J Mol Evol (1996) 42:201–210

JOURNAL OF MOLECULAR EVOLUTION

The Evolution of Stramenopiles and Alveolates as Derived by "Substitution Rate Calibration" of Small Ribosomal Subunit RNA

Yves Van de Peer, Gert Van der Auwera, Rupert De Wachter

Departement Biochemie, Universiteit Antwerpen (UIA), Universiteitsplein 1, B-2610 Antwerpen, Belgium

Received: 26 June 1995 / Accepted: 15 September 1995

Abstract. The substitution rate of the individual positions in an alignment of 750 eukaryotic small ribosomal subunit RNA sequences was estimated. From the resulting rate distribution, an equation was derived that gives a more precise relationship between sequence dissimilarity and evolutionary distance than hitherto available. Trees constructed on the basis of evolutionary distances computed by this new equation for small ribosomal subunit RNA sequences from ciliates, apicomplexans, dinoflagellates, oomycetes, hyphochytriomycetes, bicosoecids, labyrinthuloids, and heterokont algae show a more consistent tree topology than trees constructed in the absence of "substitution rate calibration." In particular, they do not suffer from anomalies caused by the presence of extremely long branches.

Key words: Small ribosomal subunit RNA — Evolutionary distance — Substitution rate calibration — Stramenopiles — Alveolates

Introduction

Since the collection of available small ribosomal subunit RNA (SSU rRNA) sequences has expanded enormously during the past few years, it becomes possible to construct large evolutionary trees, including hundreds of sequences. However, for tree construction methods such as maximum parsimony and maximum likelihood this is problematic due to the combinatorial explosion of the number of possible tree topologies. Some distance methods, like clustering or neighbor-joining, do not suffer from this problem. The neighbor-joining algorithm (Saitou and Nei 1987), which is very fast and thus allows the construction of trees including several hundreds of sequences, also appears to be very effective in recovering the true tree topology (Sourdis and Nei 1988; Saitou and Imanishi 1989).

Nevertheless, the reliability of phylogenetic trees based on distance data depends greatly on the accurate estimation of the evolutionary distances from the observed sequence dissimilarities. The dissimilarity between two aligned sequences is the number of observed substitutions divided by the number of compared positions, whereas the distance is the number of substitutions that actually occurred during divergence divided by the number of compared positions. The latter quantity is larger than the former, except for very small distances. Conversion of dissimilarity into distance can be based on different evolutionary models. The equation of Jukes and Cantor (1969) assumes equal probability for all substitutions. Other equations (e.g., Kimura 1980; Tajima and Nei 1984) take into account the frequencies of different substitution types, but this does not usually have a large influence on tree topology. An important drawback of these models is that they do not take into account differences in evolutionary rate among the different sites of a molecule. In the case of SSU rRNA, the substitution rate was estimated (Van de Peer et al. 1993a) to vary by a factor of more than 1,000 from the least variable to the most variable site. Olsen (1987) had already demon-

Correspondence to: R. De Wachter

strated that application of the Jukes and Cantor correction to sequences composed of sites with unequal substitution rates leads to underestimation of large evolutionary distances. He proposed a different evolutionary model and assumed a log-normal distribution of substitution rates over the sequence positions. Jin and Nei (1990) followed a similar approach but assumed that there is a gamma distribution of substitution rates over the sequence positions. In a recent study of our research group (De Rijk et al. 1995), based on SSU rRNA and large ribosomal subunit RNA (LSU rRNA) sequences, trees based on the latter model of evolution were indeed more consistent with evolutionary hypotheses than trees based on a Jukes and Cantor correction.

Still, none of the known equations for estimating the evolutionary distance is based on an actual measurement of the substitution rates of the different alignment positions. In a previous paper (Van de Peer et al. 1993a), a method was developed to calculate the relative substitution rate of the alignment positions. Nucleotides were then subdivided into subsets, each subset showing a much narrower range of substitution rates than the entire sequence. Dissimilarities were then computed for each subset of positions and converted into evolutionary distances by means of an equation in which a specific parameter valid for that particular subset, and estimated from the data, was introduced. In the present study, we make a new estimate of the substitution rates of individual nucleotides on the basis of a more extensive sequence alignment of eukaryotic SSU rRNAs, and from the resulting rate distribution we derive an equation that gives a more precise relationship between sequence dissimilarity and evolutionary distance. This new method is applied to the evolution of stramenopiles and alveolates.

The stramenopiles, as defined by Patterson (1989), comprise species that either possess evenly spaced tripartite tubular hairs attached to the flagellum or other parts of the cell surface, or species that have been derived from organisms having these structures. Taxa presently included in the stramenopiles are the bicosoecids, labyrinthuloids, oomycetes, hyphochytriomycetes, diatoms, chrysophytes, phaeophytes, xanthophytes, eustigmatophytes, and synurophytes (Leipe et al. 1994). Alveolates (Gajadhar et al. 1991; Patterson and Sogin 1992) comprise the ciliates, dinoflagellates, foraminifers, and apicomplexans and are characterized by the possession of alveoli, which are membrane-bound flattened vesicles or sacs underlying the plasma membrane (Preisig et al. 1994). Phylogenetic reconstructions based on SSU rRNA regularly suggest stramenopiles and alveolates to be sister groups (e.g., Wolters 1991; Wainright et al. 1993; Van de Peer et al. 1993b; Leipe et al. 1994), although bootstrap support for this grouping is usually low. A study combining SSU and LSU rRNA sequence data into an alignment containing over 5,290 positions (Van der Auwera et al. 1995) supported the sisterhood of both groups at a bootstrap level of >95%. Together with the green plants, green algae, fungi, animals, and red algae, stramenopiles and alveolates make up the so-called "crown" (Knoll 1992) of eukaryote evolution.

