Copyright
by
Teofil Nakov
2014

The Dissertation Committee for Teofil Nakov
certifies that this is the approved version of the following dissertation:

Studies of phylogenetic relationships and evolution of functional traits in diatoms

Committee:

Edward C. Theriot, Supervisor

Robert K. Jansen

Beryl B. Simpson

John W. La Claire II

David C. Cannatella

Studies of phylogenetic relationships and evolution of functional traits in diatoms

by

Teofil Nakov, B.S.

DISSERTATION
Presented to the Faculty of the Graduate School of
The University of Texas at Austin
in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN
May 2014

Dedicated to my parents, Todorka and Vane Nakovi, and my inspirational brother Novica Nakov.

Acknowledgments

I am indebted to my advisor, Ed Theriot, for extending the freedom to make my own mistakes. In retrospect, it seems I made many, but the hope is I learnt from each. I also thank the rest of my committee, Beryl Simpson, Bob Jansen, John La Claire II, and David Cannatella for their guidance throughout my time in graduate school.

My longstanding lab mates, Elizabeth Ruck and Matt Ashworth endured many discussions and rants, were collaborators on various projects and great companionions thoughout my time in the Theriot lab. Sandra Pelc and Ginnie Morrison (a.k.a "the Linder twins") made graduate school and living in Austin all the more eventfull. Mariska Brady, Anna Yu, the Jansen, and Simpson lab members offered a thought-provoking environment on many Friday seminars in Biolabs and beyond (HiTW). The Competitive Excluders soccer team were a relaxing outlet during weeks of lab and computer work. Elena Jovanovska, Alexandar Pavlov, Zlatko Levkov, and Svetislav Krstic, my collaborators from the Institute of Biology, Faculty of Natural Sciences in Skopje, Macedonia, kept on reminding me of the beauty of simply looking at diatoms. I thank Zlatko Levkov in particular, for introducing me to the world of algae and his guidance through my early days in science.

Andy Alverson (University of Arkansas) inflenced my research from a distance and was a collaborator on Chapter 3 of this Dissertation. Sarah Spaulding (USGS), Marina Potapova (ANSP), and Yuri Galachyants & Co. (LIN, Irkutsk) collaborated for Chapters 1 and 2. Much of this work would not have been possible without Andy's, Sarah's, Marina's or Yuri's kind contributions.

To dear my friends, Daniela Nedelkova, Vlado Božinovski, Igor Šterbinski, Maru Fuentes Hernandez Šterbinski, Goran Velkovski (known colectivelly as the Hromagnons) and Bojan Šutinoski, in Macedonia and wherever you globetrot, thank you for being a positive influence throughout my life. Thank you also to Peter Ruck for his boundless energy and creativity in a friendship outside the world of science. Sofía Rodríguez Brenes, the epitome of peacefulness and compassion, helped in many subtle ways.

I end this acknowledgment with my parents, Todorka and Vane, and my brother, Novica. They have been a source of love and encouragement throughout graduate school and before. To Novica, especially: thank you. You continue to inspire me.

Studies of phylogenetic relationships and evolution of functional traits in diatoms

Publication No. ____________________

Teofil Nakov, Ph.D.
The University of Texas at Austin, 2014

Supervisor: Edward C. Theriot

The research presented here deals with inferring phylogenetic trees and their use to study the evolution of functional traits in diatoms (Heterokontophyta: Bacillariophyceae). Two chapters are concerned with the phylogeny of a mainly freshwater group, the Cymbellales, with a convoluted taxonomic history and classification. I generated a multi-gene dataset to test the monophyly of the Cymbellales and reconstruct the relationships within the order. The molecular data were equivocal with respect to the monophyly of the Cymbellales, especially when taking into account some problematic taxa like *Cocconeis* and *Rhoicosphenia*. Aside from the problem with their monophyly, my work shows that the current genus- and family-level classification of the Cymbellales is unnatural, arguing for the need of nearly wholesale re-classification of the group. The two following chapters make use of phylogenetic trees to model the evolution of functional traits. I explored the evolution of cell size across the salinity gradient finding that the opposing selective forces exerted by marine and fresh waters select for different

optimal cell sizes – larger in the oceans and smaller in lakes and rivers. Thereafter, I modelled the evolutionary histories of habitat preference (planktonic-benthic) and growth form (solitary-colonial) across the diatoms. These traits exhibit markedly different evolutionary histories. Habitat preference evolves slowly, is conserved at the level of large clades, and its evolution is generally uniform across the tree. Growth form, on the other hand, has a more dynamic evolutionary history with frequent transitions between the solitary and colonial growth forms and rates of evolution that vary through time. I hope that these empirical studies represent an incremental advancement to the understanding of the evolution diatom species and functional diversity.

Table of Contents

Acknowledgments v

Abstract vii

List of Tables xii

List of Figures xiv

Chapter 1. Testing the monophyly of the Cymbellales (Bacillariophyceae): The phylogenetic placement of *Rhoicosphenia* in relation to cymbelloid, gomphonemoid, and monoraphid diatoms 1

- 1.1 Abstract . . . 1
- 1.2 Introduction . . . 2
- 1.3 Material and Methods . . . 5
  - 1.3.1 Cultures and DNA methods . . . 5
  - 1.3.2 Dataset assembly and sequence alignment . . . 6
  - 1.3.3 Phylogenetic analyses . . . 7
  - 1.3.4 Hypothesis testing . . . 8
  - 1.3.5 Gene and partition-specific tree topologies and branch lengths 11
  - 1.3.6 Long branch attraction . . . 11
  - 1.3.7 Site-heterogeneous models of evolution . . . 12
- 1.4 Results . . . 14
  - 1.4.1 Phylogenetic relationships . . . 14
  - 1.4.2 Gene and partition-specific branch lengths and topologies . 16
  - 1.4.3 Comparison of topological hypotheses . . . 21
  - 1.4.4 Long branch attraction . . . 25
  - 1.4.5 Site-heterogeneous models of evolution . . . 26
- 1.5 Discussion . . . 31
- 1.6 Acknowledgements . . . 35

Chapter 2. Molecular phylogeny of the Cymbellales (Bacillariophyceae) with a comparison of models for accommodating rate-variation across sites 36

2.1 Abstract 36
2.2 Introduction 37
2.3 Material and Methods 40
  2.3.1 Taxon sampling 40
  2.3.2 Cell cultures and DNA methods 40
  2.3.3 Sequence alignment 41
  2.3.4 Estimation of tree-independent evolutionary rates 43
  2.3.5 Phylogeny inference 44
2.4 Results 47
  2.4.1 Tree-independent evolutionary rates 47
  2.4.2 The credible set of evolutionary models 47
  2.4.3 Phylogenies 50
  2.4.4 The likelihood of family- and genus-level classification . . . 55
2.5 Discussion 56
  2.5.1 Rates, models, and partitions 56
  2.5.2 Phylogenetic relationships 60
    2.5.2.1 Monophyly of the Cymbellales and early evolution of the group 61
    2.5.2.2 Monophyly of and relationships within the Cymbellaceae and Gomphonemataceae 62
    2.5.2.3 *Geissleria* and *Placoneis* 65
  2.5.3 Concluding remarks 65
2.6 Acknowledgements 68

Chapter 3. Using phylogeny to model cell size evolution in marine and freshwater diatoms 69

3.1 Abstract 69
3.2 Introduction 70
3.3 Material and Methods 72
  3.3.1 Cell volume 72
  3.3.2 Phylogeny 74

3.3.3 Tests for phylogenetic signal and size differences between marine and freshwater taxa 75

3.3.4 Adaptive models for cell size evolution in marine vs. freshwater environments 77

3.4 Results 80

3.4.1 Cell size in marine vs. freshwater Thalassiosirales 80

3.4.2 Ornstein-Uhlenbeck models for cell size evolution in marine vs. freshwater Thalassiosirales 81

3.5 Discussion 83

3.6 Acknowledgements 90

Chapter 4. Evolution of habitat preference and growth form across diatoms **91**

4.1 Abstract 91

4.2 Introduction 92

4.3 Material and Methods 95

4.3.1 Trait data and phylogenetic trees 95

4.3.2 Individual traits 96

4.3.3 Combined traits 99

4.4 Results 100

4.4.1 Trees and trait distribution 100

4.4.2 Habitat preference 100

4.4.3 Growth form 103

4.4.4 Combined traits 106

4.5 Discussion 109

4.5.1 Traversals between habitats are rare, but asymmetric in lineages with colonial growth form 109

4.5.2 Variable pace of growth form evolution across the diatom phylogeny 113

4.5.3 Concluding remarks 115

4.6 Acknowledgements 117

Bibliography **118**

Vita **135**

List of Tables

1.1 Phylogenetic placement of *Rhoicosphenia* and *Cocconeis* and corresponding clade support values. 18

1.2 Ratio of the branch lengths leading to *Cocconeis* and *Rhoicosphenia* against the tree average. 21

1.3 Maximum likelihood topology tests for the phylogenetic position of *Rhoicosphenia* in relation to the cymbelloids and monoraphid diatoms. 23

1.4 Bayes factor topology tests for the phylogenetic position of *Rhoicosphenia* in relation to the cymbelloids and monoraphid diatoms. 24

1.5 Posterior frequency topology tests for the phylogenetic position of *Rhoicosphenia* in relation to the cymbelloids and monoraphid diatoms. 25

1.6 Posterior predictive and cross-validation tests for the adequacy and goodness-of-fit of the GTR, CAT and QMM models. 31

2.1 The number of rate parameters (substitution types) in the credible models for the data partitioned by-gene-by-codon and by-rate. . . 50

2.2 Posterior probability of the monophyly of various taxon combinations under different models. 56

3.1 Comparison standard and phylogenetic one-way ANOVA for the size difference between marine and freshwater Thalassiosirales. . 81

3.2 Model selection for the evolution of maximum and minimum cell volume in the Thalassiosirales. 83

3.3 Maximum likelihood estimates of cell size optima (θ), strength of selection (α), and rate of stochastic motion (σ^2) for freshwater and marine species of Thalassiosirales. 88

4.1 Time-homogeneous and time-heterogeneous models for the evolution of habitat preference across diatoms. 102

4.2 Time-homogeneous and time-heterogeneous models for the evolution of growth form across diatoms. 105

4.3 Two trait models for the evolution of the combined growth form and habitat preference phenotype across diatoms. 109

List of Figures

1.1 Placement of *Rhoicosphenia* among cymbelloid and monoraphid diatoms. 15
1.2 Maximum likelihood gene trees 17
1.3 Bayesian inference gene trees 19
1.4 Bayesian inference with partition-specific branch lengths and/or topologies 22
1.5 Long branch extraction 27
1.6 Random sequence attraction 28
1.7 Phylogenetic mixture models 29

2.1 Distribution of tree-independent evolutionary rates by gene. 49
2.2 Phylogenetic relationships within the Cymbellales. 52
2.3 Topology and clade support of the CCDE clade under different models for accommodating rate variation. 53
2.4 Topology and clade support of the GG clade under different models for accommodating rate variation. 54
2.5 Light and scanning electron microscopy of *Geissleria decussis*. . . 66

3.1 Phylogenetic relationships within the diatom order Thalassiosirales 76

3.2 Cell volume evolution in the Thalassiosirales through time and cell size optima for marine and freshwater lineages 85

4.1 Distribution of habitat preference and growth form across the diatom phylogeny 104

4.2 Maximum parsimony, maximum likelihood and Bayesian reconstruction of states shifts 107

4.3 Transition probabilities for shifts in growth form under time-homogeneous and time-heterogeneous models 108

4.4 Evolution of the combined habitat + growth form phenotypes. . . 110

Chapter 1

Testing the monophyly of the Cymbellales (Bacillariophyceae): The phylogenetic placement of *Rhoicosphenia* in relation to cymbelloid, gomphonemoid, and monoraphid diatoms

1.1 Abstract

I performed phylogenetic analyses focused on the placement of *Rhoicosphenia* in relation to the Cymbellales and monoraphid diatoms. In the most strongly supported trees, *Rhoicosphenia* was recovered as sister to *Cocconeis* thus questioning the monophyly of the Cymbellales as currently circumscribed. Even though *Rhoicosphenia* + *Cocconeis* was the favored arrangement, topological hypothesis testing under maximum likelihood showed that an alternative where *Rhoicosphenia* is sister to, or within, the cymbelloid diatoms cannot be rejected. In the Bayesian framework, hypothesis testing argued strongly in favor of the *Rhoicosphenia* + *Cocconeis* hypothesis and against the monophyly of the Cymbellales. Analyses of gene– and partition–specific branch lengths and topologies showed that the *rbcL* data, and especially third codon positions, are driving this grouping. When trees were estimated in the absence of either *Cocconeis* or *Rhoicosphenia* the remaining long branch attached to a distant position in the tree, providing some support for long-branch attraction (LBA). However, neither *Cocconeis* nor *Rhoicosphenia* were

attracted to an introduced random sequence appreciably more than other taxa in the tree. These results, combined with support for the *Cocconeis* + *Rhoicosphenia* clade across analyses of $\leq$81%, suggested that this association is likely due to stochastic rather than systematic errors (i.e. LBA). Site-heterogeneous phylogenetic models, appropriate for cases when LBA is suspected, separated *Cocconeis* and *Rhoicosphenia*, but failed to identify supported alternative positions for these taxa which were placed in a polytomy. The phylogenetic position of *Rhoicosphenia* and the monophyly of the Cymbellales, therefore, remained ambiguous highlighting the need for additional data and improved taxon sampling.[1]

1.2 Introduction

In this study I set out to clarify the phylogenetic placement of the diatom genus *Rhoicosphenia* Grunow (Bacillariophycidae, raphid pennates) with respect to cymbelloid and monoraphid diatoms and by extension the monophyly of the Cymbellales D.G. Mann in Round et al. (1990). Historically, *Rhoicosphenia* was combined with monoraphid diatoms (e.g. Patrick & Reimer 1966). This grouping was suggested because of *Rhoicosphenia*'s flexed frustules and reduced raphe system on the convex valve – characters shared with representatives of the monoraphid group

---

[1] This chapter was submitted for peer review as Nakov, Potapova, and Theriot: Testing the monophyly of the Cymbellales (Bacillariophyceae): "The phylogenetic placement of *Rhoicosphenia* in relation to cymbelloid, gomphonemoid, and monoraphid diatoms". Author contribution: Marina Potapova kindly shared unpublished sequences for monoraphid diatoms; Edward Theriot, supervisor.

(e.g. *Achnanthidium* Kützing; summarized in Kociolek & Stoermer 1986). After extensive study of the natural history and morphology (Mann, 1982b,a, 1984), D.G. Mann separated *Rhoicosphenia* from the monoraphid diatoms into its own family Rhoicospheniaceae D.G. Mann (Mann, 1984). Kociolek and Stoermer's (1986) finding that *Rhoicosphenia* is more closely allied to *Gomphonema* Ehrenberg than "Achnanthes" or *Cocconeis* Ehrenberg corroborated Mann's (1984) view. Not long after, the Rhoicospheniaceae were combined within the Cymbellales together with the Anomoeoneidaceae, Cymbellaceae and Gomphonemataceae (Round et al. 1990).

Although the placement of *Rhoicosphenia* in the Cymbellales was proposed based on corroborating evidence from morphology and phylogenetic trees (Mann, 1982b,a, 1984; Kociolek & Stoermer, 1986), the few subsequent morphological phylogenetic analyses with adequate taxon sampling have questioned the monophyly of the order. Jones and coworkers found *Rhoicosphenia* outside the remaining Cymbellales, with the latter recovered as sister to the Lyrellales (Jones et al., 2005), and Cox & Williams (2006) recovered *Rhoicosphenia* as part of a polytomy with a number of naviculoid genera. These results hinted that Mann's (1982a) assessment about the monoraphids that "*Rhoicosphenia* is no more like the monoraphids than many naviculoids" might be appropriate for *Rhoicosphenia*'s alliance with the Cymbellales as well.

The problem with the phylogenetic position of *Rhoicosphenia* seems to be exacerbated by the lack of a synapomorphy for the Cymbellales. The mode of sexual reproduction has featured heavily in the interpretation of the evolutionary history of the order (Kociolek & Stoermer, 1988; Mann & Stickle, 1995). Where known, cymbelloid diatoms undergo physiological anisogamy. *Rhoicosphenia*, on the other hand, is isogamous (Mann 1982b; Mann and Stickle 1995). Identifying a synapomorphy from the pool of characters associated with frustule morphology has proven difficult. Indeed, Cox and Williams (2006) coded 25 frustule-related characters for a diversity of raphid diatoms including 14 cymbelloids and found the Cymbellales polyphyletic. It seems that the only character that might be shared and derived among the four families of the Cymbellales to the exclusion of the monoraphids and the Lyrellales is the morphology of the chloroplast. Taxa in the Cymbellales possess a single chloroplast with its centre against the girdle of the secondary valve side and lobes extending beneath both valves (Cox and Williams 2006; Kociolek and Stoermer 1988; Mann and Stickle 1995). However, even chloroplast morphology varies. In *Placoneis*, the chloroplast centre is in the middle of the cell away from the copulae (Cox, 1987; Mann & Stickle, 1995). As it stands, there is no formal synapomorphy for the Cymbellales and phylogenetic studies have suggested nonmonophyly of the order (Bruder and Medlin 2007; Cox and Williams 2006; Jones et al. 2005).

In this study, I evaluated competing hypotheses about the phylogenetic placement of *Rhoicosphenia* with respect to cymbelloid and monoraphid diatoms. To this end, I sequenced the nuclear encoded small ribosomal subunit rDNA (SSU) and the chloroplast encoded ribulose bisphosphate carboxy-oxigenase large subunit (*rbcL*) for a diverse set of taxa including representatives from all four families of the Cymbellales and major genera of monoraphids. I was concerned with one particular question: Is the placement of *Rhoicosphenia* statistically significantly better outside the Cymbellales (Jones et al. 2005; Cox and Williams 2005) than within the Cymbellales (Kociolek and Stoermer 1986; Round et al. 1990). To address this, I employed an array of analyses including maximum likelihood (ML) and Bayesian phylogeny inference (BI), hypothesis testing, long-branch attraction tests, and phylogenetic mixture models combined with analyses of goodness-of-fit.

1.3 Material and Methods

1.3.1 Cultures and DNA methods

Cultures of the Cymbellales were grown in biphasic soil + water media (Czarnecki, 1987) at room temperature on a windowsill. *Rhoicosphenia abbreviata* (Agardh) Lange-Bertalot was grown in COMBO media (Kilham et al., 1998) at room temperature. DNA was extracted using the Power-soil kit (MoBio) following manufacturers instructions. For some species I was unable to establish monoclonal cultures so DNA was extracted from single cells using the Chelex-100 protocol (Richlen & Barber, 2005). Primers used, polymerase chain reaction (PCR) conditions and sequencing followed Ruck & Theriot (2011). Monoraphid diatoms were grown in Diatom Medium (Cohn et al., 2003) at 15 °C on a 14:10 light-dark cycle. DNA was extracted according to manufacturer's instructions with the DNeasy Plant MiniKit (Qiagen, Hilde, Germany). Primers and PCR conditions for SSU gene were those recommended by Elwood et al. (1985) and Medlin et al. (1988); for *rbcL* those by (Daugbjerg & Andersen, 1997) and Jones et al. (2005). Newly generated sequences are available from Genbank accession numbers: XXXX - XXXX.

1.3.2 Dataset assembly and sequence alignment

I started with the alignment of Ruck and Theriot (2011) which has relatively good sampling of raphid pennate diatoms including representatives of all but two families of raphid pennates. I enriched this alignment with new data for representatives from the monoraphids (15 acessions) and the Cymbellales (19 acessions) including*Rhoicosphenia* (Supplementary Table 1). Since data for the photosystem II cp43 protein (*psbC*; used in Ruck and Theriot [2011]) are currently unavailable for *Rhoicosphenia*, this gene was removed from the alignment. Publicly available SSU or *rbcL* data from members of the Lyrellales (AY571756, AY571755, AY571747, AJ535149, AJ544659, AY571757) and *Dickieia* Berkeley ex Kützing (AY485462) were added because these taxa have been shown to be phylogenetically close to the Cymbellales (Bruder & Medlin, 2007; Jones et al., 2005). I aligned *rbcL* by hand. For the alignment of SSU, I used a covariance model secondary structure alignment (Cannone et al., 2002; Nawrocki, 2009; Nawrocki et al., 2009; Theriot et al., 2009). The SSU alignment was masked to exclude columns for which reliable covarying nucleotides were difficult to locate (Nawrocki 2009).

I concatenated *rbcL* and SSU into a single matrix (10% missing data) and partitioned it by codon positions within *rbcL* and stem vs. loop regions within SSU. Preliminary analyses of this all-raphid supported a clade composed of the Berkeleyacae, Cymbellales, monoraphids, and Lyrellales (Bayesian posterior probability (BPP) =0.95). Therefore, I removed all taxa outside this clade retaining a sample of 46 accessions relevant to the position of *Rhoicosphenia* and monophyly of the Cymbellales.

1.3.3 Phylogenetic analyses

For maximum likelihood (ML) I used serial and threaded executables of RAxML v.7.4.2 (Stamatakis 2006; downloaded February 13 2013). For each ML analysis I performed 1008 optimizations starting from a parsimony tree and GTR+Γ+I model of evolution for each partition. The tree with highest log likelihood from 1008 optimizations was taken as the "best" tree and used for subsequent analyses and discussion. Clade support was assessed from 10^3 nonparametric bootstrap replicates under GTR+Γ+I using the rapid bootstrap algorithm (Stamatakis et al., 2008).

For Bayesian inference I used parallel MrBayes v.3.2 (Altekar et al., 2004; Ronquist et al., 2012). Standard Markov chain Monte Carlo (MCMC) simulations were used to obtain posterior distributions of trees and clade support values. For each analysis I ran 6 MCMC runs with one cold and one heat chain each for 2×10^7 retaining one out of 1000 samples. The model had Γ−distributed rate variation across sites and a proportion of invariable sites (I). I did not fix the substitution matrix. Instead, I performed reversible-jump MCMC that samples, in proportion to their posterior probability, any model from the GTR (reversible) family (Huelsenbeck et al., 2004). Substitution model uncertainty was therefore incorporated in the analysis and the resulting topologies, branch lengths, and parameter estimates were averaged over the credible set of substitution models. Convergence and stationarity of MCMC samplers was first assessed through the output from MrBayes. I ensured that for all analyses the average standard deviation of split frequencies at the end of the run was <0.05 and the potential scale reduction factor was <5% different than 1.0. Posterior distributions of parameters were also visually checked in Tracer (Rambaut & Drummond, 2007). Topological convergence and stationarity was assessed using the "compare", "cumulative" and "var" routines in AWTY (Nylander et al., 2008). Specifically, I ensured that the post-burnin posterior distributions between runs had highly correlated posterior probabilities of splits ("compare"), that the posterior probabilities of splits along 10 increments had a linear trend without major fluctuations ("cumulative), and that the topological differences within and among MCMC runs were similar (var). MCMC chains that failed the convergence diagnostics were removed before summarizing the posterior.

1.3.4 Hypothesis testing

I used three approaches to phylogenetic topology hypothesis testing. First I performed maximum likelihood searches with an unconstrained topology and a topology constrained to monophyly of the Cymbellales sensu Round et al. (1990). I created a hard constraint forcing *Rhoicosphenia* + the remaining cymbelloids to be monophyletic leaving all but one bipartition to be estimated from the data. The best trees among 1008 optimizations for the unconstrained and constrained searches were compared using the approximately unbiased test (AU) and Shimodaira-Hasegawa test (SH) calculated with Consel v.0.2 (Shimodaira, 2002). I used the default multiscale bootstrap settings of Consel (10 sets of 10^4 replicates) and confirmed that the resulting p-values have adequate standard errors (Shimodaira & Hasegawa, 2001).

