Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 7732–7736, July 1996
Evolution

Substitution rate calibration of small subunit ribosomal RNA identifies chlorarachniophyte endosymbionts as remnants of green algae

(long branch attraction/dissimilarity/evolutionary distances)

Y. VAN DE PEER*, S. A. RENSING†, U.-G. MAIER†‡, AND R. DE WACHTER*§

*Departement Biochemie, Universiteit Antwerpen (UIA), Universiteitsplein 1, B-2610 Antwerp, Belgium; and †Institut für Biologie II, Zellbiologie, Universität Freiburg, Schänzlestrasse 1, D-79104 Freiburg, Germany

Communicated by Lynn Margulis, University of Massachusetts, Amherst, MA, February 26, 1996 (received for review November 29, 1995)

ABSTRACT **Chlorarachniophytes are amoeboid algae with chlorophyll a and b containing plastids that are surrounded by four membranes instead of two as in plants and green algae. These extra membranes form important support for the hypothesis that chlorarachniophytes have acquired their plastids by the ingestion of another eukaryotic plastid-containing alga. Chlorarachniophytes also contain a small nucleus-like structure called the nucleomorph situated between the two inner and the two outer membranes surrounding the plastid. This nucleomorph is a remnant of the endosymbiont's nucleus and encodes, among other molecules, small subunit ribosomal RNA. Previous phylogenetic analyses on the basis of this molecule provided unexpected and contradictory evidence for the origin of the chlorarachniophyte endosymbiont. We developed a new method for measuring the substitution rates of the individual nucleotides of small subunit ribosomal RNA. From the resulting substitution rate distribution, we derived an equation that gives a more realistic relationship between sequence dissimilarity and evolutionary distance than equations previously available. Phylogenetic trees constructed on the basis of evolutionary distances computed by this new method clearly situate the chlorarachniophyte nucleomorphs among the green algae. Moreover, this relationship is confirmed by transversion analysis of the *Chlorarachnion* plastid small subunit ribosomal RNA.**

Algae that have plastids surrounded by four membranes, such as diatoms, chrysophytes, phaeophytes, haptophytes, chlorarachniophytes, and cryptomonads, are believed to have acquired these organelles by engulfing another eukaryotic alga. In two groups of algae, cryptomonads and chlorarachniophytes, a remnant of this endosymbiotic event can be found. Between the two inner and the two outer plastid membranes of these algae, a narrow cytoplasmic compartment exists that contains a nucleus-like structure, thought to be the vestigial nucleus of the phototrophic eukaryotic endosymbiont. It was previously demonstrated in both cryptomonads (1–3) and chlorarachniophytes (4) that the nucleomorphic DNA encodes a small subunit ribosomal RNA (SSU rRNA). Because SSU rRNA sequences are known for many eukaryotic taxa (5), it should be possible to trace the evolutionary origin of the nucleomorphs. In this study, we are particularly interested in the phylogenetic origin of the chlorarachniophyte endosymbiont. Although the plastids of *Chlorarachnion* contain chlorophyll a and b, previous analyses on the basis of SSU rRNA failed to demonstrate the descent of the endosymbionts from a green alga, as traditionally proposed (6, 7). Some of these analyses strongly suggest that the *Chlorarachnion* endosymbiont shares a common origin with the endosymbionts of the cryptomonads and that both are probably related to the red algae (8, 9); others are unable to ally the endosymbiont to any extant group of eukaryotes (4, 10, 11). The origin of the chlorarachniophyte endosymbiont may possibly be obscured by the high evolutionary rates of the nucleomorph SSU rRNA sequences. It is generally known that high evolutionary rates can introduce errors in a tree topology (10, 12–14). In distance matrix methods, errors are introduced by underestimation of evolutionary distances when an unrealistic model of nucleotide substitutions is assumed (13, 15). In a recent paper (15), a method was developed to compute the evolutionary distance between SSU rRNA sequences that takes into account differences in the substitution rate of the nucleotides. The substitution rates of the individual nucleotides of the molecule were estimated on the basis of an extensive sequence alignment of 750 eukaryotic SSU rRNAs. From the resulting rate distribution an equation was derived that discriminates more selectively between sequence dissimilarity and evolutionary distance and that estimates large evolutionary distances more accurately. To elucidate the phylogenetic position of the *Chlorarachnion* endosymbionts, neighbor-joining (16) trees were constructed on the basis of distances computed by means of this new equation.

