MOLECULAR PHYLOGENETICS AND EVOLUTION
Vol. 6, No. 3, December, pp. 391–407, 1996
ARTICLE NO. 0088

Evolution of the Diatoms (Bacillariophyta)

IV. A Reconstruction of Their Age from Small Subunit rRNA Coding Regions and the Fossil Record

WIEBE H. C. F. KOOISTRA AND LINDA K. MEDLIN[1]

Alfred Wegener Institute for Polar and Marine Research, Am Handelshafen, 12, D-27570 Bremerhaven, Germany

Received August 28, 1995; accepted March 22, 1996

Small subunit ribosomal RNA (ssu rRNA) coding regions from 30 diatoms, 3 oomycetes, and 6 pelagophytes were used to construct linearized trees, maximum-likelihood trees, and neighbor-joining trees inferred from both unweighted and weighted distances. Stochastic accumulation of sequence substitutions among the diatoms was assessed with relative rate tests. Pennate diatoms evolved relatively slowly but within the limits set by a stochastic model; centric diatoms exceeded those limits. A rate distribution test was devised to identify those taxa showing an aberrant distribution of base substitutions within the ssu rRNA coding region. First appearance dates of diatom taxa from the fossil record were regressed against their corresponding branch lengths to infer the average and earliest possible age for the origin of the diatoms, the pennate diatoms, and the centric diatom order Thalassiosirales. Our most lenient age estimate (based on the median-evolving diatom taxon in the maximum-likelihood tree or on the average branch length in a linearized tree) suggests that their average age is approximately 164–166 Ma, which is close to their earliest fossil record. Both calculations suggest that it is unlikely that diatoms existed prior to 238–266 Ma. Rate variation among the diatoms' ssu rRNA coding regions and uncertainties associated with the origin of extant taxa in the fossil record contribute significantly to the variation in age estimates obtained. Different evolutionary models and the exclusion of fast or slow evolving taxa did not significantly affect age estimates; however, the inclusion of aberrantly fast evolving taxa did. Our molecular clock calibrations indicate that the rRNA coding regions in the diatoms are evolving at approximately 1% per 18 to 26 Ma, which is the fastest substitution rate reported in any pro- or eukaryotic group of organisms to date.

[1] To whom reprint requests should be addressed. Fax: 49-471-4831-425. E-mail: lmedlin@awi-bremerhaven.de.

INTRODUCTION

The diatoms are one of the best described microalgal groups and are renowned for their siliceous cell walls with a myriad of ornamentations and perforations. These cell wall patterns are so precise that they form the basis for diatom taxonomy and systematics (Round *et al.*, 1990). Traditionally, two groups have been recognized. Centric diatoms have radial symmetry of their cell wall patterns, whereas pennate diatoms are bilaterally symmetrical (Round *et al.*, 1990). Further, raphid pennate diatoms have a slit (raphe) in the cell wall for movement; the araphid pennates lack this slit and are nonmotile. The siliceous cell walls preserve remarkably well, and consequently, diatoms have a detailed fossil record (Forti and Schulz, 1932; Strel'nikova and Martirosjan, 1981; Gersonde and Harwood, 1990; Harwood and Gersonde, 1990).

Many workers have speculated on the origin of the diatoms. Diatoms may have been derived from a spherical uniformly scaled monad (Round, 1981; Round and Crawford, 1981, 1984) with an anterior flagellum (Cavalier-Smith, 1986) or from a cyst-like form similar to the extant Parmales in the chrysophyte algae (Mann and Marchant, 1989). The presence of scales on diatom zygote cell walls indicates that a scaly ancestor existed at some point in their phylogeny. Recent phylogenies based on nuclear-encoded small-subunit ribosomal RNAs (ssu rRNA) place the diatoms within the pigmented heterokont algal lineages (Bhattacharya *et al.*, 1992; Leipe *et al.*, 1994) and specifically related to a select group of these algae with a reduced flagellar apparatus that lacks the two central microtubules (Saunders *et al.*, 1995).

Although the first reliable diatom fossil record is of a centric form from the early Jurassic (ca. 185 Ma; Rothpletz, 1896), it is generally believed that diatoms originated further back in time than their fossil record indicates (Round *et al.*, 1990). Early records are thought to have been destroyed by diagenesis (Round *et al.*, 1990). After the early Jurassic record, no diatoms are recovered until the Lower Cretaceous where highly

1055-7903/96 $18.00

diversified centric diatoms occur in many deposits believed to have originated from shallow-water marine environments (115–110 Ma; Forti and Schulz, 1932; Strel'nikova and Martirosjan, 1981; Gersonde and Harwood, 1990; Harwood and Gersonde, 1990). A period of poorly preserved deposits occurs between the Lower Cretaceous (110 Ma) and the Upper Cretaceous (87–65 Ma). After that, diatomaceous deposits are widespread and more or less continuous until the present day. Although most Lower Cretaceous taxa are extinct by the Upper Cretaceous (Harwood, 1988; Strel'nikova, 1990), some taxa survive this period and are clearly related to extant diatoms. At this time the first pennate diatoms, the araphids, appear in the fossil record (Moshkovitz *et al.*, 1983). The first raphid pennate diatoms appear at ca. 50 Ma (Strel'nikova, 1990). Today, they represent the most diverse group in the diatoms (Round *et al.*, 1990).

A comparison of ssu rRNA coding regions from 30 diatoms (Medlin *et al.*, 1996a,b) recovered a basic divergence within the group that disagrees with their current systematics. None of the features delimiting the three diatom classes (the centrics, the araphid pennates, or the raphid pennates) could be correlated with the two clades recovered in the ssu rRNA tree. Instead, the two clades in the tree are correlated with morphological features present in certain Lower Cretaceous taxa and with organelle arrangement in extant taxa. Both the centric and the araphid pennate diatoms were shown to be paraphyletic.

In this paper we calculate the evolutionary age of the diatoms, the pennate diatoms, and one order of centric diatoms, the Thalassiosirales, by reconstructing a time scale for their evolution based on a regression of first appearance dates of diatom taxa from the fossil record with their branch lengths in the ssu rRNA phylogeny. These data are used to estimate the average and earliest possible age for these lineages (Hillis and Moritz, 1990).

To address biases that can affect the accuracy and precision of these calculations, we have investigated the variation in the sequence substitution rate using the relative rate tests of Wu and Li (1985), Tajima (1993), and Li and Bousquet (1992) and designed a test to recognize taxa showing an aberrant distribution of base substitutions within the ssu rRNA coding regions. We investigated the effect of different evolutionary models (linearization, maximum likelihood, and neighbor joining), the effect of different sequence positional weighting schemes, the exclusion of taxa on the tree topology, and the age estimates for the diatoms. We assessed the influence of gaps in the fossil record on our age estimates by predating the origin of taxa first appearing after a gap in the fossil record to various times within the gap. We also reevaluated the use of extant morphological features that define extant taxa in determining the first appearance dates for individual lineages.

MATERIAL AND METHODS

Rate Calculations

Small subunit rRNA coding regions from 30 diatoms (Medlin *et al.*, 1996a,b), three oomycetes (Neefs *et al.*, 1991), and six pelagophytes (Saunders *et al.*, 1995) were used in this study (Table 1). Unalignable positions from helices 10, E10-1, E21-1, 41, and 47 (Neefs *et al.*, 1991) were removed prior to analysis. The final data set

TABLE 1

Sources of Small Subunit rRNA Sequences Used in This Study

Taxa	GenBank Accession No.
Oomycetes	
Lagenidium giganteum Couch	X54266
Phytopthora megasperma Drech.	X54265
Achyla bisexualis Coker	M32705
Pelagophyceae	
Apedinella radians (Lohm.) Campb.	U14384
Dictyochoa speculum Ehr.	U14385
Pelagococcus subviridis And.	U14386
Pseudopedinella elastica Skuja	U14387
Rhizochromulina cf. *marina* Hibb. & Crét.-Din.	U14388
Pelagomonas calceolata And. & Saund.	U14389
Bacillariophyceae	
Centric diatoms	
Stephanopyxis cf. *broschii* Grun.	M87330
Rhizosolenia setigera Brightw.	M87329
Coscinodiscus radiatus Ehrenb.	X77705
Corethron criophilum Castracane	X85400
Actinocyclus curvatulus Janisch	X85401
Aulacoseira distans (Ehrenb.) Sim.	X85403
Aulacoseira ambigua (Grun.) Sim.	X85404
Melosira varians C. Ag.	X85402
Chaetoceros didymus Ehrenb.	X85392
Chaetoceros rostratus Lauder	X85391
Cymatosira belgica Grun. in Van Heurck	X85387
Papiliocellulus elegans Hasle, v. Stosch & Syv.	X85388
Eucampia antarctica (Castr.) Mangin	X85389
Streptotheca thamesis Shrubsole	X85385
Ditylum brightwellii (West) Grunow	X85386
Thalassiosira eccentrica (Ehrenb.) Cleve	X85396
Thalassiosira rotula Meunier	X85397
Skeletonema costatum (Grev.) Cl.	X52006
Skeletonema costatum (Grev.) Cl.	X85395
Skeletonema pseudocostatum Medlin	X85393, X85394
Porosira glacialis (Grun.) Jørgensen	X85398
Lauderia borealis Gran	X85399
Thalassionema nitzschioides (Grun.) V.H.	X77702
Rhaphoneis cf. *belgica* (Grun.) Grun.	X77703
Raphid pennate diatoms	
Nitzschia apiculata (Greg.) Grun.	M87334
Bacillaria paxillifer (Müll.) Hend.	M87325
Cylindrotheca closterium (Ehrenb.) Reim. & Lewin	M87326

contained 1711 of a possible 1800 positions. Differences in the substitution rates between lineages were tested according to Li and Bousquet (1992) and among all sequences according to Wu and Li (1985) and Tajima (1993) using distances generated from the Jukes and Cantor (1969) evolutionary model.

To determine if taxa with a significantly elevated substitution rate were evolving as to be expected for ssu rRNA coding regions, we tested the observed distribution of sequence differences along their ssu rRNA coding region against an expected distribution of base differences from a revised model of ssu rRNA coding region evolution by van de Peer *et al.* (1993 and pers. comm.). Their model makes use of the fact that the substitution probability is not equally distributed over the entire sequence but depends on its site in the sequence. Sequence positions can be ordered into four classes, each with its own average variability (v_i). The observed number of base differences in variability class i along a test taxon's sequence ($m_{i,\mathrm{obs}}$) is obtained by recording all positions belonging to variability class i in van de Peer's model where the base is different from that in the test taxon's nearest neighbor *and* its next nearest neighbor. Positions containing a gap in any of the three sequences were excluded from the analysis. The class with the highest variability from van de Peer's model was also excluded because some of the sequence positions belonging to this class could not be reliably aligned with our data set.

The expected distribution of observed base differences over the variability classes in this test taxon is dependent upon each class size and its average variability and is calculated as follows. The observed fraction of differences, $f_{s,\mathrm{obs}}$, in all variability classes along the sequence is

$$f_{s,\mathrm{obs}} = \sum_i (m_{i,\mathrm{obs}})/N_s, \tag{1}$$

where N_s is the sum of the included positions in all variability classes,

$$N_s = \sum_i N_i, \tag{2}$$

and N_i is the number of positions in variability class i.

