Journal of Biodiversity, Bioprospecting and Development

Medlin, J Biodivers Biopros Dev 2014, 1:3
DOI: 10.4172/2376-0214.1000129

Research Article Open Access

Evolution of the Diatoms: VIII. Re-Examination of the SSU-Rrna Gene Using Multiple Outgroups and a Cladistic Analysis of Valve Features

Linda K. Medlin*

Marine Biological Association of the UK, The Citadel, Plymouth PL1 2PB, UK

Abstract

The resolution of the SSU rRNA gene for phylogenetic analysis in the diatoms has been evaluated by Theriot et al. who claimed that the SSU rRNA gene could not be used to resolve the monophyly of the three diatoms classes described by Medlin and Kaczmarska. Although they used both only bolidomonads and heterokonts as outgroups, they did not explore outgroups further away than the heterokonts. In this study, the use of the multiple outgroups inside and outside the heterokonts with the rRNA gene for recovering the three monophyletic clades at the class level is evaluated. Trees with multiple outgroups ranging from only bolidophytes to Bacteria and Archea were analyzed with Bayesian and Maximum Likelihood analyses and two data sets were recovered with the classes being monophyletic. Other data sets were analyzed with non-weighted and weighted maximum parsimony. The latter reduced the number of clades and lengthened branch lengths between the clades. One data set using a weighted analysis recovered the three classes as monophyletic. Taking only bolidophytes as the only outgroup never produced monophyletic clades. Multiple outgroups including many heterokonts and certain members of the crown group radiation recovered monophyletic clades. The three classes can be defined by clear morphological differences primarily based on auxospore ontogeny and envelope structure, the presence or absence of a structure (tube process or sternum) associated with the annulus and the location of the cribrum in those genera with loculate areolae. A cladistic analysis of some of these features is presented and recovers the three classes.

Keywords: Diatoms, Phylogeny; SSU rRNA gene; Outgroups

Introduction

The diatoms (Bacillariophyta) are arguably one of the most successful microalgal groups in both aquatic and terrestrial habitats with over 10,000 described species and potentially many more cryptic species [1]. Their architecturally complex siliceous cell walls (valves and girdle bands) are unique among the algae. Since the 19th century, their classification system has been based on the intricate designs of their cell walls (see review of the diatom classification history in [2]). Their unusual pattern of cell size reduction of one of the daughter cells following mitosis result in a population of cells of smaller sizes, which, normally, can only be restored to the cell's maximum cell size following sexual reproduction [3,4]; this is also unique among the algae.

In 1993, Medlin et al. [5] produced the first phylogeny of the diatoms that showed two of the morphological groups of the diatoms (centric and araphid pennates) were not monophyletic and this conflicted with the accepted perception that the group was divided into two natural groups: the centrics and the pennates, the latter divided into araphid and raphid diatoms. In Medlin et al.'s analyses, the centric diatoms fell into two clades with high bootstrap support and the araphid pennates were also in two groups but within the pennate diatoms. At that time, Skeletonema costatum representing the genus Skeletonema Greville was the only centric species that did not clade with the other centrics. As more centrics were added to the tree [6,7], it became clear that there were two groups of centric diatoms: the radial centrics and the bipolar centrics with the radial Thalassiosirales falling into the bipolar clade. This result has been very difficult for diatomists to accept using the morphology of the valves as the first criterion to separate taxa at higher taxonomic levels. The same applied to the araphid pennate diatoms, which did not clade together but fell into two groups, later termed the basal and the core araphids [8]. Instead, it was necessary to use sexual reproductive features and internal morphological features to find support for all clades [9,10] with limited support from valve morphology. However, more support from valve morphology has recently been documented [11,12]. The presence of a process inside the annulus (point of initiation of silicification in forming valves), the properizonial auxospore structure (except for the loss of properizonial bands in the Thalassiosirales) and, where present, loculate areolae with internal cribra (a few exceptions in the Eupodiscaceae) support the Mediophyceae as a monophyletic group. The Coscinodiscophyceae are united by their auxospore structure with only scales, the absence of any process within their annulus, and loculate areolae with an external cribrum, although some genera with pseudo-loculate areolae have an internal cribrum. The Golgi arrangement also provides some support for the two clades, although there are several exceptions in each group [10,13].

Medlin [12] has provided arguments to suggest that coalescent models, which are normally used for shallow branches in any tree at the species level, can also be used to provide support for ancient polymorphisms and incomplete lineage sorting at greater depths in the tree to explain why there are so many paraphyletic taxa in the diatoms and why the Thalassiosirales are embedded in the bipolar centric diatom clades. The sexual reproduction typical of the bipolar centrics involves auxospores that contain bands that mould the auxospores into a bi-multi polar shape in addition to the scales that surround the auxospores of the radial centrics. Essentially, the Thalassiosirales have lost the ability to make bands to squeeze the auxospores into a bipolar shape and thus have reverted to the ancestral state of a radial

***Corresponding author:** Linda K. Medlin, Marine Biological Association of the UK, The Citadel, Plymouth PL1 2PB, UK, Tel: +6012 369 9854; E-mail: lkm@mba.ac.uk

Received July 11, 2014; **Accepted** September 23, 2014; **Published** September 26, 2014

Citation: Medlin LK (2014) Evolution of the Diatoms: VIII. Re-Examination of the SSU-Rrna Gene Using Multiple Outgroups and a Cladistic Analysis of Valve Features. J Biodivers Biopros Dev 1: 129. doi:10.4172/2376-0214.1000129

centric with only scales on its auxospores. The production of scales on the auxospores is the ancestral state because nearly all diatoms produce them regardless of whether they have added additional bands to modify the original spherical shape of the initial cell or not. The Thalassiosirales are found in a derived position in the bipolar centrics in most trees [10,15-19].

Medlin and Kaczmarska [10] revised the systematics of the diatoms and their revision was the result of nearly 20 years of consistent recovery of mismatch between molecular (SSU rRNA gene) and the accepted diatom systematics of the time [20,21 p. 48-53, 22 p. 126-129], which either segregated diatoms into two major orders, centrics and pennates, or into one class of centrics and two classes of pennates, respectively. Medlin and Kaczmarska established three classes in two subphyla: Coscinodiscophytina with the single class Coscinodiscophyceae (radial centrics) and the Bacillariophytina with two classes, the Mediophyceae (bipolar centrics plus the radial Thalassiosirales) and the Bacillariophyceae (the pennate diatoms). Medlin [12] has further revised the classification of the pennate diatoms to recognize formally the basal and core araphids as subclasses. The basal araphids contains the araphids with both perizonial and properizonial bands on the auxospores, whereas the core araphids have only perizonial bands like the raphid pennate diatoms. The araphid diatoms differ from the raphid diatoms in the following points: 1) the raphid diatoms have a slit in the valve for movement and the araphids do not and 2) where sexual reproduction has been studied, the araphid diatoms possess a filament attached to the male sex cell that attaches to the female cell to draw them together for fusion and one or both gametes are released from the gametangia [8], whereas both gametes of the raphid diatoms are not released from the gametangia and conjugation between mating types occurs by juxtaposition of the two gametangia [22]. The filament has been documented from four araphid genera (see references in [8] and amoeboid gametes are known in two other genera. Grammatophora Ehrenberg and Rhabdonema Kützing do not release their female gametes from gametagania like the raphid diatoms [23,24] and in this respect, it may not be so unusual that Rhabdonema has often been recovered at the base of the raphid diatom clade (see results below).

All molecular analyses since the proposed re-classification have been controversial. Theriot et al. [16] concluded that the new system for higher level systematics of diatoms proposed by Medlin and Kaczmarska [10] was premature because their extensive re-analysis rejected their results, albeit very weakly. Using a cladistic analysis of auxospore envelope features, they [16, figure 9] claimed that ontogenetic and morphological characters associated with the post-sexual cell of diatoms (auxospore) do not support the molecular division of diatoms because the order Thalassiosirales possesses auxospore characters of a class to which molecular analyses do not assign them. Medlin [12] refuted this argument claiming that the presence of only scales in the Thalassiosirales is a loss of the band character and reversion to an ancestral polymorphism and the formal cladistic analysis, presented here, uses other valve features that pull the Thalassiosirales into the Mediophyceae in contrast to the analysis preformed by Theriot et al..

The question then arose as to whether or not the SSU rRNA gene was the appropriate gene for resolving the issue of the two monophyletic centric classes instead of a grade of centric clades. Theriot et al. [16] formally assessed the resolution of the SSU rRNA genes based on the methods used for analysis and rejected it as the sole use for phylogenetic analyses and directed their later research into multi-gene analyses.

Multiple gene data sets are now starting to appear [8,14,17-18,25-29]. Whether or not the addition of more genes/different genes provides additional support for the two centric classes is controversial. The work by Bruder and co-workers has concentrated primarily on the pennate genera and thus has not helped the resolution of the centric problem. The multi-gene araphid analysis by Medlin and co-workers continues to show that there are two groups of araphid diatoms, thus supporting the SSU rRNA tree. Using slightly different analytical methods and fewer outgroups, Theriot et al. [17] did not, at first, recover the same classes as in Medlin and Kazcmarska [10] and instead recovered a grade of centric diatom clades with a monophyletic pennate diatom clade. Later, Ashworth et al. [18] with the same taxon sampling but with more genes recovered a monophyletic bipolar centric clade, the Mediophyceae, as described by Medlin and Kaczmarksa [10] but their analyses continued to recover a grade of radial centric diatoms (Class Coscinodiscophyceae), whereas Medlin and co-workers continue to recover monophyletic clades for the three classes using single genes [30,31]. In a larger data set, Ashworth et al. [19] have also recovered a monophyletic Mediophyceae (support 81,100) and 4 clades of coscinodiscophyte centric diatoms. In some, but not all of the four gene analyses by Sato [8] and with multiple outgroups, the two centric classes are monophyletic.

Medlin and Kaczmarska [10] have proposed that the recovery of the two centric clades as monophyletic groups is highly dependent on an alignment based on the secondary structure of the SSU rRNA gene and the use of multiple outgroups. The effect of the secondary structure alignment on the topology of the rRNA tree has been documented in several studies [5,11,32-33]. In this study, the use of multiple outgroups from bolidomonad to Archea for the resolution of the centric clades is addressed here using a single gene, the SSU rRNA gene and the results obtained compared using a Bayesian analysis (MB), Maximum Likelihood analysis (ML) and a weighted and non-weighted maximum parsimony analysis (MP). Using a single representative of any outgroup would result in the erroneous phylogenies as discussed by Theriot et al. [16] resulting from long-branch attraction. Instead, for each outgroup taxon, e.g. haptophytes, a minimum of 2-3 species in the large scale phylogenetic analyses were taken so as to ensure that any effect from long branch attraction from the outgroup is absent or negligible [34-37].

Materials and Methods

Taxon sampling

The data base includes all diatom sequences that were in Genbank as of 2009 plus unpublished clone library sequences from Thomas Friedl and new strains from Belgium (Drs. K. Sabbe and W. Vyverman), which are shown in Supplementary Table 1.

Data analysis

New rRNA sequences from the diatoms were uploaded and aligned to the SIlVA SSU rRNA sequence alignment in the ARB program using maximum primary and secondary structural similarity in the ARB program (Technical University of Munich, Germany). The ARB data base release used in these analyses contained over 325,362 eukaryotic and prokaryotic sequences. Bases were aligned with one another based on their pairing across a helix. The ARB program generates a maximum parsimony (MP) tree from all sequences and all positions in the database as its reference tree, using a filter based on 50% base frequency across all group. Subsets of these sequences were downloaded for further analyses using a positional variability by 50% parsimony filter for all eukaryotes or prokaryotes within ARB or no filter at all.