MATERIALS AND METHODS

Substitution Rate Calibration and Computation of Evolutionary Distances. Evolutionary distances were estimated as described (14, 15). In short, this goes as follows. For an alignment of 750 eukaryotic SSU rRNA sequences, the relative variability or relative substitution rate v_i is determined for every alignment position in which more than 25% of the sequences have a nucleotide. This rate v_i is a parameter in the equation

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

which expresses the probability p_i that an alignment position i contains a different nucleotide in two sequences, as a function of the evolutionary distance d separating these sequences. After estimation of all v_i values as described earlier (14, 15), alignment positions are grouped into sets of similar variability. A spectrum of relative nucleotide substitution rates is thus obtained. Once the shape of this spectrum is known, it is possible to derive the following equation for the dissimilarity (fraction of observed substitutions), f, between two sequences as a function of the evolutionary distance d separating them (15):

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

Abbreviations: SSU rRNA: small subunit ribosomal RNA; cpDNA, plastid DNA.

‡Present address: Lehrstuhl für Pflanzensystematik, Universität Bayreuth, D-95440 Bayreuth, Germany.
§To whom reprint requests should be addressed.

$$f = \frac{3}{4}\left\{1 - \exp\left[-\frac{4}{3}p \ln\left(1 + \frac{d}{p}\right)\right]\right\}. \quad [2]$$

Parameter p is a value that depends on the shape of the substitution rate spectrum, and a value of $p = 0.26$ was derived for eukaryotic SSU rRNA (15). The inverse of Eq. **2**,

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

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

Construction and Drawing of Evolutionary Trees. 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 rrna.uia.ac.be or by World Wide Web (WWW) at URL http://rrna.uia.ac.be/rrna/(5). All the sequences are aligned on the basis of similarity in primary and secondary structure by using the DCSE sequence editor (17). Evolutionary trees were constructed by the neighbor-joining method (16). The construction and drawing of evolutionary trees were done with the software package TREECON for Windows (18), which runs on IBM-compatible computers and is available from the authors upon request. More information on TREECON can also be found at URL http://www.uia.ac.be/u/yvdp/. The calibration method (15) is currently being incorporated in this software package and will be available within the near future.

RESULTS AND DISCUSSION

The Nuclear Tree. Fig. 1 shows a neighbor-joining tree for 107 SSU rRNA sequences of representatives of the so called "crown taxa" (19). Although the exact phylogenetic position of *Dictyostelium* is still being discussed (e.g., ref. 20), this organism was chosen as outgroup since it branches off relatively close to the "crown" taxa but is always found separate from these in trees based on SSU rRNA sequences. Evolutionary distances were estimated by using Eq. **3**. Bootstrap analysis (21), involving the computation of 200 trees from resampled data, was also performed. As can be seen, the SSU rRNA sequences of the *Chlorarachnion* nucleomorphs are clustered with those of the green algae (top of the tree). The latter form a monophyletic cluster with the green plants, as expected. As for the cryptomonad nucleomorphs, these are clustered with the red algae. The bootstrap value on the branch leading to green algae and chlorarachniophyte nucleomorphs is low (*viz.* 58%), although the branch uniting all chlorophytes is supported at a bootstrap level of 73%. However, if the cryptomonad nucleomorph sequences are omitted from the analysis, the cluster of green algae and chlorarachniophyte nucleomorphs is supported at a bootstrap level of 89%. Moreover, the branch leading to green plants, green algae and chlorarachniophyte nucleomorphs is then supported at a bootstrap level of 95%, instead of 73%. Apparently, there is a certain attraction between the nucleomorphs of *Chlorarachnion* and nucleomorphs of the cryptomonads. If trees are constructed using the correction of Jukes and Cantor (22) or Kimura (23) to convert dissimilarity into evolutionary distance, chlorarachniophyte nucleomorphs cluster usually with cryptomonad nucleomorphs, sometimes with high bootstrap support (e.g., ref. 8). The same is true for trees inferred by maximum parsimony and maximum likelihood (refs. 8, 9, 11, and 24; our unpublished results). Furthermore, both groups of nucleomorphs are then often clustered with the red algae. We believe this clustering to be artificial and caused by the underestimation of evolutionary distances. As demonstrated previously (13, 15), higher evolutionary rates tend to pull sequences closer to the base of the tree. Moreover, due to the serious underestimation of large evolutionary distances, distant species seem closer to one another than they actually are, which often results in artificial clustering of long branches (15).

