M (1994b) Unbiased estimates of the number of nucleotide substitutions when substitution rate varies among different sites. J Mol Evol 38:295–299

Saitou N, Nei M (1987) The neighbor-joining method: a new method for reconstructing phylogenetic trees. Mol Biol Evol 4:406–425

Sidow A, Thomas WK (1994) A molecular evolutionary framework for eukaryotic model organisms. Curr Biol 4:596–603

Sleigh M (1989) Protozoa and other protists. Edward Arnold, New York

Smothers JF, Dohlen CDV, Smith LHJ, Spall RD (1994) Molecular evidence that the myxozoan protists are metazoans. Science 265: 1719–1721

Sogin ML (1989) Evolution of eukaryotic microorganisms and their small subunit ribosomal RNAs. Am Zool 29:487–499

Tajima F (1993) Unbiased estimate of evolutionary distance between nucleotide sequences. Mol Biol Evol 10:677–688

Uzzell T, Corbin KW (1971) Fitting discrete probability distribution to evolutionary events. Science 253:1503–1507

Van de Peer Y, Neefs J, De Rijk P, De Wechter R (1993) Evolution of eukaryotes as deduced from small ribosomal subunit RNA sequences. Biochem Syst Ecol 21:43–55

Wainright PO, Hinkle G, Sogin ML, Stickel SK (1993) Monophyletic origins of metazoa: an evolutionary link with fungi. Science 260: 340–342

Wakeley J (1993) Substitution rate variation among sites in hypervariable region I of human mitochondrial DNA. J Mol Evol 37:613–623

Woese CR (1987) Archaebacteria. Sci Am 260:98–122

Wolters J (1991) The troublesome parasites–molecular and morphological evidence that Apicomplexa belong to the dinoflagellate-cliate clade. Biosystems 25:75–83

Zharkikh AA, Rzhetsky A, Morosov PS, Sitnikova TL, Krushkal JS (1991) VOSTORG: a package of microcomputer programs for sequence analysis and construction of phylogenetic trees. Gene 101: 251–254