Table 1. SSU rRNA sequences used in the present study to compute the relative substitution rate of alignment positions, as distributed over the main eukaryotic taxa

Taxon[a]	Number of sequences
Metazoa	169
Fungi[b]	166
Green plants	89
Green algae	95
Red algae	35
Heterokont algae[c]	26
Ciliates	38
Apicomplexa	45
Other protoctists[d]	87

[a] These taxa correspond to major clusters discernable in phylogenetic trees based on SSU rRNA sequence alignments
[b] Included are chytridiomycetes, zygomycetes, ascomycetes, and basidiomycetes
[c] Included are phaeophytes, chrysophytes, bacillariophytes, xanthophytes, and eustigmatophytes
[d] Included are dinoflagellates, kinetoplastids, slime moulds, microsporidia, and diplomonads

Material and Methods

An alignment of SSU rRNA sequences is maintained in our research group, regularly updated, and made publicly available by electronic file transfer by anonymous ftp on host uiam3.uia.ac.be (Van de Peer et al. 1994). In March 1995, this database comprised about 890 complete or nearly complete sequences of eukaryotes. Seven hundred fifty of these sequences were used in the present study to compute the relative substitution rate of the alignment positions. Table 1 shows their distribution over the main eukaryotic taxa. The reduction from about 890 sequences to 750 was due to the elimination of duplicate sequences belonging to the same species and to the elimination of sequences belonging to closely related species of the same genus. All the sequences were aligned on the basis of similarity in primary and secondary structure using the DCSE sequence editor (De Rijk and De Wachter 1993).

Nucleotide substitution rates were estimated as described in Van de Peer et al. (1993a) for each of the alignment positions that are not absolutely conserved and that contain a nucleotide in at least 25% of the aligned sequences. This was done on a VAX-station 3100 (Digital). The inference and drawing of evolutionary trees constructed by neighbor-joining (Saitou and Nei 1987) were done with the software package TREECON (Van de Peer and De Wachter 1993, 1994), which runs on IBM-compatible computers.

Results

Dissimilarity as a Function of Evolutionary Distance in Eukaryotic SSU rRNA

After sequence alignment, the first step in the construction of a distance tree consists of computing a distance

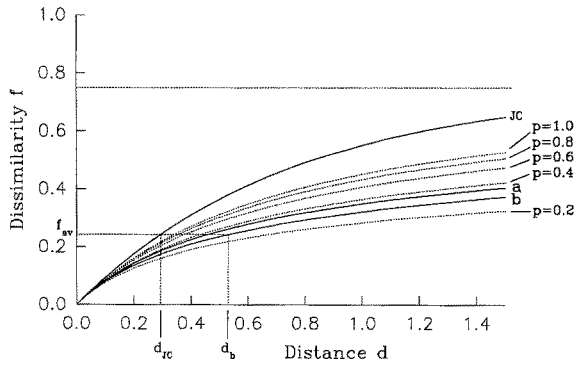

Fig. 1. Graphic representation of functions (1), (3), and (4). Curve *"JC"*: function (1). Curve *"a"*: function (3) computed for the substitution rate spectrum of Fig. 2**a.** Curve *"b"*: function (3) computed for the substitution rate spectrum of Fig. 2**b.** *Dotted curves:* function (4) with values 0.2 to 1.0 for parameter *p.* The average dissimilarity between members of the genus *Plasmodium* and the other Hematozoa is indicated as f_{av}. The corresponding distance is d_{JC} if the relation between dissimilarity f and distance d is expressed by equation (1) (curve *"JC"*), and d_b if the relation between f and d is expressed by equations (3) or (4) (curve *"b"*). See Results for details.

matrix. As a first approximation, the dissimilarity, or fraction of substitutions observed when two nucleotide sequences are compared, increases as a function of their evolutionary distance d according to the equation

$$f = \frac{3}{4}\left[1 - \exp\left(-\frac{4}{3}d\right)\right] \qquad (1)$$

The inverse of equation (1) is the Jukes and Cantor equation used to convert the dissimilarity observed in an alignment into the evolutionary distance listed in distance matrices. The shape of function (1) is shown as curve "JC" in Fig. 1. However, as stated in the Introduction and explained in detail in the Appendix, this equation does not remain applicable if the individual nucleotides of the molecular clock show considerable differences in evolutionary rate, such as is the case for SSU rRNA. In this situation, a different relationship between dissimilarity and distance exists, which can be approximated quite closely, provided that the substitution rate of each nucleotide, relative to the substitution rate of the entire molecule, is known. As demonstrated previously (Van de Peer et al. 1993a), the relative substitution rates of individual nucleotides in SSU rRNA can indeed be estimated from a distance matrix derived from an extensive sequence alignment representing a sufficiently diverse species assortment. In the latter study an alignment of 205 eukaryotic SSU rRNAs was used. For the purpose of the present study, these rates were estimated anew on the basis of a more extensive alignment (Van de Peer et al. 1994) of 750 eukaryotic SSU rRNA sequences (Table 1).