When the branch lengths are correctly estimated, making the long branches even longer, their artifactual attraction disappears. As can be seen in Fig. 1, the branch leading to the chlorarachniophyte nucleomorphs is exceptionally long. Also the evolutionary rate of the SSU rRNA sequences of cryptophyte nucleomorphs and red algae is higher than the average of the other eukaryotic SSU rRNA sequences. The "substitution rate calibration" we applied here computes the evolutionary distance more accurately and is less sensitive to the "long branch attraction" effect (15). In this respect, it is also interesting to point out the phylogenetic position of the genus *Plasmodium* among the apicomplexans. In the tree of Fig. 1, this fast evolving genus is clustered with the other hematozoan genera *Theileria* and *Babesia*, where it should belong on the basis of morphological characteristics (25). However, in most studies which do not take into account the substitution rates of individual nucleotides, *Plasmodium* is clustered erroneously (e.g., refs. 26 and 27). Recently published trees (28) on the basis of SSU rRNA and only including apicomplexans, dinoflagellates, and ciliates, showed *Plasmodium* to be clustered with the other apicomplexans but without statistical reliability and separated from the other hematozoans by the coccidians. Considering transversions only, *Plasmodium* was clustered with the other hematozoans, but also without statistical reliability (28). When "nucleotide substitution rate calibration" is applied, *Plasmodium* is clustered consistently with the other Hematozoa (ref. 15; this study).

If the tree of Fig. 1 is assumed to be correct, chlorarachniophytes and cryptomonads have obtained their endosymbionts separately. This hypothesis is in accordance with the recent finding that their respective plastid genomes harbor radically different ribulose bisphosphate carboxylase (*rbcL*) genes (10), making a common origin extremely unlikely. The endosymbionts of the cryptomonads seem to cluster with the red algae. Although not supported by bootstrap analysis in the tree of Fig. 1, this relationship has been suggested previously on the basis of SSU rRNA trees (1, 27, 29, 30). Furthermore, like the red algae, cryptomonads contain phycobilins as accessory pigments.

In the phylogenetic tree of Fig. 1, the chlorarachniophytes (host cells) are clustered at a high bootstrap level (98%) with the Euglyphina, which are rhizopods with hyaline, filiform pseudopodia. This is consistent with previous analyses on the basis of SSU rRNA sequences (11). The cryptomonads (host cells) are loosely grouped with the glaucophytes. This relationship, although not supported by bootstrap analysis, has also been suggested previously on the basis of SSU rRNA (9). Overall, nuclear SSU rRNA data point to a polyphyletic origin of algae that contain plastids surrounded by four membranes as well as to multiple, independent secondary endosymbioses.

The Plastid Tree. Fig. 2 shows a neighbor-joining tree based on the cpDNA (plastid DNA)-encoded SSU rRNA sequences of land plants, green algae, euglenophytes, red algae, and algae containing chlorophyll a and c (chromophyte algae). cpDNA-encoded SSU rRNA sequences of *Chlorarachnion* and of the cryptomonads are also included. The tree is rooted with the cyanobacterium *Nostoc*. The AT content in cpDNA-encoded SSU rRNA sequences differs substantially (10), and it has been shown that such biases can be the cause of serious artifacts in tree topology (31–33). It has also been shown that transversion analysis (34), where transitions are not taken into account in computing the evolutionary distance, is less sensitive to substitutional bias (31, 35). For this reason, the distance tree in Fig. 2 was also constructed by using transversion analysis, and dissimilarity was converted into evolutionary distance by using the following equation (34):