The Jukes–Cantor distance (*op. cit.*), $d_{s,\mathrm{obs}}$, of the observed fraction of differences in all variability classes is

$$d_{s,\mathrm{obs}} = -(3/4)/v_s * \ln (1 - 4/3 * f_{s,\mathrm{obs}}), \tag{3}$$

where v_s is the average variability over the included variability classes,

$$v_s = \left[\sum_i (v_i * N_i)\right]/N_s, \tag{4}$$

and where v_i is the average variability of the positions in variability class i as calculated by van de Peer *et al.* (1993 and pers. comm.). The incorporation of v_s into Eq. (3) is needed if certain sequence positions are excluded from the test. With all sequence positions included, v_s should be ca. 1. The expected fraction, $f_{i,\mathrm{exp}}$, of base differences in variability class v_i is obtained from the inverse of the Jukes–Cantor equation,

$$f_{i,\mathrm{exp}} = (3/4) * [1 - e^{(-4/3 * v_i * d_{s,\mathrm{obs}})}]. \tag{5}$$

The expected number of sequence differences, $m_{i,\mathrm{exp}}$, in variability class v_i is then

$$m_{i,\mathrm{exp}} = f_{i,\mathrm{exp}} * N_i. \tag{6}$$

The equality of the expected and the observed distribution is evaluated using a χ^2 test (Rohlf and Sokal, 1981).

Tree Construction Methods

Maximum-likelihood (ML) trees (Felsenstein, 1981) were constructed using fastDNAml (Olsen, *et al.*, 1994). Neighbor-joining (NJ) trees were constructed from Jukes–Cantor (JC) distances and from both weighted and unweighted Kimura-two-parameter (K-2-p) distances (Kimura, 1980) using PHYLIP (Felsenstein, 1993). Linearized trees were constructed from unweighted K-2-p distances and their resulting NJ tree using the method developed by Takezaki *et al.* (1995) to correct for unequal rates of evolution. Weighted distances were calculated by using revised weighting classes from the model of van de Peer *et al.* (1993 and pers. comm.), by weighting each position inversely proportional to the number of base substitutions at that position in our own data set (MacClade; Maddison and Maddison, 1992) and by using the gamma distribution (Jin and Nei, 1990) option in MEGA (Kumar *et al.*, 1993). Taxa included in each of these tree construction algorithms are listed below.

Molecular Clock Calibrations

A regression of first appearance dates of diatom taxa from the fossil record against branch lengths of taxa in all trees was performed. The average age of any undated node was determined by multiplying the length of its median or average branch by the regression coefficient. The earliest possible age for any undated node is the upper 95% confidence limit around the age estimate given the distance of its median or average branch (Hillis and Moritz, 1990). Average and earliest possible ages were also given for the shortest and longest branches in that lineage. For the linearized trees, standard errors (SE's) associated with the Kimura-two-parameter distances between neighbor lineages were calculated according to Li and Bousquet (1992), using the next nearest neighbor's taxon with the median branch length as reference. The SE's associated with the height of internal nodes in the trees were obtained by dividing each SE by 2. The 95% upper and

lower limits around the average estimate of the origin were obtained by multiplying the SE/2 by the appropriate critical value of the Student's t distribution corresponding to $n - 1$ taxa in that lineage.

RESULTS

Rate Calculations

Differences in branch lengths can be observed among all diatom lineages, regardless of the tree building algorithm used (Figs. 1 and 2). Because significant substitution rate differences may seriously bias clock calculations, we tested for rate variation between pairs of lineages and among all taxa. The relative rate test of Wu and Li (1985) was applied to all pairwise combinations of diatoms with each of eight reference taxa (six pelagophytes: *Dictyocha speculum, Rhizochromulina* cf. *marina, Pseudopedinella elastica, Apedinella radians, Pelagococcus subviridis,* and *Pelagomonas calceolata* and two oomycetes: *Achlya bisexualis* and *Lagenidium giganteum*) to identify significantly fast or slow evolving taxa. Table 2 shows the average and standard deviation of the relative rate of evolution for each pairwise combination of the diatoms with the eight reference taxa. At the 95% confidence level, *Eucampia, Chaetoceros sp., Chaetoceros didymus, Porosira, Lauderia, Coscinodiscus, Actinocyclus,* and *Melosira* accumulated base substitutions significantly faster than all other diatoms. In addition, *Skeletonema spp.* and *Chaetoceros rostratus* evolved significantly faster than the slowest pennate diatom species. Although the pennates evolved relatively slowly, their substitution rates were comparable to those of many centric taxa. Results from the relative rate test according to Tajima (1993) were not significantly different from those obtained according to Wu and Li (1985) (data not shown).

The relative rate test of Li and Bousquet (1992) was used to assess the rate difference between entire clades or lineages within these clades (Fig. 1A). With all taxa included, the difference in the average distance (D) between clades 1 and 2 in Fig. 1 was 0.0183. The hypothesis that these clades evolve at the same speed was rejected ($P < 0.0001$). With all significantly fast evolving taxa (Table 2) removed, the hypothesis that the two clades evolved at the same speed was not rejected ($D = 0.0092$, $P > 0.05$). Using *Stephanopyxis* as reference taxon, the pennates evolve significantly slower than all Thalassiosirales ($D = 0.0271$, $P < 0.0001$). With the fast evolving *Porosira* and *Lauderia* removed from the Thalassiosirales, the hypothesis that the two lineages evolved at the same speed was rejected at the 0.05 probability level ($D = 0.0118$).

To determine if the taxa with a significantly elevated substitution rate (Table 2) were evolving according to a

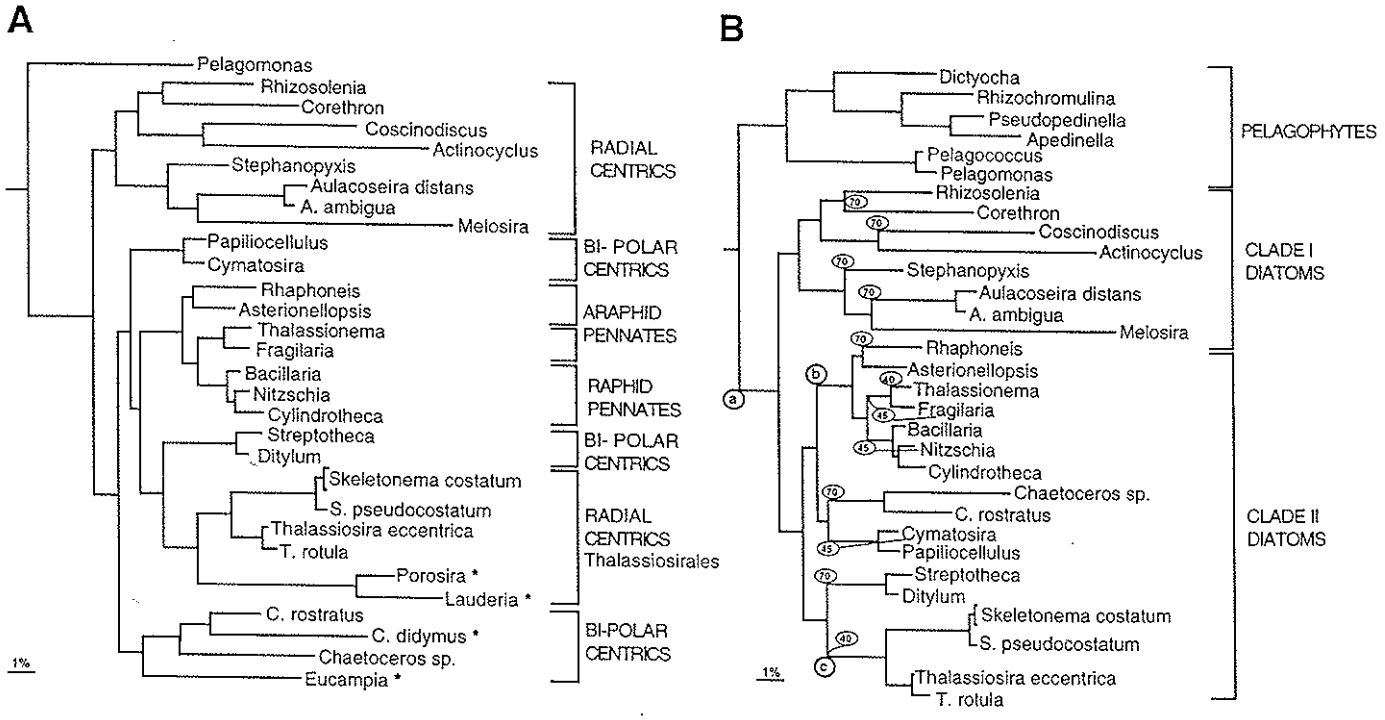

FIG. 1. Phylogeny of the diatoms inferred from a maximum-likelihood analysis (FastDNAml, Olsen *et al.,* 1994). (A) Analysis including 30 diatoms, 1 pelagophyte, and 3 oomycetes, the latter pruned from the left of the tree. (B) Analysis including 6 pelagophytes, 3 oomycetes and 26 diatom taxa; the aberrantly evolving diatom taxa (see Table 3) were excluded. The oomycetes were pruned from the left of the tree. Encircled numbers indicate first appearance (in Ma) in the fossil record of the genus or family to which the taxa to the right belong. Encircled letters indicate (a) origin of the diatom lineage, (b) origin of the pennate lineage, and (c) origin of the Thalassiosirales.

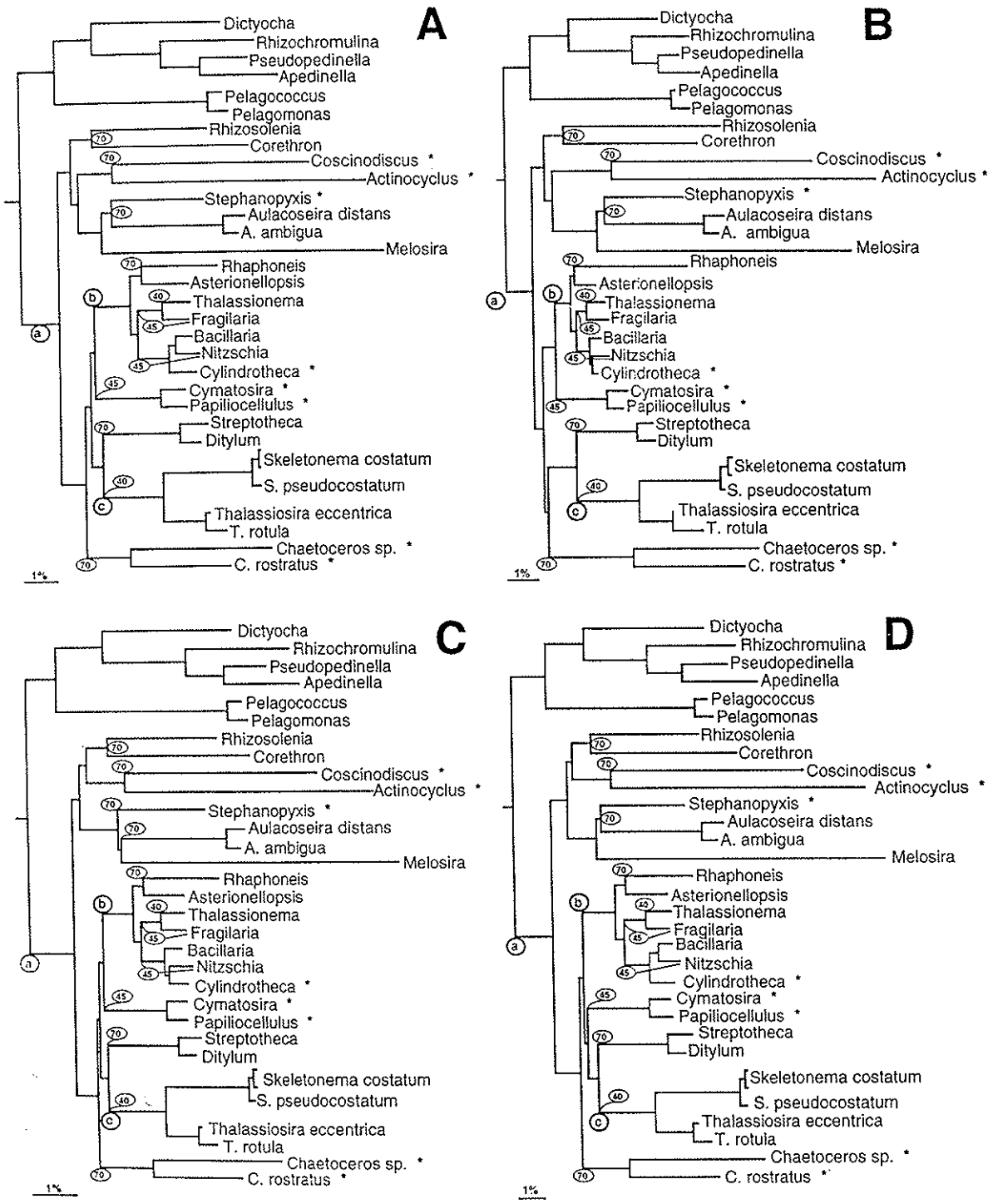

FIG. 2. Phylogeny of the diatoms inferred from a neighbor-joining analysis using Kimura-two-parameter distances under various evolutionary models: (A) unweighted; (B) positional weighting according to van de Peer *et al.* (1993 and pers. comm.); (C) positional weighting inversely proportional to the number of base substitutions at that position, inferred using MacClade (Maddison and Maddison, 1992); and (D) positional weighting using the gamma distribution (Jin and Nei, 1990) option in MEGA (Kumar *et al.*, 1993) with $\alpha = 1$. Encircled numbers indicate first appearance (in Ma) in the fossil record of the genus or family to which the taxa to the right belong. Encircled letters indicate (a) origin of the diatom lineage, (b) origin of the pennate lineage, and (c) origin of the Thalassiosirales. Taxa marked with an asterisk change position in the various analyses.