Second, I performed Bayesian topological hypothesis testing using Bayes factors (BF). To obtain unbiased estimates of the marginal likelihood I used stepping-stone MCMC (ssMCMC) sampling (Fan et al., 2011; Xie et al., 2011) as implemented in MrBayes v.3.2 (Ronquist et al., 2012). The following settings were used for all analyses described bellow. I divided the ssMCMC runs into 50 steps and ran the samplers for a total of 2×10^7 generations retaining every 100th sample. The initial burnin was one step long (392100 generations) and in each subsequent step the first 980 samples were discarded as burnin. Each step of the ssMCMC routine was 294100 generations long and parameter estimates from each step were based on a sample size of 2941 (after thinning and burnin). I first replicated the ML analyses by performing a simulation where (i) all topologies had equal prior probability (unconstrained) and (ii) enforcing a hard constraint for the monophyly of *Rhoicosphenia* + the cymbelloid diatoms (Hypothesis H0). In an additional analysis I enforced a hard constraint for the monophyly of *Rhoicosphenia* + *Cocconeis* (Hypothesis H1). Bayesian topological hypothesis tests as described above, contrasting an unconstrained and hard-constraint analyses, tend to be biased towards supporting the constrained topology even when the data were better explained by the alternative, unconstrained topology (Bergsten et al., 2013). To accomodate for this bias I performed additional analyses in which clades whose monophyly was relatively well supported in the unconstrained ML and standard MCMC analyses were constrained to monophyly using a partial constraint. With a partial constraint, I fulfilled the criterion that a set of taxa is monophyletic to the exclusion of another set of taxa. Taxa left out of any partial constraint, however, were allowed to "float" in and out of partial constraints. This ensured that the MCMC did not sample trees where, for example, a species of *Achnanthidium* fell inside the Cymbellales or vice versa. With respect to the floating taxa (e.g. *Rhoicosphenia*) trees were sampled irrespective of their placement but proportional to their posterior probability. I again compared three hypotheses. The first analysis enforced six partial constraints, monophyly of: the outgroup taxa (Berkeleyaceae), Lyrellales, *Planothidium*, *Lemnicola*, achnanthidoids (*Achnanthidium*, *Rossithidium*, *Pssamothidium*, and *Pauliela*), and cymbelloids (*Anomoeoneis*, *Cymbella*, *Cymbopleura*, *Didymosphenia*, *Gomphonema*, and *Gomphoneis*). *Rhoicosphenia*, *Cocconeis*, and *Dickieia* were kept out of any partial constraint and allowed to float in and out of constraints or attach anywhere along the tree. The above set up is analogous to the unconstrained analysis in that it does not alter the prior probability of trees with respect to *Rhoicosphenia* and *Cocconeis* specifically. The other two analyses kept the same setup with the addition of hard constraints for the monophyly of *Rhoicosphenia* + cymbelloids (H0) and the monophyly of *Rhoicosphenia* + *Cocconeis* (H1). Competing hypotheses were compared using BFs calculated as twice the difference in the natural logarithm of the marginal likelihood with values greater than 10 taken as strong support for the better hypothesis (Kass & Raftery, 1995).

In the third approach to topological hypothesis testing I assessed the posterior probability of a particular hypothesis by calculating the frequency of trees consistent with hypothesis H0 or H1 (as above) in the post-burnin posterior distribution of standard MCMC analyses. I performed the test for the unconstrained and the partial constraint analyses with *Rhoicosphenia*, *Cocconeis*, and *Dickieia* allowed to float.

1.3.5 Gene and partition-specific tree topologies and branch lengths

To investigate the topologies and branch lengths estimated from different portions of the data I performed (i) separate analyses for each gene (both ML and BI), (ii) analyses with joint topology but separate estimates of branch lengths for each partition (BI only) and (iii) analyses with separate branch lengths and topologies for all partitions (BI only). As before, the SSU alignment was partitioned into stem and loop positions and *rbcL* into codon positions. Thereafter, RAxML and MrBayes (standard MCMC only) analyses proceeded as described above.

1.3.6 Long branch attraction

I carried out two tests to investigate whether the recovered monophyly between *Rhoicosphenia* and *Cocconeis* could be due to long-branch attraction (LBA). First I performed long-branch extraction (LBE) (Pol & Siddall, 2001; Siddall & Whiting, 1999) where the tree is optimized from datasets in which one of the taxa suspected to attract is removed. The logic of the LBE test is, if the two taxa attract, then by removing one of them, the other will be free to remain in the same position or attract to another, perhaps distant, branch in the tree. A position distant from the original placement of the kept long-branch taxon, supports the suspicion that a branch is long enough to attract. The same procedure can be repeated for the other long-branch taxon and a similar pattern is viewed as corroborating evidence that the association between the two long-branch taxa could be due to LBA artefacts.

Second I performed a "random sequence attraction" test where a sequence in the dataset is replaced with a randomly generated sequence with the same base frequencies to assess the behavior of the branches suspected to be subject to long-branch artefacts (Sullivan & Swofford, 1997). I produced 10^3 datasets in which *Cocconeis* was replaced with a random sequence with the same nucleotide frequencies as *Cocconeis*. For each dataset I found the best tree by ML out of 12 optimizations. I recorded the number of times when *Rhoicosphenia* was found sister to the introduced random sequence. If *Rhoicosphenia* fell sister to the random sequence more than expected by chance alone then I took this as evidence that the *Rhoicosphenia* branch is long enough to attract. I repeated the test for *Cocconeis* replacing the *Rhoicosphenia* entry with a random sequence with the same nucleotide frequencies as *Rhoicosphenia*.

1.3.7 Site-heterogeneous models of evolution

I employed two models of site-specific heterogeneity available in PhyloBayes 3.3 (Lartillot & Philippe, 2004). The first model is a nonparametric approach that combines the alignment columns into categories based on the profile of equilibrium frequencies (CAT model). The second model is a mixture of GTR matrices

(QMM model). These models differ in that, under CAT, the stationary frequencies of nucleotides differ across the alignment but exchangeability rates are constant, whereas under QMM, the exchangeability rates are also allowed to differ. The rate matrix in both models was GTR coupled with a discretized Γ distribution with 4 rate categories.

In PhyloBayes I ran 3 MCMC chains for $5 - 5.5 \times 10^4$ cycles. Convergence and stationarity of the chains was assessed after discarding the first $1 - 5 \times 10^3$ cycles as burnin and thinning the posterior to every 10th sample. I ensured that in each analysis effective sample sizes for parameter estimates were >300 and discrepancies between chains, the ratio of twice the difference between means over the sum of standard deviations for a parameter, were $\leq$0.1.

I assessed the goodness-of-fit of the CAT and QMM models using posterior predictive simulation and cross-validation as implemented in PhyloBayes. For posterior predictive tests, I simulated a dataset based on the parameters from every 10th cycle from the posterior after burnin. I asked if the models adequately describe (i) the nucleotide diversity per site ("div" test in PhyloBayes) and (ii) the compositional heterogeneity ("comp" test in PhyloBayes). In these tests, a test statistic calculated from the observed data is compared to a distribution of the test statistic from the simulated datasets. Strong discrepancies between the observed and simulated test statistics indicate failure of the model to adequately describe the structure of data.

For cross-validation, I estimated parameters on a learning set (9/10 of the alignment) and used these parameters to calculate the likelihood of the test set (the remaining 1/10 of the dataset). The procedure was repeated ten times and performance of the models was assessed based on the average log-likelihood scores across replications. Cross-validation was performed with topology and branch lengths fixed to the half-compatible majority rule tree of the QMM model. For comparison, the posterior predictive and cross-validation tests were also performed for a site-homogeneous GTR-only model.

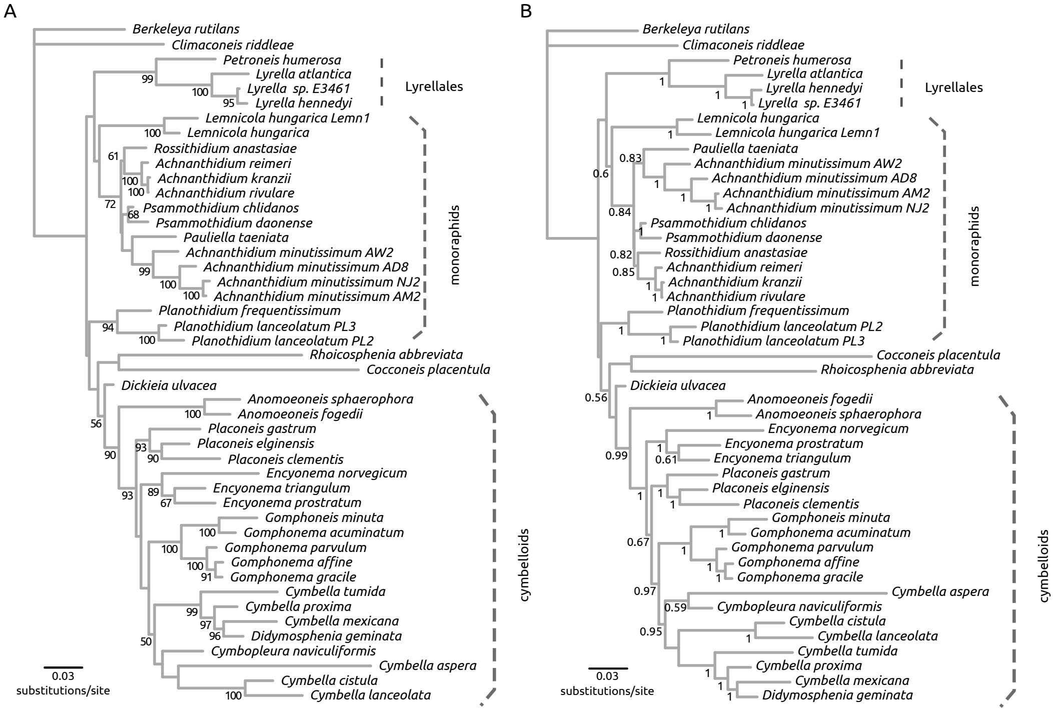
1.4 Results

1.4.1 Phylogenetic relationships

Inference under maximum likelihood placed *Rhoicosphenia* as a poorly supported sister to *Cocconeis* (bootstrap proportion (BP) <50%; Fig. 1.1A). *Dickieia* was the only branch separating the *Rhoicosphenia* + *Cocconeis* clade from the well supported lineage composed of members of the Anomoneidaceae, Cymbellaceae, and Gomphonematacae (Fig. 1.1A). The monoraphid diatoms were thus polyphyletic. *Planothidium* was a poorly supported sister to the lineage ((*Rhoicosphenia*, *Cocconeis*), (*Dickieia*, cymbelloids)), and the achnanthidioids + *Lemnicola* were sister to the Lyrellales (BP<50). At the genus level, *Achnanthidium* and *Cymbella* were polyphyletic while *Gomphonema* was paraphyletic with respect to *Gomphoneis* Cleve (Fig. 1.1A).

Bayesian inference painted much of the same picture described above. In the maximum *a posteriori* phylogeny the *Rhoicosphenia* + *Cocconeis* clade (posterior probability (PP) <0.5) was sister to a clade composed of *Dickieia* and the cymbelloids (PP<0.5; Fig. 1.1B). The remaining monoraphids were polyphyletic:

Figure 1.1: Maximum likelihood (A) and Bayesian maximum a posteriori (MAP, B) phylograms of the relationships between cymbelloid and monoraphid genera, and the Lyrellales inferred from the combined SSU+*rbcL* alignment. Support values bellow 50% or 0.5 are omitted. Branch lengths of the MAP topology were re-estimated in RAxML with a partitioned GTR+Γ+I model.



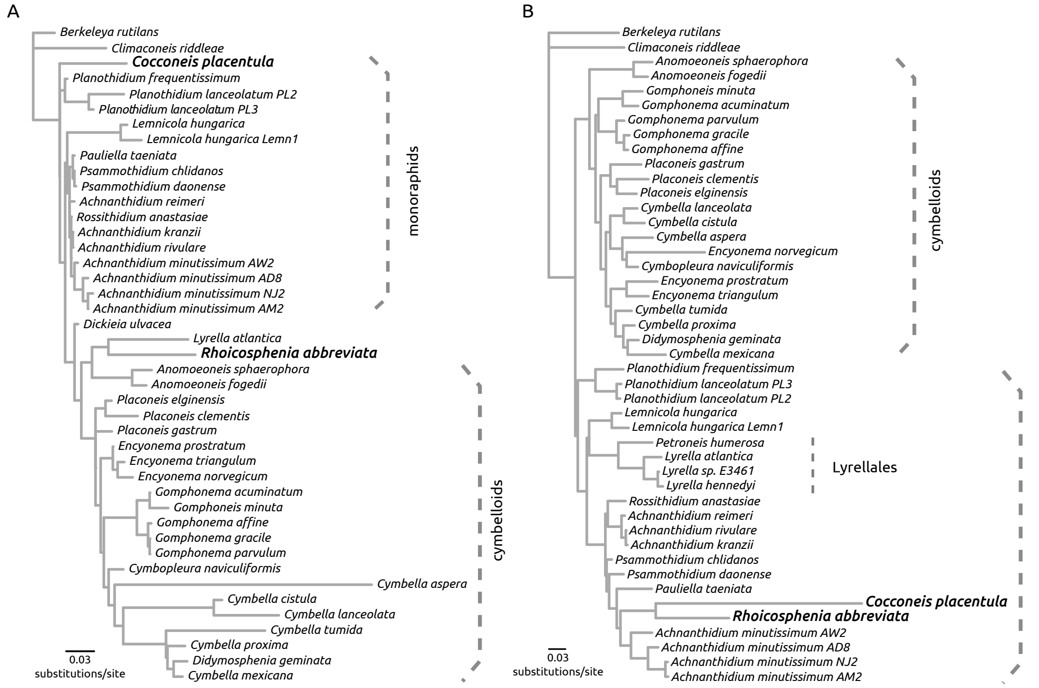
Planothidium was sister to the above clade while the achnanthidioids + *Lemnicola* (PP=0.6) were sister to the Lyrellales (Fig. 1.1B). The half-compatible majority rule consensus summarizing the posterior dissolved the poorly-supported backbone in the ingroup placing *Cocconeis* and *Rhoicosphenia* in a large polytomy.

1.4.2 Gene and partition-specific branch lengths and topologies

Under ML, the SSU gene tree recovered *Rhoicosphenia* as sister to the only *Lyrella* with available SSU sequence (Table 1.1; Fig. 1.2A). The *Rhoicosphenia* + *Lyrella* clade was poorly supported as sister to *Anomoeoneis* making the Cymbellales paraphyletic. *Cocconeis* was recovered as sister to all other ingroup taxa (Fig. 1.2A). In contrast, the *rbcL* gene tree (ML) recovered a sister relationship between *Cocconeis* and *Rhoicosphenia* (Table 1.1; Fig. 1.2B). Moreover, in the *rbcL* gene tree both *Cocconeis* and *Rhoicosphenia* display substantially longer branches compared the remaining taxa – a result not observed with SSU-inferred branches (Fig. 1.2B).

Under BI, based on SSU, *Rhoicosphenia* was sister to Lyrella (PP = 0.7; Table 1.1; Fig. 1.3A) while *Cocconeis* was sister to *Planothidium* (PP = 0.52; Table 1.1; Fig. 1.3A). The half-compatible majority rule consensus tree of the SSU-only alignment showed considerable uncertainty with a number of taxa resolved in a large polytomy (Fig. 1.3A). Despite having more structure, the *rbcL* gene tree nonetheless had a seven-way polytomy in the backbone. As in the ML analyses, *Cocconeis* and *Rhoicosphenia* were found as sister taxa with PP=81 (Table 1.1; Fig. 1.3B). The branches leading to these taxa, especially *Cocconeis*, were longer than the remain-

Figure 1.2: Maximum likelihood phylograms of the SSU (A) and *rbcL* (B) gene trees. Bootstrap support values are omitted.



der (Fig. 1.3B).

Table 1.1: Phylogenetic placement of *Rhoicosphenia* and *Cocconeis* and corresponding clade support values.

analysis	*Rhoicosphenia* sister to	Support[a] (%)	*Cocconeis* sister to	Support (%)
ML SSU gene tree	*Lyrella*	61	In polytomy	NA
ML *rbcL* gene tree	*Cocconeis*	70	*Rhoicosphenia*	70
BI SSU gene tree	*Lyrella*	70	*Planothidium*	51
BI *rbcL* gene tree	*Cocconeis*	81	*Rhoicosphenia*	81
BI separate partition branch lengths joint topology	*Cocconeis*	65	*Rhoicosphenia*	65
BI separate partition branch lengths and topology[b]	*Cocconeis*	79	*Rhoicosphenia*	79

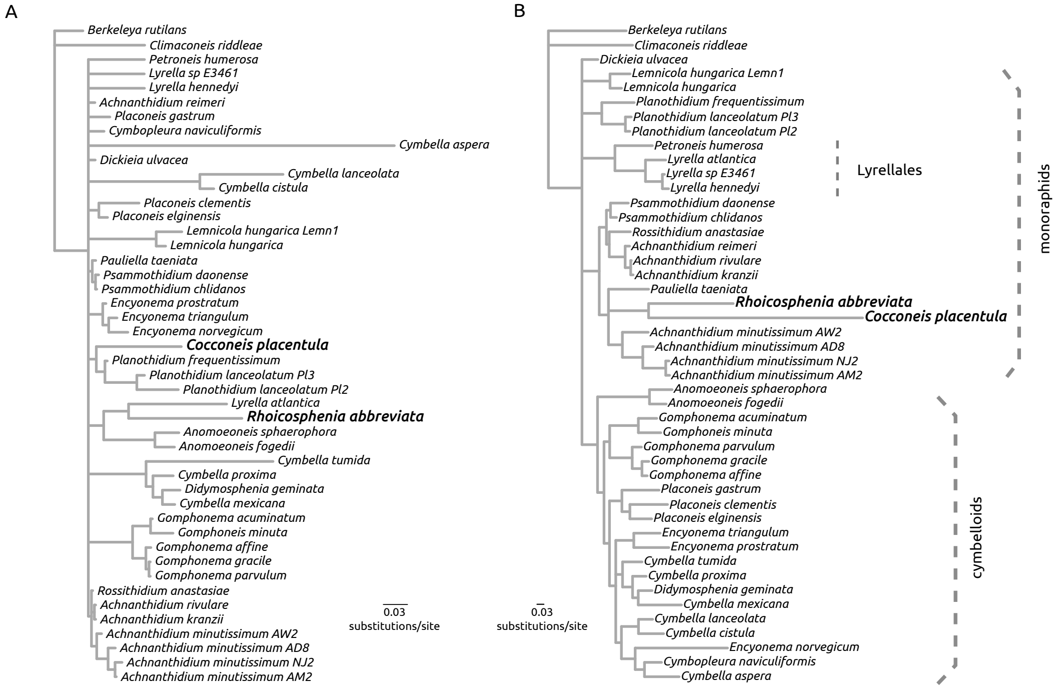
---

[a] Both bootstrap support and posterior probability are shown as percent.

[b] Results from the partition of 3rd codon positions of *rbcL*.

I also performed analyses of the concatenated dataset with separate estimates of branch lengths for each partition under a joint topology and with separate estimates of both branch lengths and topologies for each partition (Table 1.1; Fig. 1.4A). When I allowed partition-specific branch lengths, I recovered a topology in which *Cocconeis* and *Rhoicosphenia* were sister taxa (PP= 0.64; Table 1.1; Fig. 1.4A). The *Cocconeis* and *Rhoicosphenia* branches across partitions were anywhere between ca. 2 to 12.5 times longer than the average branch length in the tree (Table 1.1A). Branches estimated from the paired sites of SSU and the second codon position of *rbcL* were similar between *Cocconeis* and *Rhoicosphenia* (Table 1.1). For the unpaired sites from SSU and first codon position of *rbcL*, the *Rhoicosphenia* branch was about twice as long as the *Cocconeis* branch. *Cocconeis* had a substan-

Figure 1.3: Bayesian half-compatible majority rule consensus of the SSU (A) and *rbcL* (B) gene trees. Posterior probabilities are omitted.



tially longer branch than *Rhoicosphenia* only for the third codon position of *rbcL* (Table 1.1).

When I unlinked both branch lengths and topologies, *Rhoicosphenia* was sister to *Cocconeis* only in the half-compatible majority rule of the third codon position of *rbcL* (PP=0.79; Table 1.1; Fig. 1.4B). The majority rule trees from the other two codon positions of *rbcL* were comb-like. The first codon position tree had *Cocconeis* as sister to *Climaconeis* and *Rhoicosphenia* in a polytomy. The second codon position tree had both taxa in a polytomy. The paired sites from SSU recovered *Rhoicosphenia* in a polytomy and *Cocconeis* as sister to *Planothidium*. The unpaired sites from SSU, on the other hand, reconstructed *Rhoicosphenia* as sister to *Lyrella* and *Cocconeis* in a polytomy. The ratios of *Cocconeis* and *Rhoicosphenia* branch lengths against the tree average were again high and similar to those from the analyses with joint topology described above (Table 1.2).

Table 1.2: Ratio of the branch lengths leading to *Cocconeis* and *Rhoicosphenia* against the tree average.

Separate branch lengths joint topology			
	Rhoic[a]	Cocco[b]	Mean[c]
SSU paired	4.69	4.49	0.96
SSU unpaired	6.32	2.78	0.95
rbcL 1st codon	5.03	2.89	0.94
rbcL 2nd codon	2.38	2.07	0.98
rbcL 3rd codon	4.16	12.59	0.85
Separate branch lengths separate topology			
	Rhoic	Cocco	Mean
SSU paired	2.25	3.18	0.98
SSU unpaired	4.71	2.29	0.95
rbcL 1st codon	4.05	2.68	0.94
rbcL 2nd codon	1.19	2.58	0.97
rbcL 3rd codon	2.97	12.09	0.86

---

[a] *Rhoicosphenia*

[b] *Cocconeis*

[c] Mean ratio after removing the *Cocconeis* and *Rhoicosphenia* branches.

1.4.3 Comparison of topological hypotheses

Maximum likelihood tests of the competing topological hypotheses, H0: monophyly of *Rhoicosphenia* + cymbelloids and H1: monophyly of *Rhoicosphenia* + *Cocconeis*, did not reject H0 in favor of H1 even though the clade *Rhoicosphenia* + *Cocconeis* is found in the most likely tree (Table 1.3; Fig. 1.1). Probability values of the AU and SH tests were >0.05 for the null hypothesis of monophyly of *Rhoicosphenia* + cymbelloids. Standard errors of the p-values were <0.01 indicating that a sufficient number of multiscaled bootstrap replicates were sampled (Table 1.3).

Figure 1.4: Bayesian half-compatible majority rule trees from the analyses with separate partition branch lengths with joint topology (A) and separate partition branch lengths and separate topology (B). In both cases branch lengths are those from 3rd codon positions of *rbcL* as is the topology in B.

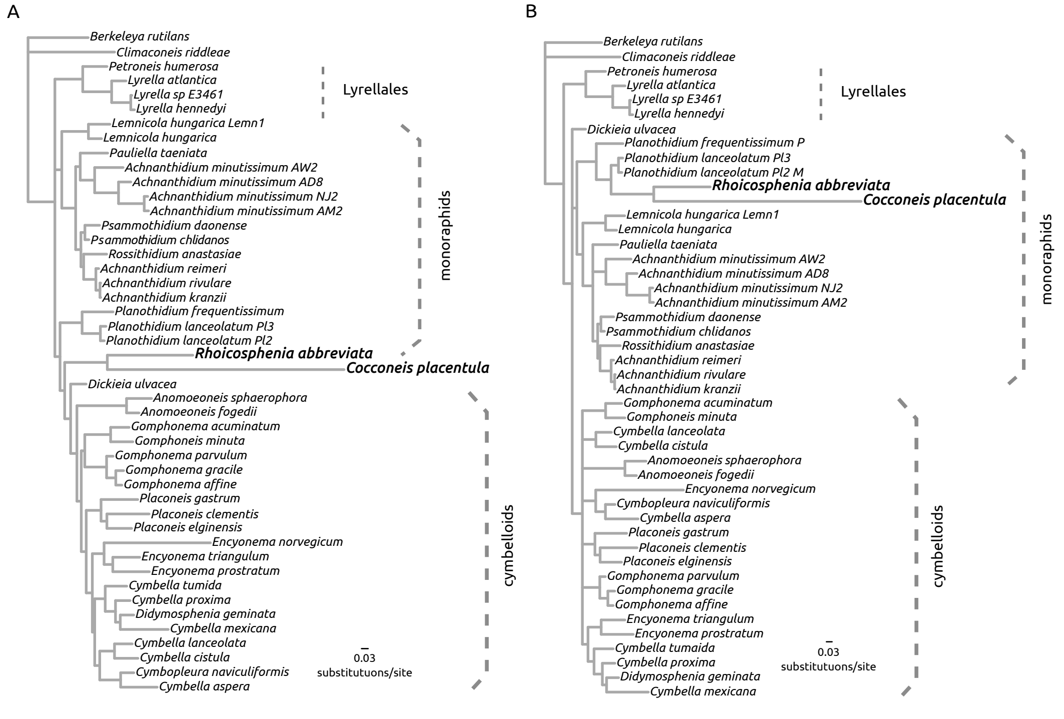


Table 1.3: Maximum likelihood topology tests for the phylogenetic position of *Rhoicosphenia* in relation to the cymbelloids and monoraphid diatoms.

Topology	AU[a] p-value	SE	SH[b] p-value	SE [c]
H1[d]: Rhoic + Cocco[e]	0.773	0.008	0.775	0.004
H0[f]: Rhoic + cymb[g]	0.227	0.008	0.225	0.004

---

[a] Approximately unbiased test

[b] Shimodaira-Hasegawa test

[c] Standard error of the probability.

[d] Alternative hypothesis of monophyly *Rhoicosphenia* and *Cocconeis*.

[e] *Cocconeis*

[f] Null hypothesis of monophyly of Cymbellales sensu Round et al. 1990.

[g] cymbelloid diatoms

Comparison of the marginal likelihoods of a model with unconstrained topology to models with hard constraints for (i) *Rhoicosphenia* + cymbelloids and (ii) *Rhoicosphenia* + *Cocconeis* showed that the null hypothesis of monophyly of the Cymbellales was favored (BF=41.46; Table 1.4). Such tests, however, are inappropriate in the Bayesian framework because they are biased in favor of the models where the prior distribution of topologies is restricted by a hard constraint (Bergsten et al., 2013). The outcome was reversed when in addition to the hard constraints we used partial constraints. The marginal likelihoods supported the monophyly of *Cocconeis* + *Rhoicosphenia* over the null hypothesis of monophyly of *Rhoicosphenia* + cymbelloids (BF=15.9; Table 1.4).