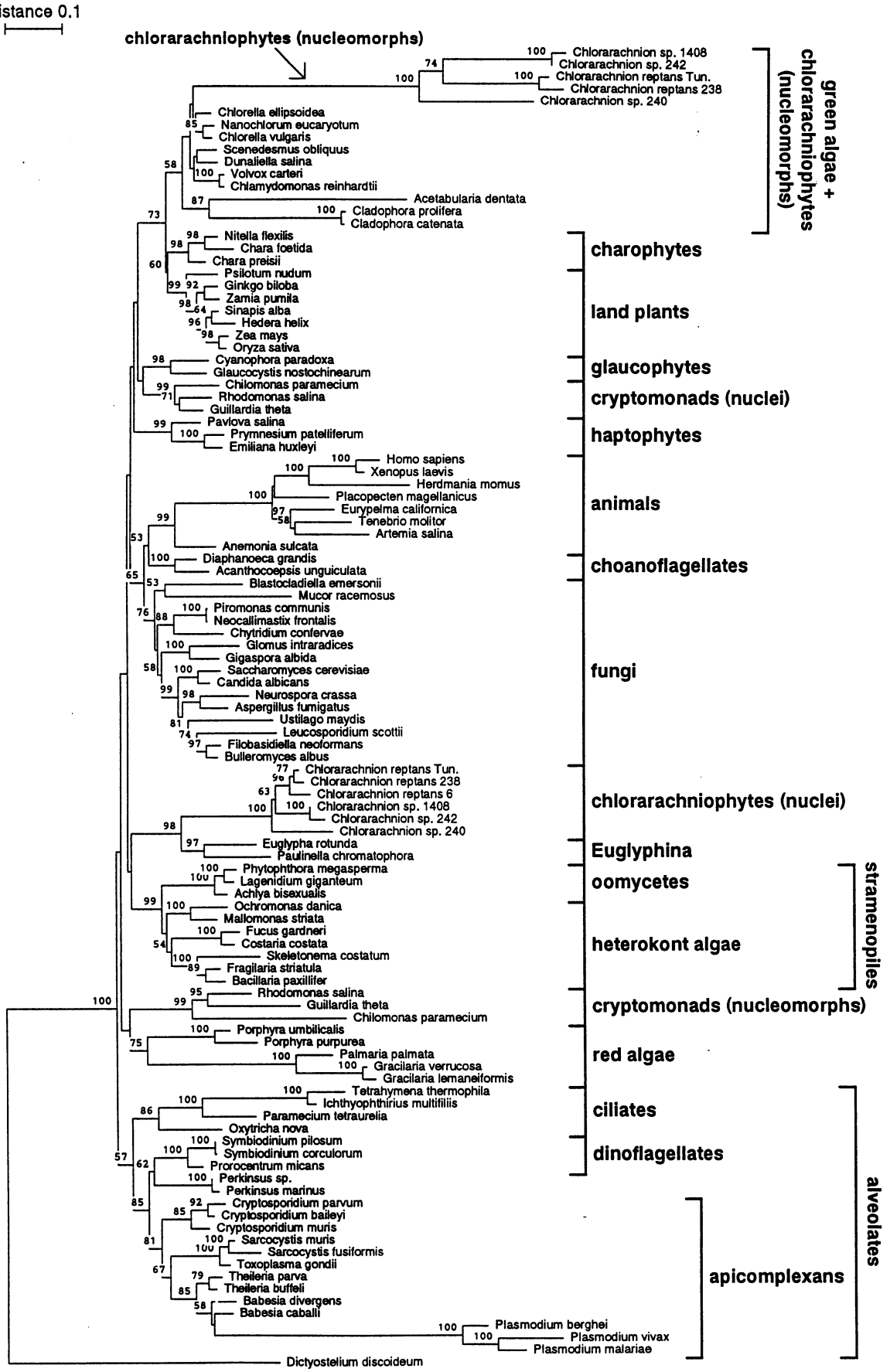

FIG. 1. (*Legend on opposite page.*)