The relative substitution rate, or relative variability, v_i, of a nucleotide at alignment position i is a parameter in the equation (Van de Peer et al. 1993a)

$$p_i = \frac{3}{4}\left[1 - \exp\left(-\frac{4}{3}v_i d\right)\right] \qquad (2)$$

which expresses the probability p_i that alignment position i contains a different nucleotide in two sequences separated by an evolutionary distance of d substitutions per nucleotide. After estimation of all v_i values as described earlier (Van de Peer et al. 1993a), alignment positions were grouped into sets of similar variability. The resulting spectrum of relative evolutionary rates as measured for eukaryotic SSU rRNA is shown in Fig. 2a. Once the shape of this spectrum is known, it is possible to derive the following more exact expression for the dissimilarity f between two sequences as a function of the evolutionary distance d separating them,

$$f = \frac{3}{4L}\sum_{i=j}^{+k} l_i\left[1 - \exp\left(-\frac{4}{3}(1+a)^i R d\right)\right] \qquad (3)$$

where L is the total number of nucleotides, l_i the number of nucleotides in set i, j the number of sets with a rate lower than the average rate of the complete molecule, k the number of sets with a higher rate, and $(1 + a)$ the ratio of the relative evolutionary rate in set i to this rate in set $i - 1$. In the distribution of Fig. 2a, the value chosen for this ratio was

$$(1 + a) = 10^{0.05} = 1.122$$

which resulted in 80 sets of nucleotides. The derivation of equation (3) is given in the Appendix, where the value of the parameter R is also listed. The shape of function (3), elaborated for the rate spectrum of Fig. 2a, is represented by curve "a" in Fig. 1.

Actually, even curve "a" in Fig. 1 is not yet a faithful reflection of dissimilarity as a function of distance in SSU rRNA, because the evolutionary rate spectrum shown in Fig. 2a was derived on the basis of a distance matrix computed by means of equation (1) (curve "JC" in Fig. 1), which only gives a first approximation of the relation between dissimilarity and distance. Therefore curve "a" reflects a better approximation, but not yet the exact relationship between the two quantities. A new distance matrix was computed by converting dissimilarities into distances on the basis of curve "a," and the relative substitution rate of each alignment position was estimated anew on the basis of this distance matrix. The resulting spectrum of evolutionary rates (not shown) is shifted somewhat with respect to the spectrum in Fig. 2a. The shape of function (3) corresponding to this shifted spectrum was found by filling in appropriate values for the parameters in equation (3). The curve representing this function (not shown) lies somewhat lower than curve "a" in Fig. 1. This process was repeated several times, each iteration resulting in a smaller change in the shape of the evolutionary rate spectrum (Fig. 2) and a slighter

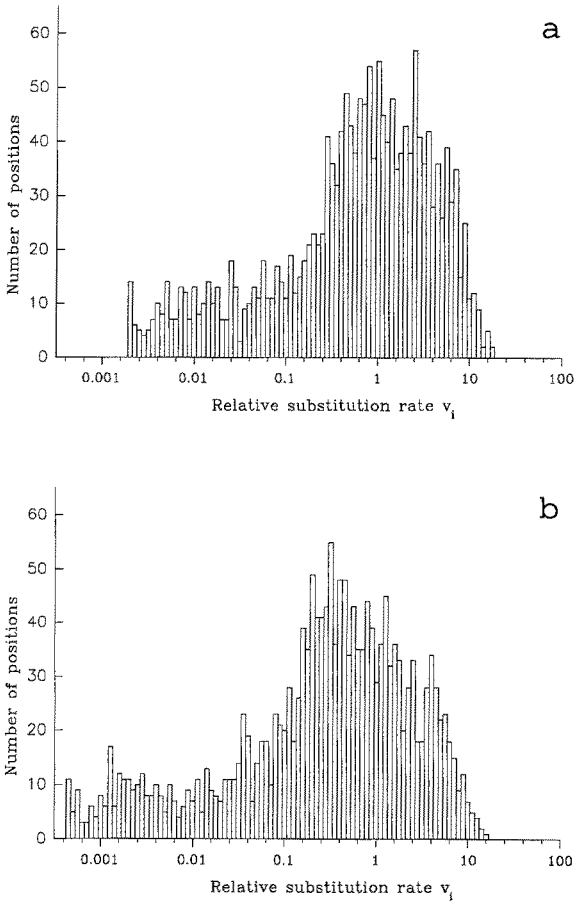

Fig. 2. Distribution of relative substitution rates of the nucleotides of eukaryotic SSU rRNA. Rates were estimated as described in Van de Peer et al. (1993a) for each of the alignment positions that are not absolutely conserved and contain a nucleotide in at least 25% of the aligned sequences. Substitution rates are measured relative to the average rate of the complete molecule. The group denoted by substitution rate 1 actually contains nucleotides that have a relative substitution rate v_i between 0.944 and 1.059, i.e., a rate span of 1.122. **a** Relative substitution rates were estimated using the relationship between evolutionary distance and dissimilarity expressed by equation (1). **b** Relative substitution rates were estimated using the relationship between evolutionary distance and dissimilarity expressed by equation (3) using parameters adapted after each iteration (see text for details).

displacement of the curve representing function (3) (Fig. 1). Five iterations were necessary for the changes to become imperceptible. The spectrum at equilibrium, shown in Fig. 2b, gives the distribution of substitution rates for 92 sets of nucleotides. The shape of function (3) corresponding to this spectrum is represented graphically by curve "b" in Fig. 1.

A disadvantage of expression (3) is that the dissimilarity corresponding to a given distance must be obtained by summation of a large number of terms, 92 in the present case. In addition, no expression can be derived for the inverse of function (3), needed to convert measured sequence dissimilarities into evolutionary distances during the computation of a distance matrix. The latter conversion was obtained during the iterative approximation of the evolutionary rate spectrum (Fig. 2b) by listing numerical values of the function for a large number of values of the argument and by interpolating linearly. This is not very practical for repeated computations of distance matrices required for the construction of evolutionary trees, especially when bootstrap analysis is performed. Fortunately, it is possible to find the following explicit expression that closely matches function (3), provided that an appropriate value is chosen for parameter *p*:

$$f = \frac{3}{4}\left\{1 - \exp\left[-\frac{4}{3}p \ln\left(1 + \frac{d}{p}\right)\right]\right\} \qquad (4)$$

Figure 1 shows the shape of function (4) for several values of p ranging from 0.2 to 1. At $p = 0.26$ the shape matches curve "b" nearly perfectly; in other words, it gives a very good approximation of function (3) when the latter is elaborated for the rate distribution of Fig. 2b. The inverse of equation (4),

$$d = p\left[\left(1 - \frac{4}{3}f\right)^{-\frac{3}{4p}} - 1\right] \qquad (5)$$

using $p = 0.26$, was used to convert dissimilarities between SSU rRNA sequences into evolutionary distances.