TABLE 2

Relative Rate Test of Wu and Li (1985) Applied to Jukes–Cantor (1969) Distances between All Pairwise Combinations of 27 Diatom Taxa[a] to a Reference Taxon

		Pennates						Centrics																		
		Aster	Rhap	Thaln	Frag	Baxil	Nitzs	Cylin	Cymat	Papit	Eucam	C.sp.	C.rost	C.didy	Strep	Dityl	Skel	Thala	Poro	Laud	Rhizo	Coret	Cosc	Acti	Steph	Melo	Aula
Pennates	Asterionellopsis		1,8	0,2	0,0	0,9	1,5	1,3	0,6	0,5	4,0	3,5	3,1	5,1	1,4	1,2	2,8	1,3	4,7	5,6	2,1	2,1	5,0	6,1	1,6	6,0	2,0
	Rhaphoneis	0,4		-1,4	-1,7	-0,8	-0,2	-0,4	-0,8	-0,9	2,7	2,2	1,6	3,8	0,0	-0,2	1,5	0,0	3,4	4,3	0,9	0,9	3,6	4,8	0,3	4,8	0,6
	Thalassionema	0,5	0,7		-0,3	0,8	1,4	1,1	0,4	0,3	3,7	3,3	2,8	4,8	1,2	1,0	2,8	1,1	4,5	5,3	2,0	1,9	4,8	5,7	1,4	5,9	1,8
	Fragilaria	0,5	0,5	0,5		1,0	1,6	1,3	0,6	0,5	3,7	3,4	3,0	4,9	1,4	1,2	2,8	1,3	4,6	5,4	2,1	2,1	4,9	5,8	1,6	6,0	2,0
	Bacillaria	0,7	0,7	0,7	0,7		1,0	0,5	-0,2	-0,3	3,3	2,9	2,3	4,5	0,7	0,5	2,2	0,6	4,0	4,9	1,5	1,5	4,3	5,3	0,9	5,5	1,3
	Nitzschia	0,7	0,6	0,6	0,7	0,4		-0,4	-0,7	-0,8	2,9	2,4	1,8	4,1	0,2	0,0	1,7	0,1	3,6	4,6	1,1	1,1	4,3	4,9	0,4	5,2	0,8
	Cylindrotheca	0,6	0,7	0,4	0,6	0,8	0,6		-0,5	-0,6	3,0	2,6	2,0	4,2	0,4	0,2	1,9	0,4	3,9	4,7	1,2	1,3	4,0	4,8	0,6	5,3	1,0
Centrics	Cymatosira	0,6	0,6	0,4	0,5	0,4	0,4	0,5		-0,2	2,8	3,0	2,6	4,6	0,9	0,6	2,4	0,8	4,0	5,0	1,7	1,6	4,6	5,3	1,0	5,4	1,4
	Papiliocellulus	0,6	0,5	0,5	0,5	0,4	0,4	0,5	0,4		3,8	3,5	2,9	4,6	1,0	0,7	2,5	0,9	4,5	5,2	1,8	1,7	4,8	5,4	1,1	5,5	1,4
	Eucampia	0,5	0,6	0,4	0,5	0,5	0,4	0,3	0,5	0,6		-0,5	-1,3	1,3	2,7	2,9	-1,2	2,7	0,9	1,8	-1,7	-1,6	1,2	2,3	-2,4	2,5	-2,0
	Chaetoceros sp.	0,4	0,5	0,5	0,4	0,7	0,6	0,5	0,7	0,7	0,5		-0,8	1,9	-2,2	-2,4	-0,7	-2,3	1,3	2,2	-1,3	-1,2	1,6	2,6	-1,9	3,1	-1,5
	C. rostratus	0,7	0,8	0,6	0,7	0,8	0,8	0,5	0,8	0,8	0,4	0,7		3,0	-1,6	-1,9	0,0	-1,7	2,0	2,8	-0,6	-0,5	2,3	3,4	-1,3	3,5	-0,8
	C. didymus	0,9	1,0	0,8	0,9	1,0	1,0	0,7	1,0	1,1	0,6	0,7	0,7		3,9	4,1	-2,4	4,0	-0,3	0,5	3,0	2,8	-0,1	1,1	3,4	1,3	3,0
	Streptotheca	0,8	0,8	0,5	0,6	0,4	0,4	0,5	0,3	0,3	0,6	0,8	0,9	1,1		-0,5	1,5	-0,1	3,0	4,4	0,9	0,9	3,6	4,7	0,2	5,0	0,6
	Ditylum	0,9	0,9	0,6	0,7	0,4	0,5	0,6	0,5	0,5	0,6	0,9	0,9	1,1	0,6		1,8	0,2	3,7	4,5	1,1	1,1	3,8	4,9	0,4	5,0	0,8
	Skeletonema	0,4	0,3	0,5	0,5	0,3	0,3	0,4	0,3	0,4	0,5	0,6	0,7	1,0	0,5	0,7		-2,1	2,2	3,0	-0,6	-0,5	2,2	3,3	-1,3	3,6	-0,8
	Thalassiosira	0,4	0,4	0,7	0,6	0,7	0,7	0,7	0,7	0,7	0,7	0,6	0,7	1,1	0,9	1,1	0,6		3,9	4,7	0,9	1,0	3,7	4,9	0,3	4,9	0,6
	Porosira	0,5	0,5	0,5	0,6	0,4	0,3	0,4	0,5	0,6	0,4	0,7	0,6	0,9	0,5	0,5	0,5	0,7		1,6	-2,6	-2,4	0,2	1,4	-3,2	1,6	-2,7
	Lauderia	0,4	0,5	0,5	0,6	0,3	0,3	0,4	0,4	0,5	0,5	0,7	0,6	0,9	0,5	0,5	0,4	0,6	0,3		-3,4	-3,3	-0,6	0,5	-4,0	0,8	-3,6
	Rhizosolenia	0,9	0,9	0,7	0,8	0,7	0,7	0,7	0,6	0,6	0,6	1,0	0,7	1,0	0,5	0,4	0,8	1,0	0,6	0,6		0,1	3,0	4,2	-0,7	4,3	-0,3
	Corethron	0,8	0,8	0,7	0,8	0,5	0,4	0,6	0,6	0,6	0,5	0,9	0,8	1,0	0,4	0,3	0,7	1,0	0,4	0,5	0,6		2,8	4,0	-0,7	4,2	-0,3
	Coscinodiscus	1,0	1,1	0,8	0,9	0,9	0,9	0,7	0,9	1,0	0,7	0,9	0,6	0,7	0,9	0,8	1,0	1,1	0,8	0,9	0,6	0,8		1,3	-3,6	1,5	-3,0
	Actinocyclus	0,6	0,8	0,5	0,6	0,7	0,7	0,5	0,6	0,7	0,4	0,5	0,3	0,4	0,7	0,7	0,7	0,8	0,6	0,7	0,7	0,8	0,5		-4,6	0,3	-4,1
	Stephanopyxis	0,4	0,4	0,4	0,5	0,2	0,2	0,4	0,3	0,3	0,5	0,6	0,8	1,0	0,5	0,6	0,2	0,5	0,4	0,3	0,8	0,6	1,0	0,7		5,2	0,5
	Melosira	0,9	0,9	0,8	0,9	0,7	0,7	0,7	0,8	0,8	0,7	1,0	0,7	0,9	0,8	0,7	0,8	0,8	0,6	0,7	0,7	0,6	0,7	0,9	0,9		4,7
	Aulacoseira	0,6	0,5	0,6	0,5	0,5	0,5	0,7	0,4	0,3	0,8	0,7	0,9	1,2	0,5	0,7	0,4	0,6	0,6	0,6	0,8	0,8	1,1	0,9	0,5	1,0	

model of evolution established for the ssu rRNA coding region (van der Peer *et al.*, 1993 and pers. comm.), we performed a rate distribution test. The observed numbers of base differences among each fast evolving taxon and its two nearest neighbors was recorded for each variability class and tested against the expected distribution over the variability classes (Table 3). The observed base differences were strongly skewed toward the classes with higher variation, i.e., the more variable positions in the ssu rRNA coding regions. Although the observed distribution of base differences along the branches leading to *Actinocyclus, Coscinodiscus, Chaetoceros sp.*, and *Melosira* and along the long internode leading to *Skeletonema spp.* was skewed, it did not deviate significantly from the expected distribution. However, the skewness of the distribution of sequence differences along the branches leading to *C. didymus, Eucampia,* and along the internode to *Lauderia/Porosira* (see Fig. 1A) did deviate from that expected.

Phylogenetic Reconstruction

The ML tree, presented in Fig. 1A, was inferred from the ssu rRNA coding regions of one pelagophyte, three oomycetes, and all diatoms (Table 1); the one presented in Fig. 1B was inferred from a data set including five additional pelagophyte sequences but excluding the aberrantly evolving diatoms (Table 3). The oomycetes were pruned from the left of the trees. Diatom phylogeny is not a star phylogeny, i.e., a rapid radiation within a short period of time relative to the length of the lineages. Diatoms diverge initially into two clades, which are labeled in Fig. 1B. The first clade contains radial centric diatoms with specialized tubes (labiate processes) located in a peripheral position through the cell wall (Round *et al.*, 1990). The second clade is composed of bi(multi) polar centric diatoms with small, usually centrally located labiate processes; the Thalassiosirales, with different centrally located tubes (strutted processes; Round *et al.*, 1990); and the pennate diatoms. In the latter group, only the araphid pennate diatoms have labiate processes, which are located apically in the bilaterally symmetrical cell. Further morphological and cytological support for the two clades is detailed in Medlin *et al.* (1996a,b). Both the centric and the araphid pennate diatoms are paraphyletic (see

TABLE 3

Test of the Observed Distribution of Sequence Differences within Three[a] of Four Variability Classes[b] as Recognized in the Nuclear ssu rRNA Molecule by van de Peer *et al.* (1993 and pers. comm.) against an Expected Distribution of an Equal Number of Sequence Differences over these Variability Classes (van de Peer *et al.* 1993 and pers. comm.)