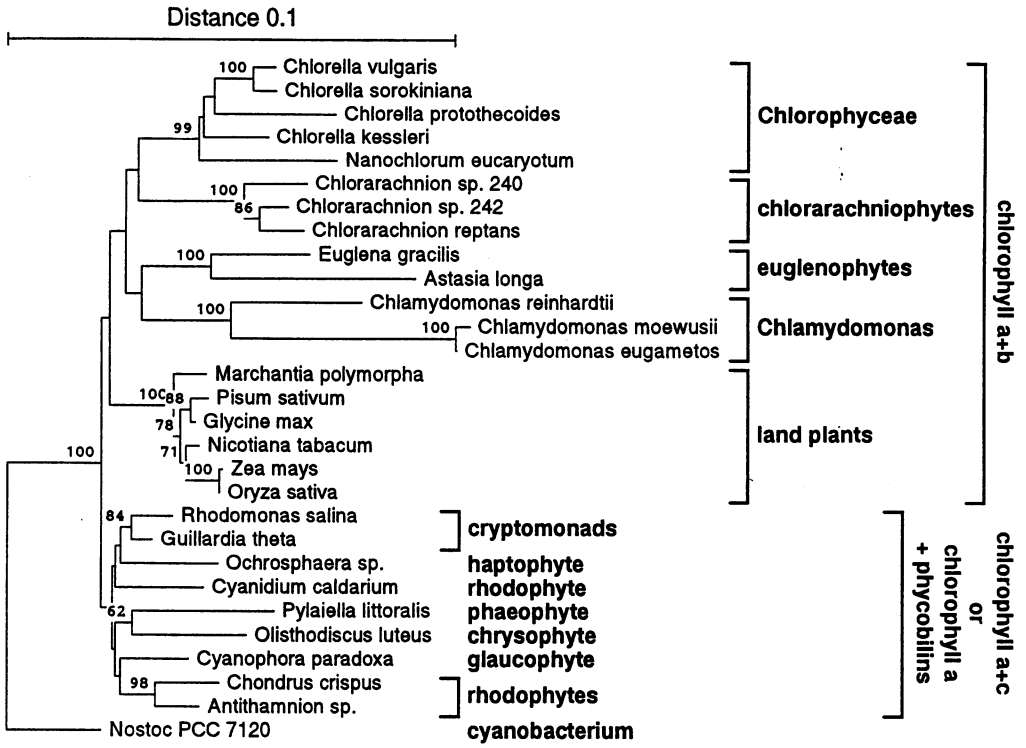

FIG. 2. Evolutionary tree of the cpDNA-encoded 16S rRNAs constructed by neighbor-joining taking into account transversions only (31). The cyanobacterium *Nostoc* was used to root the tree. Conventions are as in Fig. 1.

$$d = -b \ln\left(1 - \frac{Q}{b}\right),$$

where Q is the fraction of transversions between two sequences, and

$$b = 1 - [(f_A + f_G)^2 + (f_C + f_T)^2],$$

where $f_A + f_G$ is the fraction of purines, and $f_C + f_T$ is the fraction of pyrimidines, computed for the complete alignment.

As can be seen, all chlorophyll a and b containing plastids are clustered together, although not supported at a high bootstrap level. The cpDNA-encoded SSU rRNAs of *Chlorarachnion* cluster with those of the green algae, which is in agreement with a tree published recently by McFadden *et al.* (10). However, instead of transversion analysis they used the LOGDET transformation method (32, 33), which was developed particularly to cope with varying substitution bias. In this respect it is also noteworthy to discuss the position of the euglenophytes. In most trees based on cpDNA-encoded SSU rRNAs and constructed with standard distance methods, maximum parsimony, or maximum likelihood, the euglenophytes are clustered with algae containing chlorophyll c or phycobiliproteins (10, 35, 36). It has been shown (33) that this position of the euglenophytes was most probably an artifact caused by a bias in the AT content. This is also suggested by our tree (Fig. 2). Transversion analysis clearly succeeds in clustering together all organisms with plastids containing chlorophyll a and b, although not supported at a high bootstrap level. The tree of Fig. 2 supports the hypothesis that euglenophytes acquired their plastids by secondary endosymbiosis of a green alga.