Table 1.4: Bayes factor topology tests for the phylogenetic position of *Rhoicosphenia* in relation to the cymbelloids and monoraphid diatoms.

Topology prior	Mean ln L[a]	SE[b]	Δ Mean ln L	2lnBF[c]
unconstrained	-15879.22	0.64	63.97	127.94
Hard= Rhoic[d] + cymb[e]	-15852.17	0.54	36.92	73.84
Hard= Rhoic + Cocc[f]	-15872.90	1.01	57.65	115.30
Partial, Rhoic float	-15822.35	1.06	7.10	14.20
Partial and Rhoic + cymb	-15823.20	0.47	7.95	15.90
Partial and Rhoic + Cocc	-15815.25	0.66	0.00	0.00

---

[a] Marginal likelihood
[b] Standard error of the mean marginal likelihood calculated from 6 ssMCMC samplers
[c] Twice the difference in mean marginal likelihood between competing hypotheses
[d] *Rhoicosphenia*
[e] cymbelloid diatoms
[f] *Cocconeis*

In accordance with the BF test, the posterior odds comparison between H0 and H1 showed that topologies where *Cocconeis* and *Rhoicosphenia* are sister taxa are about 10-11 times more likely than the null hypothesis (Table 1.5). Within a posterior distribution of 6×10^4 trees, only 1700-1800 (2.8-3%) were consistent with H0 versus nearly 19000 (30-31%) consistent with H1. The results are almost identical for the cases where the analysis was run unconstrained or with a partial constraint allowing *Rhoicosphenia* and *Cocconeis* to float in and out of partial constraints (Table 1.5).

Table 1.5: Posterior frequency topology tests for the phylogenetic position of *Rhoicosphenia* in relation to the cymbelloids and monoraphid diatoms.

Analysis	Hypothesis	# trees[a]	Posterior frequency
Unconstrained	H0[b]	1710	0.028
Unconstrained	H1[c]	18893	0.315
Partial backbone constraint	H0	1801	0.030
Partial backbone constraint	H1	18666	0.311

---

[a] Number of trees consistent with a particular hypothesis from a posterior sample of 6×10^4 trees

[b] Null hypothesis of monophyly of Cymbellales sensu Round et al. 1990

[c] Alternative hypothesis of monophyly *Rhoicosphenia* and *Cocconeis*

1.4.4 Long branch attraction

Since *Cocconeis* and *Rhoicosphenia* sit on very long branches (Fig. 1.1, 1.2; Table 1.2), I tested whether their sister relationship could be due to LBA. I first performed long-branch extraction (LBE) assessing the placement of *Cocconeis* in the absence of *Rhoicosphenia* and vice versa. When *Rhoicosphenia* was removed from the dataset, *Cocconeis* was recovered as sister to *Planothidium* (BP=55) and the monoraphid diatoms were monophyletic (BP<50%; Fig. 1.5A). The Lyrellales were sister to the *Dickieia* + the cymbelloids clade. On the other hand, when *Cocconeis* was removed from the dataset, *Rhoicosphenia* fell as the sister-group to a clade of *Achnanthidium* species (BP<50%; Fig. 1.5B). The monoraphids were thus paraphyletic with respect to *Rhoicosphenia*. In both cases, removal of one of the long branches, resulted with a distant, albeit poorly supported, placement of the remaining long branch taxon: *Cocconeis* became sister to *Planothidium* and *Rhoicosphenia* to species of *Achnanthidium* (Fig. 1.5). Therefore the LBE tests lended some support that *Rhoicosphenia* and *Cocconeis* sit on branches long enough to attract (Siddall and Whiting 1999).

Results from the "random sequence attraction" test did not support the suspicion of LBA between *Cocconeis* and *Rhoicosphenia*. Let us assume that each taxon in the dataset is equally likely to be attracted to the introduced random sequence. Then by chance alone we expect that each taxon will be attraced to the random sequence in about $10^3/45 = 22.2$ trees (2%) trees). When *Rhoicosphenia* was replaced with a random sequence, *Cocconeis* was attracted to it 5.6% of times. The same outcome was observed for *Rhoicosphenia* when *Cocconeis* was replaced with a random sequence. Compared to the mean number of trees in which the remaining taxa were attracted to the random sequence, *Cocconeis* and *Rhoicosphenia* are nearly three times more likely to be attracted to a random sequence. However, in both cases there were several other taxa that were attracted to the random sequence more frequently than *Cocconeis* and *Rhoicosphenia* (Fig. 1.6). Thus, while both *Rhoicosphenia* and *Cocconeis* are prone to attract to an introduced random sequence this is not appreciably more than other taxa in the tree and LBA need not be invoked.

1.4.5 Site-heterogeneous models of evolution

I employed two site-heterogeneous models of evolution, a mixture of base frequency profiles (CAT) and a mixture of rate-matrices (QMM), that have been shown to fare better than site-homogenous models with respect to systematic errors such

Figure 1.5: Maximum likelihood phylograms of the long-branch extraction test. (A) *Rhoicosphenia* removed and (B) *Cocconeis* removed. Bootstrap proportions bellow 50% are omitted.

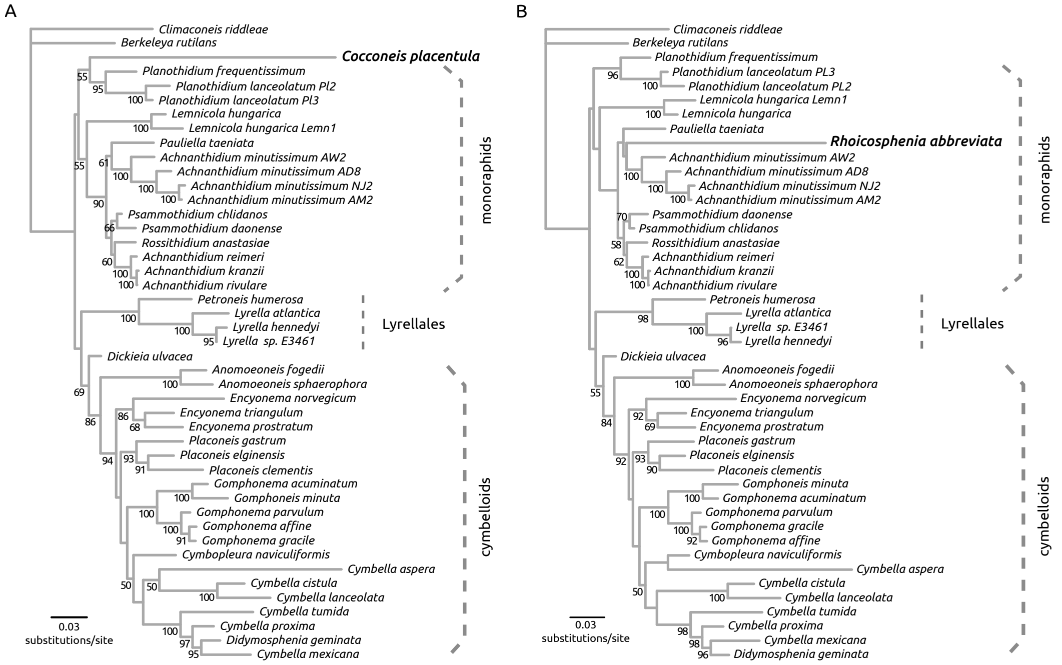
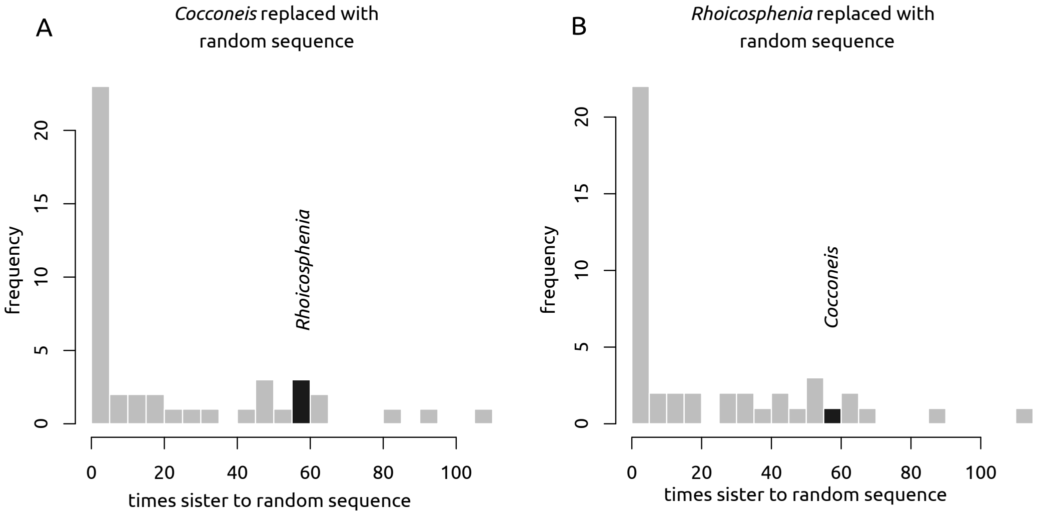


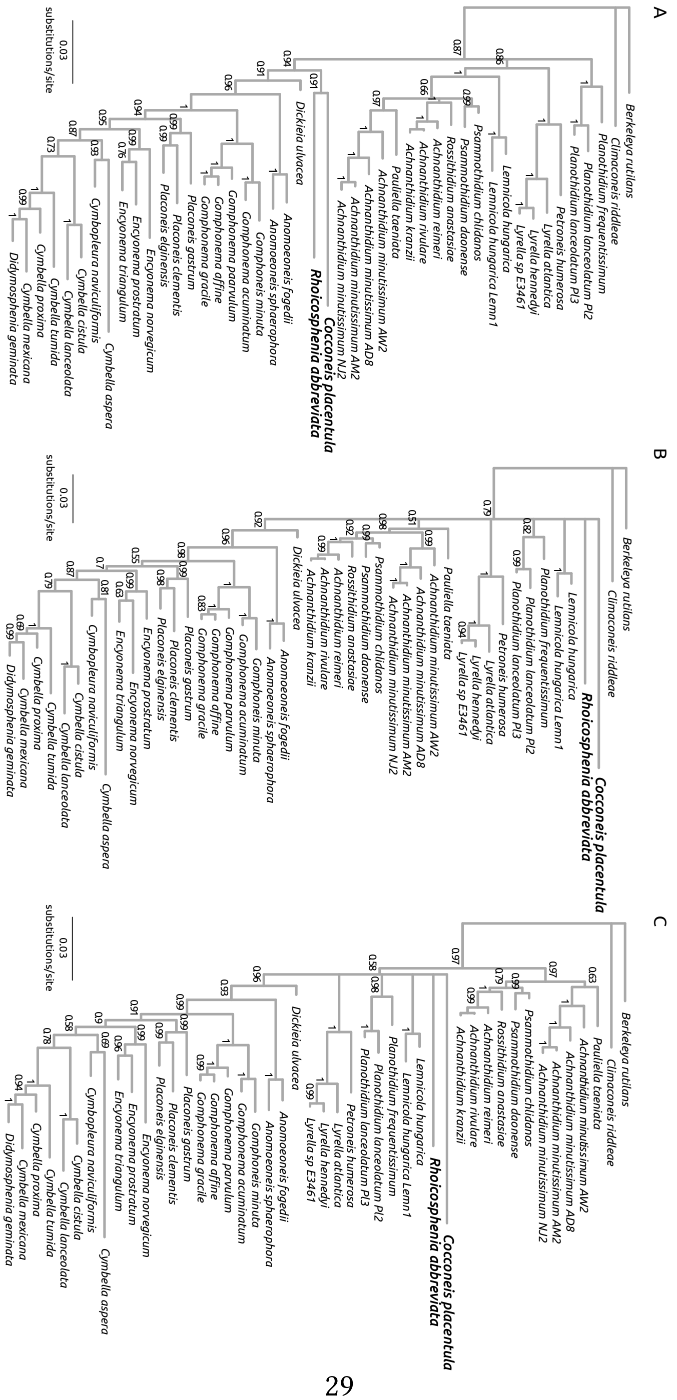
Figure 1.6: Neither *Rhoicosphenia* (A) nor *Cocconeis* (B) were attracted to a random sequence appreciably more than other taxa in the dataset.



as LBA (Lartillot et al., 2007). Since the likelihoods of MrBayes and PhyloBayes analyses are not comparable, I also performed a PhyloBayes analysis using a single-matrix GTR model. The unpartitioned GTR model recovered monophyly of *Cocconeis* + *Rhoicosphenia* (PP=0.91; Fig. 1.6A). The mixture models (CAT and QMM) separated the two taxa but placed them in a polytomy with the Cymbellales, Lyrellales and monoraphids failing to identify alternative position for either species with confidence (Fig. 1.6B, C). Aside from the different outcomes with respect to the two focal taxa, the GTR, CAT, and QMM trees were topologically similar (Fig. 1.6).

I performed posterior predictive simulations to assess the models' performance given heterogeneities in the nucleotide diversity and taxon-specific nu-

Figure 1.7: Majority rule consensus (50%) summaries from the PhyloBayes analyses. (A) the single-matrix GTR model and (B) the CAT mixture and (C) the QMM mixture.



cleotide composition. The diversity test compared the observed mean number of residues per site to the same statistic calculated from datasets simulated using model parameters from the thinned posterior. The test statistic of the heterogeneity test is the maximum square deviation between global and taxon-specific nucleotide frequencies. Analyses showed that the GTR model inadequately described the structure of the present dataset both in terms of nucleotide diversity per site and taxon-specific biases in composition (Table 1.6). In contrast the CAT and QMM models performed adequately (Table 1.6).

To examine the goodness-of-fit for these tree models, I employed a cross-validation procedure on a fixed topology in which parameters estimated from a training set (9/10 of the data) were used to calculate the likelihood of a test set (1/10 of data). As expected given the results of the posterior predictive tests, the CAT and QMM models outperformed GTR (Table 1.6). The performance of CAT (favored in 6 of 10 replicates) and QMM was, however, similar (Table 1.6).

Table 1.6: Posterior predictive and cross-validation tests for the adequacy and goodness-of-fit of the GTR, CAT and QMM models.

Nucleotide diversity per site			
Model	Observed	Mean predicted	p-value[a]
GTR	1.33	1.43	0.00
CAT	1.33	1.34	0.14
QMM	1.33	1.33	0.26
Taxon-specific compositional heterogeneity			
Model	Observed	Mean predicted	p-value
GTR	0.00013	0.00009	0.024
CAT	0.00013	0.00011	0.16
QMM	0.00013	0.00011	0.10
Pairwise cross-validation (CV) score comparison			
comparison[b]	Δ CV scores[c]	SD[d]	# favored[e]
GTR-CAT	62.0	8.66	10
GTR-QMM	60.5	5.98	10
CAT-QMM	-1.44	6.80	4

---

[a] Nonsignificant p-values indicate the model adequately captures nucleotide diversity and taxon-specific heterogeneity in data.

[b] In pairwise comparisons the first model is the reference.

[c] Positive CV score indicates model performs better than reference.

[d] Standard deviation of the CV score.

[e] Number of times a model is preferred compared to reference out of 10 replicates.

1.5 Discussion

The genus *Rhoicosphenia* is peculiar in that it shares similarities with both cymbelloid diatoms (e.g. chloroplast morphology) and monoraphid diatoms (e.g. reduced raphe system of the convex valve). Although assigned to the Cymbellales, recent morphological phylogenies have placed *Rhoicosphenia* outside the order questioning the current classification of these freshwater diatoms (Cox and Williams 2006; Jones et al. 2005). Due to the absence of sequence data, the phylogenetic placement of *Rhoicosphenia* has thus far not been assessed with molecular techniques. I sequenced *rbcL* and SSU from a clone of *Rhoicosphenia abbreviata* from Waller Creek, Texas, and used it to test hypotheses about its phylogenetic affinity with respect to the cymbelloid and monoraphid diatoms. Our tests included the most appropriate taxon sampling for the problem to date including representatives from the remaining three families of the Cymbellales and major genera of monoraphid diatoms. Molecular data placed *Rhoicosphenia* outside the Cymbellales as sister to the monoraphid diatom *Cocconeis placentula* Ehrenberg (Fig. 1.1). This arrangement was favored by both ML and BI hypotheses tests (Tables 1.3,1.4,1.5) although ML analyses did not reject the monophyly of the Cymbellales as currently defined (Table 1.3).

The branches leading to *Cocconeis* and *Rhoicosphenia* are longer than most others on the tree (Figs 1.1, 1.4; Table 1.2) raising suspicion about possible long-branch attraction between these two taxa. Indeed, when trees were built in the absence of one of these two long branches, both *Cocconeis* and *Rhoicosphenia* acquired alternative placement along the phylogeny (Fig. 1.5). Gene trees and partition-specific branch lengths and topologies showed that the association between *Cocconeis* and *Rhoicosphenia* is largely driven by the *rbcL* sequences, and in particular the 3rd codon positions (Figs 1.2, 1.3, 1.4; Tables 1.1,1.2). Regardless, this potentially artefactual relationship was never recovered with very strong support, generally <81% and 91% only when the data was analyzed unpartitioned (Table 1.1;

Fig. 1.7) and other taxa were attracted to a random sequence more than either *Cocconeis* or *Rhoicosphenia* (Fig. 1.6). Therefore, it is likely that the recovery of a sister relationship between *Cocconeis* and *Rhoicosphenia* is due to stochastic rather than systematic error (i.e. LBA).

When we suspect that analyses might be mislead by LBA, it seems logical to turn to other sources of information as a way of corroborating or refuting our suspicion. Morphology offers an alternative character set that is resistant, or at least very unlikely to be prone, to LBA-type artefacts (Bergsten 2005). The morphological matrix of Kociolek and Stoermer (1986) resembles, however vaguely, the molecular dataset analyzed here. Lets assume that Kociolek and Stoermer's *Gomphonema* represents our Cymbellales and "Achnanthes" our monoraphids, then the two datasets differ by the absence of *Mastogloia* in the molecular data and the absence of the Lyrellales, Berkelyaceae, and *Dickieia* in the morphological matrix. Scoring the morphological matrix on a tree that has the clade *Rhoicosphenia* + *Cocconeis* offered a substantially less parsimonious answer (27 steps) than Kociolek and Stoermer's original reconstruction of *Rhoicosphenia* as sister to *Gomphonema* (19 steps). Comparing the present dataset to the morphological matrix of Kociolek and Stoermer (1986) seems a bit like comparing apples to oranges, but it does offer another line of support that in the molecular phylogeny the association between *Cocconeis* and *Rhoicosphenia* could be artefactual.

The state space of nucleotide data is small (four states: A, C, G, T). It is therefore relatively easy for two taxa to acquire the same state in their nucleotide sequence through parallel or convergent change. This is one of the major contributors to LBA artefacts and the main reason why probabilistic tree inference methods that can accommodate unobserved substitutions perform better than parsimony when dealing with LBA artefacts (Huelsenbeck 1995; Huelsenbeck and Hillis 1993). Nonetheless, in certain conditions even probabilistic methods can be prone to LBA and especially when the model inadequately describes some aspect of the evolutionary process. This arises in cases where the employed evolutionary model fails to account for a major property of the data (e.g. failing to properly account for invariable sites or across-site rate variation; Huelsenbeck 1997; Kolaczkowski & Thornton 2009; Lartillot et al. 2007; Sullivan & Swofford 1997). I therefore explored the effects of phylogenetic mixture models that have been shown to fare better with respect to heterogeneities across sequences, taxon-specific biases in nucleotide composition, and LBA artefacts (Lartillot et al. 2007). When applied to this dataset, the mixture models substantially reduced the probability of the *Cocconeis* + *Rhoicosphenia* association, but failed establish supported alternative placement for either of these taxa (Fig. 1.7). Moreover, the mixture models performed adequately capturing the heterogeneity in nucleotide diversity across the alignment and composition across taxa (Table 1.6). A single-matrix GTR model failed at this task recovering *Cocconeis* and *Rhoicosphenia* as sister taxa (PP=0.91; Fig. 1.7; Table 1.6. Despite the improvements introduced by the mixture models seen in their better cross-validation scores (Table 1.6), however, the position of *Rhoicosphenia* remained ambiguous prolonging the uncertainty surrounding the monophyly and classification of the Cymbellales.

Going forward the most promising approach is to focus on targeted taxon sampling such that the added taxa will break up long branches. The obvious place to start would be other species of *Cocconeis* and *Rhoicosphenia* as well as genera from the Cocconeidaceae and Rhoicospheniaceae (e.g. *Campyloneis* Grunow, *Campylopyxis* Medlin, etc.). Additional molecular data could also help especially unlinked nuclear and mitochondrial protein coding genes. Finally, combining molecular and morphological data will undoubtedly offer a more refined and complete picture of the evolutionary history of the Cymbellales, monoraphid diatoms, and their allies.

1.6 Acknowledgements

This study was partially funded by NSF EF 062941, the Blumberg fellowship, and USGS/NPS NRPP 141338. TN acknowledges the Plant Biology Graduate program (UT Austin) for research scholarships. Analyses were performed at the Texas Advanced Computing Center (http://www.tacc.utexas.edu).

Chapter 2

Molecular phylogeny of the Cymbellales (Bacillariophyceae) with a comparison of models for accommodating rate-variation across sites

2.1 Abstract

I reconstructed the phylogeny of representatives from nine genera and three families of the Cymbellales using two nuclear and three chloroplast genes. After rooting with *Anomoeoneis*, *Placoneis* was found as sister to a clade composed of *Cymbella*, *Cymbopleura*, *Encyonema*, *Gomphonema*, and *Gomphoneis*. The latter group was divided into mainly heteropolar and dorsiventral lineages. The data and chloroplast morphology also supported a close relationship between *Geissleria decussis* and *Placoneis*. Expectedly, the sequenced genes exhibited substantial across-site rate variation (ASRV) which prompted us to assess the stability of the inferred relationships in the face of different approaches for modeling ASRV. While the overall topology remained stable across alternate analyses, relationships between *Cymbella* and *Cymbopleura* and within one clade of *Gomphonema* varied dependent on the employed model. In some cases a strongly supported relationship under one model was not recovered by another model that differed solely in how the data was partitioned. These topological fluctuations appeared in areas of the tree with the least balanced taxon sampling and they altered the outcomes

of phylogenetic hypotheses tests of monophyly. Assessing how different models for ASRV affect tree topology and clade support values therefore seems important in cases of sparse or unbalanced taxon sampling or when assessing the phylogenetic affinity of previously unsampled taxa when lineage-specific biases in base composition or evolutionary rate are more difficult to detect.[1]

2.2 Introduction

The Cymbellales are a predominantly freshwater group of raphid diatoms comprising 28 genera and over 1100 species (McGuiry and McGuiry 2013). The order is one of few groups of raphid diatoms with chloroplasts positioned against the valves and encapsulated gametangia (Jones et al., 2005). The closest phylogenetic neighbors of the Cymbellales include the Lyrellales and some monoraphid diatoms. This grouping appears in molecular phylogenies (Ruck & Theriot, 2011) and in analyses of physiological characters associated with the mode of sexual reproduction and protoplast organization (Jones et al., 2005). Genera within the Cymbellales vary with respect to symmetry of the frustule (symmetric, dorsiventral, heteropolar), growth habit (free-living, stalked, tube-dwelling), and mode of sexual reproduction (isogamous, anisogamous). The current classification of the order reflects the variation in symmetry – the families Anomoneidaceae, Cymbel-

---

[1] This chapter is accepted for publication in Phycologia as Nakov, Ruck, Galachyants, Spaulding, and Theriot: Molecular phylogeny of the Cymbellales (Bacillariophyceae) with a comparison of models for accommodating rate-variation across sites. Author contributions: Elizabeth Ruck helped with field and lab work; Yuri Galachyants helped with field work in Siberia; Sarah Spaulding contributed field collections and funding; Edward Theriot, supervisor.

laceae, Gomphonemataceae, and Rhoicospheniaceae, roughly correspond to the overall shape of the silica frustule (Round et al., 1990).

Aspects of the phylogeny of the Cymbellales have been investigated using phylogenetic methods for nearly 30 years. Historically, investigations have addressed the relationship between the Cymbellales and monoraphid diatoms (Kociolek & Stoermer, 1986), the evolution of apical pore field bearing taxa (Kociolek & Stoermer, 1988), and interrelationships within the Gomphonemataceae (Kociolek & Stoermer, 1989, 1993). The incorporation of molecular data helped narrow down on the sister group to the Cymbellales (Jones et al., 2005), and in combination with broadly sampled morphological phylogenies (Cox & Williams, 2006), suggested that the Cymbellales as defined by Round et al. (1990) may not be monophyletic (Bruder & Medlin, 2007)[2]. Some low level relationships within the Cymbellales are consistently recovered and highly supported by morphological and molecular data (e.g. the monophyly of *Gomphonema* and *Gomphoneis*; Kociolek & Stoermer 1988; Kermarrec et al. 2011.) Higher-level relationships, however, remain unclear and contrasting phylogenetic placement has been inferred for the species-rich genera *Encyonema*, *Cymbella*, *Gomphonema*, and *Placoneis* (Kociolek & Stoermer, 1988; Jones et al., 2005; Cox & Williams, 2006; Bruder & Medlin, 2007; Kermarrec et al., 2011). *Placoneis*, for example, has been found in a polytomy with a number of other raphid pennates (Cox & Williams, 2006), as sister to the remaining Cymbellaceae and Gomphonemataceae (Kociolek & Stoermer, 1988), as sister to *Gomphonema* (Jones et al., 2005), and as sister to *Cymbella* (Bruder & Medlin, 2007). Overall, we

---

[2] Chapter 1 of this dissertation arrives at a similar conclusion.

remain far from a phylogenetic hypothesis that covers the breadth of family- or genus-level diversity within the group.