Depending on the filter and the outgroups used, the number of

Data set	No. of taxa	No. of bases	No. out-groups	Outgroups	Comments
1	429	2901	10	4 Bolidomonads, pelagophytes, synurophytes, thaustochyrids	Figure1E pennates within Cos clade
2	438	3700	16	As DS 1 plus euglenophytes, cryptophytes, haptophytes	outgroups inside the diatoms except for Euglena
3	435	1558	6	cryptophytes, 4 boldiomonads	Rhabdonema at base of diatoms, bolidomonds inside
4	436	1504	9	Euglena, haptophtyes, cryptomonads, 4 bolidomonads	Cryptomonads and haptophytes inside the diatoms
5	436	1495	7	4 bolidomonads, dinoflagellates, cryptomonads, haptophytes	Pennates diverge before centrics
6	438	3151	4	cryptomonads 1 bolidomonad	Figure 1D multiple clades
7	438	1494	4	cryptomonads, 3 bolidomonads	Pennates diverge before centrics, bolidomonads not monophyletic
8	422	1573	3	ciliate and 2 bolidomonads	bolidomonads inside diatoms
9	422	3371	5	3 bolidomonads, ciliate and haptophyte	Pennates diverge before centrics Bolidomonads, haptophytes inside diatoms
10	422	1502	4	ciliate and 3 bolidomonads	Cos sister to pennates, 2 clades of meds, bolidomonads inside cos
11	427	2994	5	2 bolidomonads + Mallomonas	Figure 1C 3 monophyletic clades
12	496	2117	30	Heterokonta	Figure 1H multiple clades
13	899	5203	27	Heterokonta	Mixed centrics and pennates
14	899	5203	2	Bolidomonas	Mixed centrics and pennates
15	667	1458	2	2 bolidomonads	Figure 1B, multiple clades
16	899	5203	4	6 bolidomonads	Figures 1A multiple clades
17	632	n/a	6	Thaustochryids, formaminfera, 1 bolidomonad, Mallomonas	Med monophyletic, cos last divergence in araphid pennates
18	859	1463	12	Cafeteria, 2 pelagophytes, 7 bolidomonads,2 dinoflagellates Excluding all coscinodiscophytes but including all partial sequences	pennates inside med, closest relative Cymatosirales and Biddulphia
20	866	n/a	3	Cafeteria + bolidomonads	Centrics and pennates mixed
21	591	828	2	haptophytes + bolidomonads	Centrics and pennates mixed
22	593	983	2	haptophytes, Cafeteria + bolidomonads	Centrics and pennates mixed
23	593	4540	2	haptophytes and synurophytes, Cafeteria	Figure 1G monophyletic mediophytes
24	596	1456	17	Ciliates, haptophytes, chlorophytes, heterokonts, 2 bolidomonads	Figure 1I, monophyletic mediophytes
25	596	1834	17	Ciliates, haptophytes, chlorophytes, heterokonts, 2 bolidomonads,	Figure 1J three monophyletic classes
26	614	4524	2	Two bolidomonads	Figures 3A,B
27	620	4550	10	Pigmented heterokonts	Figures 3C,D
28	648	4522	11	Two bolidomonads	Figures 3E,F
29	667	6098	11	Pigmented heterokonts + Cafeteria	Figures 3G,H
30	702	1842	25+	Bacterial root, all phyla	Figures 3I,J, Figure 1K
31	629	933	2	Bolidomonads	Mediophytes before coscinodisophytes
33	690	4543	40	Pigmented heterokonts + Cafeteria +crown group	Figure 1L
34	677	1493	11	Heterokonta + Cafeteria + 2 bolidomonads	Figure 1F, multiple clades

Cos = Cocinodiscophyceae, Med = Mediophyceae. n/a = not available, nexus file not saved.

Table 1: Summary of the major features and results from the dataset generated in this study. Data sets discussed in the text are shown by their respective figure numbers.

bases included in the analysis varied. The number of taxa also varied because I tested various combinations of outgroups. The various outgroup taxa are shown in Supplementary Table 1 and different combinations of these taxa were selected for an in-depth analysis of the effect of different multiple outgroups on the branching pattern within the diatoms. 34 data sets were generated consisting of varying numbers of taxa using two or three representatives from each major pro- and eukaryotic groups as outgroups that could be used in a ML analysis using the 50% base frequency filter or no filter at all (Table 1).

In the ARB program release used here (Version 5.5), a maximum likelihood analysis (RAxML) can be performed within ARB and prior to that release, the bioinformatics department at the Alfred Wegener Institute, Bremerhaven, Germany adapted the primary source code of ARB so that Bayesian Inference (BI) analyses could be performed within ARB at first with only single runs and later with parallel runs with single processors and later with multiple processors. Because these data sets were analyzed over a period of several years, the analysis details changed as the program availability/capabilities changed. Data sets 1-10, 12-16 were analyzed using a BI initiated from within ARB with a single run with 5,000,000 generations. Data set 11 was run from within ABR using parallel runs with a single processor and ran for over three months. Data sets 17-18, 20-26, 33, 34 were analyzed from within ARB with two parallel runs with multiple processors and allowed to run until the two runs converged on the same tree, with the assumption that the analysis had completed when the two trees converged. The analysis used the default settings in the program and all data sets had their posterior probabilities calculated from the last 1000 trees. The conversion of the two runs was also tested using the compare option in the AWTY program [38].

Data sets 26-32 were also run within ARB using the RAxML (ML) function within ARB. Bootstrap ML analyses were run on the CIPRES website (http://www.phylo.org) for data sets 25 and 30.

A final comparison was also made between weighted and non-weighted parsimony analyses for a selected of the data sets. Data sets 26-32 were exported from ARB and MP analyses were implemented with PAUP*. Introduced gaps were treated as missing data, and informative characters were treated as multistate and unordered. Unweighted MP trees were obtained using the tree-bisection reconnection (TBR) branch-swapping option and a heuristic search with random additions of the taxa. Weighted MP analyses were also performed in PAUP using a rescaled consistency index as the weighting method. MacClade (Ver. 4, [39]) was used to generate the figure showing the rate variation among the sites in the rRNA molecule, which is the basis for the weighted parsimony analysis.

Stylized trees are presented with the major clades condensed in order to compare the analyses from the BI and ML analyses. In these trees, the pennate clade is colored yellow, the mediophycean clades blue, the coscinodiscophycean clades green and the bolidomonad sister group red. In the MP analysis, to illustrate the differences between the weighted and non-weighted analyses, the taxon labels were removed and the branches were colored in a similar fashion: Pennates (orange), Mediophytes (blue), Coscinodiscophytes (green) and Bolidomonads (red). MP trees were treated differently because the weighting had a pronounced affect on the branch lengths as well as the number of clades, whereas in the BI and ML analysis, the number of clades was more important, hence the clades were collapsed. Data Set 11 and 25 are discussed in detail elsewhere with regards to taxon relationships (Medlin submitted). Tree files were entered into FigTree (Ver.1.4) to produce the stylized trees.

Cladistic analysis

The data set from Medlin et al. [11] was entered into MacClade and valve features were encoded and the data set exported and analyzed in PAUP using a cladistic analysis.

Results

Bayesian analysis

For the BI analysis, the number of outgroups ranged from only two bolidophyte taxa to 45 taxa covering all major eukaryote groups with a bacterial/archeal root (Supplementary File Table 1). Not all of the analyses are illustrated but in Table 1, the results for each analysis are listed. Some combinations resulted in incomprehensible relationships, such as centric and pennates mixed. In others, the pennates diverged before the centrics. These abnormal results, which are contrary to either the fossil record or well established/accepted relationships among the diatoms, were regarded as artifacts of the combination of outgroups and bases selected for analysis and were not investigated further. In Figure 1, with different multiple outgroups, the resolution of the centric clades differs. Using only bolidomonads as the outgroup, resulted in a grade of clades for the centrics (Figure 1A with two bolidomonad taxa, Figure 1B with many bolidomonad taxa). Adding additional outgroups reduced the number of grades in the centrics in most cases. Adding one Mallomonas sp. plus many bolidomonads recovered monophyletic coscinodiscophytes, mediophytes and pennates (Figure 1C) but increasing the synurophytes with many bolidomonads created multiple clades (Figure 1D). Including more members of the pigmented heterokonts created a monophyletic mediophyte clade and four clades of coscinodiscophytes (Figure 1E). Including both heterotrophic and pigmented heterokonts with only two bolidophytes (Figure 1F) only reduced the coscinodiscophyte clades by one, and left the mediophytes with two clades, whereas multiple bolidophytes with heterotrophic and pigmented heterokonts (Figure 1G) maintained a monophyletic mediophyte clade and a two clade coscinodiscophyte clade. Another analysis with the same outgroups but with fewer diatoms and more bases resulted in more clades (Figure 1H). Taking selected crown group representatives as outgroups without heterokonts but including two bolidomonads proved to be too distant to the ingroup and abnormal results were often obtained (data sets 2-10). However, when heterokonts with ciliates, haptophytes and chlorophytes were included, one analysis recovered three monophyletic clades (Figure 1J) and the same data set with fewer bases recovered two clades in the coscinodiscophytes and a monophyletic mediophyte clade (Figure 1I). Selecting a representation of all crown group taxa, plus pigmented heterokonts and Cafeteria

Source	Coscinodiscophyceae	Mediophyceae	Bacillariophyceae	**Basal +Core araphids**	No. of outgroups
Alverson *et al.*, [16], fig. 3	**M**	4	**M**	yes	2 bolidomonads
Alverson *et al.*, [16], fig. 4	**M**	3	**M**	yes	2 bolidomonads
Alverson *et al.*, [16], fig. 5	2 plus *Ellerbeckia*	2	**M**	Basal plus 4 core	Multiple
Alverson *et al.*, [16], fig. 6	3 plus *Ellerbeckia*	1 + 5 clade polytomy	**M**	Basal plus 5 core	Multiple
Ashworth et al. [18], Fig. S1	5	**M** (75)	**M**	2 basal plus one core	1 bolidomonad
Cavalier-Smith & Chao [55]	2	**M**	**M**		Multiple
Choi *et al.*, [54]	**M**	3	**M**		Multiple
Ehara *et al.* [45]	**M**	**M**	**M**		Single
Lee *et al.* [56]	2	2	n/a	One Pennate	
Medlin and Kaczmarska, [10], Fig. 1	**M** (63)	**M** (98)	**M** (100)	Not shown	Multiple
Medlin and Kaczmarska, [10], Fig. 3	4	4	**M** (100)		Mulitple bollidomonads
Medlin et al., [5]	**M** (68)	M (95)	**M** (100)	yes (53+93)	Multiple
Medlin et al., [6]	**M** (91)	3	**M** (100)	yes (93+100)	Single
Medlin et al. [7]	**M**	3	**M**	yes (100 + 3 clades)	Multiple
Medlin et al., [9]	**M** plus *Ellerbeckia*	4	**M** (91)	yes (90+ <50)	Single
Medlin et al., [11]	**M** (100)	**M** (100)	**M** (99)	**yes (100+97)**	
Sato [8], fig 4 publication 10	**M**(< 50)	**M** (51)	**M** (91)	yes (76+ <50)	Multiple
Sims *et al.,* [21]	**M** (100)	**M** (100)	**M** (100)	**yes (100 +100)**	Multiple
Sorhannus [52], Fig. 1	3	3	**M**	yes	Single
Sorhannus [52], Fig. 2	3, polytomy	**M**	**M**	yes	
Sorhannus [52], Fig. 3	**M**	4	**M**	Not shown	
Sorhannus, [53]	3 plus *Ellerbeckia*	8	**M**	yes	Single
Theriot *et al.* [15]	2 plus *Corethron*	2	**M** (100)	only basal mutiple core	Single

Table 2: Summary of the results from major studies on the evolution of the diatoms and whether or not the classes were recovered as monophyletic (**M**) and if not how many clades could be assigned to each class for the centric diatoms. For the araphid diatoms, the recovery of basal and core araphids is indicated. Numbers in brackets refer to the highest bootstrap or posterior probability support for the clade, even if multiple types of analyses were conducted. The number of outgroups used in each study is also indicated because this will have an effect on the monophyly of the clades.