The cpDNA-encoded SSU rRNAs of the cryptomonads are grouped with those of algae containing chlorophyll c or phycobiliproteins, which is also in agreement with previous trees based on SSU rRNA (29, 35, 37). Since the cpDNA-encoded SSU rRNA of *Chlorarachnion* does not cluster with the cpDNA-encoded SSU rRNA of cryptomonads, the hypothesis that the two groups of algae obtained their endosymbionts independently, is further supported. Moreover, the plastid tree of Fig. 2 corroborates the nuclear tree based on substitution rate calibration analysis (Fig. 1) not only in identifying the chlorarachniophyte endosymbionts as related to the green algae, but even to the point of identifying the *Chlorella/Nanochlorum* clade as their closest relatives, although this relationship is not supported by bootstrap analysis. A close evolutionary relationship between the chlorarachniophyte endosymbionts and green algae has also been proposed on the basis of chloroplast morphology and pigment composition (38–40). Furthermore, the telomere sequences of the chlorarachniophyte nucleomorph chromosomes (41) and *Chlorella* nuclear chromosomes (42) are very similar and differ only by one transition.

In conclusion, we want to point out the importance of nonmolecular data. On the basis of most standard tree inferring methods, the nucleomorph and plastid rRNA phylogenies gave misleading results, sometimes very convincing, regarding the phylogenetic position of the nucleomorph endosymbiont. If there had not been good nonmolecular reasons to question these molecular results, rate calibration and transversion analysis might never have been applied in this case. It is probable that many more such misleading results are present in eukaryotic trees based on SSU rRNA sequences. Substitution rate

FIG. 1. Evolutionary tree constructed by the neighbor-joining method (14) from a distance matrix of 107 eukaryotic SSU rRNA sequences and based on "substitution rate calibration." The slime mold *Dictyostelium discoideum* was used to root the tree. 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 50% are shown at the internodes. Taxon designations are placed to the right of the corresponding clusters. Note that the cryptomonad *Rhodomonas salina* was previously named *Pyrenomonas salina*, and that the sequence formerly deposited as *Cryptomonas phi* in fact belongs to *Guillardia theta*.

calibration should contribute to the elimination of artifacts caused by fast-evolving sequences.

We thank W. Martin, G. McFadden, S. Douglas, and three anonymous referees for critical comments on the manuscript. 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; by the National Fund for Scientific Research; and by the Deutsche Forschungsgemeinschaft. This study was performed in the framework of the Institute for the Study of Biological Evolution of the University of Antwerp. Y.V.d.P. is Research Assistant of the National Fund for Scientific Research.