I addressed the phylogeny of the Cymbellales using a newly generated five-gene nuclear and chloroplast gene dataset covering nine genera representing the Anomoeoneidaceae, Cymbellaceae, and Gomphonemataceae. Our main focus was on the phylogenetic relationships recovered with the newly-generated data and their correspondence to the current classification of the Cymbellales. I was also interested in the variation of evolutionary rates and nucleotide composition across the sequenced genes. In particular, I investigated how different approaches to accommodate variation in rate or base frequency affect the inferred phylogenetic relationships. To this end, I compared the results of analyses with different alignment partitioning schemes and methods that do not partition the data a priori, but build a tree while simultaneously estimating heterogeneity in the data. The resulting phylogenies were generally congruent across analyses and were largely similar to previously constructed morphological and molecular phylogenies. However, certain phylogenetic relationships and support values for particular nodes were susceptible to the chosen data partitioning scheme or method for modeling rate-variation. In accordance, phylogenetic hypothesis tests (e.g. the probability of monophyly of groups) performed with different phylogenetic models, had substantially different outcomes despite being based on the same data. I discuss cases where exploration of data-partitioning and modeling rate-variation might be of importance and show that *Geissleria decussis* (Østrup) Lange-Bertalot et Metzeltin is closely related to *Placoneis*.

2.3 Material and Methods

2.3.1 Taxon sampling

Taxon sampling included 63 accessions from nine genera representing all families of the Cymbellales except the Rhoicospheniaceae (GenBank accessions: KJ011555-KJ011855). Representatives of *Rhoicosphenia* were excluded in this dataset because in the context of broader taxon sampling encompassing the Cymbellales, Lyrellales, and monoraphid diatoms, *Rhoicosphenia* falls outside the Cymbellales (Chapter 1). I followed the classification of Round et al. (1990). Anomoneidaceae are represented with three species of *Anomoeoneis*. From the Gomphonemataceae I sampled *Gomphonema*, *Gomphoneis*, and *Didymosphenia* (23 accessions combined). From the Cymbellaceae I sampled *Cymbella*, *Cymbopleura*, *Encyonema*, *Encyonopsis*, and *Placoneis* (36 accessions combined).

2.3.2 Cell cultures and DNA methods

A number of cultures used in this study originate from Dr. David Czarnecki's algal culture collection (Loras College, Dubique, Iowa). Many of these cultures are housed at the UTEX Culture Collection (UT Austin) although some have not survived in culture since the time of DNA extraction. Taxonomic names for these strains are as originally identified by Dr. Czarnecki. For newly collected strains, single cells were isolated with capillary pipetting and used to inoculate freshwater mediums, either biphasic soil + water (Czarnecki, 1987) or COMBO (Kilham et al., 1998). For cultured material I extracted DNA using the Power-soil kit (MoBio, Carlsbad, CA, USA) following manufacturer's instructions. As a backup, DNA was extracted from single or multiple wild cells using a Chelex-100 (Bio-Rad, Hercules, California, USA) method (Richlen & Barber, 2005). Primers, PCR, and sequencing protocols for the nuclear encoded SSU and large ribosomal subunit rDNA (LSU) and the chloroplast encoded *rbcL* gene are as described in Alverson et al. (2007) and Ruck & Theriot (2011). For chloroplast encoded photosystem I and II genes, *psaB* and *psbA*, sequences were obtained using newly designed primers under similar PCR conditions to the rbcL gene. Primer sequences used for amplification of the photosystem genes were: psab22F 5'-TTTAGCCCAGCYCTWGCACA-3' and psaB2000R 5'-CAATTAATTCTTGCCARTAACCAC-3' (for *psaB*), and psbA7 8F 5'-CCGTTTATACATCGGTTGGTTYGG-3' and psbA997R 5'-GGGAAGTTGTGCGC GTTACGTTC-3' (for *psbA*).

2.3.3 Sequence alignment

Ribosomal RNA gene fragments were aligned using the SSU-ALIGN package (Nawrocki, 2009; Nawrocki et al., 2009). SSU-ALIGN performs secondary structure alignments using the covariance model (CM; Cannone et al. 2002; Nawrocki 2009; for diatom examples see: Alverson et al. 2007; Theriot et al. 2009. For SSU, the raw sequences were directly aligned to the consensus CM model of Eukarya available as part of the SSU-ALIGN package. Alignment columns with low posterior probability (PP) were removed using the SSU-MASK routine from the SSU-ALIGN package.

Masked columns are ones in large loops for which positional homology and covarying nucleotides are difficult to assign (Nawrocki, 2009; Nawrocki et al., 2009). The total alignment after masking was 1525 nucleotides (nt) long and a total of 356 nt were masked.

A consensus CM for LSU is not available as part of the SSU-ALIGN package. However, given a reference CM the SSU-ALIGN and SSU-MASK routines can be performed for any RNA molecule (Nawrocki, 2009; Nawrocki et al., 2009). To create a reference secondary structure CM for LSU and align the sequenced LSU fragment to this CM, I downloaded the primary alignment for full length LSU sequences from the Comparative RNA Website (http://www.rna.icmb.utexas.edu/). This is a curated alignment by the Guttel lab (UT Austin) containing 117 full length eukaryotic LSU sequences. From this alignment I extracted the sequences of 6 heterokont taxa to use as a seed for the construction of a consensus secondary structure CM for heterokonts: *Skeletonema pseudocostatum* Medlin (Y11511), *Ochromonas danica* Pringsheim (Y07976), *Nannochloropsis salina* Hibberd (Y07974), *Phytophthora megasperma* Drechsler (X75631), *Scytosiphon lomentaria* (Lyngbye) Agardh (D16558), and *Tribonema aequale* Pascher (Y07978). The SSU-ALIGN package provides a routine to build a covariance model (SSU-BUILD) given a set of aligned sequences and consensus secondary structure annotation. I used the LOCARNA package (Will et al., 2007) to construct a preliminary secondary structure alignment and consensus annotation for two sequences (*S. costatum* and *O. danica*). This alignment was used as input for the building of a CM for heterokont LSU. Using SSU-BUILD I first created a CM for the alignment of *S. costatum* and *O.*

danica. To this CM, I aligned all 6 seed heterokont sequences using SSU-ALIGN. This alignment was used as input for the building of another CM, this time based on 6 instead of 2 full length LSU sequences. I "refined" this 6-taxon CM by re-aligning the 6 sequences and rebuilding the secondary structure CM. Finally, the newly generated LSU partial sequences (D1-D2 region) from the Cymbellales were aligned to this seed CM of 6 full length LSU sequences. After masking alignment columns with low posterior probabilities (68 nt total) the matrix was 479 nt long.

Chloroplast genes were aligned by hand after color-coding the nucleotide alignment by amino acids. The nuclear and chloroplast gene matrices were thereafter concatenated.

2.3.4 Estimation of tree-independent evolutionary rates

I estimated tree-independent evolutionaty rates through a pairwise comparison of the pattern of character states (site partition) of each site to every other column in the alignment (Cummins & McInerney, 2011). In this approach, a character that shares its site partition with many other characters in the matrix, is considered a slow-evolving site with rate near 1. Conversely, if a character's state partition is unique or shared with a small number of sites, its rate is considered high (near 0). The rates obtained by this approach are relative and unitless. I used this approach because tree-based generation of evolutionary rates is dependent on the topology and branch lengths of a, presumably, robust phylogeny. One possible advantage of this approach is that sites are combined in rate-based partitions irrespective of compartment or gene. The rates were estimated after concatenating the alignments. The tree-independent rates estimated here are dataset dependent, such that the same position in the context of a single-gene alignment or a combination of other genes might have a different rate estimate.

2.3.5 Phylogeny inference

I compared the results of three models that accommodate across site rate variation (ASRV) differently. While they all incorporate a discretized Γ distribution for ASRV, they differ with respect to partitioning or treatment of the substitution rate matrix.

The first model partitioned the data by gene. Chloroplast genes were further divided into codon positions and nuclear genes were partitioned into stem and loop regions based on the secondary structure alignment. This division yielded a model with 13 partitions dissecting the alignment along the conventional boundaries of fast and slowly evolving positions across the gene. In preliminary analyses, this scheme had the best likelihood from among a set of 12 less complex partitioning schemes that used genes, codons, and secondary structure information. I will refer to this partitioning scheme as the "by-gene-by-codon" model.

The second model partitioned the data by tree-independent evolutionary rate (Cummins & McInerney, 2011). For partitioning, I split the distribution of tree-independent evolutionary rates into 10 subsets. The slowly-evolving subsets 2-6 together had < 100 characters total and were combined with the slowest subset (invariant sites) yielding a model with 5 partitions. I will refer to this partitioning scheme as the "by-rate" model.

The above described partition models were ran in MrBayes v.3.2 (Ronquist et al., 2012). I ran 6 Metropolis-coupled Markov chain Monte Carlo (MCMC) samplers with one cold and one heated chain each. The simulations ran for 2×10^7 generations with a thinning rate of 10^{-3} and a 0.25 burnin fraction. Partitions were allowed to evolve at different relative rate using a variable rate prior. Model parameters (state frequencies, rate matrix, shape of Γ distribution for ASRV and the proportion of invariable sites [I]) were unlinked between partitions. The substitution matrix was not fixed. Instead, I ran reversible-jump MCMC across all models of the GTR family simultaneously accommodating model selection uncertainty (Huelsenbeck et al., 2004). With this routine, for each partition the MCMC visits the matrices with highest posterior probabilities. The topologies, branch lengths, and model parameters are therefore averaged over the credible set of models.

Initial analyses under both partition models resulted in total tree length considerably longer than the tree lengths inferred from Maximum Likelihood (ML) analyses. This is a well-documented problem in Bayesian phylogenetics whereby MCMC samplers can get "trapped" in a posterior distribution of extremely long trees (Brown et al., 2010; Marshall, 2010). One of the ways to circumvent sampling long trees is to modify the prior distribution of branch lengths (MrBayes default: exponential with rate, λ=10 and mean, μ=0.1). I set the mean and standard deviation of the prior on branch lengths to =0.037 for the by-gene-by-codon and =0.05 for the by-rate partition model. These values were calculated based on the average branch lengths of the phylogeny estimated with ML (Brown et al., 2010).

Convergence and stationarity of runs, after discarding the burnin, was assessed directly from the output by inspecting if the average standard deviation of split frequencies (ASDSF) were <0.01 and if the potential scale reduction factor (PSRF) was <5% different from 1.0 for estimated parameters. The "cumulative", "compare", and "var" routines from Are I there yet? (Nylander et al., 2008) were used to assess convergence of the distributions of topologies sampled by the MCMC chains.

The third model does not partition the data a priori, rather it allows for multiple substitution rates matrices in the phylogenetic model based on the variation present in the dataset. Starting with a calculation of a profile of equilibrium frequencies for each site in the alignment, the dataset is divided into categories of sites with similar base-frequency profiles. Then, the substitution process is modeled with the same rate-matrix for each category where exchangeabilites are multiplied by the category-dependent nucleotide frequencies. In addition, ASRV in each category is accommodated with the discretized Γ distribution. This multi-matrix model (called GTR+Γ+CAT) is implemented in PhyloBayes (Lartillot et al., 2013). In PhyloBayes I ran 3 MCMC samplers for 5.6×10^3 cycles for a posterior distribution of 4.6×10^3 trees (after discarding the first 18% of samples as burnin). I assessed convergence of the MCMC samplers using the "bpcomp" and "tracecomp" routines provided with PhyloBayes. I ensured that parameters had effective sample sizes > 300 and discrepancies between runs were <0.1 for both parameter estimates and tree bipartitions frequencies. The discrepancy is defined as the ratio between twice the difference in mean values between chains over the sum of their standard deviations for a particular parameter in the MCMC. I refer to this model as the "multi-matrix" model.

2.4 Results

2.4.1 Tree-independent evolutionary rates

Tree-independent evolutionary rates across the concatenated alignment and within each gene display bimodal distributions (Fig. 2.1). The majority of characters (4493) were assigned to the slowest rate bin. These are invariant or nearly invariant sites that convey little information about the branching of the phylogeny. A total of 1625 characters were pooled in the 4 fast-evolving bins (out of 10 bins) and 627 sites were placed in the highest rate bin. Fewer than 100 alignment columns have moderate evolutionary rate (Bins 2-6; Fig. 2.1). In the context of the entire alignment, each gene was characterized with bimodal distribution of evolutionary rates (Fig. 2.1). Among the five sequenced genes, *rbcL*, *psbA*, and SSU have similar counts of fast- and moderately fast-evolving sites (Fig. 2.1; Bins 7-10). LSU and *psaB* on the other hand exhibit high counts in the range of very fast-evolving sites (Fig. 2.1, Bins 9-10). For example, in the LSU alignment, 205 out of 479 characters (43%) are binned in the two highest rate bins with only 46 moderately evolving sites remaining (Bins 2-8).

2.4.2 The credible set of evolutionary models

When I partitioned the dataset by codon positions and secondary structure, the GTR rate matrix with six exchangeabilities was within the credible set of models (those with PP >0.05) only for the third codon partition of *rbcL* (Table 2.1). For the second codon partitions of *rbcL* and *psbA* there were no models sampled at frequency >5%. Overall, the uncertainty in model selection, as reflected by the number of models in the credible set, can be high with as many as seven credible models for certain partitions (Table 2.1). When the data was partitioned by rate, GTR was credible for two partitions and models with two rate categories (e.g. HKY) were plausible in three cases. Model selection in the fast-rate partitions, where rate-matrices with higher number of parameters are favored, is uncertain (Table 2.1).

Figure 2.1: Tree-independent evolutionary rates by gene. A. *rcbL.* B. *psaB.* C. *psbA.* D. SSU. E. LSU. Left side shows the variability of rates across each gene. Right side summarizes the rates in a histogram. Dashed lines denote partition cut-off values for the by-rate model. Rates near 1.0 are slow.

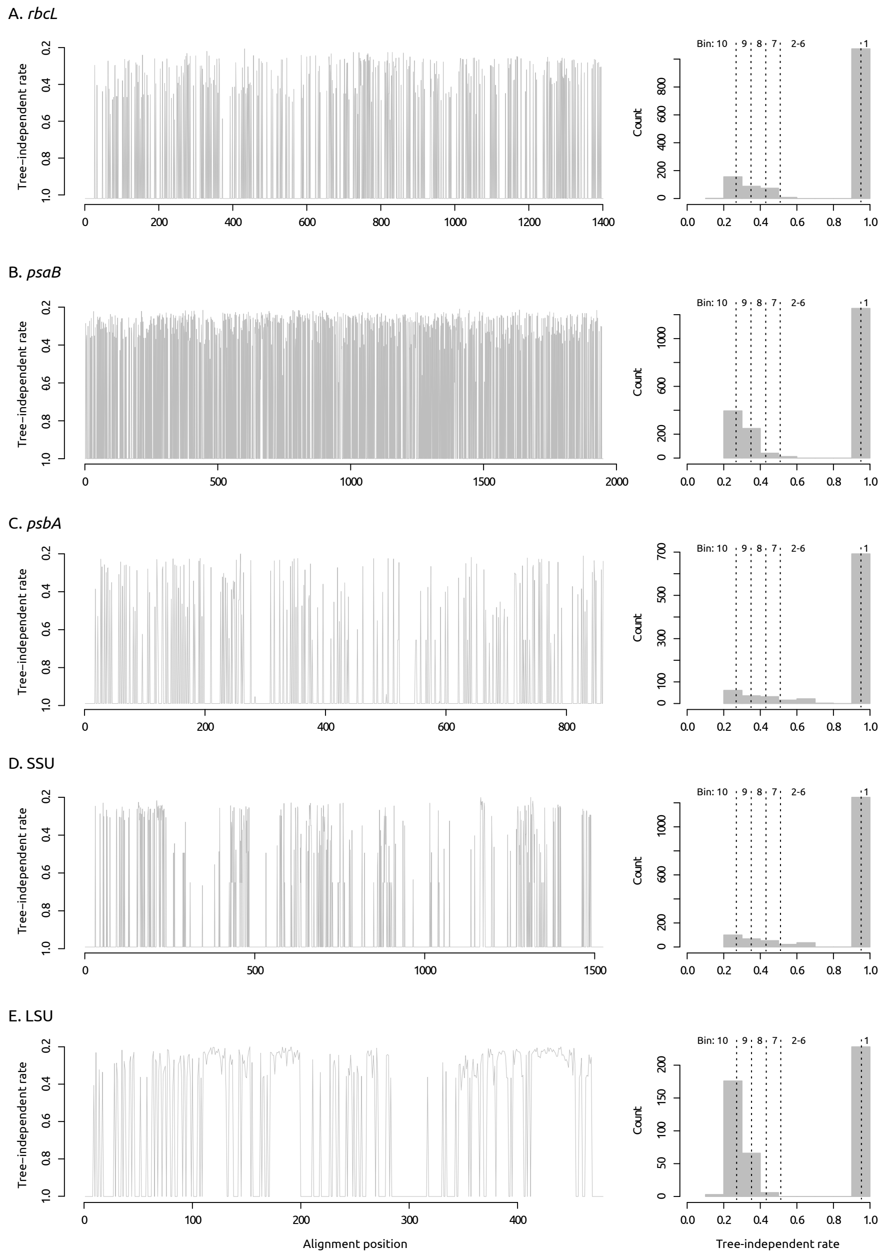


Table 2.1: The number of rate parameters (substitution types) in the credible models for the data partitioned by-gene-by-codon and by-rate.

Partition model	partition	# models with PP >0.05	# rate parameters[a] 2	3	4	5	6
by-gene-by-codon	*rbcl* 1st	3		■	■		
	rbcl 2nd	0					
	rbcl 3rd	5			■	■	■
	psaB 1st	7		■	■	■	
	psaB 2nd	3		■	■		
	psaB 3rd	4			■	■	
	psbA 1st	2		■	■	■	
	psbA 2nd	0					
	psbA 3rd	5			■		
	SSU stem	5		■	■	■	
	SSU loop	5		■	■		
	LSU stem	7		■	■	■	
	LSU loop	4		■	■		
by-rate	Bin 1-6	2	■	■			
	Bin 7	3	■	■			
	Bin 8	6	■	■	■		
	Bin 9	6		■	■	■	■
	Bin 10	5		■	■	■	■

---

[a] GTR, a model with 6 types of substitutions, can be denoted with the string "123456" where each number represents a rate parameter. The string "121121" corresponds to the HKY model with two rate parameters and "111111" is the Jukes-Cantor model.

2.4.3 Phylogenies

In previous studies with wider taxon sampling that included the Lyrellales and monoraphid diatoms, *Anomoeoneis* was recovered as the earliest diverging taxon among the sampled lineages sampled (Chapter 1). Therefore I rooted the trees at the branch between *Anomoeoneis* and the rest of the tree. The three models resulted with the same overall topology: (*Anomoeoneis*, (*Placoneis*, ((*Gomphonema*, *Gomphoneis*), (*Encyonema*, (*Cymbella*, *Cymbopleura*))))) (Fig. 2.2). The monophyly of *Placoneis* + the remaining taxa was strongly supported (PP=1.00) as was the monophyly of *Gomphonema* + *Gomphoneis* (GG) and *Encyonema* + the lineage of *Cymbella* + *Cymbopleura* + *Didymosphenia* + *Encyonopsis* (CCDE) (Figs 2.2,2.3; PP=1.00). The sister relationship between *Encyonema* and the CCDE clade received weaker support (0.58<PP<0.75).

Discrepancies between the phylogenies produced by the three models were recovered mainly with respect to the placement of members of *Cymbopleura*. In the consensus of the by-gene-by-codon partition model, *Cymbella aspera (Ehrenberg) Cleve* was recovered as sister to a group of *Cymbopleura* and the sole representative of *Encyonopsis* (PP=0.97; Fig. 2.3), while a member of *Cymbopleura inaequalis* Krammer group was sister to the remaining *Cymbella* and *Didymosphenia* (PP=0.97; Fig. 2.3. The by-rate partition model recovered *C. inaequalis* as a weakly supported sister (PP=0.54) to a clade of *C. aspera*, *Encyonopsis* and the remaining *Cymbopleura*. The latter three groups formed a polytomy (PP=0.84; Fig. 2.3). The remaining *Cymbella* + *Didymosphenia* were sister to the lineage of *C. aspera* and *Cymbopleura* (Fig. 2.3). In the consensus of the site-heterogenous multi-matrix model the entire CCDE clade was recovered as a five-way polytomy (Fig. 2.3). Similar fluctuations in topology across phylogenetic models were also recovered for low level relationships within one of the three clades of *Gomphonema* (Fig. 2.3).

Figure 2.2: Phylogenetic relationships within the Cymbellales inferred using different models for accommodating rate variation. The topology from the three analyses was almost identical. Bayesian posterior probabilities are shown as: by-gene-by-codon | by-rate | multi-matrix. Asterisks denote nodes where each model provided PP>0.95. Node "CCDE" leads to the lineage of *Cymbella*, *Cymbopleura*, *Didymosphenia*, and *Encyonopsis* (see Fig. 2.3). Node "GG" leads to the *Gomphonema* + *Gomphoneis* clade (see Fig. 2.4).

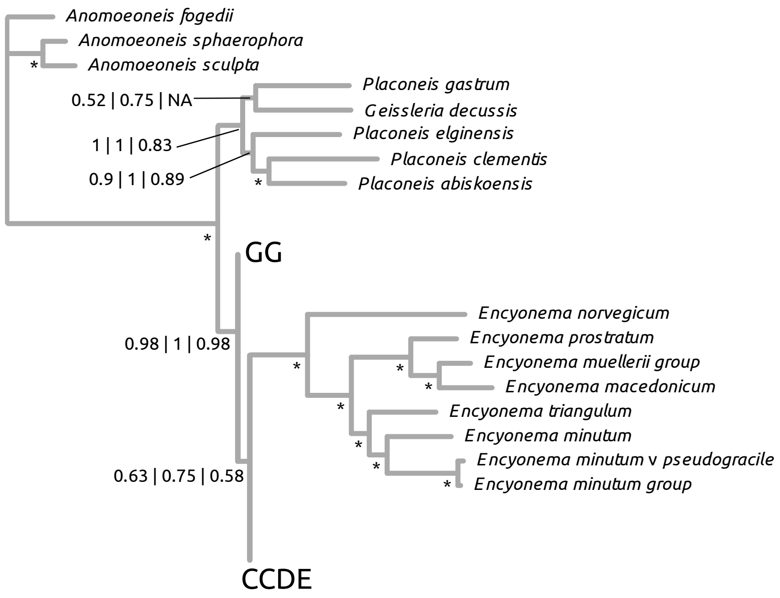


Figure 2.3: Topology and clade support of the CCDE clade under different models for accommodating rate variation: A –by-gene-by-codon partition model, B –by-rate partition model, and C –multi-matrix model. "CCDE" denotes the attachment to the phylogeny in Fig. 2.2.