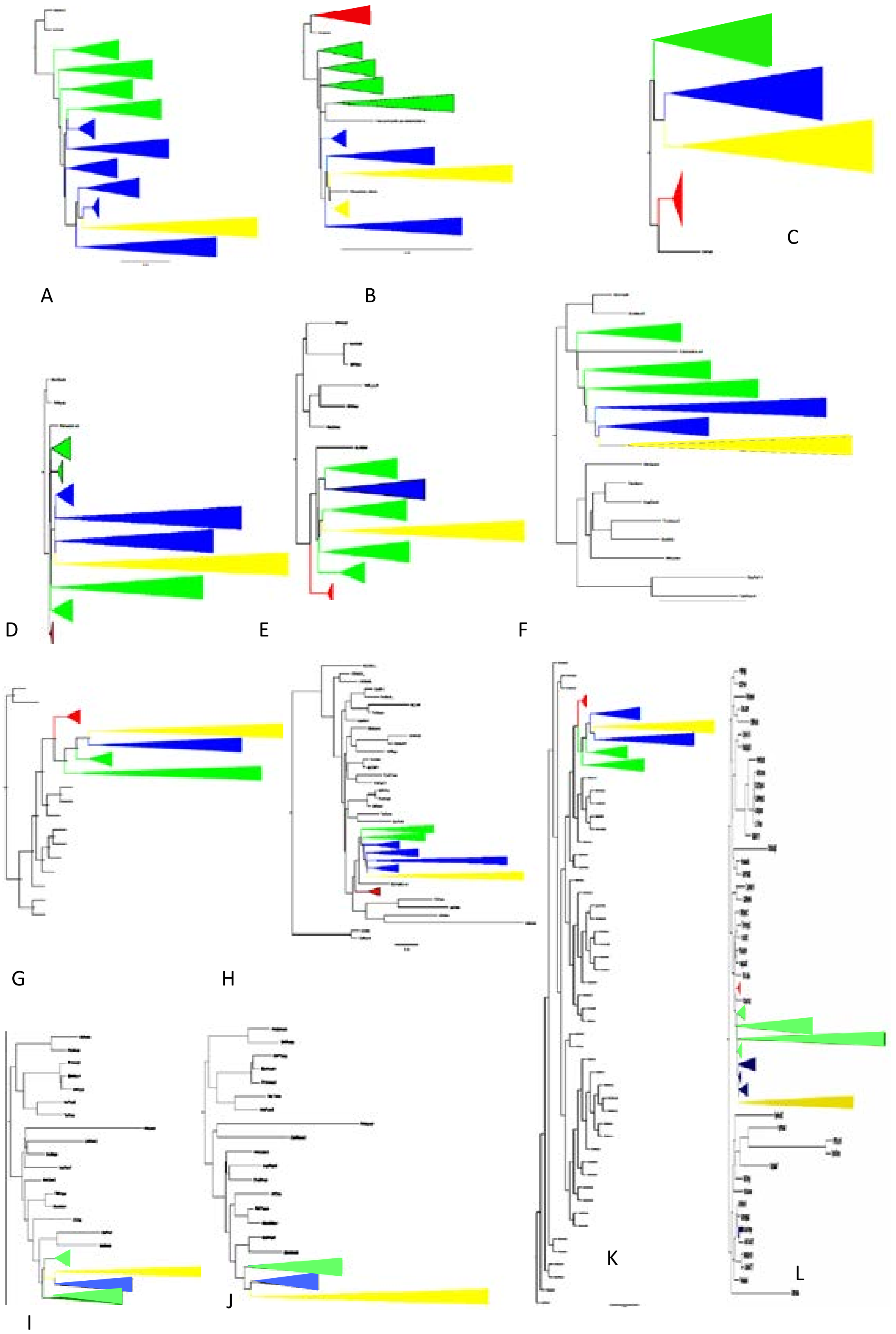

Figure 1: Last fully resolved BI tree from selected data sets in Table 1. Clades are collapsed for ease of assigned them to various diatom classes, Green = Coscinodiscophytes, Blue = Mediophytes, Yellow= Bacillariophytes, Red = multiple bolidomonad outgroups.

Fensome and Patterson resulted in multiple clades for both centric classes (Figure 1L). However, when using the maximum number of outgroups (45) with a bacterial/archeal root (Figure 1K), the number of grades was reduced to two inside each class. With these analyses, in only two data sets (11 and 25, Figures, 1C, J, respectively) were three monophyletic clades recovered. Ellerbeckia sol was pulled out into its own clade and in all but one at the base of the other clades in five data sets (Figures1A,E, F,H and L).

The number of taxa included in the data sets ranged from 422 to 899. The largest data set included all of the uncultured clone library sequences from Genbank whose identity had placed them within the diatoms with the ARB database release used here. Most of these are relatively short, ca. 250 bp and their inclusion in the analyses often distorted the trees (Table 1). Decreasing the number of bases and the number of taxa taken for the analysis altered the results, usually increasing the number of clades (compare Figure 1D tree inferred from 5203 bases with Figure 1B from 1458 bases). In one case (Figure 1J), the inclusion of more bases recovered the three classes as monophyletic clades and fewer bases induced multiple clades (Figure 1I). The datasets with fewer bases were generated with the 50% positional parsimony filter, whereas the latter has been generated with no filter. The justification for using all of the bases in the alignment is that because the alignment is based on the secondary structure of the molecule, all bases are aligned correctly and there is no reason to reject them based on an ambiguous alignment.

Maximum parsimony analysis

The rRNA molecules have a secondary structure that must be maintained for the molecule to be functional in the ribosome. Therefore, there are some bases that cannot change, whereas others must have a compensatory base change across a helix to maintain the secondary structure and functionality of the molecule. Only those bases in loops are free to mutate randomly. Thus, some positions that mutate have a higher mutation rate than others and those whose mutations are restricted by the secondary structure impart more phylogenetic signal than those that do not. The data set from Medlin et al. [11] was used to illustrate the base change frequency across the SSU rRNA gene. The data set was entered into MacClade and a total frequency base change map was generated (Figure 2). It is immediately evident that some positions (Figure 2A) have a higher frequency of change than others. These are the positions in the loops and those in the most variable regions of the molecule, with compensatory base changes, whereas those with no base changes are the most conserved positions that never change. A reweighting of the positions in MacClade results in the plot shown in Figure 2B, where the base change frequency is more evenly distributed along the molecule. Data sets 27-32 were analyzed using a RAxML analysis within ARB and exported from ARB and reanalyzed with PAUP using an unweighted and weighted MP analysis. The results of the RAxML analysis are presented in Table 1 and the results of the MP analysis are presented in Figure 3, where the unweighted MP analysis is presented alongside the weighted MP for each data set illustrated. Only the weighted MP branch lengths are colour-coded and all taxon labels are removed to illustrate the effect that the weighting has on branch lengths in the analysis. The effect of the weighting analysis is immediately seen with regards to the branch lengths separating the major clades. In each case, these branch lengths are elongated along the backbone of the tree because additional weight has been placed on the positions defining these branch lengths. Also, there are more clades in the unweighted analysis than there are in the weighted analysis and this is because the weighting has forced a consolidation of the clades.

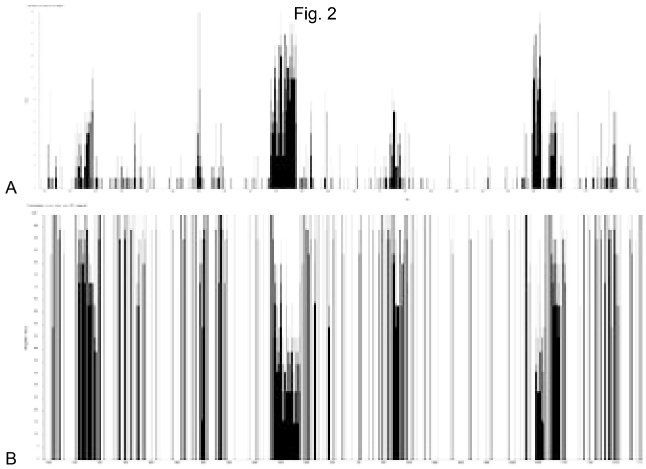

Figure 2: Comparison of the number of bases changes (steps) plotted for the M54 data set from base 1 to base 1250, which includes two hypervariable regions of the SSU. A. Map of unweighted frequency changes, where sites that are saturated are given the same weight as those that change infrequently. B. Map weighted frequency changes, where hypervariable regions are down-weighted and rarely changing position are up-weighted. The weighting mask obtained in B can be exported as a nexus file in MacClade for further analyses.

The weighting also reduced the length of any taxa with long branches in most cases because these multiple base changes are down-weighted primarily because they occur in loops, which are free to mutate. In the data sets tested, only dataset 29 recovered the three monophyletic clades with the exception that one bolidomonad taxon was inside the mediophytes (Figure 3H).

Cladistic analysis

Whether or not structures (tubes) are inside the annulus is one of the characters that has been used to separate the two centric classes. A formal cladistic analysis of this character was made using the data set from Medlin et al. [11] and compared to the formal cladistic analysis made by Theriot et al. [16] for the features of the auxospore. Figure 4 shows that the Thalassiosirales are inside the Mediophyceae, whereas in figure 9 of Theriot et al. [16] they are grouped with the Coscinodiscophyceae because they based on their cladistic analysis using auxospore features. They commented that the auxospore structure cannot be used to define the major clades of the diatoms. This is only true for the Thalassiosirales who have lost the ability to make bands and thus have reverted to the ancestral state of having only scales on the auxospores and thus the cladistic analysis will place them in the coscinodiscophyte clade making the mediophytes paraphyletic. Aside from this one exception, the auxospore features clearly separate the three classes. Using other valve features, the classes as defined by Medlin and Kaczmarska [10] are recovered (Figure 4).

Discussion

Taxon sampling

In this study, analyses with partial and full length sequences were included. The inclusion of partial sequences (mostly originating from clone library samples) is difficult to analyze and often render the results uninterruptable. In cases where full length or nearly full length sequences were used, reasonable clades corresponding to order and family levels in the diatoms have been recovered. The number of the clades in each of the trees where the three classes were not

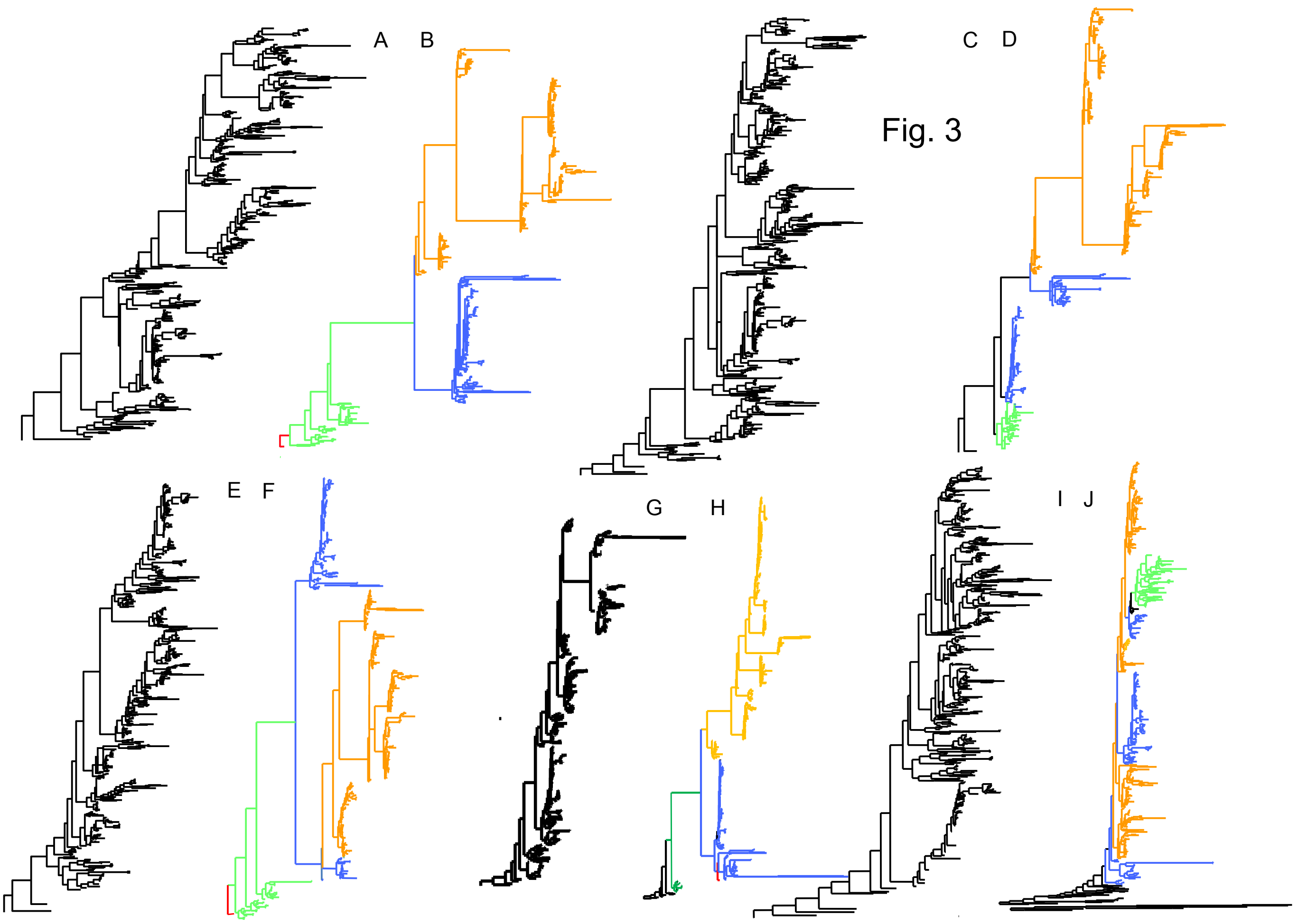

Figure 3: Comparison of weighted and unweighted MP analyses from selected data sets in Table 1. All taxon labels are removed to illustrate changes in branch lengths and number of clades. Green = Coscinodiscophytes, Blue = Mediophytes, Orange = Bacillariophytes, Red = Bolidomonads. A) data set 26 no weighting, B) data set 26 weighting, C) data set 27 no weighting, D) data set 27 weighting, E) data set 28 no weighting, F) data set 28 weighting, G) data set 29 no weighting, H) data set 29 weighting, I) data set 30 no weighting, J) data set 30 weighting.

monophyletic was highly dependent on the number of outgroups and the phylogenetic affinity of the outgroups. The unreliability of short partial sequences being placed in their correct phylogenetic position and/or distorting other phylogenetic relationships is problematic for many of the new generation sequencing to identify their samples.