Evolution of Alveolates and Stramenopiles

Figure 3 shows a neighbor-joining tree of all hitherto known stramenopile and alveolate SSU rRNA sequences, constructed from a matrix of distances computed using equation (5). Bootstrap analysis (Felsenstein 1985), involving the computation of 500 trees from resampled data, was also performed. On the basis of the presumed sister relationship of stramenopiles and alveolates (see Introduction), the root was placed between these two taxa. This tree will be further referred to as the "calibrated" tree. As can be seen, the alveolates are subdivided into three main groups, viz. the apicomplexans, dinoflagellates, and ciliates. The foraminifer *Ammonia* forms the fourth evolutionary lineage within the alveolates. A study based on the comparison of partial LSU rRNA sequences (Pawlowski et al. 1994) suggested that the foraminifers branch closely to plasmodial and cellular slime molds, thus prior to the divergence of stramenopile and alveolate taxa. However, using the complete SSU rRNA sequence of *Ammonia*, Wray et al. (1995) decided that the foraminifers most probably are an "alveolate" lineage, which is consistent with our findings.

In the calibrated tree, the Apicomplexa are subdivided into two primary groups, corresponding to two different classes, viz. Hematozoa and Coccidia (Vivier and Desportes 1989). Whereas the monophyly of the Hematozoa seems rather well established in the calibrated tree, the

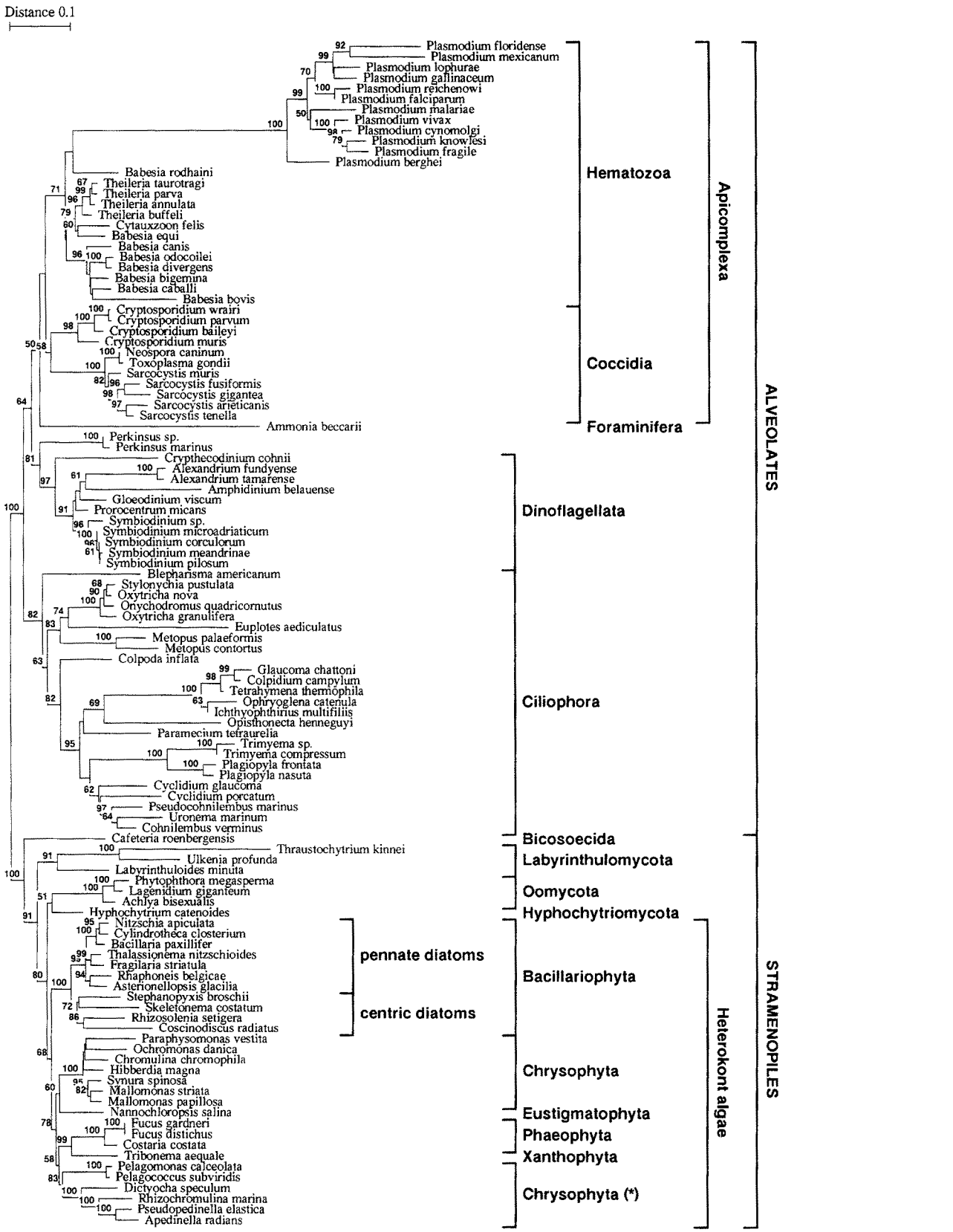

Fig. 3. Evolutionary tree of 112 SSU rRNA sequences of stramenopiles and alveolates based on "substitution rate calibration." The evolutionary distance between two organisms is obtained by summing the lengths of the connecting branches along the horizontal axis, using the *scale* on top. Bootstrap values above 250 (50%) are shown in *percentages* at the internodes. Taxon designations are placed to the right of the corresponding clusters. The classification of Chrysophyta indicated by (*) is subject to discussion (see text).