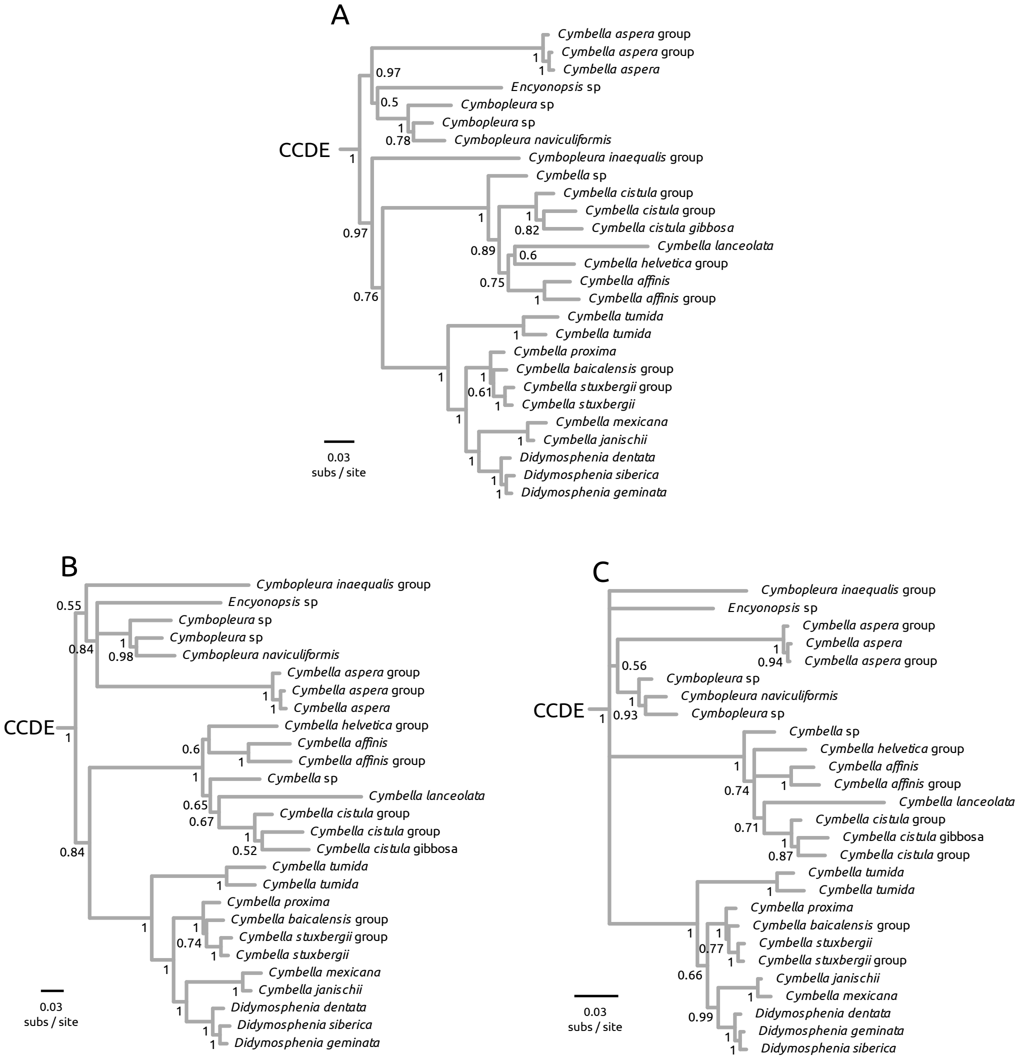
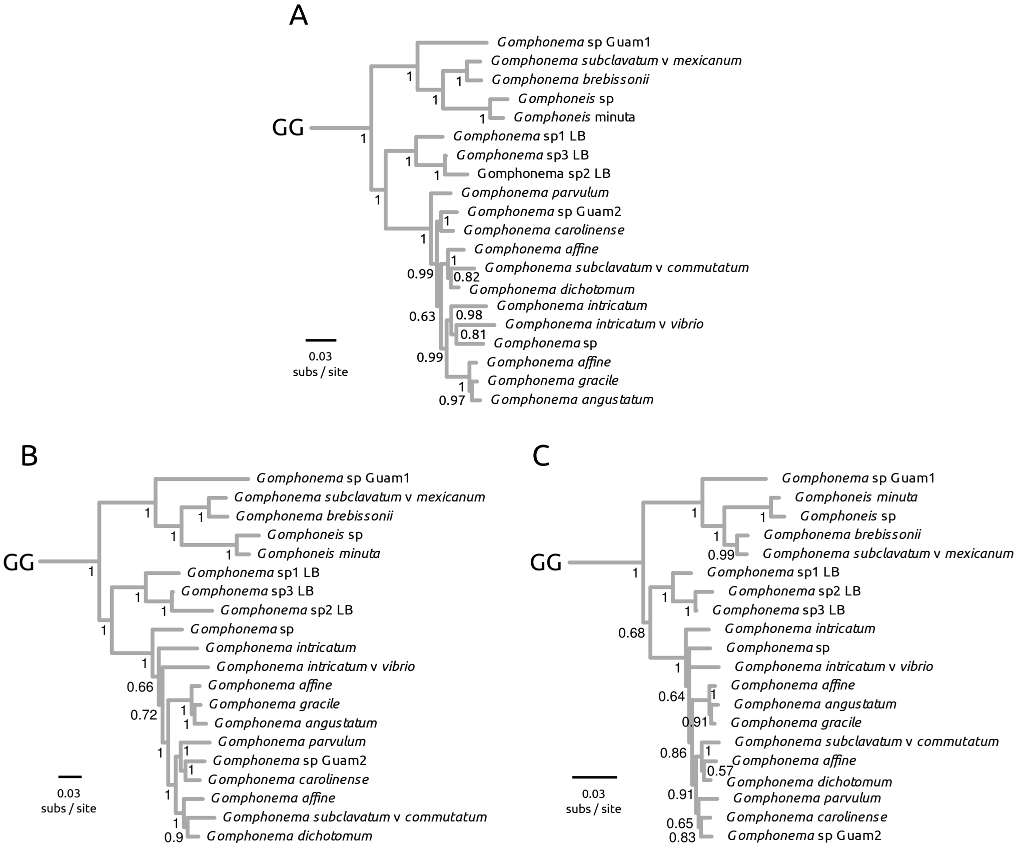


Figure 2.4: Topology and clade support of the GG clade under different models for accommodating rate variation: A –by-gene-by-codon partition model, B –by-rate partition model, and C –multi-matrix model. "GG" denotes the attachment to the phylogeny in Fig. 2.2.



2.4.4 The likelihood of family- and genus-level classification

The likelihood of monophyly of the Cymbellaceae, Gomphonemataceae, *Gomphonema*, and *Cymbella* is essentially zero. Irrespective of partitioning strategy, these three taxa do not appear as monophyletic in the posterior distribution of topologies. This comes as a result of the 1.00 PP for the monophyly of the *Gomphonema brebissonii* clade + *Gomphoneis* (Fig. 2.4) and the monophyly of the *Cymbella mexicana* clade + *Didymosphenia* (Fig. 2.3). Even when we take the sister relationship between the *C. mexicana* clade + *Didymosphenia* into account, *Cymbella* (this time with *Didymosphenia*) is not monophyletic (posterior frequencies $\leq$0.005) due to the interspersed lineages of *Cymbopleura* (Fig. 2.3). *Cymbopleura* alone, although still poorly sampled, is polyphyletic and the posterior probabilities for the monophyly of the four accessions range from 0.001 in the by-gene-by-codon model to 0.075 in the by-rate model. Similarly low probabilities are observed for the monophyly of the four *Cymbopleura* + *Encyonopsis* (Table 2.2). With respect to the Cymbellaceae I tested three combinations of taxa: the family as currently defined by Round et al. (1990) (CCDE + *Encyonema* + *Placoneis*), the family less *Placoneis* (CCDE + *Encyonema*), and the family less *Encyonema* (CCDE + *Placoneis*). The posterior probabilities of these groups across phylogenetic models were 0.001-0.022, 0.41-0.75, and 0.0-0.008, respectively (Table 2.2). The data supports a grouping of the CCDE clade and *Encyonema* to the exclusion of *Placoneis*, but the monophyly of the entire Cymbellaceae and the monophyly of the CCDE clade + *Placoneis* were rejected ($PP<0.05$). A grouping of the CCDE clade and *Gomphonema* + *Gomphoneis* is better supported ($PP=0.18$-0.31) than the monophyly of the Cymbellaceae as currently defined (Table 2.2).

Table 2.2: Posterior probability of the monophyly of various taxon combinations under different models.

Group	Posterior probability[a]		
	by-gene-by-codon	by-rate	multi-matrix
Cymbopleura	0.001	0.075	0.065
Cymbopleura + *Encyonopsis*	0.002	0.050	0.039
CCDE + *Encyonema* + *Placoneis*	0.022	0.001	0.003
CCDE + *Encyonema*	0.640	0.757	0.411
CCDE + *Placoneis*	0.008	<0.001	0.005
CCDE + *Gomphonema*/*Gomphoneis*	0.308	0.217	0.186

---

[a] Frequencies were calculated from 3×10^4 trees for the by-gene-by-codon and by-rate models and 2.1×10^4 trees for the multi-matrix model.

2.5 Discussion

2.5.1 Rates, models, and partitions

The sequenced markers show substantial rate variation. Bimodal distributions of tree-independent evolutionary rates with high counts of sites in both the slow- and fast-evolving sites display this pattern neatly (Fig. 2.1). Of interest is how to model this variation and do different approaches to model ASRV produce the different phylogenetic relationships.

Although accommodating ASRV is always recommended, how to best model this variation is an area of active research (Cummins & McInerney, 2011; Rajan, 2013; Stabelli et al., 2012; Sperling et al., 2009). Modeling ASRV with a discretized Γ distribution and allowing for invariant sites in the alignment (I) are common approaches that seem essential in many cases (Yang, 1996; Sullivan & Swofford, 1997). Partitioning the alignment into subsets expected to have more homogeneous rate distributions and estimating parameters (including Γ and I) from each partition separately is another important approach (Brown & Lemmon, 2007). In many cases, however, it is not clear what is the best partitioning strategy; i.e. one that captures the variation in the data but at the same time does not overinflate the number of partitions and estimated parameters. Methods to select the most appropriate partition scheme have been developed, however, these are limited to a predefined set of partitioning strategies requiring *a priori* decisions on how the data is dissected (Lanfear et al., 2012).

An obvious strategy, employed frequently in diatom phylogenetics, is to partition the data along conventional boundaries of rate variation like codon positions in protein coding genes and secondary structure information for RNA molecules. For the present dataset, the favored conventional partitioning scheme is one that accounts for codon position in each chloroplast gene and for paired and unpaired sites in both RNA genes. This scheme yields a model with 13 partitions (nine for each codon in three chloroplast genes and four for stems and loops in SSU and LSU) and requires the estimation of 130 free parameters of the evolutionary model (assuming GTR+Γ+I for each partition). In some cases, for example, shorter protein-coding genes like *psbA* here, estimation proceeds from relatively small partitions which can result with high uncertainty in parameter estimates. It is easy to see that as datasets become larger both in terms of genes and taxa such partitioning becomes increasingly less tractable.

An alternative approach employed here is to characterize rates across the alignment before building a tree and group characters into partitions based on how fast they evolve (Cummins & McInerney, 2011; Rajan, 2013). Such approaches obviate the need for *a priori* decisions with respect to where partition boundaries are drawn and have the added benefit that characters are grouped on individual basis – based on their rate rather on their position in the gene. This means that a fast evolving site residing in a first codon position might be grouped with a site from a third codon position and that sites from different genes can be grouped in the same partition on the grounds that they evolve with a similar rate. A reasonable rate-based partitioning strategy for the five-gene dataset analyzed here is one that divides the alignment in five partitions. Although a much simpler model with 80 parameters less than dissecting by codons and secondary structure, this partitioning scheme improved the likelihood by some 1800 likelihood units (difference in mean harmonic means of log-likelihoods of by-gene-by-codon and by-rate models). Note that the number of rate-based partitions used (five in our case) is arbitrary and that it can be further optimized across a range of coarser- or finer-grained dissection of the distribution of rates. Another consideration is that with this approach the number of partitions does not necessarily need to increase as fast as it would for conventional partitioning schemes. If we were to add another gene in the dataset, we could still use the same rate cut-off values to produce a five-partition model whereas in the by-gene-by-codon approach the addition of a new protein coding gene might require three additional partitions (16 total) to accommodate each codon position.

One can also forgo partitioning in the strict sense and employ a multi-matrix model that categorizes rate variation and builds a phylogeny at the same time (Lartillot et al., 2013; Lartillot & Philippe, 2004). Here of interest are the base frequency profiles of sites and how can sites with similar profiles be grouped into categories. Each category is then modeled with a separate rate-matrix. For the present dataset, using the multi-matrix model I found a posterior mean of 130 categories indicating substantial heterogeneity in terms of equilibrium frequencies and the rate-matrix. The multi-matrix approach provides an advantage in that rate-heterogeneity or base composition differences across the gene(s) can be modeled with a minimal investigator input and potential bias. Using this model, the uncertainty of the relationships within the CCDE clade becomes even more apparent as there is no prevailing topology in the posterior distribution of trees. Contrary to reconstruction of *C. inaequalis* as sister to the remaining *Cymbopleura* + *C. aspera* + *Encyonopsis* or as sister to the another clade of *Cymbella*, the CCDE clade is a five-way polytomy in the consensus of the multi-matrix model (Fig. 2.2). This approach has the added benefit that it is more robust to long-branch attraction artefacts (Lartillot et al., 2007) (see also Chapter 1) making it an appealing alternative to partitioning.

Although the number of estimated parameters may seem less important than the inferred phylogenetic relationships, the intricacies of the phylogenetic model become more important when different models produce different topologies or when clade support values change. For the dataset analyzed here, the trees estimated from the three models are largely topologically congruent (Figs 2.2,2.3,2.4).

However, different approaches to modeling ASRV seem to have bearing on the relationships between *Cymbella* and *Cymbopleura*. Using the conventional partitioning strategy, I found that *C. inaequalis* was a strongly supported sister to a clade of *Cymbella* (Fig. 2.3). When I partitioned by tree-independent evolutionary rate, on the other hand, the same taxon was sister to a clade composed of the remaining *Cymbopleura*, *C. aspera*, and *Encyonopsis* (Fig. 2.3). In accordance, these topological fluctuations were accompanied by a drastic change in the posterior probability for the monophyly of *Cymbopleura* (Table 2.2). While the two models generally agree –there is not much evidence supporting the monophyly of *Cymbopleura*– this discrepancy in posterior probability highlights that the posterior distributions of topologies can be substantially different when rate-variation is accommodated in different ways. It is worth noting that this topological lability is associated with the area of the tree with the longest branches (*C. inaequalis*, *C. aspera*) and likely poorest taxon sampling. Appropriate modeling of rate-variation seems especially important in conditions of sparse taxonomic coverage when lineage-specific biases in base frequencies or evolutionary rate are difficult to accomodate.

2.5.2 Phylogenetic relationships

In contrast to many other groups of raphid pennate diatoms, the Cymbellales have a decent history of phylogenetic study with morphological (Kociolek & Stoermer, 1986, 1989, 1988, 1993) and molecular data (Bruder & Medlin, 2007; Kermarrec et al., 2011). In addition, phylogenetic studies of the raphid pennates as a whole (Cox & Williams, 2006; Ruck & Theriot, 2011) and groups related to the Cymbellales (Jones et al., 2005) have provided useful insight into the phylogenetic position of the Cymbellales among raphid pennates and some relationships within the order (see also Mann & Stickle 1995). This discussion sets aside species-level relationships and focuses on the monophyly of the Cymbellales, early evolution of the group, and the monophyly and relationships within the Cymbellaceae and Gomphonemataceae.

2.5.2.1 Monophyly of the Cymbellales and early evolution of the group

The Cymbellales lack a synapomorphy (Jones et al., 2005; Cox & Williams, 2006). Thus, morphological cladistics so far do not support their monophyly. In molecular phylogenies, the monophyly of the Cymbellales seems marker-dependent. SSU trees place *Anomoeoneis* as sister to *Lyrella* (Bruder & Medlin, 2007) or *Lyrella* + *Rhoicosphenia* (Chapter 1). Trees inferred from *rbcL* and combined datasets, on the other hand, place *Anomoeoneis* as strongly supported sister to the Cymbellaceae + Gomphonemataceae (Chapter 1). In addition, although *Rhoicosphenia* is susceptible to long-branch attraction making its phylogenetic placement difficult to ascertain, this genus cannot be rejected as sister to a lineage of *Anomoeoneis* + Cymbellaceae + Gomphonemataceae (Chapter 1). The molecular data, therefore, seems to support, or at least does not reject, the monophyly of the Cymbellales. It remains to be seen whether the morphological or molecular hypothesis will stand with improved taxon and character sampling. Combining the two types of data would be the most satisfactory approach.

In early studies, *Anomoeoneis* and *Placoneis* have been used as outgroups for morphological character coding (Kociolek & Stoermer, 1988). Thus, their position as sister to the cymbelloid and gomphonemoid taxa, and as early diverging lineages, has been implied but not formally tested (Kociolek & Stoermer, 1988). Mann & Stickle (1995) studied a suite of morphological and physiological characters and suggested that *Anomoeoneis* and *Rhoicosphenia* are early diverging taxa in the Cymbellales, whereas *Placoneis* sits higher up in the tree associated with members of the Cymbellaceae. This view seemed to gain some support from previous molecular trees where *Placoneis* was inferred as sister to *Cymbella* (Bruder & Medlin, 2007). However, it turned out that this result depends on taxon and gene sampling because a later study found *Placoneis* sister to *Gomphonema* (Kermarrec et al., 2011). The molecular phylogenies presented here (see also Chapter 1) support a slightly different scenario with respect to the early diverging lineages: *Anomoeoneis* diverges first and *Placoneis* second before the lineage of asymmetric taxa (Fig. 2.2). With respect to the early evolution of the group, therefore, our five-gene phylogenies are more akin to the cladogram of (Kociolek & Stoermer, 1988) where the early diverging taxa are naviculoid while asymmetric forms appear later in the evolutionary history of the group.

2.5.2.2 Monophyly of and relationships within the Cymbellaceae and Gomphonemataceae

The monophyly of the Cymbellaceae and Gomphonemataceae is not supported by either morphological or molecular data (Chapter 1; Figs 2.2,2.3,2.4; Table 2.2). The Gomphonemataceae currently includes six genera all of which have at some point been included in cladistic studies (Kociolek & Stoermer, 1988, 1989, 1993). *Gomphonema* and *Gomphoneis* are consistently recovered in the same clade and in the majority of cases *Gomphonema* is paraphyletic with respect to *Gomphoneis* (Fig. 2.4; see also Kociolek & Stoermer 1993; Kermarrec et al. 2011). Morphology places *Gomphopleura* Reichelt ex Tempere as sister to species of *Gomphoneis* Kociolek & Stoermer 1993, while molecules place *Reimeria* Kociolek & Stoermer as sister to species of *Gomphonema* (Kermarrec et al., 2011). *Didymosphenia* is the only genus consistently recovered outside the Gomphonemataceae (Fig. 2.3; see also Kociolek & Stoermer 1988; Kermarrec et al. 2011). Moving *Didymosphenia* to the Cymbellaceae, therefore, seems to provide a resolution for the problem. The Gomphonemataceae would appear to be monophyletic. Note, however, that no phylogenetic study thus far has included all genera assigned to the Gomphonemataceae together. Therefore, the validity of the Gomphonematacae, even after the necessary transfer of *Didymosphenia* to the Cymbellaceae, cannot be ascertained at the present time.

The nonmonophyly of the Cymbellaceae, at least with the limited taxon sampling available, stems from three results. First is *Didymosphenia* (see above), second is the placement of *Placoneis*, and third is the ambiguity concerning the sister taxon to the CCDE clade. Concerning *Placoneis*, the sister relationship between the *Gomphonema* + *Gomphoneis*, *Encyonema* and the CCDE clade is strongly supported as is the placement of *Placoneis* (+ *Geissleria*, see below) as their sister (Fig. 2.2). Moreover, topological hypothesis tests reject the monophyly of the Cymbellaceae as currently circumscribed as well as a hypothesis in which *Placoneis* is sister to the CCDE clade to the exclusion of *Encyonema* (Table 2.2). It seems that a transfer of *Didymosphenia* to the Cymbellaceae and of *Placoneis* outside the

Cymbellaceae, perhaps to a new family, will solve these two issues. The problem, however, might not end there. The sister relationship between *Encyonema* and the CCDE clade received equivocal support and a clade composed of the CCDE clade and *Gomphonema* + *Gomphoneis* cannot be ruled out (Fig. 2.2; Table 2.2). Dependent on the resolution of these two nodes after additional taxon and gene sampling, the restructuring of the Cymbellaceae might also involve reconsideration of *Encyonema* as a member of the family.

Cymbopleura and its relationship to *Cymbella* warrants a closer look. *Cymbopleura* was first recognized as subgenus of *Cymbella* (Krammer, 1982) and has a negative description. Krammer (2003) described it as a genus that contains slightly asymmetric or symmetric species that posses *Cymbella*-type raphe, but unlike *Cymbella* sensu stricto, lacks apical pore fields. *Cymbopleura* might be helpful in organizing Cymbellaceae species with different levels of dorsiventral asymmetry. However, since its creation, *Cymbopleura*'s position relative to *Cymbella* has been unclear and it has remained without a defining character. Whenever *Cymbopleura* is included in a phylogenetic study, both *Cymbella* and *Cymbopleura* are not monophyletic (Fig. 2.3; see also Bruder & Medlin 2007; Kermarrec et al. 2011). From a classification point of view, such relationships argue either for the creation of one morphologically extremely variable genus that will include *Cymbella*, *Cymbopleura*, and potentially *Encyonopsis* (cf. Krammer & Lange-Bertalot 1986), or further subdivisions of *Cymbella* and *Cymbopleura* into smaller genera representing more cohesive lineages. Additional taxon sampling from *Cymbella* and especially *Cymbopleura*, *Encyonopsis*, *Afrocymbella* Krammer, and so on, will likely suggest a better phylogeny-based classification scheme for *Cymbella* sensu lato.

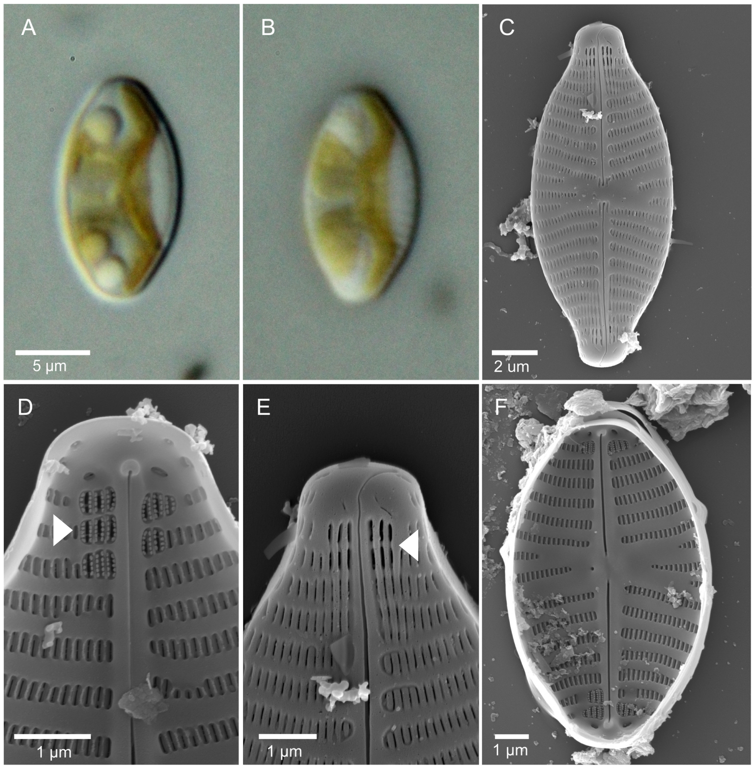
2.5.2.3 *Geissleria* and *Placoneis*

Geissleria was erected from members of *Navicula* Bory based on the morphology of the valve apices (Lange-Bertalot & Metzeltin, 1996). In *Geissleria*, unlike *Navicula sensu stricto*, the last few striae can have larger, more elongate or more densely spaced pores with inwardly pointed silica outgrowths called "annulae." The organization of the live cell, including the chloroplast structure, had not been characterized at the time of its separation from *Navicula*. Cox (1987) discussed *G. decussis* as one of the *Navicula* taxa that shares a number of similarities with *Placoneis*, but differs mainly by the striae arrangement and the presence of annulae. Similar to *Placoneis*, *Geissleria* has a single chloroplast with medially positioned center and lobes beneath both valves (Fig. 2.5). Our phylogenies showed that *G. decussis* falls inside *Placoneis* and electron microscopy confirmed that this a *Geissleria* species with typical annulae (Fig. 2.5). Given the current taxon sampling, the placement of *Geissleria* renders *Placoneis* paraphyletic (Fig. 2.2). It remains to be seen, after additional *Geissleria* species have been sequenced, if *Placoneis* will remain paraphyletic or *Placoneis* and *Geissleria* represent reciprocally monophyletic sister lineages.

2.5.3 Concluding remarks

With representatives from three families and nine genera, the phylogeny presented here is the most comprehensive effort to date to elucidate the relationships within the Cymbellales. The overall tree topology is stable across different ways

Figure 2.5: Light and scanning electron microscopy (SEM) of *Geissleria decussis* (strain: UTEX FD50). A,B. Live cell at two focal planes showing chloroplast morphology. Differential interference contrast, x1600, scale bar=5 μm. C-F. SEM of acid-digested material showing the external (C, E) and internal structure of the silica cell wall (D, F). Arrowheads in D and E point to the internal and external view of the annulae. Scale bar= 2 μm for C and = 1 μm for D-F.



to account for the heterogeneity present in the dataset. Nonetheless, alternative approaches to model the rate-variation or base frequencies across the alignment can result with different topologies. Sometimes, as in the case of *C. inaequalis*, a strongly supported placement under one partition model is shown to be potentially artefactual by a simpler, yet better-fitting, partitioning scheme (Fig. 2.3). Ways in which rate-variation is accommodated can also have implications for topological hypothesis testing as exemplified by the change in posterior probability for the monophyly of *Cymbopleura* (Table 2.2). Exploration of different ways to model the variation in the substitution process or nucleotide composition between and within markers seems of particular importance when taxon sampling is sparse like, for example, when the goal is to place a previously unsampled lineage on a phylogenetic tree.