Using too small data set has often been evoked to explain discrepancies in phylogenetic trees when the analysis results go against conventional wisdom. This was the first criticism by Alverson and Theriot [40] of Kooistra et al.'s [41] finding that Toxarium was not a pennate diatom but belonged to the bipolar centrics and concluded that the elongate pennate shape had evolved twice [42]. Alverson and Theriot [40] suggested that Kooistra et al. [41] analysis was flawed by poor taxon sampling. However, when they repeated the analysis and found the same results, their criticism changed to suggest that elongated valve shape had evolved once in the bipolar centrics too [42] and did not evolve twice because the elongated shape was present in many bipolar centrics. It took a formal cladistic analysis to prove that the highly elongated valve shape found in true pennates had evolved for a second time in the highly elongated bipolar centrics Toxarium,

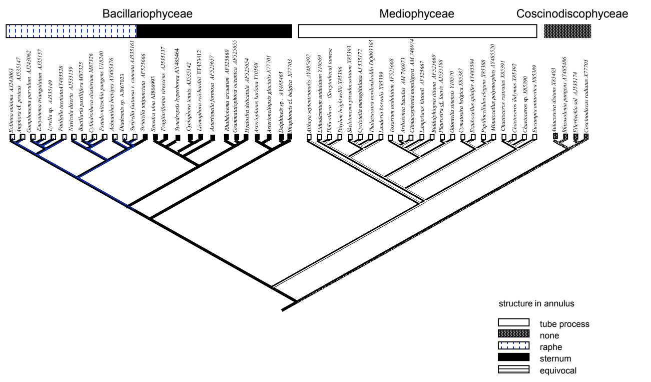
Bacillariophyceae
Mediophyceae
Coscinodiscophyceae
Eolimna minima AJ243063
Amphora cf. proteus AJ535147
Gomphonema parvulum AJ243062
Encyonema triangulatum AJ535157
Lyrella sp. AJ535149
Pauliella taeniataAY485528
Navicula diserta AJ535159
Bacillaria paxillifera M87325
Cylindrotheca closterium M87326
Pseudo-nitzschia pungens U18240
Achnanthes brevipes AY485476
Diadesmis sp. AJ867023
Surirella fastuosa v. cuneata AJ535161
Striatella unipunctata AF525666
Synedra ulna AJ866993
Fragilariforma virescens AJ535137
Synedropsis hyperborea AY485464
Cyclophora tenuis AJ535142
Licmophora reichardtii EF423412
Asterionella formosa AF525657
Rhabdonema arcuatum AF525660
Grammatophora oceanica AF525655
Hyalosira delicatula AF525654
Asterioplanus kariana Y10568
Asterionellopsis glacialis X77701
Delphineis sp. AY485465
Rhaphoneis cf. belgica X77703
Attheya septentrionalis AY485492
Lithodesmium undulatum Y10569
Helicotheca = (Streptotheca) tamese
Ditylum brightwellii X85386
Skeletonema pseudocostatum X85393
Cyclotella meneghiniana AJ 535172
Thalassiosira nordenskioldii DQ093365
Lauderia borealis X85399
Toxarium undulatum AF525668
Ardissonea baculus AM 746973
Climacosphenia moniligera AM 746974
Lampriscus kittonii AF525667
Biddulphiopsis titiana AF525669
Pleurosira cf. laevis AJ535188
Odontella sinensis Y10570
Cymatosira belgica X85387
Extubocellus spinifer AY485504
Papillocellulus elegans X85388
Minutocellus polymorphus AY485520
Chaetoceros rostratus X85391
Chaetoceros didymus X85392
Chaetoceros sp. X85390
Eucampia antarctica X85389
Aulacoseira distans X85403
Rhizosolenia pungens AY485486
Ellerbeckia sol AJ535174
Coscinodiscus radiatus X77705
structure in annulus
tube process
none
raphe
sternum
equivocal

Figure 4: Cladistic analysis of the Medlin et al.2008 data set using the morphological feature of presence of a structure associated with the annulus of the diatoms. States are 1) nothing present 2) tube process of some nature 3) annulus fills with silica to form sternum 4) annulus fills with silica to form sternum, the raphe slit appears in sternum. Monophyletic Coscinodiscophyceae, Mediophyceae, and Bacillariophyceae are recovered.

Climacosphenia Ehrenberg, and Ardissonia De Notaris [11]. The same conclusion can be reached by mapping valve outline on phylogenetic trees (see trees in [9,10,43]), and illustrated in Kaczmarska and Medlin) [44].

In Theriot et al. [16], the number of taxa in their largest SSU dataset was 1336 of which 673 were diatoms. They analyzed their two data sets (one with only seven bolidomonads as an out group and one with all remaining stramenopiles that were in Genbank in 2007 at their time of analysis and still they recover multiple clades: 2 coscinodiscophycean and 3 mediophycean. All data sets in their paper were analysed with MP and BI except for the multiple stramenopile dataset, which was only analyzed by unweighted MP. As shown in Figures 2 and 3, MP needs to be weighted to achieve an optimal analysis. Their grade of clades in their stramenopile rooted data base tree (their figure 1) is likely caused by their unweighted maximum parsimony analysis. The data set in this study that corresponds most closely to their multiple stramenopile outgroup analysis is data set 29, in which with the WT MP analysis, monophyletic classes were recovered with the following exception: one bolidomonad was outside the diatoms as its sister group and the other fell into the mediophytes (Figure 3H). The unweighted analysis recovered two coscinodiscophycean and three mediophycean clades as did the Theriot et al. [16] analysis.

Theriot et al. [16] discounted the recovery of the monophyletic classes using the cox1 gene [45] stating that it was based on too small a data set (note Theriot et al. record the dataset as having only four species but in fact nine taxa were used in that analysis). This gene has not been further explored with regards to the phylogeny of the diatoms, likely because of the large intraspecific and intrageneric level of divergence in this gene [45-49], making it difficult to amplify across many taxa and it is likely this gene is just as saturated as rbcL [8].

Grade of clades vs. monophyly

The data sets analyzed here clearly show that increasing the number of outgroups and the identity of the outgroup have an effect on the number of clades recovered in the diatom phylogeny. Using one or multiple bolidomonads as outgroup, although they are the correct sister group for the diatoms [50], will result in numerous clades in both centric groups (see all of the early work by Theriot).

The addition of multiple genes with this single outgroup has recovered a monophyletic mediophyte clade but left the coscinodiscophyte group with multiple clades [16,18,19]. The analysis by Sato [8] with multiple genes and multiple outgroups has recovered both groups monophyletic but only with the ML analysis.

Thus, if the use of multiple outgroups and not just the bolidomonads affects both single gene and multiple gene analyses and their ability to recover monophyletic classes, then it is highly recommended that multiple outgroups be taken for all future analyses with either single or multiple genes.

The question of grades of clades vs. monophyly is the subject of much active research and debate among diatomists. Theriot et al. [16] commented that most analyses recover grades of clades rather than monophyletic ones and provided the following references to support their statement: [5-7,9-10,42,45,51-53].

Some of these references clearly show a different interpretation in some of the trees presented than that claimed by Theriot et al. [16] because in some of them there are clear monophyletic groups or small numbers of clades and certainly not what one would interpret as a grade of clades, which implies in its strictest sense, many, multiple clades (Table 2). In Medlin and Kaczmarksa [10], their figure 3 shows the Coscinodiscophyceae composed of two small clades (2 species) and a large polytomy of the remaining species and the Mediophyceae composed of four small clades. But, figure 1 in the same paper shows all three classes to be monophyletic. In Medlin et al. (figure 3) [5], which is the only tree in that paper produced from a secondary structure alignment of 9 diatom taxa, there is a monophyletic Coscinodiscophyceae and the Mediophyceae is represented by one taxon as sister to a monophyletic Bacillariophyceae.

Further, in Medlin et al. [6,7] with 29 diatom taxa, there is a monophyletic Coscinodiscophyceae, Mediophyceae represented by three clades and a monophyletic Bacillariophyceae. In Medlin et al. [9], there is Paralia plus a monophyletic Coscinodiscophyceae and Mediophyceae composed of four clades. The Coscinodiscophytina are monophyletic in Sorhannus [51], (figure 3) but with Ellerbeckia pulled out of the remaining clade), and in Choi et al. [53] (figure 7) and composed of only two clades in Cavalier-Smith and Chao [52]. The Mediophyceae are monophyletic in Sorhannus [51] (figure 2) Cavalier-Smith and Chao [52] (figure 3) and composed of three clades in Choi et al. [53]. In Sorhannus [54], the Coscinodiscophyceae comprise one very large clade (42 species) plus three minor ones (5 species) and the Mediophyceae contain three minor clades plus a polytomy of all remaining species. In Alverson et al. [15, figures 3 and 4], the Coscinodiscophyceae are monophyletic and the Mediophyceae are composed of four clades. In their figures 5 and 6, the Coscinodiscophyceae are composed of two clades and the Mediophyceae of two to six unresolved polytomies, respectively. Using another gene, in Ehara et al. [45] with nine taxa, all three classes are monophyletic. Using three genes, Lee et al. [55] using one pennate as an outgroup recovered two clades of Coscinodiscophyceae and two clades of Mediophyceae, one of which was sister to one of the coscinodiscophycean clades.

However for the statement regarding 'grade of clades' to be true, one should find many multiple grades of clades from centrics to pennates in all of the supporting references that they cite. Instead, the contrary is found where one or both of the centric classes is monophyletic or the non-monophyletic centric class is composed of only a few clades often associated with a large polytomy, certainly not a grade of clades following a strict interpretation of the meaning of the word gradation to imply several or multiple, not just one or two.

Simonsen [21] (figure 3, page 44), frequently cited by Theriot and co-workers as showing a grade of clades from centric to pennates, does, to the contrary, show a deep basal dichotomy in the diatoms that does not represent segregation into centrics vs. pennates evident in his taxonomic system [21, p. 48-53]. Simonsen did advocate a formal separation of centrics and pennates at the order level (Centrales and Pennales), beneath the Class Bacillariophyceae [21, p.11]. However in his phylogenetic tree, he illustrates two separate lineages within the centrics: the radial and non-radial ones. Recognition of two separate groups of centrics (radial and non-radial) plus pennates shown on that figure is very similar to the basal dichotomy Medlin and Kaczmarksa [10] recovered between the subdivisions Coscinodiscophytina and Bacillariophytina. The only exception to this is Simonsen's suborder Coscinodisciineae includes the Thalassiosiraceae, which we now know from molecular data belongs to the bipolar centric group not the radial centric group. Thus, except for the Thalassiosiraceae, the first divergence in the tree by Simonsen is essentially that of the proposed subdivision Coscinodiscophytina. Furthermore, Simonsen [21] also defines Coscinodisciineae by one of the characters that Medlin and

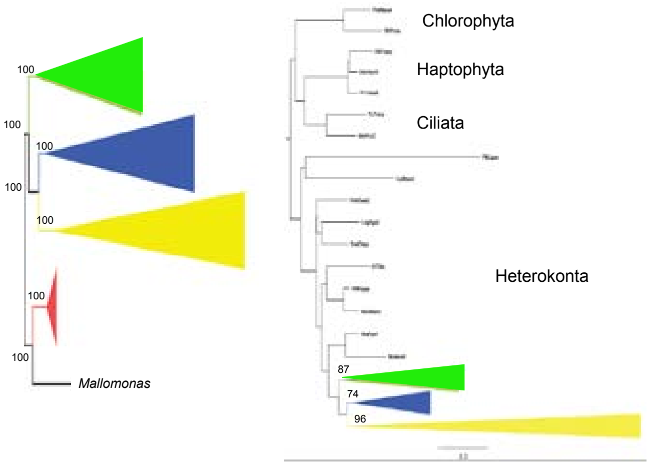

Figure 5: Detail of trees obtained by data set 11 (left) and 25 (right) showing the three monophyletic diatoms classes with multiple outgroups.

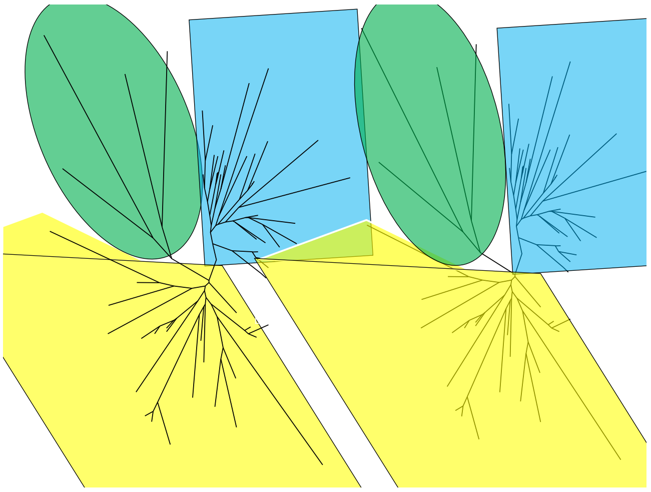

Figure 6: Final consensus trees obtained from two parallel Bayesian analyses for the M54 data set in Medlin et al. [11] (2008). Taxon names have been removed for ease of comparison of the branching patterns and branch lengths, which are identical in the two trees because the two independent analyses converged on the same tree. Eclipse surrounds the Coscinodiscophyceae; the square the Mediophyceae and the rectangle the Bacillariophyceae.