	$m_{observed}$[c]				$m_{expected}$[d]				
Taxon	m_1	m_2	m_3	m_s	m_1	m_2	m_3	χ^2[e]	P
Actinocyclus	10	16	39	65	12.7	18.0	34.3	1.498	n.s.
Coscinodiscus	9	21	39	69	13.5	19.1	36.4	1.955	n.s.
Melosira	10	12	40	62	12.1	17.2	32.7	3.611	n.s.
Chaetoceros sp.	6	10	25	41	7.9	11.3	21.8	1.093	n.s.
C. didymus	4	12	33	49	9.5	13.5	26.0	5.322	$P < 0.1$
Eucampia	7	15	42	64	12.5	17.7	33.7	4.959	$P < 0.1$
Lauderia	3	5	17	25	4.8	6.9	13.4	2.168	n.s.
Porosira/Lauderia	8	13	47	68	13.3	18.9	35.8	7.592	$P < 0.025$
Skeletonema spp.	1	6	16	23	4.4	6.3	12.3	3.764	n.s.
Rhaphoneis	7	9	8	24	4.6	6.6	12.8	4.017	n.s.

[a] Part of the sequence positions belonging to variability class 4 of van de Peer (pers. comm.) could not be reliably aligned in our data set and were excluded from the test. N_1, N_2, and N_3 contained 764, 350, and 334 positions, respectively.

[b] Positions in variability classes v_1, v_2, and v_3 have an average variability of 0.18, 0.57, and 1.18, respectively (van de Peer, pers. comm.).

[c] The observed number of base differences in each of the three variability classes along a lineage (m_{1obs}, m_{2obs}, m_{3obs}) is obtained as follows: all positions with a base difference defining an end branch are recorded; these are the positions where the base is different from both the reference taxon's nearest neighbor and its next nearest neighbor. Gaps in any of the three sequences are omitted. Each recorded position is assigned to one of the four variability classes (or omitted when in class 4) of van de Peer (pers. comm.).

[d] The expected distribution of base differences in each of the four variability classes along the same lineage is determined as follows: The total number of recorded differences in the four variability classes ($m_{s,obs}$) is divided by the total number of positions in the four variability classes ($N_s = \Sigma N_i$). From the obtained dissimilarity $f_s = [(m_{s,obs})/N_s]$, the Jukes and Cantor (1969) distance is calculated according to equation 4 in van de Peer *et al.* (1993), where $v_s = [\Sigma(v_i * N_i)]/N_s$. The expected dissimilarity for each of the variability classes (v_1, v_2, and v_3) is then calculated according to Eq. (3b) in van de Peer *et al.* (1993), where $f_i = (3/4) * [1 - \exp(-4/3 * v_i * d_s)]$ and the expected number of base differences in a variability class v_i is $m_{i,exp} = f_i * N_i$.

[e] The equality $m_{obs} = m_{exp}$ is tested using $\chi^2 = \Sigma[(m_{i,obs} - m_{i,exp})^2/m_{i,exp}]$ with three degrees of freedom.

TABLE 2. *Note.* The test statistic is the distance from the reference taxon to taxon 1 minus the distance from the reference taxon to taxon 2, divided by the square root of the variance[b] along these distances. This test statistic follows the standard normal distribution.[c] Entries above the diagonal are averages based on eight outgroup taxa.[d] Entries below the diagonal are standard deviations associated with these averages. Values shaded with darker stippling indicate that the taxon in the column evolves significantly faster than the taxon in the row. Values shaded with lighter stippling are those where the taxon in the column evolves significantly slower than the taxon in the row.

[a] The taxa *Aulacoseira distans, Thalassiosira rotula, Skeletonema costatum* (Atlantic isolate), and *S. pseudocostatum* were excluded from the table because the results are almost identical to those of *A. ambigua, T. rotula,* and *S. costatum* (Pacific isolate), respectively.

[b] Covariances were calculated according to the footnote of Table 2 in Li and Tanimura (1985). Note the typo error in P_{03} of Li and Tanimura (1985) (4/3 should be −4/3).

[c] Double sided. $P(2.576) = 0.01$.

[d] The pelagophytes *Dictyocha sp., Rhizochromulina sp., Pseudopedinella sp., Apedinella sp.,* and *Pelagomonas calceolata* and the oomycetes *Achyla bisexualis* and *Lagenidium giganteum* were used as reference taxa. The average value was tested.

labels in Fig. 1A). Relationships within each of the clades correspond well to ordinal level in current diatom systematics (Round *et al.*, 1990).

Neighbor-joining trees resulting from the unweighted analysis and the three different sequence positional weighting strategies are presented in Fig. 2. All of them show the same basic ramification into the two clades as in the ML tree (Figs. 1A and 1B), with minor changes in the positions of some taxa within each clade and some differences in branch lengths. These positional changes are indicated by an asterisk beside the relevant taxa in Fig. 2 and only affect the branch lengths associated with some dating points.

The relative rate test (Table 2) and the varying branch lengths seen in the trees (Figs. 1 and 2) indicate large rate differences among the taxa, which can distort tree topology. We tested the effect of the removal of various taxa on tree topology. If the aberrantly fast evolving *Eucampia, C. didymus, Porosira,* and *Lauderia* (Table 3) were included in the analysis, tree topology did not change significantly (compare Figs. 1A and 1B). The most noticeable effect was that *Eucampia* and *Chaetoceros* spp. were grouped together, but this is not taxonomically unsound. If the significantly fast evolving *Melosira, Actinocyclus, Coscinodiscus,* and *Chaetoceros sp.* (Table 2) were also eliminated from the analysis, then the branching order of the resulting NJ tree was similar (data not shown). If the slow evolving *Cymatosira, Papiliocellulus, Fragilaria,* and *Asterionellopsis* were eliminated from the analysis, the resulting tree topology also did not change (data not shown).

A linearized NJ tree derived from Kimura-two-parameter distances is presented in Fig. 3. All significantly faster evolving taxa were excluded from this analysis. The linearization effectively eliminates any other rate variation among the taxa once the significantly fast evolving taxa are eliminated (Takezaki *et al.*, 1995). In this tree, the taxa belonging to Clade 2 (see above, i.e., the pennates, the Thalassiosirales, and three lineages leading to the bipolar centric diatoms) collapse into a polytomy. The branch leading to Clade 2 also collapses into a polytomy with two lineages containing Clade 1 taxa.

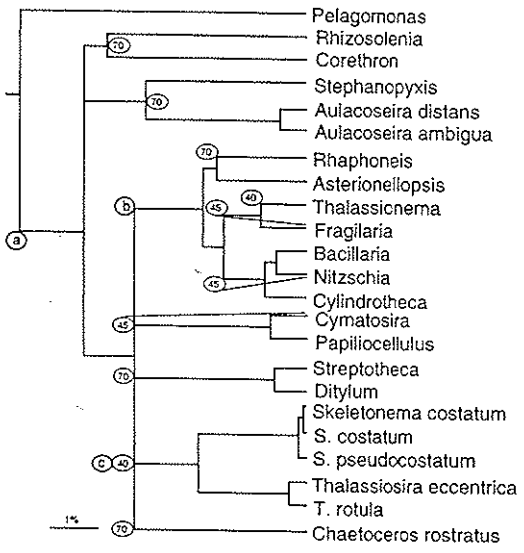

FIG. 3. Linearized neighbor-joining tree constructed from unweighted Kimura-two-parameter distances and the unlinearized NJ tree according to Takezaki *et al.* (1995). All significantly faster evolving taxa were excluded from the analysis.

The Fossil Record

Times of origin for extant taxa were obtained from their first appearance dates (see encircled numbers at the internal nodes in Figs. 1B, 2, and 3) with the following exceptions. The diatom fossil record shows an extensive gap from 185 to 115 Ma and a period of poorly preserved diatom deposits from 110 to 70 Ma. These latter deposits occur at a time when diatom wall morphology changes dramatically (Gersonde and Harwood, 1990) and certain morphological features currently used in diatom systematics, e.g., the strutted or labiate processes, had not yet developed. The taxa that appear prior to the second gap are very distinct from those postdating the gap, with few clear indications of taxa that are related. Thus, taxa with a first appearance in the fossil record at 70 Ma may have originated anytime between 115 and 70 Ma or before. Therefore, we formulated four hypotheses:

Hypothesis A—taxa that first appear at 70 Ma are 70 Ma old.

Hypothesis B—taxa that first appear at 70 Ma have their origin in the middle of the period of poor preservation, i.e., at 90 Ma.

Hypothesis C—the same as hypothesis B but *Aulacoseira, Stephanopyxis,* and *Chaetoceros* are assumed to appear at 115 Ma. The overall cell wall structure of *Aulacoseira, Stephanopyxis,* and *Chaetoceros* is very similar to that of three extinct taxa in the 110–115 Ma deposit: *Archepyrgus, Amblypyrgus,* and *Calyptosporium,* respectively (see illustrations in Gersonde and Harwood, 1990). However, the extinct taxa lack any special processes through the cell wall that define our assumed modern descendants. Nevertheless, we have assumed that these three extinct taxa are ancestors of the three extant genera based on overall valve structure and/or linking mechanisms.

Hypothesis D—the same as hypothesis C but the Thalassiosirales are assumed to appear at 115 Ma. If we restrict the definition of the strutted process as the morphological trait determining the origin of the Thalassiosirales, this group appears first at 40 Ma. However, fossil taxa believed to be direct ancestors of the Thalassiosirales, e.g., *Thalassiosiropsis* (Hasle and Syvertsen, 1985) and *Praethalassiosiropsis* (Gersonde and Harwood, 1990), appear at 70 and 115 Ma, respectively.

These taxa possess tubes that are likely progenitors of the strutted processes.

We have calculated time estimates under the above four hypotheses for the origin of the diatoms, the pennates, and the Thalassiosirales, first with a linearized tree with all of the fast taxa excluded and then with trees obtained under different evolutionary models and with different sets of taxa.

Time Estimates

The linearized tree. Branch lengths from the linearized unweighted NJ tree with all significantly faster evolving taxa omitted (Fig. 3) were regressed against first appearance dates from the fossil record under each hypothesis. The effect of the linearization on the age estimates is presented in Table 4. Time estimates under hypothesis C were also calculated for an unlinearized tree with all fast taxa removed. Age estimates based on the upper and lower 95% confidence limits of the branch lengths of lineages as well as those based on the longest and shortest branch of lineages are reported for comparison.

The average age of the diatoms (Fig. 3, point a) under each hypothesis is younger than their earliest fossil record, whereas the earliest possible age (except for hypothesis A) predates this record. Hypotheses C and D suggest that the earliest possible age is ca. 231–238 Ma and does not differ significantly from those obtained from a nonlinearized tree, either with all fast taxa eliminated (228–230 Ma, Table 4) or just the aberrant ones (225 Ma, Table 5). The average age of the pennates (Fig. 3, point b) falls in the period of poor preservation of diatom fossils, after which the first record of a pennate appears. Estimates for the origin of the pennate lineage are older in the linearized tree than in the nonlinearized tree because the length of the internode leading to slowly evolving pennates is lengthened by the linearization procedure. Time estimates suggest

TABLE 4

Estimates of the Origin of the Diatoms, the Pennates, and the Thalassiosirales According to Hillis and Moritz (1990)

			Origin of the diatoms based on all lineages			Origin of the Pennates based only on the pennate lineage			Origin of Thalassiosirales based only on this lineage		
	S(Ma,d)[a]	Rate[b]	95%[c] upper limit	Average	95%[c] lower limit	95% upper limit	Average	95% lower limit	95% upper limit	Average	95% lower limit
Linearized NJ tree from Kimura distances											
Hypothesis A[d]	9.3	17.9 ± 2.9	128 (183)	113 (164)	98 (146)	78 (121)	70 (110)	62 (99)	85 (130)	70 (110)	55 (90)
Hypothesis B[e]	11.7	21.7 ± 3.7	156 (224)	137 (201)	119 (179)	95 (148)	85 (135)	75 (122)	104 (159)	85 (135)	66 (110)
Hypothesis C[f]	14.4	24.1 ± 4.6	173 (257)	152 (231)	132 (205)	106 (171)	94 (156)	83 (141)	115 (183)	94 (156)	74 (127)
Hypothesis D[g]	13.0	26.2 ± 4.0	188 (264)	166 (238)	144 (210)	115 (174)	103 (159)	90 (143)	125 (187)	102 (159)	80 (129)
NJ tree from Kimura distances											
Hypothesis C	19.4	21.4 ± 6.0	145 (255)	127 (230)	109 (204)	80 (160)	69 (144)	58 (127)	113 (210)	94 (183)	76 (155)
			Longest branch	Median branch	Shortest branch	Longest branch	Median branch	Shortest branch	Longest branch	Median branch	Shortest branch
Hypothesis C	19.4	21.4 ± 6.0	161 (277)	126 (228)	110 (205)	84 (167)	66 (140)	64 (136)	107 (201)	107 (201)	72 (149)

Note. Branch lengths are obtained from neighbor-joining trees generated from Kimura-two-parameter distances.[h] The trees were either linearized according to Takezaki *et al.* (1995) or unlinearized. Figures are estimates of the average time of origin of a lineage in Ma, obtained by multiplying the distance by the regression coefficient ("a" in Fig. 3). Figures in parentheses are estimates of the earliest possible time of origin of a lineage based on the upper 95% confidence limit of a time estimate given the distance. All significantly faster evolving taxa (Table 2, *Actinocyclus, Chaetoceros, sp., C. didymus, Coscinodiscus, Eucampia, Lauderia, Melosira,* and *Porosira*) were excluded from the analyses.