I have shown that the current family- and genus-level classification of the Cymbellales is unnatural –the spicies-rich genera *Cymbella*, *Cymbopleura*, *Gomphonema*, and *Placoneis* are para- or polyphyletic. The observed topology is generally consistent with morphological phylogenies with comparable taxon sampling. A basal position of *Anomoeoneis* followed by the divergence of *Placoneis* and the remaining genera divided into sister lineages of mainly heteropolar and mainly dorsiventral taxa was also recovered using morphology (Kociolek & Stoermer, 1988). The molecular data, however, seems somewhat equivocal with respect to relationships within the large clade of cymbelloid and gomphonemoid species. Although cymbelloids and *Encyonema* are recovered as sister taxa, an association between cymbelloids and gomphonemoids to the exclusion of *Encyonema* cannot be discounted.

I have included merely 30% of the named genera in the Cymbellales falling short of an exhaustive sampling of lineages relevant to the early branching order and higher classification of the Cymbellales. Our phylogeny, however, provides a platform for future work in the Cymbellales as it points to areas of the tree that need attention, e.g. *Cymbopleura* and *Encyonopsis*, and identifies relationships congruent between molecules and morphology that can be used to establish a natural classification system for the Cymbellales.

2.6 Acknowledgements

I thank the members of the Department of Cell Ultrastructure at the Limnological Institute, Siberian Branch of the Russian Academy of Sciences (Irkutsk, Russia) for their warm welcome and invaluable assistance in the field and laboratory. I am especially grateful to Drs Michael A. Grachev and Yelena V. Likhoshway for their logistic support of this study. Tanya and Sergey Shishlyannikov, Asya Kostykovskaya, and Volodya Bulatov offered companionship and field assistance around Lake Baikal. Matt Ashworth and Sandra Pelc helped with scanning electron microscopy and data collection, respectively. The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources. This study was partially funded by NSF EF 062941, Blumberg fellowship, and USGS/NPS NRPP 141338.

Chapter 3

Using phylogeny to model cell size evolution in marine and freshwater diatoms

3.1 Abstract

Strategies for optimizing fitness in a dilute, competitive, and changing environment are thought to underlie cell size evolution in phytoplankton. Support for cell size as an adaptive trait comes from observed shifts in cell size distributions in response to environmental cues at geologic time scales and across environmental gradients. Physico-chemical differences between marine and fresh waters are thought to drive diatom cell size evolution in opposite directions, with larger sizes conferring benefits in marine habitats and small sizes in fresh waters. I tested this hypothesis in one lineage of diatoms, the Thalassiosirales, that spans marine and freshwaters, has a well-supported phylogeny, and whose members are relatively homogenous with respect to cell shape, growth habit, and habitat preference. A comparison of adaptive models for cell size evolution supports the hypothesis for different cell size optima between marine and freshwater habitats. The data are best explained by a model with separate selective regimes for marine and freshwater lineages. However, a scenario of stabilizing selection towards a single global cell size optimum irrespective of habitat cannot be completely discounted. Understanding the processes that shape cell size evolution in phytoplankton would ben-

efit from models that incorporate phylogeny, intrinsic properties of species (e.g., cell shape, colony formation, and motility), more specific habitat characterization, as well as genetic and genomic properties of different phytoplankton groups.[1]

3.2 Introduction

Cell size plays a central role in nearly all aspects of phytoplankton physiology, ecology, and evolution (Finkel et al., 2010). Traits ranging from nutrient acquisition to photophysiological characteristics all scale with cell size (Key et al., 2010; Finkel et al., 2010; Edwards et al., 2012), compelling researchers to acknowledge it as a "master" trait in phytoplankton (Litchman et al., 2010). In general, smaller cells have higher rates of cell division and increased efficiency in nutrient acquisition, so unicellular algae are expected to evolve picoplanktonic ($\leq 2\ \mu$m) dimensions, thereby maximizing their surface:volume ratio (Raven, 1998; Jiang et al., 2005). Paradoxically, however, phytoplankton cell size varies by some nine orders of magnitude (Litchman et al., 2009; Finkel et al., 2010). Large-celled phytoplankton are thought to have evolved in response to grazing pressure or selection for increased nutrient storage in fluctuating environments (Jiang et al., 2005; Litchman et al., 2009). Allometric models also show that large cell size can evolve in a grazer-free stationary environment when nutrient assimilation, rather than nutrient uptake, is limiting (Verdy et al., 2009). A strategy that jointly optimizes resource uptake and predator defense by using a non-limiting resource (e.g., silica

---

[1] This chapter was publshed as Nakov, Theriot and Alverson, 2014, Limnol. Oceanogr. 59: 79-86. Author contributions: Andrew Alverson helped draft the manuscript; Edward Theriot, supervisor.

in diatoms) to increase cell size could also contribute to the vast size variation in diatoms (Thingstad et al., 2005). Non-adaptive hypotheses suggest that intrinsic properties, such as genome size, could underlie some of the observed variation in cell size (Connolly et al., 2008), even in the face of extrinsic forces selecting for smaller cells, when the size of intracellular compartments restricts further cell size reduction (Raven et al., 2005).

Shifts in cell size distributions of phytoplankton communities coincident with changing environmental conditions set the foundation for adaptive models of cell size evolution (Finkel et al., 2005, 2007; Litchman et al., 2009). For example, physico-chemical differences between marine and freshwater environments are thought to play a determinant role in the evolution of phytoplankton cell size between these two habitats (Litchman et al., 2009). Indeed, marine diatoms are, on average, an order of magnitude larger than their freshwater counterparts (Litchman et al., 2009). To account for this difference, a model was developed that accounted for cell sinking rate and empirical allometries of nitrogen (N) and phosphorous (P). Under this model, the disparity in cell size between marine and freshwater diatoms was thought to reflect differences in: (i) prevailing nutrient limitation between the two environments (N in marine, P in freshwater), (ii) frequency of nutrient fluctuations, with intermittent N pulses in marine environments selecting for large cell size, and (iii) differences in mixed-layer depth that, together with sinking rate, select for smaller size in freshwater diatoms (Litchman et al., 2009). The model also predicts that in conditions of intermediate frequency N pulses, the marine environment can support multiple evolutionarily stable cell size optima. That is, large-

and small-celled species can coexist by adopting different strategies for nutrient uptake and utilization. Following a pulse of N, small-celled species respond with rapid population growth, whereas the storage ability of large-celled species allows them to sustain growth for longer periods of time (Litchman et al., 2009).

I used phylogenetic comparative methods to study the evolution of cell size in the diatom order Thalassiosirales, a lineage that spans marine and freshwater habitats and has a well-supported phylogenetic hypothesis (Alverson et al., 2007) and chronology (Alverson, 2013). I found that the observed difference in size between marine and freshwater diatoms (Litchman et al., 2009) applies even at this more narrow phylogenetic scale. The data are best fit by a model with separate selective regimes in which marine and freshwater lineages have experienced similar strengths of selection, driving cell size towards their respective size optima. Aside from incorporating phylogeny, I identify several additional factors that merit consideration when modeling the evolution of cell size in phytoplankton, especially diatoms.

3.3 Material and Methods

3.3.1 Cell volume

I restricted our analyses to the diatom order Thalassiosirales for two main reasons. An important sampling bias became evident after compiling a cell volume dataset for taxa included in large-scale reconstructions of the diatom phylogeny (Theriot et al., 2010). Namely, freshwater species were over-represented in the pennate lineage, and marine taxa were over-represented in the paraphyletic "centric" portion of the tree. This is problematic because centric diatoms are generally larger than pennates (Finkel et al., 2005). It is possible that cell shape constrains overall cell volume in some way (Naselli-Flores et al., 2007), so the overall morphological (cylindrical) and habitat (planktonic) uniformity within Thalassiosirales allowed us to control (more or less) for these variables and focus more specifically on cell volume.

I compiled cell volumes for a total of 52 species, a subset of those represented in published phylogenies (Alverson, 2013). As appropriate for phylogenetic comparative analyses, in the case of multiple conspecific accessions I removed all but one of these accessions. Gathering reliable size information for unidentified taxa was problematic therefore these were also removed from the analyses. For marine species, I compiled cell volume data from the Helsinki Commission Phytoplankton Expert Group (HELCOM PEG) dataset (http://www.helcom.fi/projects/on_going/peg/en_GB/biovolumes/) or (Leblanc et al., 2012). For freshwater species, I used data from the National Water Quality Assessment (NAWQA) dataset (http://diatom.ansp.org/autecology/) assembled from rivers across the United States. For the remaining species, cell volume was calculated using size ranges reported in the primary literature. The pervalvar axis length (the height of a cylindrical cell) is rarely documented, making it difficult to calculate cell volume using real measurements. I therefore adopted a "hidden dimension" approach where cell height was assigned a value relative to the diameter. A prevailing ratio used to calculate cell height in a number of Thalassiosira species is $height = diameter \times 0.5$

(Leblanc et al., 2012). I applied this same ratio to calculate the pervalvar height of all *Cyclotella*, *Thalassiosira*, and *Stephanodiscus* species for which cell volumes were not available. The relationship between cell diameter and cell height varies considerably across *Skeletonema* species, with larger heights than diameters in some species, and the opposite pattern in others (Sarno et al., 2005, 2007). I therefore made a simplifying assumption that cell height equals cell diameter in *Skeletonema* species for which pervalvar measurements were unavailable. Considering all cells as cylinders, I calculated the minimum and maximum cell volumes for each species. Although analyses based on average cell volumes would have been preferable, I were unable to do so for two reasons. Cell volume data from the Leblanc et al. (2012) dataset, and the additional calculations performed here, were based on the minimum and maximum reported diameters. In the HELCOM PEG dataset, volumes were split into size classes with no information as to how many individuals per size class were measured. Averaging was therefore impossible in both of these cases, so our analyses were based on minimum and maximum cell volumes for each species. Cell volumes were $\log_{10}$-transformed prior to all statistical analyses (O'Meara et al., 2006).

3.3.2 Phylogeny

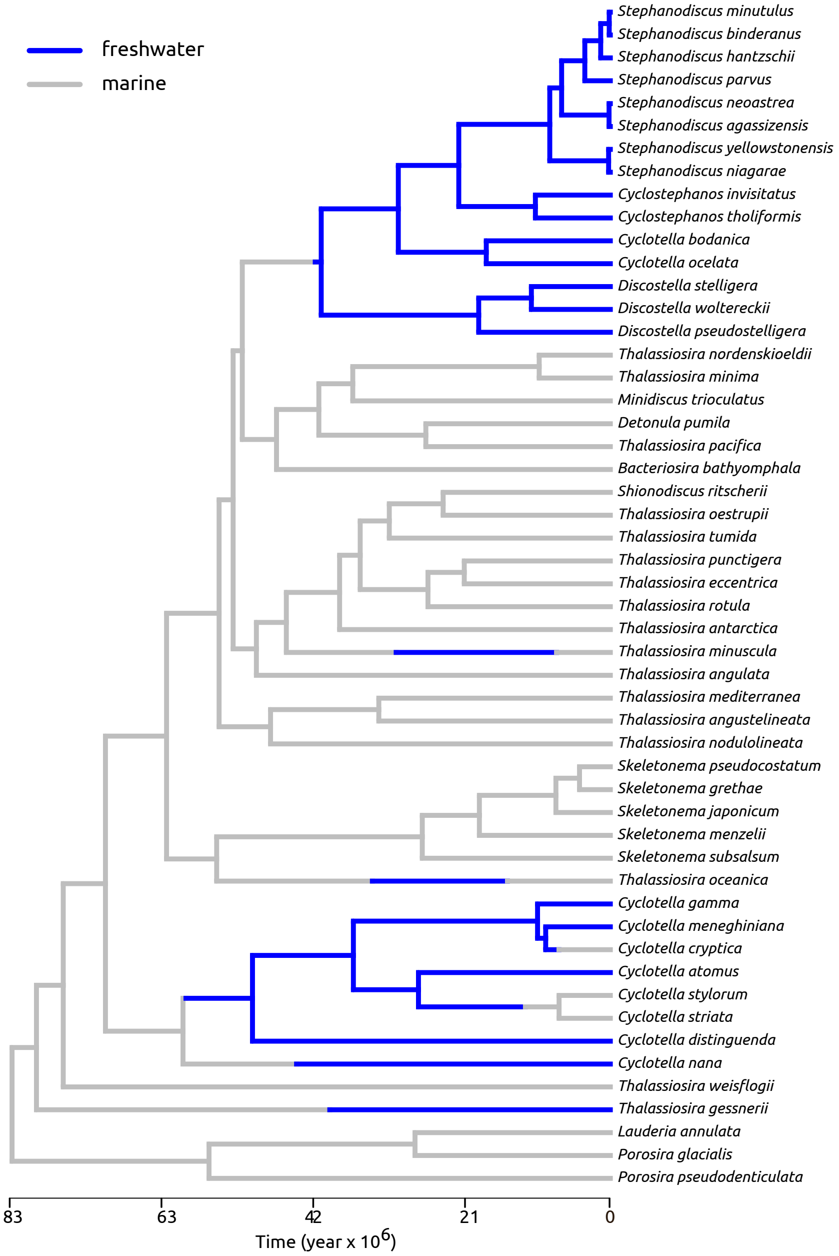
All comparative analyses of cell volume data used the time-calibrated phylogeny from Alverson (2013). The original tree topology was based on Bayesian analysis of a combined dataset of two plastid and two nuclear markers (Alverson et al.,

2007) and showed strong support for: (i) a marine common ancestor for the entire lineage, (ii) two freshwater colonizations that led to substantial species diversifications, and (iii) several reverse, freshwater-to-marine transitions within one of the freshwater lineages (Fig. 3.1). The time-calibrated phylogeny used here showed that the two major freshwater colonizations likely occurred in series, first in the Paleocene then later in the Eocene Alverson (2013). Complete analytical details for the original phylogenetic tree are available in (Alverson et al., 2007), and details of the molecular clock analyses are available in Alverson (2013).

3.3.3 Tests for phylogenetic signal and size differences between marine and freshwater taxa

To assess phylogenetic signal in minimum and maximum cell volumes I used a likelihood ratio test to determine whether the branch-scaling parameter (λ) was significantly different from zero (Pagel, 1999; Revell, 2010). The λ parameter scales the internal branches of a phylogeny to reduce, in effect, the overall amount of shared evolutionary history between species. The optimal value of λ is obtained by fitting a Brownian motion (BM, random walk with a single mean and variance) model of evolution to the trait (cell volume) and phylogeny. Maximum likelihood estimates (MLEs) of λ that are not significantly different from zero indicate that the data are best explained by an unresolved phylogeny, i.e., the species' trait values can be treated as independent data points. At the opposite extreme, when the MLE of λ is not significantly different from one, the trait data are best explained by the hierarchical structure of the phylogeny with unscaled branches (Pagel, 1999; Revell, 2010). For cases in which λ is significantly greater than zero, the correla-

Figure 3.1: Phylogenetic relationships within the diatom order Thalassiosirales, modified from Alverson (2013). Branch colors correspond to the selective regimes for the two-optimum OU models and are derived from stochastic character mapping of habitat (freshwater-marine. This is 1 of 492 stochastic maps used in the analyses.



tion between species' trait values is greater than expected by chance alone, i.e., the trait exhibits phylogenetic signal and this must be accounted for in the model. To assess the effect of phylogenetic signal on the comparison of cell size distributions between marine and freshwater diatoms, I fitted two ANOVAs, testing both minimum and maximum cell volume for marine vs. freshwater species. One model assumed independence of data points, modeling cell size evolution as a random walk on a star phylogeny with separate means for marine and freshwater taxa (λ=0; Star_{min} and Star_{max}; equivalent to a Student's t-test). The other incorporated the tree and the MLE of λ value simultaneously accounting for phylogenetic signal (λ=MLE; Tree_{min} and Tree_{max}). I used the small-sample Akaike Information Criterion (AICc) to compare the two models, penalizing for the increased number of parameters in the phylogenetic model.

3.3.4 Adaptive models for cell size evolution in marine vs. freshwater environments

I used methods based on the Ornstein-Uhlenbeck (OU) process (Hansen, 1997) to model the evolution of cell size in an adaptive framework. In a phylogenetic context, the OU process is often applied to model stabilizing selection towards the optimal value of a trait (Hansen, 1997). In these models, the change of a trait through time is controlled by a constant representing the "pull" (α) of a trait value towards its optimum value (θ), also interpreted as the strength of selection or rate of adaptation (Hansen, 1997; Butler & King, 2004; Beaulieu et al., 2012). Another constant captures the deviation of the trait value from the optimum (σ^2), which is interpreted as the rate of stochastic motion or, more simply, the rate of evolution

(Hansen, 1997; Butler & King, 2004; Beaulieu et al., 2012). The latter is equivalent to the variance of a Brownian motion (random walk) process. Under the OU model, a quantitative trait evolves in small increments determined by variance in the random walk process (σ^2), but selection simultaneously pulls the trait value towards an optimum (θ) with the strength of that pull determined by an attractor constant (α). In a phylogenetic framework, OU methods have been extended to allow modeling scenarios that include multiple selective regimes and corresponding trait optima (Butler & King, 2004). This is achieved by a priori "painting" the branches of a phylogeny based on hypothesized location(s) of shifts in the selective regime. One can also relax the assumptions of a single rate of stochastic motion (σ^2) and adjust the strength of selection (α), permitting species in each selective regime to proceed towards an optimum trait value at their own pace (Beaulieu et al., 2012).

For the evolution of cell size in marine and freshwater diatoms, I compared the fits of two BM and five OU models:

– BM –null BM model with a single trait mean and variance of random walk (σ_g^2);

– BMS –BM model that divides the chronogram (Fig. 1) into marine and freshwater clades, allowing a different variance of random walk for each (σ_m^2, σ_f^2) (O'Meara et al., 2006);

– OU –single-optimum OU model with one parameter for the variance of random walk (σ_g^2) and strength of selection (α_g) towards a global optimum (θ_g);

– OUM –OU model with separate cell size optima for marine and freshwater lineages (θ_m, θ_f) but global σ_g^2 and α_g parameters for the different selective regimes (Butler & King, 2004);

– OUMV –two-optimum OU model (θ_m, θ_f) with separate random walk variances for marine and freshwater selective regimes (σ_m^2, σ_f^2) and one global selection parameter (α_g);

– OUMA –two-optimum OU model (θ_m, θ_f) with a separate strength of selection parameter in each selective regime (α_m, α_f) and a global random walk parameter (σ_g^2), and;

– OUMVA –two-optimum OU model with separate σ^2 and α for each selective regime.

The indices in model names refer to: M –means, V –variances, and A –attractors. So the OUM model has separate means, or trait optima, for each selective regime; and the OUMVA model has separate trait optima, variances and attractors for each selective regime. The indices in parameter names refer to: m –marine, f –freshwater and g –global, so θ_m, θ_f, θ_g represent the marine, freshwater, and global cell size optima, respectively.

Selective regimes (i.e., marine vs. freshwater) for internal nodes on the phylogeny were assigned using stochastic character mapping (Huelsenbeck et al., 2003) as implemented in the R package "phytools" (Revell, 2012). I sampled 500 stochastic maps with a model of equal rates of marine-to-freshwater and freshwater-to-marine transitions. To accommodate uncertainty in ancestral state reconstruction, the BM and OU models described above were fit to each of the sampled character histories. Optimization failed in 8 cases, so I present averages of parameter estimates and likelihoods from 492 stochastic maps. Models were fit with the R package "OUwie" (Beaulieu et al. 2012). Data manipulation and additional analyses were done with functions from the R packages "ape", "geiger", and "phytools" (Paradis et al., 2004; Harmon et al., 2008). Models were compared using AICc scores.

3.4 Results

3.4.1 Cell size in marine vs. freshwater Thalassiosirales

Across all species, minimum and maximum cell volumes varied by five and six orders of magnitude, respectively (Fig. 3.2). Save some large *Stephanodiscus* and *Cyclotella* species, freshwater taxa (n=21) were concentrated towards the lower end of the size range for both minimum and maximum cell volume (Fig. 3.2). Marine taxa (n=31) showed the opposite trend (Fig. 3.2). Nevertheless, cell volume distributions for marine and freshwater taxa were broadly overlapping, especially for minimum cell volume (Fig. 3.2). For minimum cell volume, the means of marine and freshwater taxa were not significantly different regardless of the test method (standard or phylogenetic one-way ANOVA; Table 3.1). Standard ANOVA detected a significant difference in maximum cell size between marine and freshwater taxa. This difference was marginally significant when I accounted for phylogenetic signal (Table 3.1).

Table 3.1: Comparison standard and phylogenetic one-way ANOVA for the size difference between marine and freshwater Thalassiosirales.

Model	Δ means	p-value	λ	k[a]	ln L[b]	AICc	ω_i[c]
Star_{max}[d]	0.83	<0.01	0	3	-77.72	161.93	0.01
Tree_{max}[e]	0.80	0.06	0.74	4	-72.21	153.28	0.99
Star_{min}	0.47	0.17	0	3	-83.23	172.96	0.03
Tree_{min}	0.76	0.11	0.66	4	-78.48	165.82	0.97

---

[a] Number of parameters in model.

[b] Natural logarithm of likelihood.

[c] Akaike weight.

[d] Based on an unresolved phylogeny.

[e] Based on tree with λ transformation.

3.4.2 Ornstein-Uhlenbeck models for cell size evolution in marine vs. freshwater Thalassiosirales

OU models, which include an adaptive component (strength of selection, α), provided a substantially better fit than single- (BM) and double-rate (BMS) random-walk models (Table 3.2). The OUM model was most commonly favored across the set of stochastic maps of habitat for both the minimum and maximum cell volume datasets (Table 3.2, Fig. 3.2). Under this scenario, marine and freshwater lineages are modeled as evolving towards separate cell size optima but with the same parameters for rate of stochastic motion (σ^2) and strength of selection (α) in both selective regimes (Table 3.3). There is, however, some uncertainty in model selection for both the minimum and maximum cell volume datasets (Table 3.2, 3.2). The average relative likelihood (or AICc weight, ω_i) of the OU model reaches 0.11 and 0.43 for maximum and minimum cell volume, respectively (Table 3.2). The single optimum model therefore has to be treated as plausible to some extent, especially for minimum cell volume.

Across the sampled character histories, optimization of more complex models that relax the assumption of common rate of stochastic motion or strength of selection (OUMV, OUMA, OUMVA) generally failed the convergence diagnostics. This was somewhat expected given the modest size of the dataset. Simulation studies have also shown difficulties in parameter optimization and large confidence intervals for small sample sizes in the more complex OU models (Beaulieu et al., 2012). Tests of the applicability of these biologically intriguing scenarios to the evolution of diatom cell size will have to wait until larger, or otherwise more suitable, cell size datasets with matching phylogenies are compiled.

Table 3.2: Model selection for the evolution of maximum and minimum cell volume in the Thalassiosirales.

Model	k[a]	Mean ln L[b]	Mean AICc	Mean ω_i[c]	% favored[d]
		Maximum cell $\log_{10}$ volume (μm^3)			
BM	2	-110.54	225.33	0.00	0.00
BMS	3	-89.22	184.94	0.00	0.00
OU	3	-79.53	165.55	0.10	4.67
OUM[e]	4	-76.19	161.24	0.90	95.33
		Minimum cell $\log_{10}$ volume (μm^3)			
BM	2	-121.04	246.34	0.00	0.00
BMS	3	-95.78	198.06	<0.001	0.00
OU	3	-82.71	171.93	0.43	0.61
OUM	4	-81.49	171.84	0.57	99.39

---

[a] Number of parameters.

[b] Natural logarithm of likelihood.

[c] Akaike width.

[d] Percent of times when a model had the lowest AICc score per stochastic map.

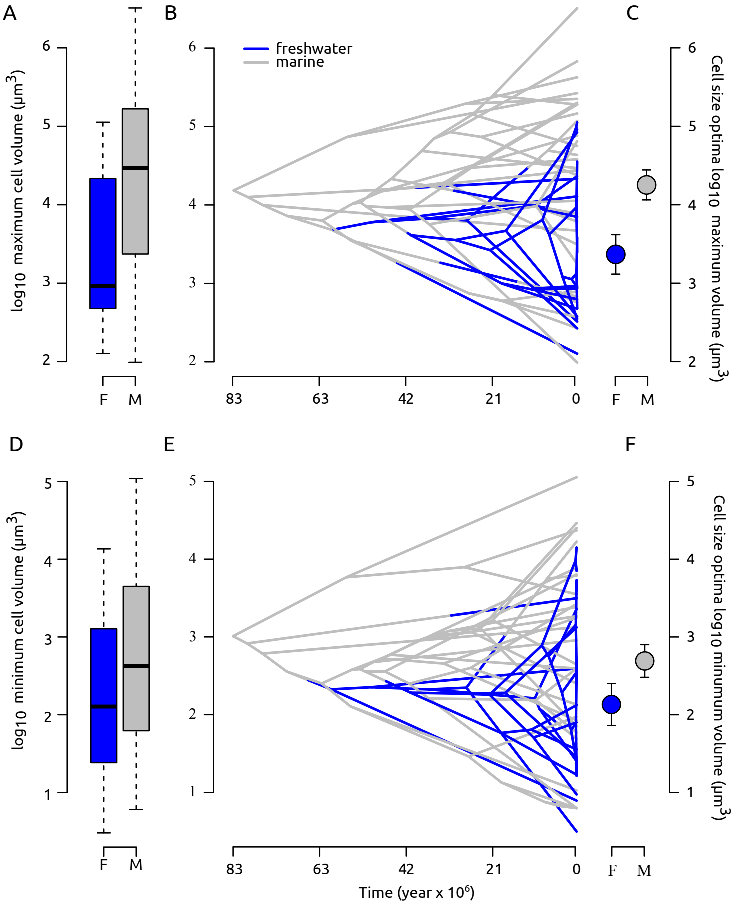
[e] Selective regimes were assigned through stochastic character mapping of habitat (freshwater-marine) over the chronogram. Values are averages from optimizations over 492 ancestral reconstructions of habitat. Results from models OUMV, OUMA, and OUMVA are omitted because of unreliable optimizations.