Kaczmarska have used, viz., the marginal ring of processes. In the second branch of the first divergence, there is a polytomy between the suborders Rhizosoleniineae and Biddulpiniineae. These groups are defined in Simonsen's key as those diatoms that do not have marginal rings of processes [21, p.45]. He would not have known as we do now from morphogenetic data that the labiate process of Rhizosolenia begins on the margin and moves to the centre, thus it was only logical for Simonsen not to include the Rhizosoleniaceae in Coscinodisciineae. Further, Simonsen's [21] Biddulpiniineae are not a grade of clades but are relegated to two clades and one of them is the Eupodiscaceae from which the pennates arise. Thus, strictly speaking, a grade of clades from centrics to pennates advocated by Theriot et al. is absent, as well in Simonsen [21]. Furthermore, consultation (LM, 1992-present) with the authors of Round et al. [22] has found that none of them maintained at the time of the writing of their book that they viewed the centrics as grading into the penates.

Methods of phylogenetic analyses of the SSU rRNA tree

Bayesian analyses and number of generations. For data sets 1-10 and 12-16, the Bayesian Inference Analyses (BI) were run for 1,000,000 generations. Twelve years ago, in 2002 when Medlin and Kaczmarska [10] was submitted, 1,000,000 was a common number of generations for small data sets that was used in Bayesian analyses, which was then newest and most advanced phylogenetic analysis available. In the manual for MrBayesV3.0, a typical MRBAYES block is given on page 8 and in that block, 1,000,000 generations is written and it is likely that most workers will have reproduced this data block identically except for modifications in file output names. In the same manual, it is stated that the only way to estimate that an analysis has run enough generations is to monitor the likelihoods of the trees to see if they have converged on a stable value (page 4 of BI v3.0 manual), which is what happens when two parallel runs are performed as done in this study. Computational times for the tree presented in Figure 3 of Medlin and Kaczmarska [10] with 500,000 generations were over 3 months as were the analyses of data sets 1-10, 12-16 done with 1,000,000 generations. These data sets were run more or less simultaneously and the results of one did not necessarily dictate the option used in the following one because of the computational time. The only statistic available for assessing the analysis at that time was the shape of the curve of the log likelihoods of each tree. When it reached a plateau, that meant that the resolution of the trees was completed or nearing completion because the variation in the log likelihoods from each new tree was minimal. In these analyses, the log likelihoods had stabilized; thus there was no indication that the analyses had not gone to completion.

Of the 30+ phylogenetic papers published in Journal of Phycology in 2004 (the same year Medlin and Kaczmarska appeared in print), only eight used a BI analysis in addition to other analyses (usually ML) and seven of the eight ran them for 1,000,000 and only one ran them for 2 million generations; this was then the most robust analysis performed in only a few labs. Nearly all other analyses published that year in Journal of Phycology (where most algal molecular phylogenies were published at that time) included only ML in addition to either MP (only one weighted) or Neighbor Joining (NJ) analyses. Longer runs with more generations were at that time too computationally intensive and furthermore, there is no indication that Medlin's trees were not optimal. As parallel processing became common place, parallel runs of the same analysis could be made and it was presumed that if the consensus trees from the two independent runs were identical, then the analysis had converged on one tree and the analysis was finished [11]. Two parallel runs were made for twelve data sets and at the end of the run and to illustrate the point that the two trees converged; Figure 7 shows the last two trees from the Medlin et al. [11] data set because that data set is small enough to be represented on one page [34]. The last two trees are identical even to their branch lengths. There no reason to believe that the analysis has not finished and this applies for all 12 data sets analyzed here.

The consensus trees from two parallel runs can be analyzed with the AWTY online analysis [38], in which the log likelihoods are plotted from one run against another run (Figure 7). If the runs are converging, the points for each tree should cluster along the regression line. In Figure 7E, the dataset 25 results from the compare option in the AWTY program show that with a 90% burn in of 1000,000 generations, the trees cluster around the regression line. Theriot et al. [16] in reanalyzing the data set from Medlin et al. [11] with 1,000,000 generations with a 90% burn in, suggested that they had not run their analysis long enough and provided a figure showing their analysis of their data set with 1,000,000 generations with 90 % burn in (Figure 7C), the same data set with 20,000,000 generations (Figure 7D) and with 50,000,000 generations (Figure 7E). I have repeated this analysis and Figure 7B with a 90% burn in of 1,000,000 generations looks very much like Theriot et al.'s with a burn-in of 90% with 20,000,000 generations of the same data set (Figure 7D) and 90% with 50,000,000 generations (Figure 7E). When the burn-in is increased to 99%, the plot shown in Figure 7A is obtained and can be compared to Theriot et al.'s plot, which they claim to be a 90% burn in at 1,000,000 generations (compare 7A to 7C). The graphs are similar but assigned different burn-in values. Theriot et al. [16] have erroneously reported a 99% burn in for a 90% burn in analysis. Thus, because of this error, Theriot et al. [16] have incorrectly claimed that Medlin et al. had not analyzed their data for long enough to obtain a correct tree. Figure 7B shows that Medlin et al. [11] analysis at 1,000,000 is not substantially different from Theriot et al.'s [16] at 20,000,000 (Figure 7D) and 50,000,000 (Figure 7E) and although running the analysis for 50,000,000 generations is optimal, the run at 1,000,000 generations was sufficient to recover the correct tree. It is therefore unlikely that Medlin et al.'s tree was an artefact of insufficient numbers of generations as claimed by Theriot et al. [16, p.288]. If indeed they have generated a plot with a 99% burn-in and presented it to be a 90% burn-in and then used that to criticize the number of generations (Theriot et al. [16, p. 288 L3, RC]), then their criticism of Medlin's et al.'s work is ill-founded. At a 90% burn-in of 1,000,000 generations with every 1000th tree saved, then the last 100 trees should be compared, yet in their plot (Figure 7C), there are less than 100 dots in the plot (at least 90 trees are missing) - clearly an error has been made in their analysis and in their conclusion.

A strategy of using 1,000,000 generations was used in all of the papers published by Medlin from 2004 to 2008 and in the first 14 data

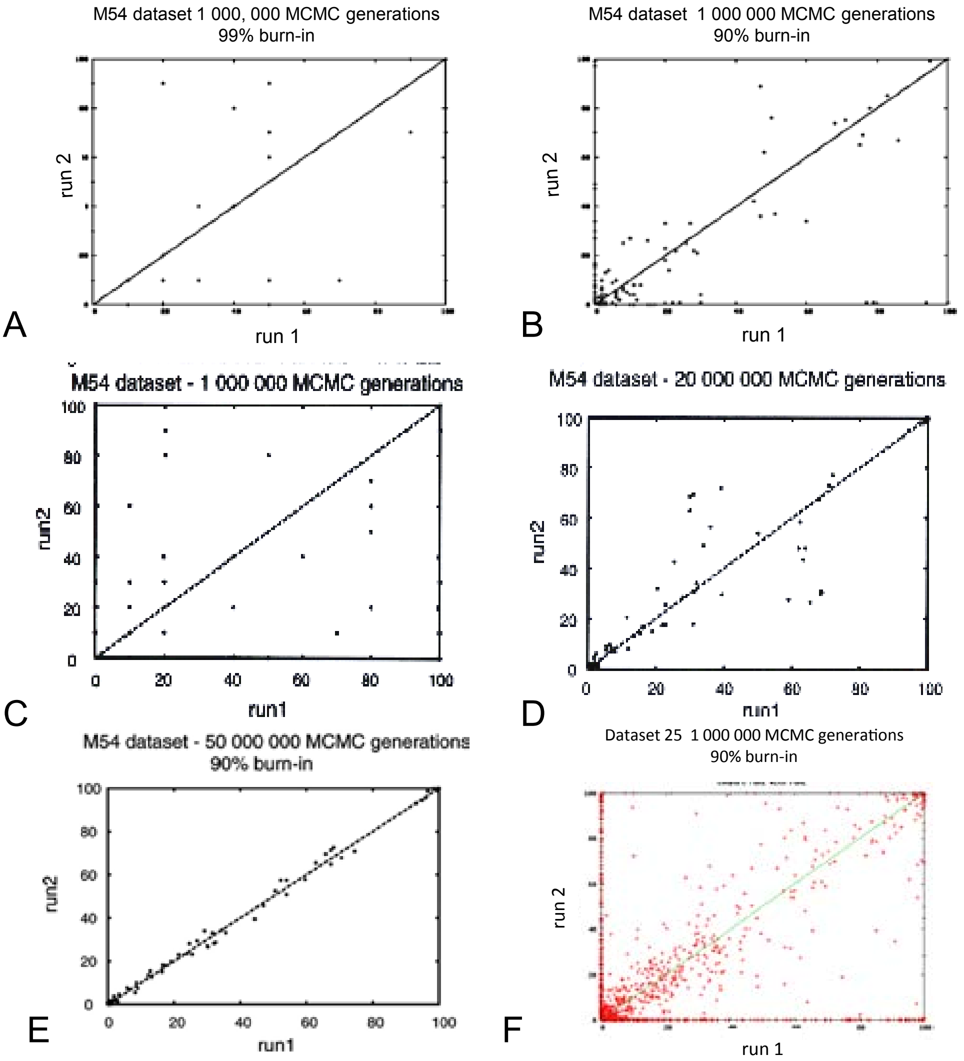

Figure 7: Comparison of trees retained by parallel Bayesian analysis using the AWTY method from the M54 data set from Medlin et al. [11] 2008 reanalyzed in Theriot et al. [16] (2009) and reanalyzed in the present study.A. 99% burn in comparison by Medlin of 1,000,000 generations. B. 90% burn in computed by Medlin of 1,000,000 generations.C. 90 % burn in comparison computed by Theriot et al. [16] of 1,000,000 generations.D. 90% burn in computed by Theriot et al., [16] (2009) of 20,000,000 generations.E. 90% burn in computed by Theriot et al., [16] (2009) of 50,000,000 generations.F. 90% burn in computed by Medlin for data set 25.Note the similarity between (A) 99% burn in calculated by Medlin and (C) 90% burn in calculated by Theriot et al. for 1,000,000 generations and the similarity between the (B) 90% calculated by Medlin for 1,000,000 generations and (D) 90% calculated by Theriot et al. [16] for 2,000,000 generations. The 90% burn in presented by Theriot et al. [16] is an error and clearly represents a 99% burn in and this error is the source of mistaken conclusions that the trees run by Medlin were not long enough.

sets analyzed in this study. It would appear at least from the AWTY analysis, that this is a sufficient number of generations to run the analyses to obtain a correct tree form the SSU gene for the diatoms. With more computer power, two parallel runs with the later data sets (17-18,20-26,33,34) were possible and Sato working on his PhD in Medlin's laboratory ran his analyses until the standard deviation values of log likelihoods of the two runs dipped below 0.01 [8]. All of these changes in my analytic methods have come gradually with improvements in the programs and greatly enhanced computer capacity.

The data shown here support that the strategy for the analysis of the rRNA gene using parallel runs until the trees converge is correct and Figure 7F confirms that the results of the present study have converged on a tree that is reasonable for the number of generations used. Of course, 50,000,000 generations will produce a slightly better tree but the computational time would be enormous.

Weighted vs. unweighted maximum parsimony analyses

Prior to 2009, Theriot and co-workers used an unweighted maximum parsimony analysis for their data analysis [16,56]. Also in their 2009 paper, the largest data set for the heterokont outgroup analysis was analysed with unweighted MP. They have never used a weighted analysis, despite acknowledging that rate heterogeneity can cause an error [57]. It is readily seen that the noise in the data from an unweighted MP analysis (Figure 2) is considerable. With such a noise level, a poor recovery of monophyletic taxa is understandable and may even be expected in unweighted MP analyses. If one does not down-weight the bases that changed more frequently and are likely to be saturated with no phylogenetic signal, i.e., noise, then the phylogenetic clades recovered are likely to be incorrect. In Figure 3, the weighted analysis reduces the number of clades, although in only one of the data sets tested with weighted MP analyses, were the three monophyletic classes recovered, although one bolidophyte was inside the mediophyte clade. The greatest distortion in the weighted MP analyses was that the mediophytes diverged before the coscinodiscophytes and the coscinodiscophytes were embedded in the pennates. Theriot and co-workers abandoned unweighted MP analyses as their preferred method after 2009 when they switched to BI analyses [16]. However, they still used an unweighted analysis for their largest data set (figure 1) with multiple heterokonts. The unweighted data set 29 (Figure 3G) recovered the same number of clades as a comparable dataset analysis by Theriot et al. [16], whereas the weighted version of the same dataset recovered three monophyletic classes.Weighted distance analyses have also been performed by other workers in other types of analyses. These were first produced by van de Peer et al. [58] for distance analyses because the rRNA molecule has positions that do not evolve at a constant rate as shown in Figure 2.