[a] S(Ma,d) is the standard deviation of the residuals of the points on the regression line through zero.

[b] Average rate in 1% per x Ma from branch point to end node ± the 95% confidence interval around the rate.

[c] Standard errors (SE) associated with the K-2-p distances between the taxa in each pair of neighbor lineages were calculated according to Li and Bousquet (1992), each with a next nearest neighbor as reference taxon. The standard error associated with the average length of the lineage was obtained by dividing the SE associated with the pairwise distances over that node (Li and Bousquet, 1992) by 2. The 95% upper and lower limits around the average estimate of the origin were obtained by multiplying the SE/2 by the appropriate critical value of the Student's *t* distribution corresponding to the n − 1 taxa in that lineage.

[d] Hypothesis A. all taxa that first appear in the fossil record at 70 Ma also have their origin at 70 Ma.

[e] Hypothesis B. all taxa that first appear at 70 Ma reset at 90 Ma; that is in the middle of this gap in the fossil record.

[f] Hypothesis C. the same as hypothesis B but with *Chaetoceros sp., C. rostratus, Aulacoseira spp.,* and *Stephanopyxis* set at 115 Ma because direct predecessor of these genera have been postulated in the fossil record at 115 Ma.

[g] Hypothesis D. the same as hypothesis C but with the Thalassiosirales dating back to 115 Ma.

[h] The neighbor-joining tree generated from K-2p distances was almost identical to the tree based on JC distances. Estimates of origin based on the JC distances differed marginally (max 5 Ma for the earliest possible time of origin values) and have been omitted from the table.

TABLE 5

Estimates of the Origin of the Diatoms, the Pennates and the Thalassiosirales According to Hillis and Moritz (1990)

			Origin of the diatoms based on all lineages			Origin of the Pennates based only on the pennate lineages			Origin of Thalassiosirales based only on this lineage		
	S (Ma,d)[a]	Rate[b]	Longest branch	Median branch	Shortest branch	Longest branch	Median branch	Shortest branch	Longest branch	Median branch	Shortest branch
Maximum likelihood tree											
Hypothesis A[d]	14.5	13.2 ± 3.8	202 (325)	101 (187)	85 (164)	58 (123)	53 (114)	48 (107)	81 (158)	79 (155)	47 (105)
Hypothesis B[e]	17.6	16.1 ± 4.6	247 (397)	123 (228)	104 (200)	71 (151)	64 (139)	59 (131)	99 (192)	97 (190)	58 (129)
Hypothesis C[f]	20.3	18.0 ± 5.3	277 (450)	138 (258)	116 (227)	80 (171)	72 (158)	66 (149)	110 (218)	108 (215)	64 (146)
Hypothesis D[g]	17.1	21.5 ± 4.5	330 (476)	164 (266)	138 (232)	95 (172)	86 (159)	78 (148)	132 (223)	129 (219)	77 (146)
Unweighted NJ tree from Kimura distances											
Hypothesis A	15.0	14.6 ± 4.1	169 (278)	90 (170)	75 (147)	57 (120)	45 (101)	44 (99)	73 (144)	72 (145)	49 (108)
Hypothesis B	17.0	17.9 ± 4.7	207 (332)	111 (202)	92 (175)	70 (142)	55 (119)	53 (117)	90 (171)	89 (171)	60 (127)
Hypothesis C	19.0	20.0 ± 5.2	232 (372)	124 (225)	103 (195)	78 (158)	62 (133)	60 (130)	100 (191)	100 (191)	67 (142)
Hypothesis D	15.6	23.6 ± 4.3	273 (388)	146 (229)	121 (197)	93 (158)	73 (131)	71 (128)	118 (192)	118 (192)	80 (140)
All significantly faster evolving taxa excluded[h]											
Hypothesis C	19.4	21.4 ± 6.0	161 (277)	126 (228)	110 (205)	84 (167)	66 (140)	64 (136)	107 (201)	107 (201)	72 (149)
Eucampia, C. didymus, Porosira, and *Lauderia* included											
Hypothesis C	21.0	15.4 ± 4.8	190 (347)	105 (221)	84 (188)	65 (156)	54 (136)	49 (127)	136 (267)	82 (184)	57 (141)
Four taxa with shortest branches excluded[i]											
Hypothesis C	20.0	20.1 ± 5.7	232 (380)	131 (240)	106 (205)	78 (163)	65 (142)	61 (135)	100 (196)	100 (196)	67 (146)
Weighted NJ tree from Kimura distances											
van de Peer[j]											
Hypothesis C	20.4	18.4 ± 5.4	223 (377)	116 (226)	54 (128)	81 (173)	30 (85)	25 (75)	94 (193)	89 (185)	59 (136)
MacClade[k]											
Hypothesis C	21.4	29.2 ± 7.0	251 (388)	127 (223)	101 (187)	78 (153)	55 (118)	53 (114)	99 (184)	100 (183)	62 (128)
Gamma[l]											
Hypothesis C	17.9	18.8 ± 4.7	245 (386)	125 (224)	103 (193)	74 (151)	60 (128)	58 (125)	100 (188)	99 (188)	66 (138)

Note. Branch lengths are obtained from a maximum likelihood tree and from a neighbor-joining tree generated from weighted and unweighted Kimura-two-parameter distances.[m] Figures are estimates of the average time of origin of a lineage in Ma, obtained by multiplying the distance by the regression coefficient ("a" in Fig. 3). Figures in parentheses are estimates of the earliest possible time of origin of a lineage based on the upper 95% confidence limit of a time estimate given the distance. Unless stated otherwise, *Eucampia, Chaetoceros didymus, Porosira,* and *Lauderia* are excluded from the analyses.

[a–g] See the footnotes in Table 4.

[m] See footnote *h* in Table 4.

[h] Apart from *Eucampia, Chaetoceros didymus, Porosira,* and *Lauderia* we also excluded *Melosira, Actinocylus, Coscinodiscus,* and *Chaetoceros sp.*

[i] *Cymatosira, Papiliocellulus, Fragilaria,* and *Asterionellopsis.*

[j] The positions in the alignment are weighted inversely to the average variability of the variability class to which they belong in van de Peer (pers. comm.) according to the procedure described in van de Peer *et al.* (1993).

[k] The positions in the alignment are weighted inversely to the number of base changes recorded for that position as observed in a most parsimonious tree (0, 1, 2, 3 . . . 14 base changes are weighted 15, 14, 13 . . . 1, respectively).

[l] The rate of nucleotide substitutions at different sites is assumed to follow the gamma distribution (Jin and Nei, 1990) with $\alpha = 1$.

that the Thalassiosirales (Fig. 3, point c) are older than their first fossil record at 40 Ma. Dates for the average and earliest possible ages of the pennates and the Thalassiosirales are identical because the linearization collapses these lineages into a polytomy (Fig. 3).

Linearization lowers the rate of evolution (Table 4) but this is an artifact of the linearization procedure because dichotomies bearing somewhat longer branches collapse into polytomies. Linearization also improves the fit of regression, hence the smaller range between the average and earliest possible time of origin under each hypothesis. Therefore, linearization increases the average time but not the earliest possible time of origin.

Time estimates based either on the average or median branch length are similar for all diatoms as well as for the pennates (Table 4). For the Thalassiosirales, these figures deviate slightly because this group contains two distinct lineages; the median branch length belongs to the faster lineage. The 95% confidence limits around the average branch lengths for each lineage are comparable to the longest or shortest branch in that lineage. Because essentially no differences exist between estimates based on either of these methods, only those based on the median branch length are discussed in our other analyses.

The maximum-likelihood tree. Branch lengths from the ML tree (Fig. 1B) were regressed against first

appearances of taxa from the fossil record under each hypothesis with only aberrantly evolving taxa eliminated. Estimates of the average and earliest possible age for the origin of the diatoms, the pennates, and the Thalassiosirales (points a, b, and c, respectively) were inferred from the longest, median, and shortest branches (Table 5).

The average age of the diatoms estimated with each hypothesis is younger than that obtained with the linearized tree and postdates their earliest fossil record (Table 5). However, their earliest possible date of origin either coincides with the earliest fossil record (hypothesis A) or predates that record (hypotheses B–D) and generally agrees with that obtained with the linearized tree. A regression plot using hypothesis D is presented in Fig. 4 and shows that with our most lenient estimates for the origin of certain lineages, it is unlikely that the diatoms existed prior to ca. 266 Ma. This roughly corresponds to a rate of evolution for the ssu rRNA coding region of 1% per 21.5 Ma. The three alternative dating hypotheses suggest that the diatoms are evolving much faster (1% per 13, 16, or 18 Ma, respectively).

The first two dating hypotheses place the average age for the pennate origin younger than their earliest fossil record (70 Ma), but hypotheses C and D suggest that the pennates originate during the period of poorly preserved diatom deposits as in the linearized tree. However, the origin of the pennates may be underestimated because the relative rate test has shown that the ssu rRNA coding regions of the pennate diatoms evolve relatively slowly.

Like the linearized tree, all age estimates for Thalassiosirales suggest that this group is older that their earliest fossil record would suggest. We believe that the dating problems associated with the Thalassiosirales are related to the assumption that the modern day strutted process defines the origin of this lineage.

The unweighted neighbor-joining trees. Unweighted NJ trees inferred from Jukes–Cantor distances were topologically identical and revealed branch lengths similar to those inferred from K-2-p distances. This was as expected because of the low JC distances among the sequences and the low transitional bias. Nevertheless, K-2-p distances were used instead of JC distances for the NJ tree construction, because the former is more likely to give the correct tree if rate differences are high (Kumar *et al.*, 1993). Using branch lengths from the unweighted NJ tree (Fig. 2A), the average and earliest possible times of origin for the three lineages are more recent than those obtained from the ML tree (Table 5) and agree with those obtained in the linearized tree (Table 4).