3.5 Discussion

Cell size is a so-called master trait underlying many important aspects of phytoplankton physiology, ecology and evolution (Finkel et al. 2010). Diatoms, and phytoplankton more generally, display striking patterns of cell size evolution through geologic time (Finkel et al. 2005, 2007) and across environments, including the marine-freshwater gradient (Litchman et al. 2009; Edwards et al. 2012). The size disparity between marine and freshwater diatoms is thought to reflect contrasting selection from the different environmental conditions in marine and fresh waters (Litchman et al. 2009). Environmental conditions in freshwaters (prevalent P limitation, shallower mixed layer depth) are thought to select for smaller overall size (Litchman et al. 2009). In contrast, the marine environment (prevalent N limitation, deep mixed layer depth) is thought to select for larger size or the coexistence of species with small and large sizes (Litchman et al. 2009). This model was based in part on the observation that marine diatoms are significantly larger than freshwater diatoms, an inference obtained with statistical tests that assumed independence of data points (Litchman et al. 2009). Across many groups of animals, the evolution of body size exhibits strong phylogenetic signal, meaning that closely related species share similar sizes based solely on their shared ancestry (Blomberg et al., 2003). If unaccounted for, phylogenetic signal has the potential to confound the strength or significance of observed trait-by-trait or trait-by-environment patterns (Whitney & Garland, 2010). Until now, phylogenetic signal in diatom cell size has not been quantified and its potential to affect inferences about the evolution of cell size across marine and fresh waters was unknown.

I used a phylogenetic framework to study the evolution of cell size in Thalassiosirales, a diatom lineage with considerable diversity in marine and freshwater habitats (Alverson et al. 2007). Similar to Litchman et al. (2009), the distribution of cell volumes in Thalassiosirales varies considerably (by >6 orders of magnitude), and marine species are, on average, an order of magnitude larger than freshwater species (Fig. 3.2). I found, however, that cell size exhibits phylogenetic signal, i.e., the shared ancestry of closely related species results in cell sizes that are more sim-

Figure 3.2: Cell volume evolution in the Thalassiosirales through time and cell size optima for marine and freshwater lineages. A –Distribution of maximum cell volume ($\log_{10}$) in marine and freshwater taxa. B –Phenogram of maximal cell volume for the Thalassiosirales. The topology, scale, and colors of the phenogram are identical to Fig. 3.1. The vertical position of each terminal branch on the y-axis represents the cell volume for that species. The vertical position of each internal node on the y-axis represents the ancestral reconstruction of cell size based on phylogenetic independent contrasts (Felsenstein, 1985). C –Maximum likelihood estimates of cell size optima (θ_f, θ_m) and standard errors from the favored OUM model averaged over 492 stochastic character maps of habitat (marine-freshwater). The marine optimum is about one order of magnitude larger than the freshwater. D, E, F –Same as A, B, C for minimum cell volume. Note that the phenogram in E represents a slightly different reconstruction of the ancestral history of habitat but is otherwise identical to B.



ilar than expected by chance alone. The phylogenetic models therefore provided a substantially better fit while retaining the power to capture the presumably adaptive difference in cell size between marine and freshwater taxa (Table 3.1). Indeed, a comparison of adaptive models for the evolution of cell size in marine and freshwater Thalassiosirales favored a model of stabilizing selection towards separate marine and freshwater optima (Table 3.2, 3.3). A simpler scenario of stabilizing selection towards a single global optimum, however, cannot be completely ruled out, especially for the minimum cell volume dataset, whereby the relative likelihood of single- and double-regime models were similar (Table (Table 3.2). This modestly sized dataset appears, therefore, to support the hypothesis of contrasting size evolution between marine and freshwater Thalassiosirales. Larger datasets will be necessary to reliably assess the applicability of more complex adaptive evolutionary scenarios that allow not only separate optimal sizes between the two selective regimes, but differing evolutionary forces operating in marine vs. fresh waters.

In addition to incorporating phylogeny, further considerations need to be made when modeling the evolution of phytoplankton cell size. In diatoms alone, a number of lineage-specific factors likely constrain, to varying extents, the overall magnitude of cell size variation. Lineages can be planktonic or benthic; live solitary or in colonies; famously take on any number of shapes; and can be actively motile or completely sessile. These traits occur in virtually all combinations and are represented in both marine and freshwater habitats. The striking contrast between, for example, a marine planktonic, solitary, sessile, and cylindrically shaped species of *Coscinodiscus* vs. a marine planktonic, spear-shaped, filamentous, and motile *Pseudo-nitzschia* illustrates the potential of intrinsic shape features to affect size, resulting in different responses to natural selection on cell size. The diversity of cell shapes and sizes in diatoms suggests that different lineages almost certainly have adopted different cell-size-related strategies to optimize their fitness.

Diatom genomes appear to evolve rapidly in many respects, and polyploidization may be common amongst closely related species (Von Dassow et al., 2008) and even within morphospecies (Koester et al., 2013). Doubling the amount of generic material generally requires an increase in the size of the nucleus and intracellular membrane system, potentially driving upward the lower size limit of a species (Raven et al., 2005). Connolly et al. (2008) found a positive correlation between cell and genome size in diatoms suggesting that non-scalable factors, like genome size, might contribute to cell size variation in diatoms. Thus, to the extent cell size correlates positively with shifts in genome size, a species' cell size will likely be balanced by neutral and/or adaptive processes between the minimum size, specified by non-scalable factors, and maximum size, driven by natural selection favoring smaller or larger size (Raven 1998; Thingstad et al. 2005; Litchman et al. 2009).

Table 3.3: Maximum likelihood estimates of cell size optima (θ), strength of selection (α), and rate of stochastic motion (σ^2) for freshwater and marine species of Thalassiosirales.

Parameter[a]	Max cell $\log_{10}$ volume (μm^3)		Min cell $\log_{10}$ volume (μm^3)	
	OU	OUM	OU	OUM[b]
θ_g	3.94 (3.94-3.94)	na[c]	2.48 (2.48 -2.48)	na
θ_f	na	3.37 (3.33-3.37)	na	2.13 (2.13-2.13)
θ_m	na	4.27 (4.27-4.34)	na	2.69 (2.69-2.69)
α_g	78.22 (78.22-78.22)	86.17 (5.8-90.13)	117.3 (116.51-118.28)	109.12 (109.09-109.14)
σ_g^2	202.97 (202.97-202.97)	191.59 (13.63-200.36)	338.68 (335.73-342.7)	300.76 (300.70-300.84)

---

[a] Averaged over 492 ancestral reconstructions of habitat (freshwater-marine). The 2.5% and 97.5% quantiles around the mean from 492 optimizations are given in parentheses. Only the results from the best two models, OU and OUM are shown.

[b] A few optimizations with lower likelihood resulted with much lower parameter estimates so median values are presented for α_g and σ_g^2.

[c] Not applicable for particular model.

Diatom lineages are nonrandomly distributed across marine and freshwater habitats. Although planktonic pennate diversity is more-or-less evenly distributed between marine and fresh waters, the non-pennate ("centric") lineages are predominantly marine. The genus *Aulacoseira* and the freshwater Thalassiosirales are the only extant centric lineages with substantial species diversity in freshwaters. Diatom-wide comparisons of cell size between the two habitats are, therefore, based on a mixed assemblage of pennate and non-pennate species in the marine habitat vs. a predominantly pennate freshwater assemblage. Illustrative of this, only 10% of the taxa in the freshwater NAWQA dataset are non-pennates, compared to 55% of the species in the marine HELCOM dataset. Within the marine environment, there is a long-term historical (i.e., throughout the Cenozoic) trend of smaller average size in pennate vs. centric diatoms (Finkel et al. 2005). Finally, it is worth noting that in some isolated cases of freshwater colonization by members of otherwise exclusively marine clades, the colonizers have retained a larger-than-average cell volume compared to other species in the freshwater community (e.g., *Pleurosira laevis*).

In this study, I attempted to control for these potentially confounding factors by choosing a model clade that is fairly uniform with respect to shape (cylindrical) and habitat (planktonic) and found cell size does exhibit phylogenetic signal. Thus, cross-species studies of phytoplankton cell size evolution that do not account for phylogeny likely violate important assumptions of statistical tests, potentially biasing the strength and significance of the inferred patterns. In Thalassiosirales, I found support for the hypothesis of separate cell size optima for marine and freshwater taxa. However, a scenario of stabilizing selection towards one global optimal size cannot be ruled out. Addition of colony formation, colony type, horizontal (coastal vs. open waters) and vertical distribution in the water column, and potentially many other factors will likely improve models of cell size evolution in this important clade of marine and freshwater diatoms, and in phytoplankton more generally.

3.6 Acknowledgements

Jeremy Beaulieu helped troubleshoot statistical analyses. Elena Litchman kindly commented on this work. This study was partly supported by NSF EF 0629410 (ECT) and the Blumberg fellowship (ECT).

Chapter 4

Evolution of habitat preference and growth form across diatoms

4.1 Abstract

I characterized the evolutionary history of habitat preference (benthic - planktonic), growth form (solitary - colonial), and their interaction on a multi-gene phylogeny encompassing a nearly complete sampling of the extant order-level diversity of diatoms. The results support markedly different evolutionary histories for these two traits. Habitat preference appears to be a slowly evolving trait, conserved at the level of large clades with infrequent traversals between the plankton and benthos. Transitions to the planktonic form seem to have been accompanied by increased morphological complexity (through the gain of projections, keels, etc.), increased cell size, and, in some cases, transition to colonial growth form – strategies that might be advantageous for life as a suspended particle. Growth form, on the other hand, has a dynamic evolutionary history, with numerous shifts between the solitary and colonial growth habits dispersed across the phylogeny. Our modeling approach revealed that the evolution of growth form is time-heterogeneous with slow transitions in some clades (e.g. *Coscinodiscus*) and fast transitions in others (e.g. pennate diatoms). Tests for coordinated evolution showed that the evolution of growth form is not dependent on habitat and that the

chances of traversing between habitats do not hinge upon species' growth form. Our findings provide a platform for future work focused on clades where transitions between phenotypes are frequent as a means of addressing the mechanisms underlying diatom species and functional diversity[1].

4.2 Introduction

Diatoms are an exceptionally diverse lineage of predominantly photoautotrophic heterokonts (Mann & Vanormelingen, 2013) responsible for substantial portions of the global primary production and atmospheric carbon removal (Nelson et al., 1995; Hopkinson et al., 2011). They have colonized the plankton and benthos, are frequently dominant in communities of lotic and lentic systems, and span the salinity barrier with substantial species diversity in each of these habitat types (Spaulding & Kociolek, 2000; Alverson et al., 2007; Vyverman et al., 2007). Their rise to dominance in many aquatic environments is a result of a combination of genetic, physiological, and morphological factors (Falkowski et al., 2004; Armbrust, 2009) shaped over an evolutionary history spanning ca. 350 million years (Brown & Sorhannus, 2010).

One remarkable feature of diatoms is their extraordinary diversity in growth form. They range from simple spheroid unicells to complex three-dimensional colonies comprised of hundreds of cells and reaching macroscopic sizes. The mechanisms of colony construction are varied as well. Diatoms form colonies

---

[1] This chapter was submitted for peer review as Nakov, Ashworth, and Theriot: Evolution of habitat preference and growth form across diatoms. Author contributions: Matt Ashworth is the principal contributor of new data for the phylogeny; Edward Theriot, supervisor.

through modified features of the silica cell wall, chitin threads, and an array of extracellular mucilaginous secretions in the form of pads, stalks, tubes, or sheets (Round et al., 1990). This diversity in growth form has a functional role. The combination of growth form (e.g. solitary or colonial) and habitat occupancy (e.g. planktonic or benthic) approximates, albeit roughly, a diatom's ecological niche. Small-celled species that grow attached to a substrate via a mucilaginous pad, for instance, are early colonizers of benthic mats and are adapted for resistance to scouring from water currents (Hoagland et al., 1982; Hoagland, 1983; McCormick & Stevenson, 1991; Johnson et al., 1997). Long filamentous or branched colonies, on the other hand, tend to establish later in the succession, when the mat is crowded and cells improve access to nutrients and light by rising above the boundary layer (Hoagland et al. 1982; Hoagland 1983; McCormick and Stevenson 1991). Growth form is similarly consequential in the plankton because species' sinking rate and vertical position in the water column is affected by colony morphology and symmetry (Padisák et al., 2003; Reynolds, 2006). Thus, the amount of light and nutrients available to a cell living in a stratified environment is at least partially influenced by the ability to form colonies and their properties. Combining these considerations with the benefit of increased organism size as a strategy for defense against predation (Yokota & Sterner, 2010), colony formation and type become chief adaptive traits with wide-ranging consequences for life in the aquatic environment.

It seems plausible, therefore, to hypothesize that the combined influence of environmental factors and species interactions have guided lineages towards alternate growth forms in the strikingly different open water versus the littoral zone habitats. Round et al. (1990, pp. 29), for example, argued: "There is also no doubt that colonial organization has been subject to strong selection in particular habitats, in relation to attachment, light and nutrient capture ...,the control of sinking rate, etc." The independent acquisition of colonial growth form, and indeed the same colony morphology, in distantly related diatom lineages can be viewed as support for this assertion. However, despite the recognized importance of studying the evolution of habitat preference and growth form, their evolutionary histories and any correlates to their distributions have seldom been investigated. Kooistra and co-workers identified lineages that transitioned to the plankton and discussed adaptations that may have accompanied such shifts (Kooistra et al., 2007, 2009). They also highlighted isogamous sexual reproduction as an obstacle for planktonic life style in pennate (bilaterally symmetrical) diatoms that has been successfully circumvented in few lineages (Kooistra et al. 2009). Research in this area of diatom evolution, however, seems to have stalled and the evolutionary histories of habitat preference and growth form have not been evaluated in a modeling framework.

Recent efforts in reconstructing the diatom phylogeny (Theriot et al., 2009, 2010; Ashworth et al., 2013) are bringing us close to a nearly complete sampling of the major extant lineages of diatoms providing the opportunity to examine functional trait evolution in previously unattainable detail. Moreover, advances in methodology of modeling discrete traits that relax assumptions of rate-constancy across a phylogeny (Beaulieu et al., 2013) allow evaluation of more realistic evolutionary scenarios. Here, I take advantage of these opportunities with the aim to characterize the evolutionary histories of habitat preference (planktonic vs. benthic) and growth form (solitary vs. colonial) across diatoms. Our work is set in a broadly sampled multi-gene phylogeny representative of 80% of order–level diversity of diatoms. A comparison of a range of models, assuming both constant and variable transition probabilities, supported a simple one-parameter model for the evolution of habitat preference and a time-heterogeneous model with variable rates across the tree for the evolution of growth form. When analyzed in combination, models of independent evolution performed better, suggesting that probabilities of traversing between phenotypes in one trait are independent of the state of the other. Our results highlight habitat preference as a slowly evolving trait conserved at the level of large clades and reveal a dynamic evolutionary history of growth form with varying pace of transitions between the solitary and colonial state.

4.3 Material and Methods

4.3.1 Trait data and phylogenetic trees

The dataset analyzed here consists of 281 diatom taxa capturing most major lineages of extant diatoms with representatives from ca. 80% of described orders. The sister lineage to diatoms, *Bolidomonas* Guillou and Chrétiennot-Dinet, was used as outgroup. I coded each species for habitat preference: planktonic (0) or benthic (1) and growth form: solitary (0) or colonial (1) as reported in the primary literature or from personal observations. I reconstructed the phylogeny of the aforementioned taxa using a three-gene nuclear and chloroplast gene dataset as in previous studies (Theriot et al. 2010; Ashworth et al. 2013). The most likely tree topology was inferred from 1008 maximum likelihood (ML) optimizations each starting from a parsimony tree in RAxML v.7.4.2 (Stamatakis, 2006). Clade support values were assessed through 10^3 nonparametric bootstrap replicates using the rapid bootstrap algorithm (Stamatakis et al., 2008). The phylogram with highest likelihood was converted to a relative-time chronogram with a root age of 100 time units using penalized likelihood as implemented in the R package "ape" (Sanderson, 2002; Paradis et al., 2004; R Development Core Team, 2013). To accommodate phylogenetic uncertainty, in addition to the "best tree", downstream analyses were also performed with 100 trees sampled at random from the 1008 optimizations. Newly-generated sequence data were deposited in GenBank (KJ577839-KJ577944).

4.3.2 Individual traits

To model the evolution of habitat preference and growth form individually I used time-homogeneous and time-heterogeneous stochastic Markov models. These two classes of models differed based on the assumptions concerning the variation of transition probabilities between character states across the phylogeny (Beaulieu et al. 2013). In the time-homogeneous models, transitions ("forward" = $0 \rightarrow 1$ and "backward" = $1 \rightarrow 0$) are constant across the entire phylogeny (Pagel, 1994). The time-heterogeneous models, on the other hand, allow different portions of the phylogeny to have different forward and/or backward transition rates. This is achieved by creating separate rate classes for slow (S) and fast (F) transition probabilities accommodating the possibility that particular lineages can have accelerated or decelerated rates of evolution relative to other portions of the tree (Beaulieu et al. 2013). Any number of rate classes is possible. However, I restricted our analyses to models with two rate classes (S and F) due to the modestly sized dataset and issues with parameter estimation from overly complex models. Heretofore, the terms "forward" and "backward" are used for convenience and do not imply transitions between ancestral and derived states.

I was interested in two types of models: those in which the forward and backward transitions are equaly probable ($0 \rightarrow 1 = 1 \rightarrow 0$, "symmetric") and those that relax this assumption ($0 \rightarrow 1 \neq 1 \rightarrow 0$, "asymmetric"). Thus, for the time-homogeneous class we have two models referred to as symmetric (number of parameters, k=1) and asymmetric (k=2). The most complex time-heterogeneous model considered here has eight parameters corresponding to the transition rates between character states in different rate classes and the transitions between rate classes in alternate character states (Beaulieu et al. 2013). For example, the gain of coloniality proceeds through two rate parameters: $0S \rightarrow 1S$ in the slow rate class and $0F \rightarrow 1F$ in the fast rate class. Transitions between rate classes are modeled analogously with $0S \rightarrow 0F$ when the lineage is solitary and $1S \rightarrow 1F$ when the lineage is colonial. I did not consider models where trait and rate class change simultaneously. From this eight-parameter model, a number of simplified models can be constructed by removing or constraining parameters to equality. To maintain reasonable model complexity, I tested models in which the probabilities of change between rate classes were symmetrical (i.e. $0S \rightarrow 0F$ = $0F \rightarrow 0S \neq 1S \rightarrow 1F = 1F \rightarrow 1S$) or equal across the entire phylogeny (i.e. $0S \rightarrow 0F$ = $0F \rightarrow 0S = 1S \rightarrow 1F = 1F \rightarrow 1S$). The symmetric models assumed that slow ↔ fast transition probabilities differ dependent on the state of growth form or habitat, while the equal model assumed that slow ↔ fast transition probabilities are constant. Focusing on testing the possibility of asymmetry in transition rates between character states within different rate classes, I compared models in which the forward and backward transitions were allowed to differ and models where these were constrained to equality. I also considered time-heterogeneous models that assessed the penalty of constraining the forward transitions to equality irrespective of the rate class while keeping the backward transitions different (i.e. $0S \rightarrow 1S = 0F \rightarrow 1F$ and $1S \rightarrow 0S \neq 1F \rightarrow 0F$), and the reverse, constraining backward transitions to equality while keeping forward transitions different (i.e. $0S \rightarrow 1S \neq 0F \rightarrow 1F$ and $1S \rightarrow 0S = 1F \rightarrow 0F$). These models were, in effect, testing the possibility that one type of transition in the trait (either the forward or backward) is constant across the phylogeny while the other varies. A total of 12 models constructed with the above reasoning were tested on the best ML phylogeny. A subset of four time-heterogeneous models that fit the trait data best as well as the two time-homogeneous models were thereafter fitted to the sample of 100 ML trees.

4.3.3 Combined traits

It is possible that habitat preference and growth form interact – colony formation might, for example, be favored in benthic species. The combination of two binary characters yields four combined phenotypes: planktonic+solitary (00), planktonic+colonial (01), benthic+solitary (10), and benthic+colonial (11). If growth form evolution depends on habitat, then the transition rate $00 \rightarrow 01$ is expected to differ from the transition rate $10 \rightarrow 11$ (Pagel, 1994; Pagel & Meade, 2006). The analogous situation for habitat is also of interest. Are transitions to the plankton dependent on growth form ($00 \rightarrow 10 \neq 01 \rightarrow 11$)? To test for interaction I fit two models: an independent model where the transitions $0 \rightarrow 1$ or $1 \rightarrow 0$ in one character were independent of the state of the other character (k= 4) and a dependent model where the probability of $0 \rightarrow 1$ or $1 \rightarrow 0$ in one character differed based on the state of the other character (k= 8). Preliminary analyses showed that a symmetric time-homogeneous model was preferred for the evolution of habitat preference when viewed individually of growth form. Knowing this, I tested an independent model where planktonic $\rightarrow$ benthic = benthic $\rightarrow$ planktonic, but solitary $\rightarrow$ colonial $\neq$ colonial $\rightarrow$ solitary. All two-trait modes were time-homogeneous. As before, analyses were performed on the 100 phylogenies.

The analyses were performed in the R packages "corHMM" and "ape" (Paradis et al. 2004; Beaulieu et al. 2013) and character state transitions were calculated in Mesquite (Maddison & Maddison, 2011). Model selection was performed using the Akaike information criterion corrected for sample size (AICc).

4.4 Results

4.4.1 Trees and trait distribution

The phylogeny used in our analyses recovered all major groups identified in recent all-diatom trees (Theriot et al. 2009, 2010) and is consistent with the current understanding of high-level relationships in diatoms (Fig. 4.1A). Among the sampled species, the ratio for habitat preference, benthic:planktonic, was about 2:1, and for growth form, solitary:colonial, was about 1:1. In the grade of clades of non-pennate diatoms, phenotypes alternate fairly often and there appears to be no bias with respect to habitat preference or growth form (Fig. 4.1A). Species in the derived lineage of pennate diatoms were predominantly benthic (Fig. 4.1A, node 2) and among them the clade of actively motile raphid pennates was dominated by solitary forms (Fig. 4.1A, node 1).

4.4.2 Habitat preference

A symmetric, time-homogeneous model performed best when fitting the habitat preference data. This model was favored in 59% of tested trees with an average Akaike weight (ω_i) =0.30 (Table 4.1). The asymmetric time-homogeneous model had the lowest AICc in 8 of the 100 trees. These two time-homogeneous models, although favored for 67% of the tested trees, had a combined average ω_i=0.53 indicating that more complex, time-heterogeneous models cannot be ruled out as plausible for the evolution of habitat preference (Table 4.1). Among the time-heterogeneous models, heavily parametrized rate matrices with >5 parameters performed poorly, suggesting that their complexity is not warranted for the present dataset. From the pool of tested time-heterogeneous models, only those with symmetric or equal transitions between rate classes (slow ↔ fast) performed comparably to the time-homogeneous models (Table 4.1). Within this set of four models, the penalty of reducing all possible shifts between rate classes to one parameter (equal model) was negligible (Table 4.1). These results are consistent with a scenario where transitions between the fast and slow rate class happen at similar rates across the entire phylogeny. Time-heterogeneous models where forward transitions in habitat (planktonic → benthic) were allowed to vary, whereas reversals (benthic → planktonic) were constrained to equality ("forward different") were preferred for 33% of trees with a combined average ω_i=0.38. These "forward different" models were better than "backward different" models where reversals, instead of forward transitions, were allowed to vary (Table 4.1). For the forward and reverse different models optimization were, at times, problematic with unreliable estimates for rate parameters.

Table 4.1: Time-homogeneous and time-heterogeneous models for the evolution of habitat preference across diatoms.

Model[a]	# rate classes	Mean lnL	k	Mean AICc	Mean ΔAICc	Mean ω_i[b]	% favored[c]
Symmetric[d]	1	-86.26	1	174.54	0.28	0.30	59
Asymmetric[e]	1	-85.52	2	175.08	0.82	0.23	8
Rates–symmetric, forward–different[f]	2	-82.54	5	175.29	1.03	0.23	23
Rates–equal, forward–different[g]	2	-84.03	4	176.21	1.95	0.15	10
Rates–equal, backward–different[h]	2	-84.78	4	177.71	3.45	0.06	0
Rates–symmetric, backward–different[i]	2	-84.75	5	179.72	5.46	0.02	0

---

[a] Models are ordered by their Akaike weights.