Therefore, one has to down-weight the positions that are most variable and up-weight the positions that seldom change in much the same way that the weighted MP data sets are achieved, only the analysis is a distance method rather than a parsimony one, so all bases are taken into consideration, whereas in the parsimony one, only the parsimonious bases are analyzed. Kooistra and Medlin [59] also used in this type of analysis of the diatoms to help recover monophyletic centric clades. Theriot et al. [16] commented on van de Peer's unusual results that his analysis produced where centrics and pennates, in a traditional fashion [sensu 22,60], were monophyletic [61], figure 3). After this publication, unpublished sequences were sent van de Peer and his tree was redone to see whether and how the new sequences affected the resulting tree. Van de Peer's weighting scheme is empirically derived from all rRNA sequences and used in distance trees, whereas the weighting schemes derived using MacClade (or within PAUP) are calculated only on the input nexus file and used in MP analyses. Clearly, the results of the two approaches would either strengthen or refute any trees. The tree produced by van de Peer with the sequences sent to him in 1997 but never published is shown in Figure 8 and the classes as Medlin and Kaczmarksa proposed are recovered by him and with his weighting scheme. Van de Peer's published tree that recovers centrics as monophyletic is an artifact of taxon sampling. It is clear that weighting of the data is essential for the SSU rRNA gene analyses.

Analyses with other genes

Theriot et al. [16] comment that the resolution of the rRNA gene was limited, suggested that multiple gene phylogenies should be embraced and commented on the recovery of the classes based on other genes. They mention that an unpublished rbcL data set does not support proposed classes. The rbcL gene is not an appropriate gene to recover deep divergences among major diatom lineages because this gene is strongly saturated at the third codon position [8,62]. Mann et al. [62] has been sequencing the rbcL gene for several years and concluded that as compared to SSU rDNA, the rbcL gene appears more suited in diatoms to studies of evolution at order to generic levels of taxonomic hierarchy. Anomalous results were also obtained in the extensive analysis of pennate species with 3 genes [25-28] where it repeatedly found that araphids were pulled inside the raphid clade when using rbcL gene alone. Even taking outgroups as far away as red algae failed to resolve this problem. The reader is referred also to the very unusual trees from rbcL produced by Fox and Sorhannus [63] and to the nearly monophyletic centrics vs. pennates produced by Choi et al., [53] (figure 7). So this gene can only be used in multiple gene analyses where its signal is compensated by the signal from other genes and, if used alone, can only be used to resolve some but not all recent divergences [64].

However the rRNA gene should not be abandoned and should be the gene of choice if single gene analyses are performed. Its alignment using a secondary structure ensures that nearly the entire full length of the molecule can be used in phylogenetic studies. Alverson et al. [15] do not use the variable V4 region of the SSU rRNA gene, which is the most variable region of the molecule and thus carries much phylogenetic information if aligned properly. The ARB program uses the van der Peer model for the secondary structure of the molecule and not the Gutell model as used by Theriot (see references in Medlin [32]). Piganeau et al. [65] have commented on the rate of evolution in the rRNA genes vs. protein genes in multicellular vs. unicellular lineages. The rate of evolution in the rRNA genes is faster than in the protein genes, making resolution better among taxa recently evolved using rRNA genes than that recovered with protein genes. They comment that a single base difference in the rRNA gene likely represents a new species and this has profound effects on the use of rRNA genes for biodiversity and barcoding studies and suggest that they should be the preferred gene for this area of resolution.

Cladistic analysis

Theriot et al. performed a cladistic analysis of the auxospore structure details as given in Table 1 of Medlin and Kaczmarska [10]. This was unnecessary because it is obvious that all radial centrics (Coscinodiscophyceae) with scaly auxospores and Thalassiosirales (as clearly stated in [66, p. 164 and 10, p. 261,267] of the Mediophyceae, will group together. In fact, Theriot et al.'s approach shown in his figure 9 is an excellent illustration of the limitations of a cladistic analysis of morphological characters when used as the sole basis for phylogenetic inferences. Theriot et al. [16] neglects to point out that

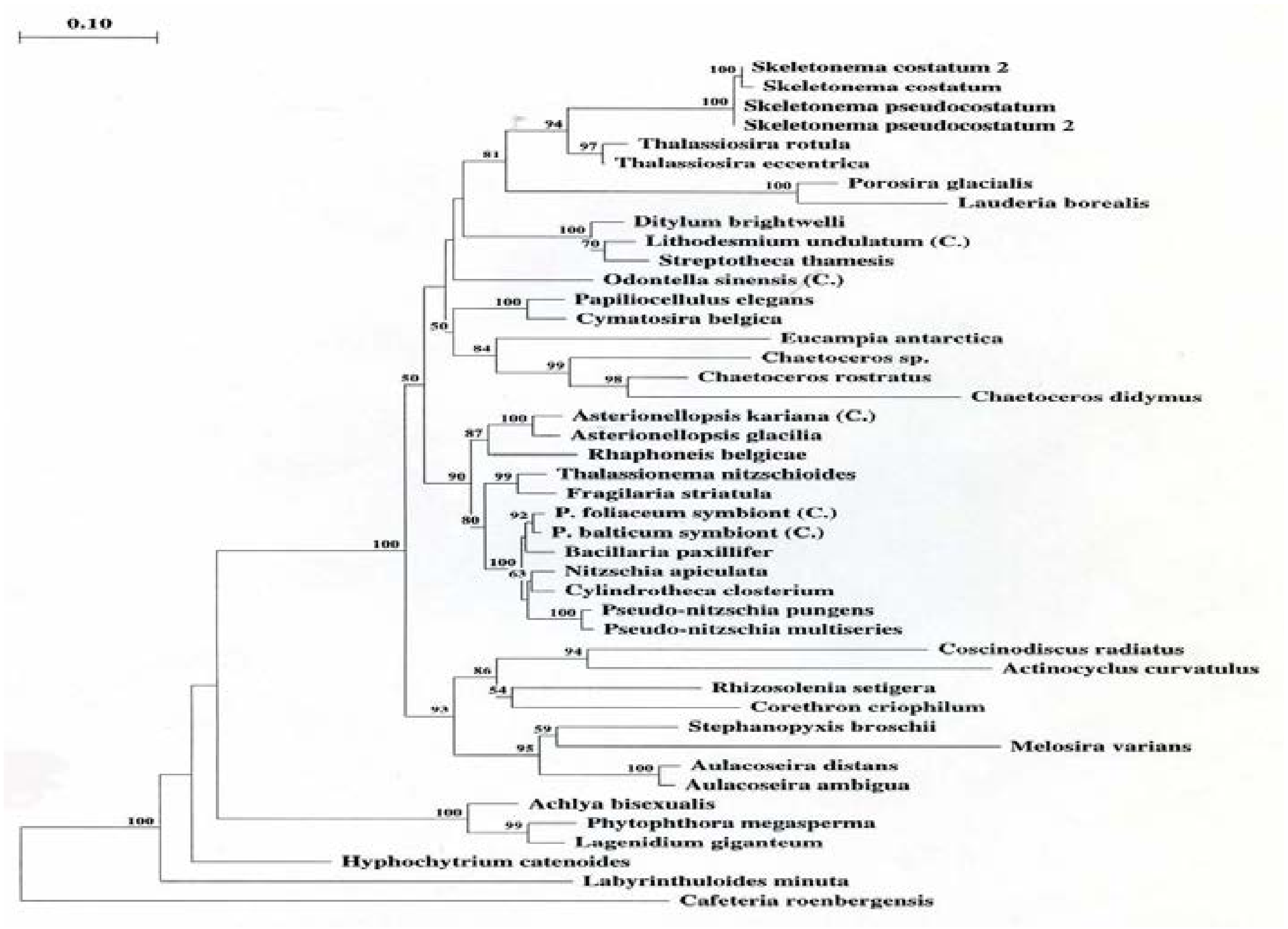

Figure 8: Weighted neighbor joining SSU tree with 100 BT replicates calculated by Van de Peer after receiving unpublished diatom sequences from Medlin.

Thalassiosirales is a consistent sister group to non-radial centric Order Lithodesmiales [6,66] (figures 1-3). Alverson et al. [67] (figures 2-4) used the Lithodesmiales as the outgroup to all Thalassiosirales so presumably, they also consider them to be their closest sister group. This relationship, consistently recovered in Theriot's lab and also shown in Theriot et al. [16] suggests that members of Thalassiosirales likely lost the properizonial bands secondarily. Kaczmarska et al. [68], Medlin and Kaczmarska [10, p. 267] and Medlin [12] state that this group maintained the ancestral character of scaly isometric auxospore in their ontogeny and morphology in the sense used in Van den Hoek et al. [69, p. 487-489] for scales in green algae shown to be dispersed throughout the molecular tree [70]. With this in mind, Theriot et al., [16] figure 9) actually supports monophyly of all three classes proposed by Medlin and Kaczmarska [10].

Theriot et al. [16] continued to review the characters that have several exceptions to their distribution across the classes and comment on the invalidity of their use to define the classes but then omitted the one character that will pull Thalassiosirales back into the Mediophyceae, as the sequence data do. That character is the presence of a structure within the annulus, which as shown in Figure 4 recovers the three monophyletic classes. The other characters used in describing the classes have some exceptions and this was all clearly stated in Medlin and Kaczmarska [10]. All of the exceptions to each of less known characters used to define the classes were noted.

Number of outgroups

These analyses have shown that the number and kind of outgroup clearly affects the recovery of the monophyletic clades. Most analyses have used only one bolidomonad as the only outgroup for all diatoms [16-19]. Although this is the correct sister group for the diatoms, the data sets analyzed here show that taking only two bolidomonads will result in multiple coscinodiscophyte and mediophyte clades. Even taking multiple bolidomonads does not improve the situation. One data set with two bolidomonads and one additional Mallomonas Perty sp. (data set 11) recovered the three monophyletic clades. Rampen et al. [71] recovered a monophyletic Coscinodiscophyte clade and two mediophyte clades using one bolidomonad and two chrysophytes. The addition of multiple outgroups within the heterokonts (pigmented only, or pigmented plus unpigmented, or only unpigmented) reduced the number of clades. The combination with the maximum of heterokonts (data set 29) recovered monophyly of both centric groups in one WT MP analysis. Taking outgroups from the crown group radiation excluding heterokonts but including bolidomonads as

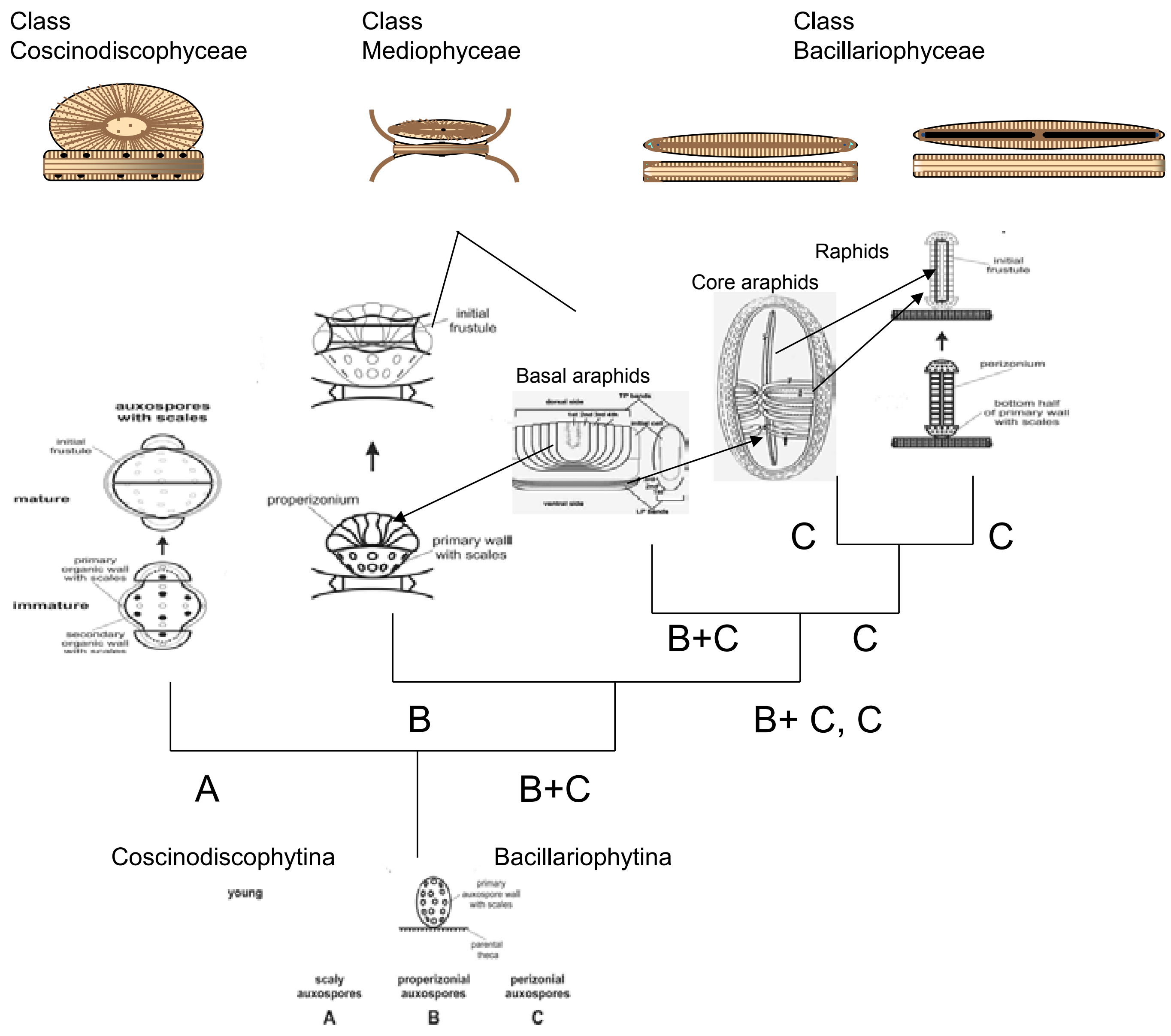

Figure 9: Scenario of our current understanding of the evolution of the auxospore and hence the evolution of the diatoms redrawn from Medlin and Sato [79]. The retention of a scaly auxospore by the Thalassiosirales in the Mediophyceae is not included in this scheme. Schematic drawings of diatom valves done by W.C.H.F. Koositra.