monophyly of the Coccidia is less well supported, which is demonstrated by the low bootstrap percentage (58%). Provisionally leaving the foraminifer *Ammonia* out of consideration, apicomplexans and dinoflagellates group together, which is also found in most "uncalibrated" evolutionary trees based on SSU rRNA (Wolters 1991; Wainright et al. 1993; Van de Peer et al. 1993b, this study) and LSU rRNA sequences (Van der Auwera et al. 1995). The possibility that dinoflagellates are derived from apicomplexans as suggested by Goggin and Barker (1993), is supported: in the calibrated tree, dinoflagellates share a common ancestor with the genus *Perkinsus,* which is usually classified as apicomplexan since it shows some distinguishing features of this taxon. As for the foraminifer *Ammonia,* its position in the present tree is poorly supported, and in trees based on different species sets it frequently is found between the ciliates and the remaining alveolates. This instability may be attributable to the availability of only one sequence.

Within the stramenopile cluster, the bicosoecid *Cafeteria* and the labyrinthuloids form the first lines of divergence, followed by the oomycetes and the hyphochytriomycetes. In accordance with a study based on LSU rRNA (Van der Auwera et al. 1995), the latter two taxa seem to share a common ancestor, although bootstrap support for this topology is low. More sequences of hyphochytriomycetes are probably needed to determine their exact phylogenetic position. The suggestion that oomycetes (and hyphochytriomycetes) have diverged shortly before the radiation of the autotrophic stramenopile taxa (Leipe et al. 1994) is also confirmed. The heterokont algae form a monophyletic cluster and apparently diverged most recently within the stramenopiles. Within the heterokont algae, four main clusters can be discerned. The largest one is formed by the diatoms or bacillariophytes. These are further subdivided into centric and pennate diatoms, both forming monophyletic clusters. The second main group is formed by the chrysophytes, which cluster with the eustigmatophyte *Nannochloropsis.* The third group is formed by the phaeophytes and the xanthophyte *Tribonema.* The fourth group, denoted with an (*) in the trees, contains organisms also classically defined as chrysophytes. However, according to a recent study by Saunders et al. (1995), these species should be reclassified, since they do not cluster with the remaining chrysophytes and are characterized by a reduced flagellar apparatus. For a more detailed discussion we refer to the latter paper. The exact divergence order among the four main clusters of heterokont algae remains dubious. What appears to be confirmed is the close evolutionary relationship between xanthophytes and phaeophytes and between eustigmatophytes and chrysophytes (e.g., Leipe et al. 1994; Saunders et al. 1995; Van der Auwera et al. 1995), although the latter relationship is not supported by bootstrap analysis.

Figure 4 shows an evolutionary tree based on the same sequences as in Fig. 3, but using the Jukes and Cantor correction (1969) to convert dissimilarity into evolutionary distance. Although both trees are similar in outline, there are notable differences, the most important of which is the position of the genus *Plasmodium.* This genus, which is separated from the other species by an extremely long branch, does not cluster with the other Hematozoa, where it should belong on the basis of morphological characteristics (Vivier and Desportes 1989) and where it is found in Fig. 3. The separation of *Plasmodium* from the other Hematozoa, also described elsewhere (e.g., Goggin and Barker 1993), is caused by an underestimation of the evolutionary distance. This is clearly visible in Fig. 1. The average dissimilarity between SSU rRNA sequences of the genus *Plasmodium* and of the other Hematozoa is 0.24 (denoted as f_{av} in Fig. 1). When this dissimilarity is converted into evolutionary distance on the basis of equation (1) (function "JC" in Fig. 1), a value of $d_{JC} = 0.29$ is obtained. However, when the dissimilarity is converted on the basis of equation (4) with $p = 0.26$ (which is virtually indistinguishable from curve "b" in Fig. 1), an evolutionary distance $d_b = 0.53$, nearly twice as large, is obtained. As a result of the serious underestimation of the evolutionary distance using the Jukes and Cantor correction, distant species seem closer to one another than they actually are, and this causes artificial clustering of long branches (Olsen 1987). In this respect, it is not unexpected to find the genus *Plasmodium* clustered with the foraminifer *Ammonia* (Fig. 4), which also exhibits a higher evolutionary rate, compared with the other alveolates. The high bootstrap value on the branch connecting *Ammonia* and *Plasmodium* (93%) should therefore be put in perspective and most probably supports an artificial clustering due to long branches attracting each other. In the calibrated tree (Fig. 3), this anomaly is not shown, suggesting that the method developed indeed computes the evolutionary distance more accurately.

Discussion

When trees are constructed on the basis of distance data, correct estimation of the evolutionary distance is crucial, and in most cases even more important than the choice of the method used to infer the tree topology. Since we do not have an exact historical record of events that took place in the evolution of our sequences, correct estimation of the evolutionary distance is not straightforward. It becomes more and more clear that the use of an unrealistic model of evolution can be the cause of serious artifacts in tree topology (Olsen 1987; Van de Peer et al. 1993a; De Rijk et al. 1995; this study). Therefore, it is very important to decipher the sequence information content as accurately as possible. In this paper, we determined the relative substitution rate of each alignment position on the basis of a large alignment of 750 eukaryotic SSU rRNA sequences. From the distribution of nucleotide substitution rates in this sequence alignment,