1. Douglas, S. E., Murphy, C. A., Spencer, D. F. & Gray, M. W. (1991) *Nature (London)* **350,** 148–151.
2. Maier, U.-G., Hofmann, C. J. B., Eschbach, S., Wolters, J. & Igloi, G. L. (1991) *Mol. Gen. Genet.* **230,** 155–160.
3. McFadden, G. I., Gilson, P. R. & Douglas, S. E. (1994) *J. Cell Sci.* **107,** 649–657.
4. McFadden, G. I., Gilson, P. R., Hofmann, C. J. B., Adcock, G. J. & Maier, U.-W. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 3690–3694.
5. Van de Peer, Y., Van den Broeck, I., De Rijk, P. & De Wachter, R. (1996) *Nucleic Acids Res.* **24,** 86–91.
6. Ludwig, M. & Gibbs, S. P. (1989) *J. Phycol.* **25,** 385–394.
7. Hibberd, D. J. (1990) in *Handbook of Protoctista*, eds. Margulis, L., Corliss, J. O., Melkonian, M. & Chapman, D. J. (Jones & Bartlett, Boston), pp. 288–292.
8. Cavalier-Smith, T., Allsop, M. T. E. P. & Chao, E. E. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 11368–11372.
9. Bhattacharya, D., Helmchen, T., Bibeau, C. & Melkonian, M. (1995) *Mol. Biol. Evol.* **12,** 415–420.
10. McFadden, G. I., Gilson, P. R. & Waller, R. F. (1995) *Arch. Protistenk.* **145,** 231–239.
11. Bhattacharya, D., Helmchen, T. & Melkonian, M. (1995) *J. Eukaryotic Microbiol.* **42,** 65–69.
12. Felsenstein, J. (1978) *Syst. Zool.* **27,** 410–410.
13. Olsen, G. J. (1987) *Cold Spring Harbor Symp. Quant. Biol.* **53,** 825–837.
14. Van de Peer, Y., Neefs, J.-M., De Rijk, P. & De Wachter, R. (1993) *J. Mol. Evol.* **37,** 221–232.
15. Van de Peer, Y., Van der Auwera, G. & De Wachter, R. (1996) *J. Mol. Evol.* **42,** 201–210.
16. Saitou, N. & Nei, M. (1987) *Mol. Biol. Evol.* **4,** 406–425.
17. De Rijk, P. & De Wachter, R. (1993) *Comput. Appl. Biosci.* **9,** 735–740.
18. Van de Peer, Y. & De Wachter, R. (1994) *Comput. Appl. Biosci.* **10,** 569–570.
19. Knoll, A. H. (1992) *Science* **256,** 622–627.
20. Kuma, K.-i., Nikoh, N., Iwabe, N. & Miyata, T. (1995) *J. Mol. Evol.* **41,** 238–246.
21. Felsenstein, J. (1985) *Evolution* **39,** 783–791.
22. Jukes, T. H. & Cantor, C. R. (1969) in *Mammalian Protein Metabolism*, ed. Munro H. H. (Academic, New York), pp. 21–132.
23. Kimura, M. (1980) *J. Mol. Evol.* **16,** 111–120.
24. Cavalier-Smith, T., Allsop, M. T. E. P. & Chao, E. E. (1994) *Philos. Trans. R. Soc. London B* **346,** 387–397.
25. Vivier, E. & Desportes, I. (1989) in *Handbook of Protoctista*, eds. Margulis, L., Corliss, J. O., Melkonian, M. & Chapman, D. J. (Jones & Bartlett, Boston), pp. 549–573.
26. Goggin, C. L. & Barker, S. C. (1993) *Mol. Biochem. Parasitol.* **60,** 65–70.
27. McFadden, G. I., Gilson, P. R. & Hill, D. R. A. (1994) *Eur. J. Phycol.* **29,** 29–32.
28. Escalante, A. A. & Ayala, F. J. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 5793–5797.
29. Maerz, M., Wolters, J., Hofmann, C. J. B., Sitte, P. & Maier, U.-G. (1992) *Curr. Genet.* **21,** 73–81.
30. Ragan, M. A., Bird, C. J., Rice, E. L., Gutell, R. R., Murphy, C. A. & Singh, R. K. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 7276–7280.
31. Woese, C. R., Achenbach, L., Rouviere, P. & Mandelco, L. (1991) *Syst. Appl. Microbiol.* **14,** 364–371.
32. Lockhart, P. J., Beanland, T. J., Howe, C. J. & Larkum, A. W. D. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 2742–2746.
33. Lockhart, P. J., Steel, M. A., Hendy, M. D. & Penny, D. (1994) *Mol. Biol. Evol.* **11,** 605–612.
34. Swofford, D. L. & Olsen, G. J. (1990) in *Molecular Systematics,* eds. Hillis, D. M. & Moritz, C. (Sinauer, Sunderland, MA), pp. 411–501.
35. Nelissen, B., Van de Peer, Y., Wilmotte, A. & De Wachter, R. (1995) *Mol. Biol. Evol.* **12,** 1166–1173.
36. Morden, C. W., Delwiche, C. F., Kuhsel, M. & Palmer, J. D. (1992) *BioSystems* **28,** 75–90.
37. Douglas, S. E. & Turner, S. (1991) *J. Mol. Evol* **33,** 267–273.
38. Hatakeyama, N., Sasa, T., Watanabe, M. M. & Takaichi, S. (1991) *J. Phycol.* **27** (Suppl.), A154 (abstr.).
39. Wilhelm, C., Müller, A. M., Borstelmann, B. & Schnetter, R. (1991) *Cryptogam. Bot.* **2,** 201–204.
40. Sasa, T., Takaichi, S., Hatakeyama, N. & Watanabe, M. M. (1992) *Plant Cell Physiol.* **33,** 921–925.
41. Gilson, P. & McFadden, G. I. (1995) *Chromosoma* **103,** 635–641.
42. Higashiyama, T., Maki, S. & Yamada, T (1995) *Mol. Gen. Genet.* **246,** 29–36.