sister produced abnormal results and only when distant crown group outgroups (cryptomonads and/or dinoflagellates) plus pigmented and non-pigmented heterokonts were included were monophyletic clades recovered in the BI analyses. Ciliates, haptophytes, chlorophytes and heterokonts seemed to provide the correct combination to recover monophyletic clades. Deeper divergences in the Eukaryotic trees, such as Giardia Kunstler and Euglena Ehrenberg produced abnormal results. The addition of bacterial and archeal outgroups to the crown group+ heterokont outgroups reduced the number of clades among the trees with reasonable results. In Medlin and Kaczmarska [10], it took 168 outgroup taxa to recover the monophyletic clades with reasonable bootstrap support.

Single or multiple genes

Irrespective of whether a single gene or multiple genes were used, I have recovered trees that consistently show the monophyly of the two centric classes (except for very early trees) and this is contingent on using multiple outgroups. Other workers have recovered one or the other class as monophyletic (Table 2). The combination of outgroups is critical and the combinations used here that resulted in monophyletic classes should be used when adding other genes. Of course, for the heterotrophic taxa, chloroplast genes will not be available so mitochondrial genes should also be a target. These are not anomalous results as Theriot et al. [16] suggest. His critique of Medlin's earlier work, while enlightened by the wisdom of hindsight and new methodology, would appear to be incorrect both in terms of data presented to support his arguments as well as their reanalysis of her data sets. Using these larger datasets with more outgroups, monophyletic classes have been recovered, but as previously emphasized that is contingent on using the full rRNA gene sequence and multiple distant outgroups. Although some of the relationships within the raphid lineage with data set 25 are not optimal, it suggests starting points for future work to address. The use of short clone library sequences appear to distort some of the relationships and should not be included in future analyses.

There is no reason why the diatom classes as defined by Medlin and Kaczmarska [10] cannot be used. If analyses continue to recover grades of clades, Medlin [12,32] have produced arguments as to why paraphyletic groups should be recognized. These and the monophyletic groups assigned to the classes can be defined by morphological features, even the araphid group now possesses positive features [72] and are no longer defined by the absence of features as Williams and Kociolek [73] denounce. Reasons given on a diatom identification website for not using the class classification proposed in Medlin and Kaczmarksa are feeble [10] (http://craticula.ncl.ac.uk/%20EADiatomKey/html/taxa.html): 'A new classification was proposed by Medlin and Kaczmarska [10], but this has not been fully evaluated and it has the practical disadvantage that two of the three major groups they recognize cannot be distinguished using light microscopic data. We have therefore retained the 1990 scheme here, though we realize its inadequacies." Not only is it mystifying as to why the radial centrics cannot be distinguished from the bipolar centrics at the light microscopic level (compare Coscinodiscus Ehrenberg to Biddulphia Gray even in a stereoscopic microscope) but also decades ago, diatom taxonomy and systematic incorporated numerous new taxa of all levels that are distinguishable only with use of electron microscopy [74] and most recently even species fully cryptic morphologically, indistinguishable using any kind of microscopy [75,76]. Also, radial Coscinodiscus with no processes in its annulus can be distinguished from radial Thalassiosira with a central strutted process, so the potential confusion of the radial Thalassiosirales being grouped with the bipolar centrics can be resolved by looking for the position of the tube processes. These distinctions are also easy to teach to newcomers in the field. There are not just centrics and pennates but 1) radial centrics with no central process, 2) bipolar centrics and radial centrics with central processes and 3) pennates; (informally already implied by Stosch, and Simonsen [21,77]. Keys to the diatoms using the morphological features known to support the three monophyletic clades are provided in Medlin [12].

Conclusion

It has been shown here that certain types of analyses with a secondary structure alignment and multiple outgroups using SSU alone do recover monophyletic classes. This has also been shown in some analyses with multiple genes and multiple outgroups. There are clear morphological differences between the classes (Figure 9) primarily based on auxospore ontogeny and envelope structure, the presence or absence of a structure (tube process or sternum) associated with the annulus and the location of the cribrum in those genera with loculate areolae. The classification of the diatoms into the 2 sub-divisions and 3 classes as proposed by Medlin and Kaczmarska [10] reflects better the diversity and molecular phylogeny of diatom lineages (Figure 9) than did the previously accepted system [22], with one centric and two pennate classes or Simonsen [21], where formally only orders Centrales and Pennales were recognized, even though Simonsen obviously had some misgivings about it as summarized in his figure 3 [21, p. 44]. The analyses shown here proves that we are not replacing one paraphyletic system with another as claimed by Williams and Kociolek [73]. In the light of errors and misrepresentations shown in Theriot et al. [16] the erection of the classes was appropriate for the time of publication and still is.

Even if it remains so that some genera of diatoms stay paraphyletic (in my analyses or in others that do not use multiple outgroups), paraphyletic taxa exist in many groups that are still in common use and in some cases recognised taxonomically, viz., prasinophytes (three clades still unnamed), birds vs. reptiles, monocots vs. dicots in green plants, Bangiophyceae (3 new classes but bangiophytes still used as a collective noun) vs. Rhodophyceae in red algae, just to name a few. Similar examples are known among diatoms, e.g., a new thalassiosiroid genus, Shinodiscus Alverson, Kang and Theriot, is deeply embedded in the genus Thalassiosira and in the clade with Thalassiosira eccentrica [42,78], figures 2-4). The same authors cannot be right in both cases; advocate strict monophyly for some groups but not for their own taxa. Medlin [32] is prepared to accept these often common examples of paraphyly in the diatom genera because paraphyly may exist for some period of time in lineage evolution until the sister taxa have gone extinct rendering the new taxon monophyletic. Thalassiosira is an excellent example of such a case illustrating how difficult it would be to split a large genus up into multiple, new genera to achieve monophyly for all distinct intra-generic clades. The erection of Conticribra certainly makes a start.

Diatomists should build upon the preliminary four gene phylogeny [8] of diatoms with multiple outgroups inside and outside the heterokonts and to test for the presence of the traditionally overlooked morphological features, such as the Golgi arrangement in the Coscinodiscophyceae and sexual reproduction in the basal araphids, which are still largely missing for a great many taxa. This will facilitate a better understanding how the features identified as being representative of the proposed classes hold for as many species whose sequences are in the trees where these data are missing. The classification system erected in Medlin and Kaczmarska [10] should be used because it better reflects the evolution of the diatoms. Theriot and his co-workers [16-19], despite their continued efforts to discredit this classification, can only very weakly reject the monophyly of both classes. They are consistently achieving a monophyletic Class Mediophyceae with high boot support [18,19] with multiple genes BUT only one bolidomonad as an outgroup. Their conclusion that SSU may be limited in resolving deep divergences in diatoms is shown here not to be the case with multiple outgroups.

Taking bolidomonads as a single outgroup has never produced all three classes monophyletic in any type of analysis; however taking multiple outgroups from within the heterokonts has in several analyses. Ciliates, haptophytes, and chlorophytes seem to be appropriate taxa for outgroups but cryptomonad, dinoflagellates, euglenoids, and Giardia are not likely because they are too divergent from the main crown group radiation. Hopefully, with time more such data will continue to become available and more outgroups inside and outside the heterokonts can be used to reinforce the monophyly of the classes with all types of analyses. Consequently, as our knowledge and available instrumentation advance, additional tests and approaches are performed, the hypotheses are updated and refined (compare changed in the trees in [5-7,9-12,14,43] as the field develops. No discussion of new ideas is possible until they are submitted to peer scrutiny. This is a healthy and normal course of research strengthening scientific discoveries in all science and evident also in diatom taxonomy; compare Simonsen [20] to Simonsen [21] and his reflection on feedback he received [21, p. 9, L10]. For this process to take its natural course, however, an accurate representation of re-analysed and criticized data is an absolute necessity (see [80]).

Acknowledgments

I am grateful to T. Friedl for allowing me to use his unpublished sequences. Drs. R. Crawford, B, Bezsteri and S. Sato critically read the manuscript. Dr. I. Kaczmarska contributed to an earlier version of this manuscript.