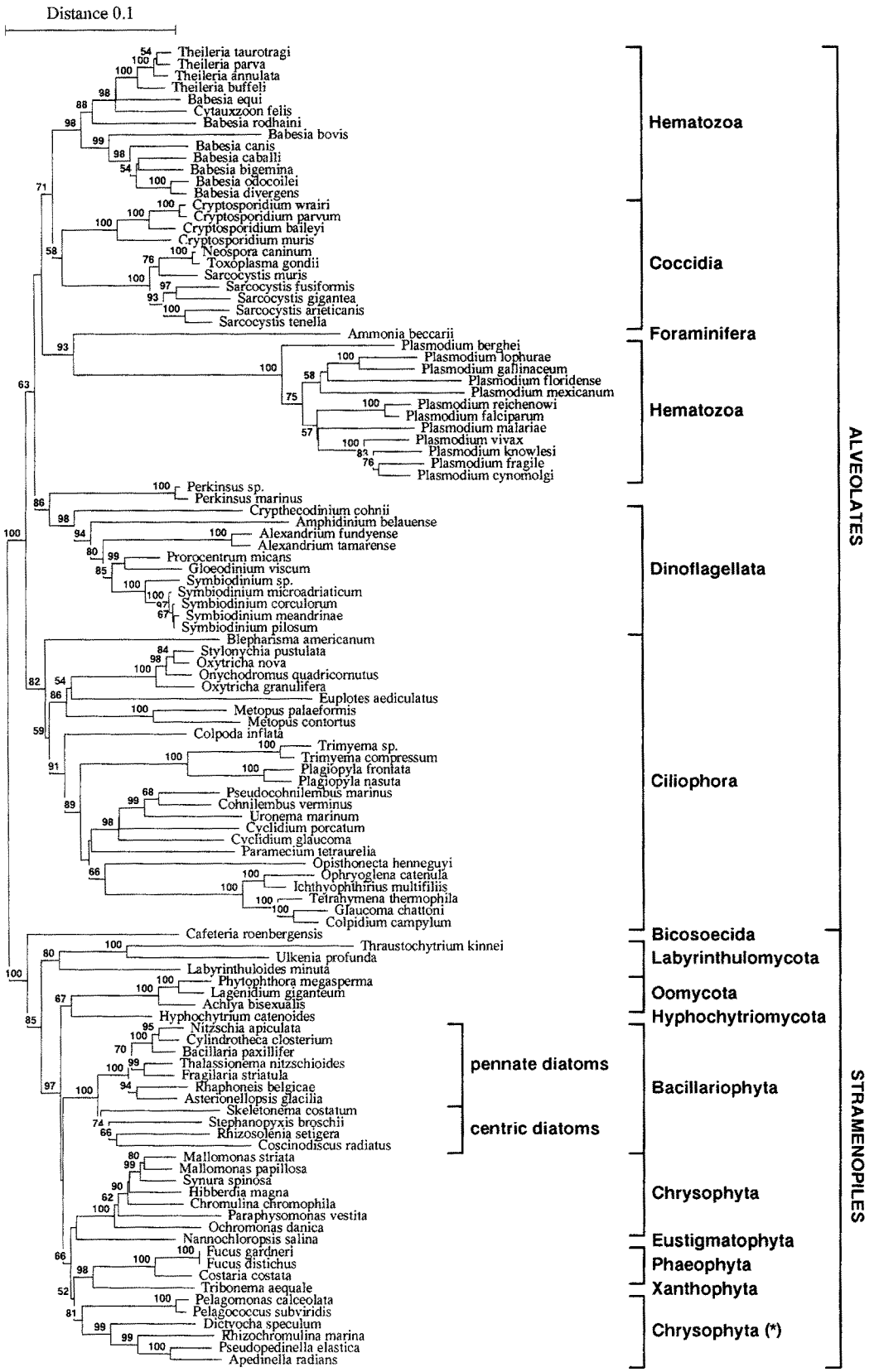

Fig. 4. Evolutionary tree including the same sequences as in Fig. 3, but using the Jukes and Cantor correction to convert dissimilarity into evolutionary distance. Conventions are as in Fig. 3.

equation (4) and the value 0.26 for parameter p could be derived. This equation gives a relationship between sequence dissimilarity and evolutionary distance specific for eukaryotic SSU rRNA. As can be seen in Fig. 1, this function deviates considerably from the relationship predicted for a molecular clock with a purely random behavior, especially at great distances. Trees constructed from distance matrices computed by means of this function are more reliable and suffer less from anomalies such as those caused by the presence of high evolutionary rates in certain branches. To our knowledge, this is the first study on ribosomal RNA sequences that takes into account the actual spectrum of substitution rates of the individual nucleotides of the molecule rather than postulating a hypothetical model for this rate spectrum.

Note Added in Proof

After this paper was accepted for publication, we learned from Dr. J. Pawlowski (personal communication) that the SSU rRNA sequence published for *Ammonia beccarii* is most probably not that of this foraminifer, but possibly that of an apicomplexan parasite of this organism.

Acknowledgments. Our research was supported by the Programme on Interuniversity Poles of Attraction (contract 23) of the Federal Office for Scientific, Cultural and Technical Affairs of the Belgian State, and by the fund for Collective Fundamental Research. Gert Van der Auwera is Research Assistant of the NFWO. We want to thank Ilse Van den Broeck and Stefan Nicolai for aligning the sequences used in this study and Dr. J. Van Casteren for assistance with the mathematical derivations. This study was performed in the framework of the Institute for the Study of Biological Evolution of the University of Antwerp.