Because the earliest possible date for the origin of the diatoms did not differ dramatically between the linearized NJ tree with all fast taxa eliminated and the nonlinearized NJ trees with only aberrantly evolving taxa eliminated, we used the data set generated under hypothesis C with the NJ tree to test the effect of taxon removal on our age estimates. The inclusion of the aberrantly fast evolving *Eucampia, C. didymus, Poro-*

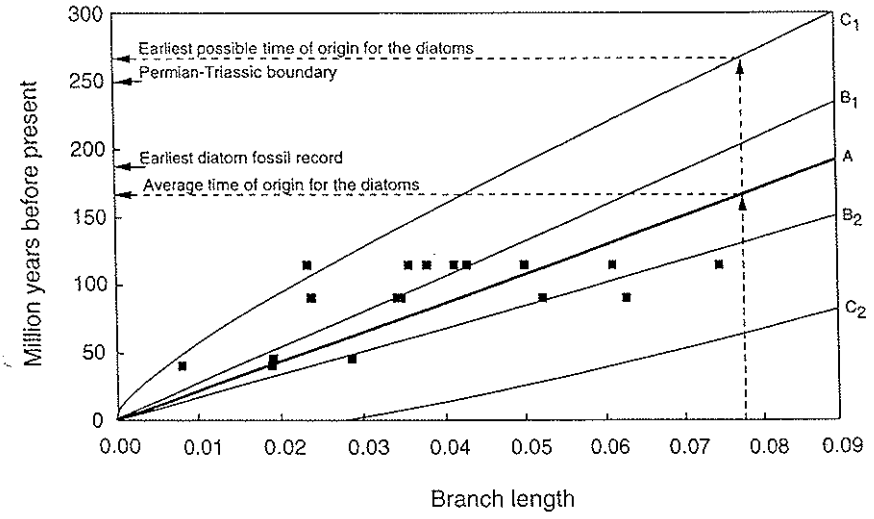

FIG. 4. Small subunit rRNA molecular clock calibration for the diatoms in which dates of first appearances from the fossil record are regressed against measured branch lengths from a maximum-likelihood tree. Hypothesis D, which predates certain taxa to have their origin before a major gap in the fossil record, i.e., at 115 Ma, was used to predate five extant taxa. (A) The regression line, constrained through the origin. Lines B_1 and B_2 are the 95% confidence limits around the regression line. Lines C_1 and C_2 are the 95% confidence limits for a new predicted value of time given the length of an undated node (Hillis and Moritz, 1990). Lower confidence limits, below zero, are reset at zero. The average time of origin of the diatoms (164 Ma) is the length of their median branch length in the maximum-likelihood tree multiplied by the regression coefficient. The earliest possible time of origin of the diatoms (266 Ma) is the date corresponding to the point on the upper 95% confidence limit given the length of the median branch length of the group. Using these calibrations, the diatoms are evolving at 1% per 21.5 Ma (from A) ± 4.5 Ma (B_1 and B_2).

ira, and *Lauderia* (Table 3) did not significantly affect tree topology (Fig. 1A), but it did substantially lower calculated ages (Table 5) because the inclusion of their long branch lengths lowers the rate of evolution. Estimates for the earliest possible times of origin for the three lineages were nearly the same with these taxa included because the new higher substitution rate was counterbalanced by the poor fit of the regression line and a new slightly longer median branch length. Removal of the remaining significantly fast evolving taxa, *Melosira, Actinocyclus, Coscinodiscus,* and *Chaetoceros sp.,* resulted in a slightly lower substitution rate per Ma. However, the average age estimates of the diatoms based on a new shorter median branch length did not change, whereas those for the pennates and the Thalassiosirales were slightly higher. When the sequences of the four slowest evolving taxa (*Cymatosira, Papiliocellulus, Fragilaria,* and *Asterionellopsis*) were removed, the substitution rate did not change. However, the estimated average ages of all diatoms and the pennates were now slightly elevated with a new, longer median branch length.

The weighted neighbor-joining trees. The effect of different weighting strategies on our age estimates under hypothesis C is presented in Figs. 2B–2D and Table 5. If sequence positions are weighted according to van de Peer *et al.* (1993 and pers. comm.), then the branch lengths separating the pennate diatoms decrease (Fig. 2B). The regression analysis gives slightly younger estimates for both the average and the earliest possible time of origin of the three lineages, especially for those of the pennate lineage. If MacClade weighting is used, all distances in the tree (Fig. 2C) are reduced, and the substitution rate is considerably lower than in all other weighting schemes (1% per 29 Ma). This is an artifact because invariant positions carry most of the weight. However, age estimates are of the same magnitude because the reduced substitution rate is counterbalanced by shorter median branch lengths (Table 5). Neither weighting methods improved the fit of the regression line. If nucleotide substitution rates were assumed to be gamma distributed (Jin and Nei, 1990, with the option $\alpha = 1$; Fig. 2D), then the fit of the regression line improved, but time estimates were not affected (Table 5).

DISCUSSION

The Evolutionary Age

Our molecular clock calculations have been used to set the time of cladogenesis of the diatoms within a more precise time frame than has been hypothesized from their fossil record. Even in the most lenient cases, we have calculated that it is unlikely that the diatoms existed prior to ca. 266 Ma. This estimate is based on a molecular clock calibration using ML branch lengths with aberrantly evolving taxa excluded and age estimates of the origin of extant taxa dated as far back as possible (hypothesis D, Tables 4 and 5). If all fast evolving taxa are eliminated, and a linearized NJ tree is constructed, then this date is slightly younger (238 Ma). These estimates for the earliest possible time of origin of the diatoms predate their earliest fossil record by ca. 80 to 50 Ma, which indicates that the diatoms are not of Proterozoic or early Paleozoic origin. Our finding contrasts with the interpretation that there is a 300 (Philippe *et al.,* 1994) to 400 Ma gap (Sörhannus, 1994) in the fossil record of the diatoms based on partial sequences from the nuclear-encoded large subunit (lsu) rRNA coding region.

The reference taxa used in this study, the pelagophytes and the oomycetes, are members of the nearest and next nearest neighbor lineages to the diatoms (Leipe *et al.,* 1994; Saunders *et al.,* 1995). These reference taxa are often so close that the distance from a fast diatom taxon to a slow diatom taxon is larger than the distance from the reference taxon to that slow taxon. Nevertheless, the trees presented in Figs. 1–3 and in Medlin *et al.* (1993, 1996a,b) demonstrate that the diatoms are monophyletic. Choosing an outgroup outside the heterokonts estimates the age of the entire heterokont lineage, including the diatoms. The 300- to 400-Ma gap in the fossil record of the diatoms suggested by Philippe *et al.* (1994) and Sörhannus (1994) has been calculated using the alveolates as nearest neighbor. The alveolates form a separate lineage in the eukaryote radiation and are not closely related to the diatoms (Bhattacharya *et al.,* 1992). The dates reported by Philippe *et al.* (1994) and Sörhannus (1994) likely correspond to the origin of the heterokont lineage and not to the origin of the diatoms, which is a later divergence within the heterokont eukaryotes.

Our regressions for the diatom lineages might also be extrapolated to similar lineages outside the diatom lineage to provide speculative starting points for estimating the ages of neighbor groups without a fossil record. To do this, their relative rates of base substitution must be comparable to those of the diatoms. Our age estimates of the diatoms could also be valid for their sister groups, i.e., the lineage containing the pelagophytes, phaeophytes, and chrysophytes (Saunders *et al.,* 1995), which is unlikely to be older than the late Permian. This extrapolation confirms the suggestion by Druehl and Saunders (1992) and Saunders and Druehl (1992) that the phaeophytes are very young in comparison to other macroalgal groups.

Our study has demonstrated that the various tree building algorithms used gave almost identical branching orders (Figs. 1 and 2). This was to be expected because simulation studies by Huelsenbeck and Hillis (1993) demonstrated that if the overall JC distances between sequences are below 0.2 and transition bias is low, then chances are high that all the algorithms we

used will find the correct tree. However, our study indicated that the tree building algorithms did influence the age estimates. Estimates of the average and earliest possible evolutionary age obtained from the ML tree were higher than those obtained from the NJ trees and likely reflect differences in the assumptions made in the evolutionary models behind these two tree-building algorithms. However, earliest possible dates of origin obtained from the linearized NJ tree were comparable to those obtained from an unlinearized NJ tree with either all fast evolving taxa excluded or just the aberrant ones. It is the inclusion of the aberrantly fast evolving taxa and not the exclusion of all fast evolving taxa that has the most dramatic effect on our age estimates, with the average age of the groups being the most seriously underestimated. The best fit of the regression among the unlinearized NJ trees was obtained with a gamma weighting of the sequence positions.

We did not use branch lengths from parsimony trees for the molecular clock calculations. Parsimony methods tend to underestimate the number of base substitutions along long branches more than NJ and ML methods in particular when rate varies by site in the sequences and among sequences (Sourdis and Nei, 1988; Jin and Nei, 1990; Zharkikh and Li, 1993; Tateno *et al.,* 1994).

The regression plot in Fig. 4 shows the scatter in our data points resulting in broad 95% probability limits around new estimates of time. This scatter can be caused by (1) rate differences among the diatoms' ssu rRNA coding regions and (2) inaccurate first appearance dates in the fossil record.

Rate differences. The relative rate tests of Wu and Li (1985), Tajima (1993), and Li and Bousquet (1992) showed a rate variation among the pennate taxa that agreed with the range of rates expected using a model of stochastically accumulating sequence substitutions. The rate differences among many centric diatoms, however, were not in agreement with such a model. These differences were not due to alignment errors because we removed unalignable regions and verified our alignment with secondary structure models (Medlin *et al.,* 1996b). Sequence-reading mistakes are an unlikely source of error because both coding and noncoding strands were completely sequenced from separate PCR (Medlin *et al.,* 1996a,b). Most sequence differences in stem regions were compensatory. Thus, the remaining long branches in our data set were the result of a higher substitution rate.

As to be expected for more or less normally distributed branch lengths in a lineage, it is insignificant whether the time estimates are inferred from the average or the median length (see Table 4, the diatoms and the pennates). If the lineage contains a few very long branches, then the estimates based on the median branch length are preferred because the deviating branch lengths do not contribute to the time estimates. To show how time estimates can be distorted if estimates are based solely on groups of rapidly or slowly evolving taxa, we also calculated the average and earliest possible time of origin for the longest and shortest branches in each lineage. The pronounced differences in these time estimates stress the fact that times of origin inferred from lineages with only one or two taxa may be grossly under- or overestimated depending on the rate of that particular taxon's ssu rRNA coding region. However, the average branch length might be preferred over the median one, if the branches in the lineage can be separated into two distinct length classes. In the Thalassiosirales (Table 4), the median length falls in the fastest class. In such a case, more taxa could be sampled prior to molecular clock analyses.

An important issue in molecular clock calculations is the removal of branches that evolve significantly differently from the average value. We argue that aberrantly evolving sequences should be removed (Table 3). However, it remains questionable whether sequences that evolve much faster or slower than their relatives but perfectly in accordance with expectations of ssu rRNA coding region evolutionary models should be removed prior to molecular clock calculations (Takezaki *et al.,* 1995). If rate variation among sequences is real, then we have to accept rate variation with time and, hence, our broad 95% confidence limits associated with our time estimates, and our imprecise age estimates. Our data suggest that the aberrantly evolving taxa have more effect on age estimates than do those taxa evolving fast but within a model of ssu evolution. When our fast evolving sequences that evolved within the limits set by the model of van de Peer *et al.* (1993) were included in the analyses, they did not distort the tree topology and had only a minor effect on the time estimates of the origin. Furthermore, the fit of the regression line was not improved by the removal of fast or slow taxa. Therefore, the incorporation of a few fast taxa in the regression does not explain the poor fit of the regression line.

Length variation in a tree is eliminated by linearization (Takezaki *et al.,* 1995). In our linearized tree (Fig. 3), phylogenetic information is lost due to the collapse of branches into polytomies. However, the results of the molecular clock calculations (Table 4) inferred from this tree show that linearization improves the fit of the regression line. Thus, rate variation *among* the sequences is at least partly responsible for the variation in time estimates we recovered.