References

1. Mann DG (1999)The species concept in diatoms. Phycologia 38: 437-495.
2. Williams DW (2007) Classification and diatom systematic: The past, the present and the future. In: Unraveling the algae: past present and future of algal systematics CRL press London 57-92.
3. Mann DG, Marchant HJ (1989) The origin of the diatom and its life cycle. In: The chromophyte alage: problems and perspectives Claredon Press, Oxford, UK.
4. Kaczmarska I, Poulickova A, Sato S, Edlund MB, Idei M, et al. (2013) Proposals for a terminology for diatom sexual reproduction, auxospores and resting stages. Diatom Research 28: 1-32.
5. Medlin LK, Williams DM, Sims PA (1993) The evolution of the diatoms (Bacillariophyta). I. Origin of the group and assessment of the monophyly of its major divisions. European Journal of Phycology 28: 261-275.
6. Medlin LK, Gersonde R, Kooistra WHCF, Wellbrock U (1996a) Evolution of the diatoms (Bacillariophyta). II. Nuclear-encoded small-subunit rRNA sequence comparisons confirm a paraphyletic origin for the centric diatoms. Molecular Biology and Evolution, 13: 67-75.
7. Medlin LK, Gersonde R, Kooistra WHCF, Wellbrock U (1996b) Evolution of the diatoms (Bacillariophyta): III. Molecular evidence for the origin of the Thalassiosirales. Desikachary Festschrift Nova Hedwigia 112: 221-234.
8. Sato S (2008). Phylogeny of araphid diatoms inferred from morphological and molecular data. PhD Dissertation. University of Bremen.
9. Medlin LK, Kooistra WHCF Schmid AMM (2000). A review of the evolution of the diatoms - a total approach using molecules, morphology and geology. In: The origin and early evolution of the diatoms: fossil, molecular and biogeographical approaches. Szafer Institute of Botany, Polish Academy of Science, Cracow, Poland: 13-35.
10. Medlin LK, Kaczmarska I (2004) Evolution of the diatoms: V. Morphological and cytological support for the major clades and a taxonomic revision. Phycologia 43: 245-70.
11. Medlin LK, Sato S, Mann DG, Kooistra, WCHF (2008b). Molecular evidence confirms sister relationship of Ardissonea, Climacosphenia, and Toxarium within the bipolar centric diatoms (Bacillariophyta, Mediophyceae), and cladistic analyses confirm that extremely elongated shape has arisen twice in the diatoms Journal of Phycology, 44: 1340-1348).
12. Medlin, LK Submitted. Coalescent models explain deep diatom divergences and argue for acceptance of paraphyletic taxa and for a revised classification for araphid diatoms. Nova Hedwigia, Under review.
13. Schmid A-MM (2001) Value of pyrenoids in the systematics of the diatoms: their morphology and ultrastructure. In: Proceedings of the 16th International Diatom. Amvrosiou Press Athens: 1–31.
14. Medlin LK, Jung I, Bahulikar R, Mendgen K, Kroth P et al. (2008a) Evolution of the Diatoms VI. Assessment of the new genera in the araphids using molecular data. Nova Hedwigia Beihefte 133: 81-100.
15. Alverson AJ, Kang SH, Theriot EC (2006b) Cell wall morphology and systematic importance of Thalassiosira ritscheri Hustedt Hasle with a description of Shinodiscus gen. nov. Diatom Research 21: 251-262.
16. Theriot EC, Cannone JJ, Gutell RR, Alverson AJ (2009) The limits of nuclear encoded SSU rDNA for resolving the diatom phylogeny. Eur J Phycol 44: 277-290.
17. Theriot EC, Ashworth M, Ruck E, Nakov T, Jansen RK (2010) A preliminary multigene phylogeny of the diatoms. Plant Ecology and Evolution 143: 278–296.
18. Ashworth A, Ruck E, Lobban C, Romanovicz R, Theriot E (2012) A revision of the genus Cyclophora and description of Astrosyne gen. nov. Bacillariophyta, two genera with the pyrenoids contained within pseudosepta. Phycologia 51: 684–699.
19. Ashworth MP, Nako T, Theriot EC (2013) Revisiting Ross and Sims (1971): toward a molecular phylogeny of the Biddulphiaceae and Eupodiscaceae (Bacillariophyceae). Journal of Phycology 49:1207–1222.
20. Simonsen R (1972) Ideas for a more natural system of the centric diatoms. Nova Hedwigia Beihefte 39: 37-54.
21. Simonsen R (1979) The diatom system: ideas on phylogeny. Bacillaria 2: 9-71.
22. Round FE, Crawford RM, Mann DG (1990) The Diatoms Biology and Morphology of the Genera. Cambridge University Press Cambridge.
23. Stosch HA von (1958). Kann die oogame Araphidee Rhabdonema adriaticum as Bindeglei zwischen den beiden grossen Diatomeengruppen angesehen warden? Berichten der Deutschen Botanischen Gellschaft 71: 241-249.
24. Magne-Simon M-F (1962). L'auxosporulation chez une Tabellariacée marine, Grammatophora marina (Lyngb.) Kütz (Diatomée). Cahiers des Biologie Marine 3: 79-89.
25. Bruder K, Medlin LK (2007) Molecular assessment of phylogenetic relationships in selected species/genera in the naviculoid diatoms Bacillariophyta I. The genus Placoneis. Nova Hedwigia 85: 331-352.
26. Bruder K, Medlin LK (2008a). Molecular assessment of phylogenetic relationships in selected species/genera in the naviculoid diatoms Bacillariophyta. II. The genus Hippondonta. Diatom Research, 23: 283-329.
27. Bruder K, Medlin LK (2008b) Molecular assessment of phylogenetic relationships in selected species/genera in the naviculoid diatoms Bacillariophyta III. Selected genera and families. Diatom Research 23: 331-347.
28. Bruder K, Sato S, Medlin LK (2008). Molecular assessment of phylogenetic relationships in selected species/genera in the naviculoid diatoms Bacillariophyta. IV. The genera Pinnularia and Caloneis. Diatom, 24: 8-24.
29. Medlin LK, Yang I, Sato S (2011) Evolution of the Diatoms VII. Four gene phylogeny assesses the validity of selected araphid genera. Nova Hedwigia, Beiheft 141: 505–514.
30. Mock T, Medlin LK (2012) Genomics and Genetics of Diatoms. In Piganeau G. (Ed.) Genomic Insights into the Biology of Algae: 245–284.
31. Bowler C, Allen AE, Badger JH, Grimwood J, Jabbari K et al. (2008) The Phaeodactylum genome reveals the evolutionary history of diatom genomes. Nature 456: 239-244.
32. Medlin LK (2010) Pursuit of a natural classification of diatoms: an incorrect comparison of published data. European Journal of Phycology 45: 155-166.
33. Rimet F, Kermarrec L, Bouchez A, Hoffmann L, Ector L et al. (2011) Molecular phylogeny of the family Bacillariaceae based on 18S rDNA sequences: focus on freshwater Nitzschia of the Lanceolatae section. Diatom Research 26: 1-20.
34. Medlin LK, Kooistra WHCF, Potter D, Saunders GW, Andersen RA (1997) Phylogenetic relationships of the 'golden algae' haptophytes, heterokonts, chrysophytes and their plastids. Plant Systematics and Evolution Supplement 11: 187-210.
35. Kuhn S, Medlin LK, Eller G (2004). Phylogenetic position of the parasitoid nanoflagellate, Pirsonia, inferred from nuclear-encoded small subunit ribosomal DNA and a description of Pseudopirsonia nov. gen. and Pseudopirsonia mucosa Drebes comb. nov. Protist 155: 143-156.
36. John U, Fensome RA, Medlin LK (2003) The application of a molecular clock based on molecular sequences and the fossil record to explain biogeographic distributions within the Alexandrium tamarense "species complex" (Dinophyceae). Mol Biol Evol 20: 1015-1027.
37. Not F, Valentin K, Romari K, Lovejoy C, Massana R et al. (2007) Picobiliphytes: a marine picoplanktonic algal group with unknown affinities to other eukaryotes. Science 315: 253-255.
38. Nylander JA, Wilgenbusch JC, Warren DL, Swofford DL (2008) AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics 24: 581-583.
39. Maddison WP, Maddison, DR (2005) MacClade. Sinauer Associates, Inc.
40. Alverson AJ, Theriot EC (2003) Taxon sampling and inferences about diatom phylogeny. Journal of Phycology, Supplement. 39: 36.
41. Kooistra WHCF, De Stefano M, Mann DG, Salma N, Medlin LK (2003b). The

phylogenetic position of Toxarium, a pennate-like lineage within centric diatoms Bacillariophyceae. Journal of Phycology, 39: 185-97.

42. Alverson AJ, Cannone JJ, Gutell RR, Theriot EC (2006a) The evolution of elongate shape in diatoms. Journal of Phycology 42: 655-68.

43. Sims PA, Mann DG, Medlin LK (2006) Evolution of the Diatoms: Insights from fossil biological and molecular data. Phycologia 45: 361-402.

44. Medlin LK, Kaczmarska I (2009) Correctly assigning original discoveries to original authors. Mol Phylogenet Evol 50: 407-408.

45. Ehara M, Inagaki Y, Watanabe KI, Ohama T (2000) Phylogenetic analysis of diatom coxI genes and implications of a fluctuating GC content on mitochondrial genetic code evolution. Curr Genet 37: 29-33.

46. Evans KM, Wortley AH, Mann DG (2007) An assessment of potential diatom “barcode” genes (cox1, rbcL, 18S and ITS rDNA) and their effectiveness in determining relationships in Sellaphora (Bacillariophyta). Protist 158: 349-364.

47. Kaczmarska I, Rei C, Moniz M (2007) Diatom taxonomy: morphology, molecules and barcodes. In: Proceedings of the 1st Central-European Diatom meeting. Berlin-Dahlem, Germany: 69-72.

48. Moniz MB, Kaczmarska I (2009) Barcoding diatoms: Is there a good marker? Mol Ecol Resour 9 Suppl s1: 65-74.

49. Moniz MB, Kaczmarska I (2010) Barcoding of diatoms: nuclear encoded ITS revisited. Protist 161: 7-34.

50. Guillou L, Chretiennot-Dine MJ, Medlin LK, Claustre H, Loiseaux-de Goer S et al. (1999) Bolidomonas: a new genus with two species belonging to a new algal class, the Bolidophyceae (Heterokonta). Journal of Phycology 35: 368–381.

51. Sorhannus U (2004) Diatom phylogenetics inferred based on direct optimization of nuclear-encoded SSU rRNA sequences. Cladistics 20: 487-497.

52. Cavalier-Smith T, Chao EE (2006) Phylogeny and megasystematics of phagotrophic heterokonts (kingdom Chromista). J Mol Evol 62: 388-420.

53. Choi HG, Joo HM, Jung W, Hong SS, Kang JS et al. (2008) Morphology and phylogenetic relationships of some psychrophilic polar diatoms (Bacillariophyta). Nova Hedwigia Beihefte 133: 7-30.

54. Sorhannus U (2007). A nuclear-encoded small-subunit ribosomal RNA timescale for diatom evolution. Marine Micropaleontology 65: 1-12.

55. Lee MA, Faria DG, Han MS, Lee J (2013) Evaluation of nuclear ribosomal RNA and chloroplast gene markers for the DNA taxonomy of centric diatoms. Biochemical Systematics and Ecology 50: 163-174.

56. Edgar SM, Theriot E (2004) Phylogeny of Aulacoseira (Bacillariophyta) based on morphology and molecules. Journal of Phycology 40: 772-788.

57. Verbruggen H, Theriot EC (2008) Building trees of algae: some advances in phylogenetic and evolutionary analysis. European Journal of Phycology 43: 229-252.

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

59. Kooistra WH, Medlin LK (1996) Evolution of the diatoms (Bacillariophyta). IV. A reconstruction of their age from small subunit rRNA coding regions and the fossil record. Mol Phylogenet Evol 6: 391-407.

60. Schutt F (1896) Bacillariales. In Die natürlichen Planzenfamilien. 1: 31-153.

61. Van de Peer Y, Van der Auwera G, De Wachter R (1996) The evolution of stramenopiles and alveolates as derived by “substitution rate calibration” of small ribosomal subunit RNA. J Mol Evol 42: 201-210.

62. Mann DG, Simpson GE, Sluiman HJ, Muller M (2001). rbcL gene tree of diatoms: a second large data-set for phylogenetic reconstruction. Phycologia, 40: 1-2.

63. Fox MG, Sorhannus UM (2003) RpoA: a useful gene for phylogenetic analysis in diatoms. J Eukaryot Microbiol 50: 471-475.

64. Amato A, Kooistra WH, Ghiron JH, Mann DG, Pröschold T, et al. (2007) Reproductive isolation among sympatric cryptic species in marine diatoms. Protist 158: 193-207.

65. Piganeau G, Eyre-Walker A, Grimsley N, Moreau H (2012) How and why DNA barcodes underestimate the diversity of microbial eukaryotes. PLoS ONE 7: 10.

66. Kaczmarska I, Beaton M, Benoit AC, Medlin LK (2006) Molecular phylogeny of selected members of the Order Thalassiosirales (Bacillariophyta) and evolution of the fultoportula. Journal of Phycology 42: 121-138.

67. Alverson AJ (2014) Timing marine-freshwater transitions in the diatom order Thalassiosirales. Paleobiology 40: 91-101.

68. Kaczmarsk I, Ehrman JM, Bates SS (2001) A review of auxospore structure, ontogeny, and diatom phylogeny. In: Proceedings of the 16th International Diatom Symposium University of Athens Press, Athens, Greece: 153-168.

69. van den Hoek C, Mann DG, Jahns HM (1995) Algae. An Introduction to Phycology. Cambridge University Press Cambridge.

70. Karol KG, McCourt RM, Cimino MT, Delwiche CF (2001) The closest living relatives of land plants. Science 294: 2351-2353.

71. Rampen SW, Schouten S, Elda Panoto F, Brink M, Andersen RA et al. (2010) Phylogenetic position of Attheya longicornis and Attheya septentrionalis (Bacillariophyta). Journal of Phycology 45: 444–453.

72. Sato S, Beakes G, Idei M, Nagumo T, Mann DG (2011) Novel sex cells and evidence for sex pheromones in diatoms. PLoS One 6: e26923.

73. Williams DM, Kociolek JP (2007) The rejection of paraphyletic taxa. European Journal of Phycology 42: 313-319.

74. Theriot E, Stoermer E (1984) Principal component analysis of Stephanodiscus: observations on two species from the Stephanodiscus niagarae complex. Bacillaria 7: 37-58.

75. Quijano-Scheggia SI, Garces E, Lundholm N, Moestrup O, Camp J (2009) Morphology physiology molecular phylogeny and sexual compatibility of the cryptic Pseudo-nitzschia delicatissima complex (Bacillariophyta) including the description of P. arenysensis sp. nov. Phycologia 48: 492-509.

76. Sarno D, Kooistra WHCF, Medlin LK, Percopo I, Zingone A (2005) Pseudocryptic diversity in the genus Skeletonema (Bacillariophyceae): Skeletonema costatum (Bacillariophyceae) consists of several genetically and morphologically distinct species. Journal of Phycology 41: 151-176.

77. Stosch HA von (1982) On auxospore envelopes in diatoms. Bacillaria 5: 127-156.

78. Alverson AJ, Jansen RK, Theriot EC (2007) Bridging the Rubicon: phylogenetic analysis reveals repeated colonizations of marine and fresh waters by thalassiosiroid diatoms. Mol Phylogenet Evol 45: 193-210.

79. Medlin LK, Sato S (2009) The biological reality of the core and basal groups of araphid diatoms. Diatom Research 24: 503-508.

80. Kaczmarska I, Medlin LK (2009) Reply to Theriot (2008; 44: 821-833: comments on Kaczmarska et al. (2006) ‘Application of phylogenetic principles’. Journal of Phycology 45: 987-994.