References

De Rijk P, De Wachter R (1993) DCSE, an interactive tool for sequence alignment and secondary structure research. Comput Appl Biosci 9:735–740

De Rijk P, Van de Peer Y, Van den Broeck I, De Wachter R (1995) Evolution according to large ribosomal subunit RNA. J Mol Evol 41:366–375

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

Goggin CL, Barker SC (1993) Phylogenetic position of the genus *Perkinsus* (Protista, Apicomplexa) based on small subunit ribosomal RNA. Mol Biochem Parasitol 60:65–70

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

Jukes TH, Cantor CR (1969) Evolution of protein molecules. In: Munro HH (ed) Mammalian protein metabolism. New York: Academic Press, pp 21–132

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

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

Leipe DD, Wainright PO, Gunderson JH, Porter D, Patterson DJ, Valois F, Himmerich S, Sogin ML (1994) The stramenopiles from a molecular perspective: 16S-like rRNA sequences from *Labyrinthuloides minuta* and *Cafeteria roenbergensis.* Phycologia 33:369–377

Olsen GJ (1987) Earliest phylogenetic branchings: comparing rRNA-based evolutionary trees inferred with various techniques. Cold Spring Harbor Symp Quant Biol LII:825–837

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

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

Pawlowski J, Bolivar I, Guiard-Maffia J, Gouy M (1994) Phylogenetic position of Foraminifera inferred from LSU rRNA gene sequences. Mol Biol Evol 11:929–938

Preisig HR, Anderson OR, Corliss JO, Moestrup O, Powell MJ, Roberson RW, Wetherbee R (1994) Terminology and nomenclature of protist cell surface structures. Protoplasma 181:1–28

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

Saitou N, Imanishi T (1989) Relative efficiencies of the Fitch–Margoliash, maximum-parsimony, maximum-likelihood, minimum-evolution, and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree. Mol Biol Evol 6:514–525

Saunders GW, Potter D, Paskind MP, Andersen RA (1995) Cladistic analyses of combined traditional and molecular data sets reveal an algal lineage. Proc Natl Acad Sci USA 92:244–248

Sourdis J, Nei M (1988) Relative efficiencies of the maximum parsimony and distance–matrix methods in obtaining the correct phylogenetic tree. Mol Biol Evol 5:298–311

Tajima F, Nei M (1984) Estimation of evolutionary distance between nucleotide sequences. Mol Biol Evol 1:269–285

Van de Peer Y, De Wachter (1993) TREECON: a software package for the construction and drawing of evolutionary trees. Comput Appl Biosci 9:177–182

Van de Peer Y, Neefs J-M, De Rijk P, De Wachter R (1993a) Reconstructing evolution from eukaryotic small–ribosomal–subunit RNA sequences: calibration of the molecular clock. J Mol Evol 37:221–232

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

Van de Peer Y, De Wachter (1994) TREECON for Windows: a software package for the construction and drawing of evolutionary trees for the Microsoft Windows environment. Comput Appl Biosci 10:569–570

Van de Peer Y, Van den Broeck I, De Rijk P, De Wachter R (1994) Database on the structure of small ribosomal subunit RNA. Nucleic Acids Res 22:3488–3494

Van der Auwera G, De Baere R, Van de Peer Y, De Rijk P, Van den Broeck I, De Wachter R (1995) The phylogeny of the Hyphochytriomycota as deduced from ribosomal RNA sequences of *Hyphochytrium catenoides.* Mol Biol Evol 12:671–678

Vivier E, Desportes I (1989) Phylum Apicomplexa. In: Margulis L, Corliss JO, Melkonian M, Chapman DJ (eds) Handbook of Protoctista. Boston: Jones and Bartlett, pp 549–573

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

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

Wray CG, Langer MR, DeSalle R, Lee JJ, Lipps JH (1995) Origin of the foraminifera. Proc Natl Acad Sci USA 92:141–145

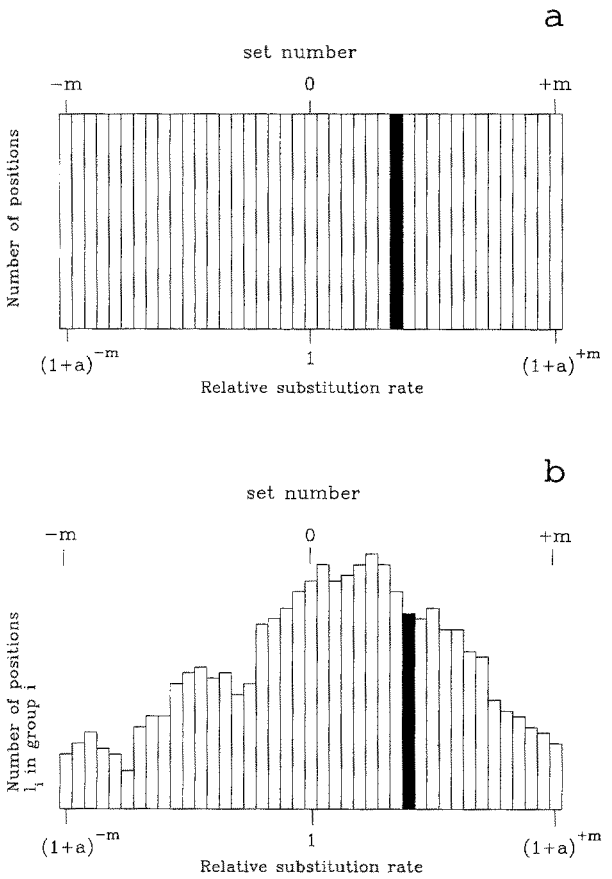

Fig. 5. Hypothetical distributions of relative nucleotide substitution rates. The *black bar* denotes the set of nucleotides with a rate equal to the average rate of the entire molecule. This rate is larger than that of set 0, the middle set, which has an evolutionary rate which is the geometric average of the rates of all sets. **a** Distribution in which all the sets contain the same number of nucleotides. **b** Distribution with unequal sets of nucleotides.