Comparisons among reported rates are highly dependent on the fractions of the sequence positions in the ssu rRNA coding region that are included in the analyses. The variability of a position in the ssu rRNA coding regions depends on its location in the secondary

structure of the rRNA molecule (van de Peer *et al.*, 1993). Therefore, the observed substitution rate depends on the number of highly variable positions that become unalignable and are removed prior to analysis. As more positions are removed, the calculated rate decreases. For this reason, rates in Tables 4 and 5 should be compared with other rates where the number of positions are reported. With our most lenient hypothesis, i.e., that in which the origin of certain taxa have been dated as far back as possible, the substitution rate in the ssu rRNA coding regions of the diatoms is ca. 1% in 21.5 to 26.2 Ma with 95% confidence limits of 4.5 to 4.0 Ma. This rate is higher than the rate observed in higher plants (ca. 1% per 70 Ma, all positions included in Savard *et al.*, 1994, or a 1% divergence in 25 Ma (=1% substitutions in 50 Ma, no positions reported, see Ochman and Wilson, 1987), in Dasycladales (1% divergence in 25 Ma, 1733 positions in Olsen *et al.*, 1994), in animals (1% divergence in 60–100 Ma, no positions reported, see Ochman and Wilson, 1987), in fungi (1% substitution in 100 Ma, based on 1589 alignable positions in Berbee and Taylor, 1993) and in bacteria (1% divergence in 50 Ma, no positions reported, see Ochman and Wilson, 1987; Moran *et al.*, 1993). Thus, the substitution rate reported here is the fastest recorded to date for an entire group and is in agreement with conclusions based on a small set of diatom taxa in Sorhannus (1994) using partial sequences of the lsu rRNA gene.

A related effect of the rate variation within the sequence is that the calculated distances may not be completely linear with elapsed time because any distance measure underestimates the actual number of substitutions when overall distances are high (Bruns and Tzaro, 1992). This effect is even more pronounced when rate variation is apparent within the molecule. In our comparisons, the three weighting schemes, which correct for site-related rate differences, had only minor effects on the topology of the NJ tree and on the time estimates (Table 5). This could be expected because the unweighted JC distances were always lower than 0.3, and the transition/transversion ratio was lower than 2 (Kumar *et al.*, 1993; Nei, 1991). However, the conspicuously short pennate branches in the van de Peer weighted NJ tree (Fig. 2B) indicate that the ssu rRNA coding regions of these taxa evolve slower than expected in the more conserved regions. Molecular clock calculations based on van de Peer and MacClade weighted distance trees gave poorer fits of the regression line in comparison to those based on gamma-weighted or nonweighted ones. Thus, site-associated rate variation *within* the ssu rRNA coding regions cannot be the reason for the scatter in our data points.

Accuracy of the first appearance dates in the fossil record. Even with the availability of first appearance dates for many diatom taxa, estimates of their origin remain inaccurate because they are strongly influenced by the interpretation of the fossil record. Tables 4 and 5 show that different treatments of the origin of extant taxa strongly influence the age estimates. Thus, the uncertainties related to the interpretation of the fossil record are the other main reason for the scatter of the points. If the first appearance of certain taxa after a gap in the fossil record is assumed to be their origin (hypothesis A), then even our earliest possible age estimates for the diatoms cannot explain their first appearance in the fossil record. Of course, the earliest fossil record (at 185 Ma) could be misdated. However, a more likely explanation is that first appearances immediately after a gap underestimate a taxon's true origin. Furthermore, the origin of a lineage should not be restricted to taxonomic delimitations of modern taxa. In our study, for example, the labiate processes, delimiting extant genera for which we have sequence data, are only known in taxa since the Cretaceous. Medlin *et al.* (1996a) showed that labiate processes, as currently defined, must be a homoplasy. If we use overall valve morphology, linking mechanisms between cells, and presence/absence of a central structure in the valve as the features considered taxonomically important, then we are able to push back the origin of four taxa from 70 Ma to 115 Ma ago. Although this approach reduces the rate and pushes the earliest possible origin of the diatoms back before its first appearance in the fossil record, it does not reduce the fit of the regression line. Only hypothesis D improves the fit of the regression line. Our clock calculations from the first three hypotheses indicate that the first appearance of the Thalassiosirales at 40 Ma might be incorrect. By redefining the morphological characters defining the Thalassiosirales, we can predate the origin of this lineage at or before 115 Ma, which greatly improves the precision of the regression and the time estimates. Thus, albeit well documented and extensive, the fossil record of the diatoms remains a major source of imprecision in the molecular clock analysis of the diatoms primarily because of the selection of morphological characters defining lineages.

Alternatively, all taxa in the 115–110 Ma fossil deposits may be representatives of a flora that went virtually extinct by the end of the Cretaceous and was replaced by the present diatom flora following a new adaptive radiation from a single surviving species into the extensive morphological spectrum we know today. Similarities between the extinct taxa at 115 Ma and the taxa that appeared from 70 Ma onward may be analogies, not homologies. This hypothesis is, however, unlikely because the phylogenetic analyses show that the diatoms diverged into the major extant lineages early in their evolutionary history, not long after their separation from the other pigmented heterokont algae.

The paucity of pre-Cretaceous diatom deposits hinders the determination of clear links in the fossil taxa with extant taxa during the two gaps in the history of

the diatoms. Pre-Cretaceous diatom fossils may be scarce because sea floor spreading has destroyed or altered most earlier ocean floor deposits. Nevertheless, other siliceous microfossils, such as sponge spicules and radiolaria, are abundant in pre-Cretaceous marine sediments, suggesting that pre-Cretaceous diatoms either were rare or were confined to habitats so adverse to preservation that they left no record (Round *et al.,* 1990). It is also likely that diatoms prior to 185 Ma ago were not as well mineralized (if at all) as they are today. Our molecular analyses provide insights into this major period of diatom radiation that cannot be reconstructed on the basis of the existing fossil record.

Biological Reasons for Differences in Rates of Evolution

The apparent rate differences among the diatoms' ssu rRNA coding regions are not unusually large because similar rate variation can be observed in the ssu rRNA coding regions from other organisms, as well as in other genes (Li and Tanimura, 1985; Wu and Li, 1985; Britten, 1986; Koop *et al.,* 1986; Powell *et al.,* 1986; Vawter and Brown, 1986; Wolfe *et al.,* 1989; Berbee and Taylor, 1993; Bakker *et al.,* 1994; McNally *et al.,* 1994; Lafay *et al.,* 1995; Russo *et al.,* 1995). Many physiological and life history factors have been proposed to explain the significant rate differences among lineages. Some examples of these factors include generation time, temperature, cell volume, body size, metabolic rates, free radical formation, etc. (Martin and Palumbi, 1993).

The substitution rate is assumed to be positively correlated with the frequency of mitotic and meiotic divisions (Martin and Palumbi, 1993). In microalgae, the potential frequency of mitotic divisions is negatively correlated with cell volume (Tang, 1995) and positively correlated with temperature (Eppley, 1972; Raven and Geider, 1988; Tang, 1995). However, ecology is also important. Centrics, which are mainly pelagic, could be considered typical *r*-strategists and spend most of their time either in unrestricted exponential growth or in zero growth. Pennate diatoms, on the other hand, are mainly benthic and are probably nearer to carrying capacity (*K*-strategists): they may grow slower (D. Mann, pers. comm.). It is significant that the ssu rRNA coding regions of the benthic centrics *Papiliocellulus* and *Cymatosira* also evolve relatively slowly, but it remains to be tested whether life history strategy is a key factor in the substitution rate.

Diatoms reproduce asexually and thereby decrease in size until they reach a particular size class, termed the reproductive window, when sexual reproduction can be initiated (Round *et al.,* 1990). Among the oogamous centrics, "male" cells divide mitotically many times within the parental cell wall, prior to meiosis, whereas "female" cells undergo only meiosis. In pennate diatoms, sexual reproduction is isogamous with no known mitotic divisions within the parental cell wall prior to gametogenesis. Only one or two gametes per cell are formed. Thus, centric diatoms undergo more mitotic divisions, at least during spermatogenesis, especially in those taxa in which the number of nuclear divisions leading to the sperm cells is large. If the substitution rate is related to the frequency of mitotic divisions, then centrics would be expected to have a somewhat faster substitution rate.

Also, the frequency of meiosis in pennates may be lower than that in centrics (Jewson, 1992a, b; Mann, 1988, pers. comm.). This agrees with our results, which show that the ssu rRNA coding regions of the pennates evolve slower than those of most centrics. This effect is particularly apparent in the van de Peer weighted tree, although this may be an artifact of this evolutionary model. These relationships between division rate, temperature, cell volume, and reproductive or life history strategies provide speculative starting points for future research.

CONCLUSIONS

Our results have shown that the ssu rRNA coding regions in most diatoms are not evolving according to a model of stochastically accumulating sequence substitutions. In particular, the centric diatoms' ssu rRNA coding regions are evolving much faster than those of the pennate taxa. The substitution rate of evolution in the small subunit rRNA coding regions of the diatoms is approximately 1% per 18 to 26 Ma, which is the fastest reported rate for any pro- or eukaryotic organism to date. Using our most lenient estimates for the first appearance dates for certain taxa, we have calculated that it is unlikely that the diatoms existed before 266 Ma ago. This age calculation is based strictly on median-evolving taxa. Aberrantly evolving taxa have the most pronounced effect on age estimates. The diatoms are a relatively young lineage, probably not appearing before the Late Permian.

ACKNOWLEDGMENTS

We thank Drs. D. Bhattacharya, J. Bousquet, D. W. Gladrow, W.-H. Li, and D. Mann for useful discussions and M. Pahlow and H. Klerk for help with the molecular clock calibrations. Dr. Y. van de Peer kindly provided access to unpublished data for the rRNA variability classes. We are indebted to Dr. R. Gersonde and Ms. P. A. Sims for providing the fossil dates for our analyses. Our research was supported by the DFG (Sm22/7-1 and Me1480/1-2). This is contribution No. 1072 from the Alfred-Wegener-Institute.

REFERENCES

Bakker, F. T., Olsen, J. L., Stam, W. T., and van den Hoek, C. (1994). The *Cladophora* complex (Chlorophyta): new views based on 18S rRNA gene sequences. *Mol. Phylogenet. Evol.* **3:** 365–382.

Berbee, M. L., and Taylor, J. W. (1993). Dating the evolutionary radiations of the true fungi. *Can. J. Bot.* **71:** 1114–1127.

Bhattacharya, D., Medlin, L., Wainwright, P. O., Ariztia, E. V.,

Bibeau, C., Stickel, S. K., and Sogin, M. L. (1992). Algae containing chlorophylls *a* + *c* are paraphyletic: Molecular evolutionary analysis of the Chromophyta. *Evolution* **46:** 1808–1817.

ritten, R. J. (1986). Rates of DNA sequence evolution differ between taxonomic groups. *Science* **231:** 1393–1398.

runs, T. D., and Szaro, T. M. (1992). Rate and mode differences between nuclear and mitochondrial small-subunit rRNA genes in mushrooms. *Mol. Biol. Evol.* **9:** 836–855.

avalier-Smith, T. C. (1986). The kingdom Chromista: Origin and systematics. *In* "Progress in Phycological Research" (F. E. Round and D. J. Chapman, Eds.), vol. 4, pp. 309–347, Bristol Biopress, Bristol.

)ruehl, L. D., and Saunders, G. W. (1992). Molecular explorations in kelp evolution. *In* "Progress in Phycological Research" (F. E. Round and D. J. Chapman, Eds.), vol. 8, pp. 47–83, Bristol Biopress, Bristol.

)ppley, R. W. (1972). Temperature and phytoplankton growth in the sea. *Fish. Bull.* **70:** 1063–1085.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum-likelihood approach. *J. Mol. Evol.* **17:** 368–376.

Felsenstein, J. (1993). "PHYLIP Manual," Version 3.5, Department of Genetics, University of Washington, Seattle.

Forti, A., and Schulz, P. (1932). Erste Mitteilungen über Diatomeen aus dem Hannoverschen Gault. *Beih. Bot. Zentralbl.* **50:** 241–246.