Appendix

Consider two nucleic acid sequences that have diverged from a common ancestral sequence and have reached, relative to each other, an evolutionary distance resulting from d substitutions per nucleotide. If the four nucleotides occur in approximately the same amounts and all substitutions are equally probable, then the expected dissimilarity f is given by the expression:

$$f = \frac{3}{4}\left[1 - \exp\left(-\frac{4}{3}d\right)\right] \tag{A1}$$

The inverse function of (A1) is the well-known equation of Jukes and Cantor (1969), often used to convert sequence dissimilarity into evolutionary distance. However, this equation is only applicable if the substitution rate of all nucleotides in the sequence is approximately the same. It has been estimated, though (Van de Peer et al. 1993a), that nucleotides in SSU rRNAs have substitution rates varying by a factor of more than 1,000 from the most conservative to the most variable. In order to derive an equation for dissimilarity as a function of distance for such a molecule, we first consider the following model illustrated in Fig. 5a.

A molecule contains n sets of nucleotides, each characterized by a different substitution rate. All sets are equally numerous and their substitution rates form a geometric series with a ratio $(1 + a)$ where a is a positive number much smaller than 1. The middle set (set 0 in Fig. 5a) has a substitution rate which is the geometric average of the rates of all sets. The ratio of the substitution rate in set i to the substitution rate in the middle set is:

$$(1 + a)^i \tag{A2}$$

When the middle sets in two diverging sequences have covered an evolutionary distance x, the i^{th} sets in the two sequences will have acquired a dissimilarity f_i

$$f_i = \frac{3}{4}\left[1 - \exp\left(-\frac{4}{3}(1+a)^i x\right)\right]$$

and the complete molecules will have acquired a dissimilarity

$$f = \frac{3}{4n}\sum_{i=-m}^{+m}\left[1 - \exp\left(-\frac{4}{3}(1+a)^i x\right)\right] \tag{A3}$$

where m is the number of sets with a lower and the number of sets with a higher substitution rate than the middle set, so $n = 2m + 1$.

However, during evolution the distance between the complete molecules, i.e., the total number of substitutions that have taken place divided by their chain length, increases more rapidly than the distance between the middle sets. Indeed, since the rates form a geometric series, the excess of substitutions in the sets with higher rates is larger than the deficit of substitutions in the sets with lower rates. As a consequence, the average distance, d, covered by the complete molecules is larger than the distance covered by the middle sets, x:

$$d = \frac{x}{n}\sum_{i=-m}^{+m}(1 + a)^i$$

The following explicit expression for d as a function of x can be found:

$$d = \frac{x}{n}\frac{(1+a)^{m+1} - (1+a)^{-m}}{a} \tag{A4}$$

From (A3) and (A4) we get dissimilarity as a function of distance for the complete molecules:

$$f = \frac{3}{4n}\sum_{i=-m}^{+m}\left[1 - \exp\left(-\frac{4}{3}\frac{na(1+a)^i}{(1+a)^{m+1} - (1+a)^{-m}}d\right)\right] \tag{A5}$$

We now consider a model that reflects more closely the actual situation in SSU rRNA illustrated schematically in Fig. 5b. The sets of nucleotides with increasing substitution rates are not equally numerous but follow a distribution that can be estimated experimentally (Van de Peer et al. 1993a) provided that an extensive collection of well-aligned SSU rRNA sequences is available. Let L be the chain length of the molecule and let there be n sets of different substitution rates, l_i being the number of nucleotides in set i. As in the previous model, the ratio of substitution rates of two successive sets is $(1 + a)$. In this case, the dissimilarity acquired by the molecules, as a function of the distance covered by the middle sets, is:

$$f=\frac{3}{4L}\sum_{i=-m}^{+m} l_i\left[1-\exp\left(-\frac{4}{3}(1+\mathrm{a})^i x\right)\right] \tag{A6}$$

The average distance between the complete molecules, as a function of the distance between the middle sets x, is:

$$d=\frac{x}{L}\sum_{i=-m}^{+m} l_i(1+a)^i \tag{A7}$$

From (A6) and (A7) we get dissimilarity as a function of distance for the complete molecules:

$$f=\frac{3}{4L}\sum_{i=-m}^{+m} l_i\left[1-\exp\left(-\frac{4}{3}(1+a)^i Rd\right)\right] \tag{A8}$$

where

$$R=\frac{L}{\sum_{i=-m}^{+m} l_i(1+a)^i} \tag{A9}$$

Actually, R is the ratio of the evolutionary rate in the middle set to the average rate of the entire molecule.

According to equation (A2), the evolutionary rates of all sets were expressed relative to the rate of the middle set, i.e., the one with the geometric average of rates. Alternatively, one can express all rates relatively to an arbitrarily chosen set different from the middle set. Equation (A8) can then be rewritten as follows:

$$f=\frac{3}{4L}\sum_{i=-j}^{+k} l_i\left[1-\exp\left(-\frac{4}{3}(1+a)^i Rd\right)\right] \tag{A10}$$

where j is the number of sets with a rate lower, and k the number of sets with a rate higher, than the rate of the reference set. The expression for R becomes:

$$R=\frac{L}{\sum_{i=-j}^{+k} l_i(1+a)^i} \tag{A11}$$

If the set chosen as reference set has the same evolutionary rate as the complete molecule, R acquires the value 1.

Expressions (A5), (A8), and (A10) can be well approximated by the explicit equation

$$f=\frac{3}{4}\left\{1-\exp\left[-\frac{4}{3}p\ln\left(1+\frac{d}{p}\right)\right]\right\} \tag{A12}$$

if the parameter p is given an appropriate value, which depends on the number of sets of different substitution rate, the distribution of the number of residues over the sets (i.e., the series l_i), and the ratio $(1 + a)$ of substitution rates between successive sets. In the case of eukaryotic SSU rRNAs, a value of $p = 0.26$ was derived as described in the Results section.