Gersonde, R., and Harwood, D. M. (1990). Lower Cretaceous diatoms from ODP Leg 113 site 693 (Weddell Sea). I. Vegetative cells. *In* "Proceedings of the Ocean Drilling Program, Scientific Results" (P. F. Barker, J. P. Kennett, *et al.*, Eds.), pp. 365–402, Ocean Drilling Program, College Station, Tx 113.

Harwood, D. M. (1988). Upper Cretaceous and lower Paleocene diatom and silicoflagellate biostratigraphy of Seymour Island, eastern Antarctic Peninsula. *In* "Geology and Paleontology of Seymour Island" (R. Feldman and M. O. Woodburne, Eds.), vol. 169, pp. 55–129, Geol. Soc. Am. Mem. Ser.

Harwood, D. M., and Gersonde, R. (1990). Lower Cretaceous diatoms from ODP Leg 113 site 693 (Weddell Sea). II. Resting spores, Chrysophycean cysts, and endoskeletal dinoflagellates and notes on the origin of the diatoms. *In* "Proceedings of the Ocean Drilling Program, Scientific Results" (P. F. Barker and J. P. Kennett *et al.*, Eds.), pp. 365–402, Ocean Drilling Program, College Station, Tx 113.

Hasle, G. R., and Syvertsen, E. E. (1985). *Thalassiosiropsis,* a new diatom genus from the fossil record. *Micropaleontology* **31:** 82–91.

Hillis, D. M., and Moritz, C. (1990). An overview of applications of molecular systematics. *In* "Molecular Systematics" (D. M. Hillis and C. Moritz, Eds.), pp. 502–515, Sinauer, Sunderland, MA.

Huelsenbeck, J. P., and Hillis, D. M. (1993). Success of phylogenetic methods in the four-taxon case. *Syst. Biol.* **24:** 247–264.

Jewson, D. H. (1992a). Size reduction, reproductive strategy and the life cycle of a centric diatom. *Phil. Trans. R. Soc. Lond.* B. **335:** 191–213.

Jewson, D. H. (1992b). Life cycle of a *Stephanodiscus* sp. (Bacillariophyta). *J. Phycol.* **28:** 856–866.

Jin, L., and Nei, M. (1990). Limitations of the evolutionary parsimony method of phylogenetic analysis. *Mol. Biol. Evol.* **7:** 82–120.

Jukes, T. H., and Cantor, C. R. (1969). Evolution of protein molecules. *In* "Mammalian Protein Metabolism" (H. N. Munro, Ed.), pp. 21–132, Academic Press, New York.

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

Koop, B., Goodman, M., Xu, P., Chan, K., and Slightom, J. (1986). Primate h-globin DNA sequences and man's place among the great apes. *Nature* **319:** 234–238.

Kumar, S., Tamura, K., and Nei, M. (1993). "MEGA: Molecular Evolutionary Genetics Analysis," version 1.0, Pennsylvania State University, University Park, PA.

Lafay, B., Smith, A. B., and Christen, R. (1995). A combined morphological and molecular approach to the phylogeny of asteroids (Asteroidea: Echinodermata). *Syst. Biol.* **44:** 190–208.

Leipe, D. D., Wainwright, P. O., Gunderson, J. H., Porter, D., Patterson, D. J., Valois, F., Himmerich, S., and Sogin, M. L. (1994). The stramenopiles from a molecular perspective: 16S-like rRNA sequences from *Labyrinthuloides minuta* and *Cafeteria roenbergensis. Phycologia* **33:** 369–377.

Li, P., and Bousquet, J. (1992). Relative rate test for nucleotide substitutions between two lineages. *Mol. Biol. Evol.* **9:** 1185–1189.

Li, W.-H., and Tanimura, M. (1985). The molecular clock runs more slowly in man than in apes and monkeys. *Nature* **326:** 93–96.

Maddison, W. P., and Maddison, D. R. (1992). "MacClade. Analysis of Phylogeny and Character Evolution," Version 3, Sinauer, Sunderland, MA.

Mann, D. G. (1988). Why didn't Lund see sex in *Asterionella?* A discussion of the diatom life cycle in nature. *In* "Algae and the Aquatic Environment" (F. E. Round, Ed.), pp. 386–412, Biopress, Bristol.

Mann, D. G., and Marchant, H. J. (1989). The origin of the diatom and its life cycle. *In* "The Chromophyte Algae: Problems and Perspectives" (J. C. Green, B. S. C. Leadbeater, and W. L. Diver, Eds.), pp. 307–323, Clarendon Press, Oxford.

Martin, A. P., and Palumbi, S. R. (1993). Body size, metabolic rate, generation time, and the molecular clock. *Proc. Natl. Acad. Sci. USA* **90:** 4087–4091.

McNally, K. L., Govind, N. S., Thomé, P. E., and Trench, R. K. (1994). Small-subunit ribosomal DNA sequence analyses and a reconstruction of the inferred phylogeny among symbiotic dinoflagellates (Pyrrophyta). *J. Phycol.* **30:** 316–329.

Medlin, L. K., Williams, D. M., and Sims, P. A. (1993). The evolution of the diatoms (Bacillariophyta). I. Origin of the group and assessment of the monophyly of its major divisions. *Eur. J. Phycol.* **28:** 261–275.

Medlin, L. K., Gersonde, R., Kooistra, W. H. C. F., and Wellbrock, U. (1996a). Evolution of the diatoms (Bacillariophyta). II. Nuclear-encoded small-subunit rRNA sequence comparisons confirm a paraphyletic origin for the centric diatoms. *Mol. Biol. Evol.* **13:** 67–85.

Medlin, L. K., Kooistra, W. H. C. F., Gersonde, R., and Wellbrock, U. (1996b). Evolution of the diatoms (Bacillariophyta). III. Molecular evidence for the origin of the Thalassiosirales. *Nova Hedwigia* **112:** 221–234.

Moran, N. A., Munson, M. A., Baumann, P., and Ishikawa, M. (1993). A molecular clock in endosymbiotic bacteria is calibrated using the insect hosts. *Proc. R. Soc. Lond.* B **253:** 167–171.

Moshkovitz, S., Ehrlich, A., and Soudry, D. (1983). Siliceous microfossils of the Upper Cretaceous Mishash Formation, Central Negev. *Israel Cret. Res.* **4:** 73–194.

Neefs, J.-M., van de Peer, Y., De Rijk, P., Goris, A., and De Wachter, R. (1991). Compilation of small ribosomal subunit RNA sequences. *Nucleic Acids Res.* **19** (Suppl.): 1987–2015.

Nei, M. (1991). Relative efficiencies of different tree-making methods for molecular data. *In* "Phylogenetic Analyses of DNA Sequences" (M. M. Miyamoto and J. Cracraft, Eds.), pp. 90–128, Oxford Univ. Press, Oxford.

Ochman, H., and Wilson, A. C. (1987). Evolution in bacteria: Evidence for a universal substitution rate in cellular genomes. *J. Mol. Evol.* **26:** 74–86.

Olsen, G. J., Matsuda, H., Hagstrom, R., and Overbeek, R. (1994). FastDNAml: A tool for construction of phylogenetic trees of DNA

sequences using maximum-likelihood. *Comput. Appl. Biosci.* **10:** 41–48.

Olsen, J. L., Stam, W. T., Berger, S., and Menzel, D. (1994). 18S rDNA and the evolution in the Dasycladales (Chlorophyta): Modern living fossils. *J. Phycol.* **30:** 729–744.

Philippe, H., Sörhannus, U., Baroin, A., Perasso, R., Gasse, F., and Adoutte, A. (1994). Comparison of molecular and paleontological data in diatom suggests a major gap in the fossil record. *J. Evol. Biol.* **7:** 247–265.

Powell, J. R., Caccone, A., Amato, G. D., and Yoon, C. (1986). Rates of nucleotide substitution in *Drosophila* mitochondrial DNA and nuclear DNA are similar. *Proc. Natl. Acad. Sci. USA* **83:** 9090–9093.

Raven, J. A., and Geider, R. J. (1988). Temperature and algal growth. *New Phytol.* **110:** 441–461.

Rohlf, F. J., and Sokal, R. R. (1981). "Statistical Tables," Freeman, New York.

Rothpletz, A. (1896). Über die Flysch-Fucoiden und einige andere fossile Algen, sowie über liasische, Diatomeen führende Hornschwämme. *Zeitschr. Deutsche. Geol. Ges.* **48:** 954–914.

Round, F. E. (1981). Some aspects of the origins of diatoms and their subsequent evolution. *BioSystems* **14:** 483–486.

Round, F. E., and Crawford, R. M. (1981). The lines of evolution of the Bacillariophyta. I. Origin. *Proc. R. Soc. London B.* **211:** 237–260.

Round, F. E., and Crawford, R. M. (1984). The lines of evolution of the Bacillariophyta. II. The centric series. *Proc. R. Soc. London B.* **221:** 169–188.

Round, F. E., Crawford, R. M., and Mann, D. G. (1990). "The Diatoms: Biology and Morphology of the Genera," Cambridge Univ. Press, Cambridge.

Russo, C. A. M., Takezaki, N., and Nei, M. (1995). Molecular phylogeny and divergence times of drosophilid species. *Mol. Biol. Evol.* **12:** 391–404.

Saunders, G. W., and Druehl, L. D. (1992). Nucleotide sequences of the small-subunit ribosomal RNA genes from selected Laminariales (Phaeophyta): Implications for kelp evolution. *J. Phycol.* **28:** 544–549.

Saunders, G. W., Potter, D., Paskind, M. P., and Andersen, R. A. (1995). Cladistic analyses of combined traditional and molecular data sets reveal an algal lineage. *Proc. Natl. Acad. Sci. USA* **92:** 244–248.

Savard, L., Li, P., Strauss, S. H., Chase, M. W., Michaud, M., and Bousquet, J. (1994). Chloroplast and nuclear gene sequences indicate late Pennsylvanian time for the last common ancestor of extant seed plants. *Proc. Natl. Acad. Sci. USA* **91:** 5163–5167.

Sörhannus, U. (1994). Relative rate tests versus paleontological divergence data for diatoms and vertebrates. *Acta Paleontol. Polonica* **38:** 199–214.

Strel'nikova, N. I. (1990). Evolution of the diatoms during the Cretaceous and Paleogene periods. *In* "Proc. 10th Int. Diat. Symp." (H. Simola, Ed.), pp. 195–204, Költz, Königstein.

Strel'nikova, N. I., and Martirosjan, G. N. (1981). Lower Cretaceous diatom algae from Stavropol. *Viestnik LGU, Ser biologiya* **3:** 52–57 [in Russian].

Sourdis, J., and 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. (1993). Simple methods for testing the molecular evolutionary clock hypothesis. *Genetics* **135:** 599–607.

Takezaki, N., Rzhetsky, A., and Nei, M. (1995). Phylogenetic test of the molecular clock and linearized trees. *Mol. Biol. Evol.* **12:** 823–833.

Tang, E. P. Y. (1995). The allometry of algal growth rates. *J. Plankton Res.* **17:** 1325–1335.

Tateno, Y., Takezaki, N., and Nei, M. (1994). Relative efficiencies of the maximum-likelihood, neighbor-joining and maximum-parsimony methods when substitution rate varies with site. *Mol. Biol. Evol.* **11:** 261–277.

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

Vawter, L., and Brown, W. M. (1986). Nuclear and mitochondrial DNA comparisons reveal extreme rate variation in the molecular clock. *Science* **234:** 194–196.

Wolfe, K. H., Sharp, P. M., and Li, W.-H. (1989). Rates of synonymous substitution in plant nuclear genes. *J. Mol. Evol.* **29:** 208–211.

Wu, C.-I., and Li, W.-H. (1985). Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc. Natl. Acad. Sci. USA* **82:** 1741–1745.

Zharkikh, A., and Li, W.-H. (1993). Inconsistency of the maximum-parsimony method: The case of five taxa with a molecular clock. *Syst. Biol.* **42:** 113–125.