[b] Average Akaike weights denoting the relative likelihood of each model.

[c] The percent of trees out of 100 for which a particular model had the lowest AICc.

[d] $0 \to 1 = 1 \to 0$

[e] $0 \to 1 \neq 1 \to 0$

[f] $0S \to 0F = 0F \to 0S \neq 1S \to 1F = 1F \to 1S \neq 0S \to 1S \neq 0F \to 1F \neq 1S \to 0S = 1F \to 0F$

[g] $0S \to 0F = 0F \to 0S = 1S \to 1F = 1F \to 1S \neq 0S \to 1S \neq 0F \to 1F \neq 1S \to 0S = 1F \to 0F$

[h] $0S \to 0F = 0F \to 0S = 1S \to 1F = 1F \to 1S \neq 0S \to 1S = 0F \to 1F \neq 1S \to 0S \neq 1F \to 0F$

[i] $0S \to 0F = 0F \to 0S \neq 1S \to 1F = 1F \to 1S \neq 0S \to 1S = 0F \to 1F \neq 1S \to 0S \neq 1F \to 0F$

Parsimony, maximum likelihood, and stochastic character mapping on the best tree, the latter two conducted with the parameter estimates from the favored symmetric model, agreed that a minimum of three plankton → benthos transitions have happened along the diatom phylogeny (Fig. 4.2A). The maximum number of these transitions was estimated as high as 12 under parsimony and 24 under stochastic mapping (Fig. 4.1A). Benthos → plankton transition happened more frequently: a minimum of 9 (under ML) and 11 times (under parsimony and stochastic mapping) and a maximum of 20 and 33 times under parsimony and stochastic mapping, respectively.

4.4.3 Growth form

A symmetric, time-heterogeneous model (k=4) performed best when fitting the growth form data. This model was favored in 74% of tested trees with an average ω_i=0.47 (Table 4.1). This model is consistent with a scenario where the transition probabilities across the tree fall either in a fast or slow rate class, but in each rate class the probabilities of gain and loss of coloniality are equal. Simplifying this model by restricting the transitions between rate classes to one parameter did not incur a substantial cost in likelihood. The resulting "rates equal traits symmetric" model had an average ΔAICc=0.96 and ω_i=0.34 (Table 4.1). Relaxing the "rates equal traits symmetric" model to allow for asymmetrical transition rates between characters states or constraining forward (or backward) transitions to equality irrespective of rate class did not offer a substantially better fit (Table 4.1). As with habitat preference, parameter optimization for forward and reverse different models was unsuccessful for some trees. Models with >5 parameters were difficult to optimize resulting in unreasonably high estimates of transition rates. In contrast to the results for habitat preference, the time-homogeneous models performed poorly, were not favored for any of the trees, and averaged $\omega_i \leq 0.01$ (Table 4.1).

Figure 4.1: A. Phylogenetic distribution of habitat preference and growth form across the diatom phylogeny. Tip labels denote species' phenotype. Growth form is first column and habitat preference second column. Branch colors denote ancestral state reconstruction of growth form estimated with the favored symmetric time-heterogeneous model. Node labels denote ancestral state reconstruction of rate classes according to the same model. B. The rate coefficients of the favored symmetric time-heterogeneous model for the evolution of growth form.

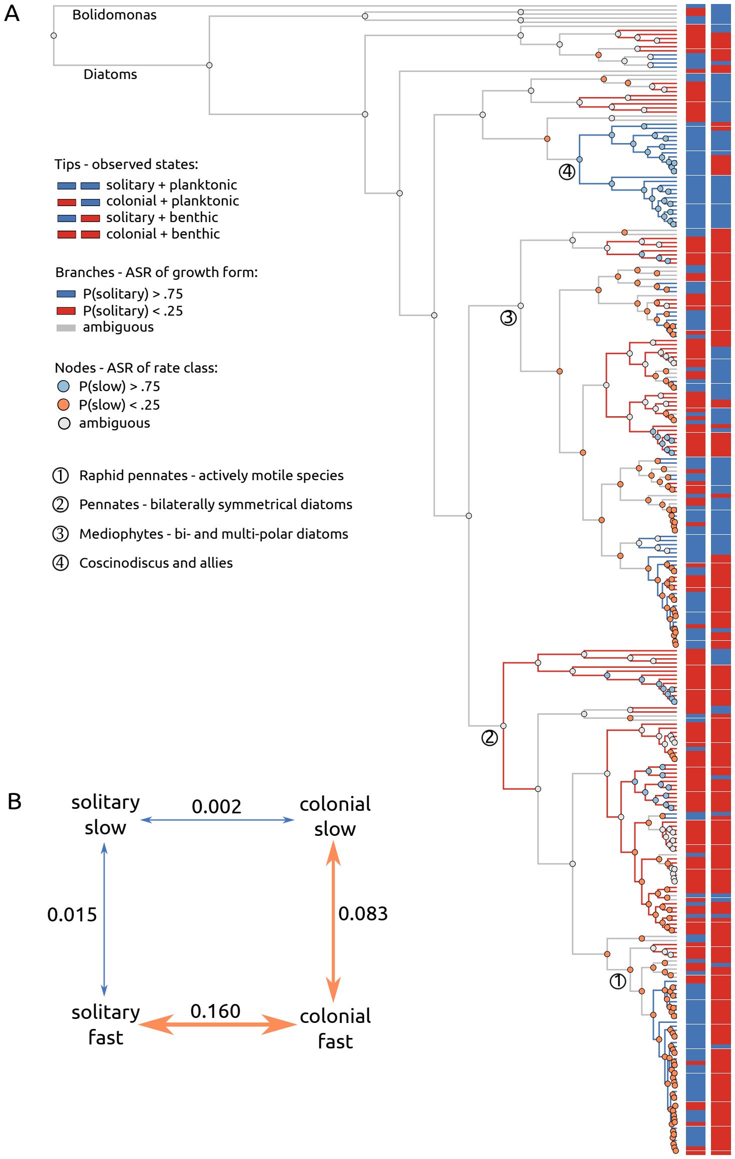


Table 4.2: Time-homogeneous and time-heterogeneous models for the evolution of growth form across diatoms.

Model[a]	# rate classes	Mean lnL	k	Mean AICc	Mean ΔAICc	Mean ω_i[b]	% favored[c]
Symmetric[d]	2	-135.49	4	279.12	0.31	0.47	74
Rates–equal, traits–symmetric[e]	2	-136.84	3	279.76	0.96	0.34	15
Rates–equal, backward–different[f]	2	-138.29	4	284.72	5.92	0.10	11
Rates–equal, traits–asymmetric[g]	2	-136.91	5	284.04	5.24	0.07	0
Asymmetric[h]	1	-141.77	2	287.58	8.78	0.01	0
Symmetric[i]	1	-143.99	1	289.99	11.19	< 0.01	0

---

[a] Models are ordered by their Akaike weights.

[b] Average Akaike weights denoting the relative likelihood of each model.

[c] The percent of trees out of 100 for which a particular model had the lowest AICc.

[d] $0S \to 0F = 0F \to 0S \neq 1S \to 1F = 1F \to 1S \neq 0S \to 1S = 1S \to 0S \neq 0F \to 1F = 1F \to 0F$

[e] $0S \to 0F = 0F \to 0S = 1S \to 1F = 1F \to 1S \neq 0S \to 1S = 1S \to 0S \neq 0F \to 1F = 1F \to 0F$

[f] $0S \to 0F = 0F \to 0S = 1S \to 1F = 1F \to 1S \neq 0S \to 1S = 0F \to 1F \neq 1S \to 0S \neq 1F \to 0F$

[g] $0S \to 0F = 0F \to 0S = 1S \to 1F = 1F \to 1S \neq 0S \to 1S \neq 0F \to 1F \neq 1S \to 0S \neq 1F \to 0F$

[h] $0 \to 1 \neq 1 \to 0$

[i] $0 \to 1 = 1 \to 0$

Parsimony, maximum likelihood, and stochastic character mapping on the best tree, the latter two conducted with the parameter estimates from the asymmetric time-homogeneous model, reconstructed a minimum of 12, 10, and 33, respectively, solitary → colonial transitions across the diatom phylogeny (Fig. 4.2B). The maximum number of solitary → colonial transitions was 20 under parsimony and 70 under the Bayesian stochastic mapping (Fig. 4.2B). Under parsimony, colonial

$\rightarrow$ solitary transitions were more frequent than the reverse with 21-29 total shifts. In contrast, under ML and stochastic mapping, the number of colonial $\rightarrow$ solitary transitions was lower (five under ML and 24-58 under stochastic mapping; Fig. 4.2B).

Overall, the distribution of growth form across the diatom phylogeny is best described by a model of evolution where some lineages transition between solitary and colonial states faster than others (Fig. 4.1, 4.3). Estimates from the favored time-heterogeneous model with symmetric rates show that the solitary $\leftrightarrow$ colonial transitions in the slow rate class (average maximum likelihood estimates [MLE]= 0.0026) are about 60 times slower than those in the fast rate class (average MLE= 0.16; Fig. 4.1B, 4.3B). Shifts in rate class tend to proceed about 5.5 faster on average in colony forming lineages (i.e. $1S \leftrightarrow 1F$) compared to the solitary species (i.e. $0S \leftrightarrow 0F$; Fig. 4.3C).

4.4.4 Combined traits

I tested for coordinated evolution between habitat and growth form by comparing the performance of a four-state model allowing transitions in one character to differ based on the state of the other character (e.g. $00 \rightarrow 01 \neq 10 \rightarrow 11$) to models where such transitions were constrained to equality (e.g. $00 \rightarrow 01 = 10 \rightarrow 11$). I assessed three cases: a dependent model where each transition rate is assigned a separate parameter (k=8), an independent asymmetric model where transitions in one character were independent of other character and forward $\neq$ backward (k=4),

Figure 4.2: Maximum parsimony (P, minimum and maximum in parentheses), maximum likelihood (L) and Bayesian stochastic character mapping (S, minimum, maximum and mean of 1000 simulations in parentheses) reconstructions of number of states shifts. A. Transitions between the planktonic and benthic habitat. B. Transitions between solitary and colonial growth form. Maximum likelihood and Bayesian optimizations were performed with the rate coefficients from the favored time-homogeneous models of evolution: symmetric for habitat and asymmetric for growth form.

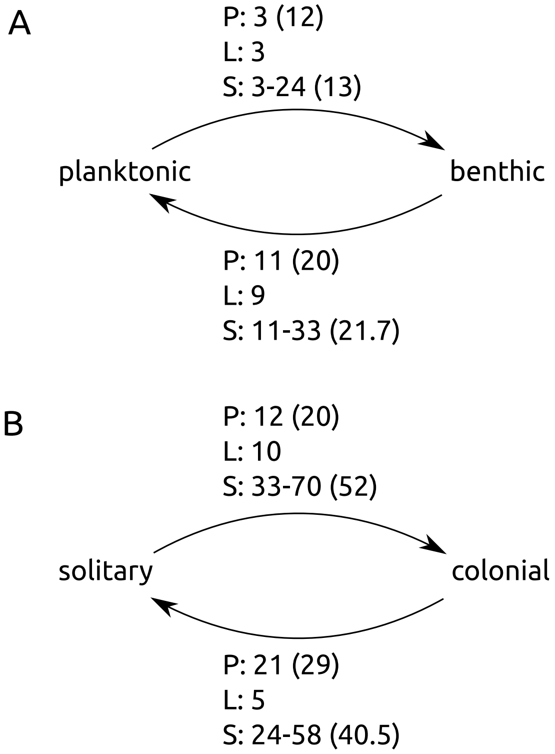
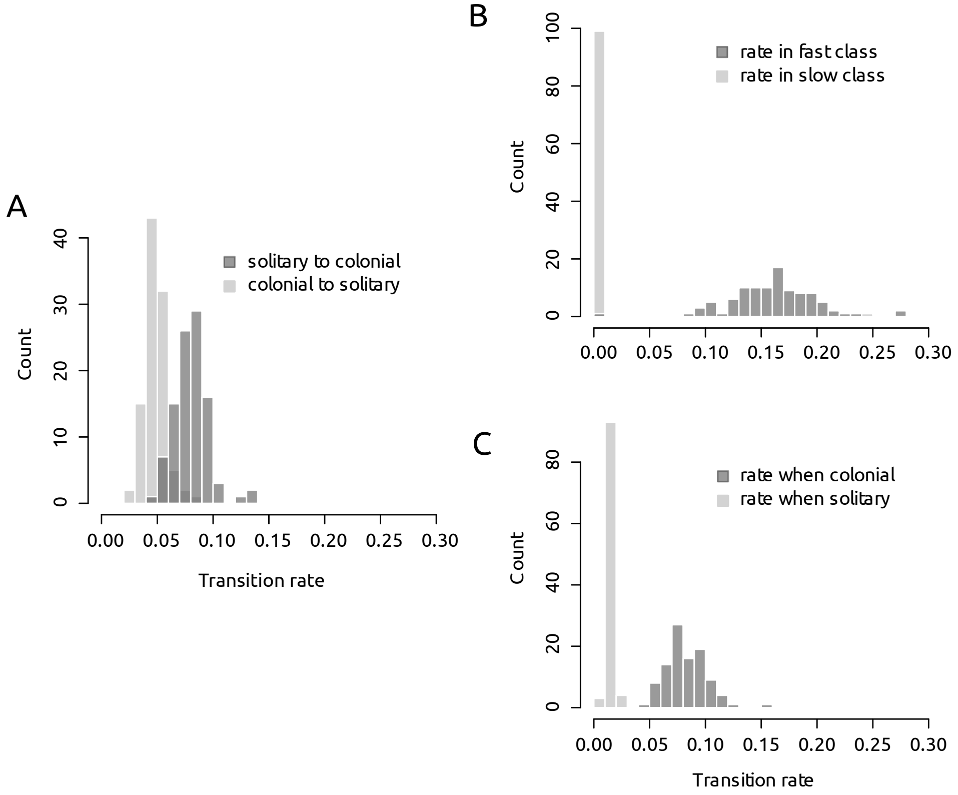


Figure 4.3: Transition probabilities for shifts between solitary and colonial growth form under best time-homogeneous and time-heterogeneous models estimated from 100 phylogenies. A. Under the asymmetric time-homogeneous model the solitary → colonial transition is much faster than the reverse, colonial → solitary transition. B. Transition probabilities for shifts between growth forms in the slow and fast rate class of the favored symmetric time-heterogeneous model. Transitions in the fast rate class are about 60 times faster than the slow rate class. C. Transition probabilities for shifting between rate classes when lineages are solitary and colonial. Colonial lineages traverse between the slow and fast rate classes faster than solitary.



and an independent constrained model where transitions between states for habitat preference were symmetric whereas transitions between states of growth form were kept asymmetric (k=3). Models in which these two traits evolve independently were favored across all trees with a combined average ω_i=0.85 (Table 4.1). Parameter estimates from the independent constrained model were very similar to those estimated from the time-homogeneous models when traits were treated separately (Fig. 4.4).

Table 4.3: Two trait models for the evolution of the combined growth form and habitat preference phenotype across diatoms.

Model[a]	Mean lnL	k	Mean AICc	Mean ΔAICc	Mean ω_i[b]	% favored[c]
Independent constrained[d]	-227.33	3	460.75	0.04	0.53	89
Independent[e]	-226.58	4	461.31	0.60	0.40	11
Dependent[f]	-224.09	8	464.70	3.99	0.07	0

---

[a] Models are ordered by their Akaike weights.

[b] Average Akaike weights denoting the relative likelihood of each model.

[c] The percent of trees out of 100 for which a particular model had the lowest AICc.

[d] $00 \rightarrow 01 = 10 \rightarrow 11 \neq 01 \rightarrow 00 = 11 \rightarrow 10 \neq 00 \rightarrow 10 = 01 \rightarrow 11 = 10 \rightarrow 00 = 11 \rightarrow 01$

[e] $00 \rightarrow 01 = 10 \rightarrow 11 \neq 01 \rightarrow 00 = 11 \rightarrow 10 \neq 00 \rightarrow 10 = 01 \rightarrow 11 \neq 10 \rightarrow 00 = 11 \rightarrow 01$

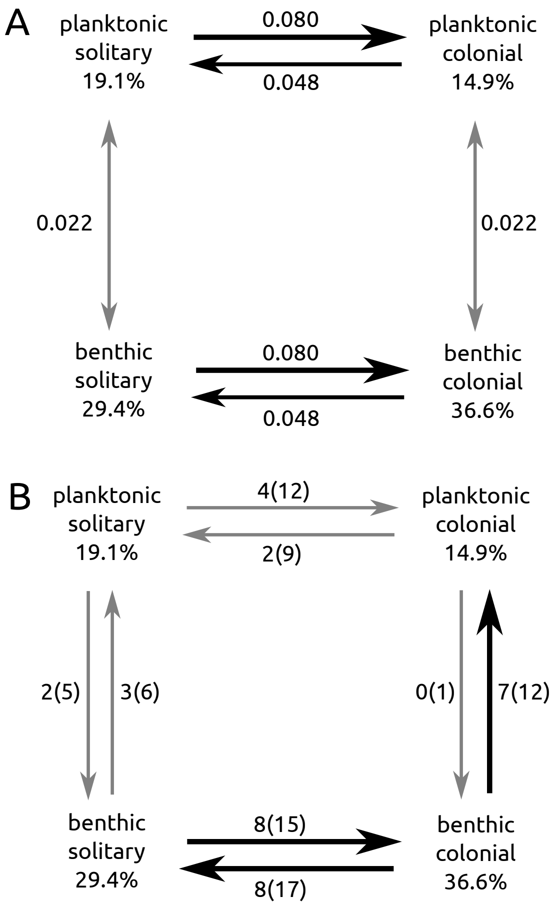
[f] $00 \rightarrow 01 \neq 01 \rightarrow 00 \neq 10 \rightarrow 11 \neq 11 \rightarrow 10 \neq 00 \rightarrow 10 \neq 10 \rightarrow 00 \neq 01 \rightarrow 11 \neq 11 \rightarrow 01$

4.5 Discussion

4.5.1 Traversals between habitats are rare, but asymmetric in lineages with colonial growth form

For a photosynthetic unicell, the plankton and benthos are diametrically different environments. Differences in physico-chemical properties of the surrounding

Figure 4.4: Evolution of the combined habitat + growth form phenotypes. A. Maximum likelihood estimates of rate coefficients from the favored independent constrained model of evolution. B. Parsimony counts of state shifts when the phenotypes were coded as a four-state character.



water, the availability of nutrients and light, and the types of microhabitats on offer require different sets of morphological and physiological adaptations (Stevenson, 1997; Reynolds, 2006). Transitions between littoral and open water habitats therefore have to be accompanied by physiological and morphological adjustments streamlining the cells for existence in the respective habitat. A morphological comparison of sister lineages differing in habitat occupancy on our tree revealed that benthos → plankton transitions tend to be accompanied by one or more of the following trait shifts: (i) increased morphological complexity of the cell, (ii) increase in cell size, and (iii) transition to colonial growth form. These trait shifts might be related to adaptations for planktonic life style (Kooistra et al. 2007, 2009). For example, increased morphological complexity of the cell, commonly achieved through the acquisition of various projections, spines, and keels, might represent a mechanism for improved boyancy. Increased morphological complexity and departure from spherical cell shape tends to increase form resistance – the difference in sinking velocity between a particle and a sphere with identical density and volume – and therefore decrease sinking velocity allowing cells to stay suspended longer (Padisak et al. 2003). The morphology of a colony can have an effect on sinking velocity as well. Tubular, spiral, or stellate arrangements that maintain colony symmetry exhibit reduced sinking relative to asymmetrical arrangements (Padisak et al. 2003; Reynolds 2006). Increase in cell size, a strategy that accompanied transitions to the plankton in two marine lineages (*Odontella* Agardh and *Trieres* Ashworth and Theriot), might be related to adaptation for higher capacity of nutrient storage given the transition to a more variable environment (Litchman et al., 2009). The repeated co-appearance of some of these phenotypes with the transition to a planktonic life style is suggestive of adaptations to life as a suspended particle. It is unlikely, however, that these traits are selected solely by the planktonic environment – large cell size, for example, is also a strategy for defense against predators (Thingstad et al., 2005; Verdy et al., 2009; Yokota & Sterner, 2010).

Estimates of the number of habitat traversals offers several insights. First, transitions to planktonic life style from a littoral habitat occurred more frequently than the opposite (Fig. 4.2). This result is consistent regardless of the method of inference (parsimony or model-based). Second, the majority of transitions to the plankton happened in lineages that had already attained a colonial growth form (Fig. 4). Third, planktonic, colony-forming lineages rarely or never transition to benthic habitats (Fig. 4.4). Taken together, these observations indicate that lineages with solitary growth form traverse the habitat boundaries rarely, but in both directions. On the other hand, when lineages are colonial, transitions between habitats become highly asymmetric (under parsimony: 7-12 benthic $\rightarrow$ plankonic vs. 0-1 planktonic $\rightarrow$ benthic). Planktonic colonial lineages can be viewed as somewhat of a "dead-end" with respect to habitat traversals. For these species, transition to the benthos would be a two-step process involving, first, loss of colonial growth habit and, second, transition to the benthos (Fig. 4.4). Overall, the benthic colonial state seems most dynamic, as changes in either trait, habitat preference or growth form, are more frequent compared to transitions to and from other states (Fig. 4.4).

These results could to some extent depend on taxon sampling. In this dataset, benthic species outnumber planktonic by factor of two and among the colony-formers, this factor increases to 2.4. The bias in favor of benthic species could be problematic, but only if it incorrectly depicts the ratio of species richness observed in nature. Estimates of species numbers in diatoms are uncertain (Guiry, 2012; Mann & Vanormelingen, 2013), but it is generally accepted that the benthos is more diverse than the plankton, especially in freshwater lakes (Mackay et al., 2010). While I cannot ascertain that the ratio of colonial+benthic : colonial+planktonic species in our dataset is a very accurate approximation of the diversity in nature, any potential bias in these data is likely in favor of the less numerous planktonic species as opposed to benthic taxa. Future studies will undoubtedly refine the findings reported here, but it is unlikely that these inferences are a result of a misrepresentation of the ratio of benthic:planktonic diversity.

4.5.2 Variable pace of growth form evolution across the diatom phylogeny

Colonial growth form independently evolved in all major lineages of photoautotrophic eukaryotes (Niklas & Newman, 2013) and diatoms are no exception (Figs 4.1, 4.2). Diatoms have two major mechanisms of aggregating into colonies. Some species achieve this through structures of the silica cell wall. In many cases these are modifications of preexisting features (e.g. enlarged costae or heavily modified marginal strutted processes refashioned to serve in valve-to-valve interlocking) or rarely structures that seem specifically acquired for cell-to-cell attachment (e.g. the periplekton of *Syndetocystis* Ralfs ex Greville). The other major mechanism is through extracellular mucilage production in form of pads, stalks, sheets, or tubes.

In many cases solitary species already possess the ability to produce mucilage and many of them use these for attachment to benthic substrata or "coccooning": surrounding itself in a sheath of mucilage. The association in a colony therefore might be a relatively simple process that requires the failure of cells to separate following mitosis and remain attached through elements of the cell wall or preexisting mucilage formations. Perhaps due to this relative simplicity and the benefits of colonial life style, colony formation has repeatedly evolved across the diatom tree. Estimates from different methods vary, but there seems to have been at least 10 (under ML) and more than 52 acquisitions (average from stochastic mapping) of the colony-forming state. The solitary growth form predominates in the large, planktonic species of *Coscinodiscus* Ehrenberg (and allies) and the benthic lineage of raphid pennate diatoms that have the ability of active movement (Fig. 4.1A). Otherwise, gains of coloniality are dispersed across the phylogeny encompassing both planktonic and benthic species across an array of cell bauplanes.

Under time-homogeneous models, there is support for asymmetry in the relative rates of transition between the solitary and colonial growth form: gains of coloniality are on average faster than losses (Fig. 4.3A). This result can be interpreted as a tendency for the acquisition of the generally beneficial colonial state. Closer examination of the data, however, revealed that this asymmetry might be a by-product of the variable rates present across the phylogeny. When I considered the possibility of time-heterogeneity in the evolutionary process there was no longer support for asymmetric transition probabilities (Table 4.2; Fig. 4.1B, 4.3). Instead, the results argue for symmetric transitions between the states of growth form, but in separate rate classes. Thus, diatoms can be (roughly) divided into two groups: lineages in which transition to and from coloniality are extremely rare where the evolution of growth form can be considered stagnant; and lineages in which this trait is labile, with frequent (60 times more probable) traversals between solitary and colonial growth forms (Fig. 4.3). Clades evolving in the slow or fast "regime" are not restricted to a particular portion of the topology, but dispersed across the tree (Fig. 4.1A). Apart from the coscinodiscoid lineage and the raphid pennates (Fig. 4.1A, nodes 1, 4), which are estimated as exclusively slow- and fast-rate class respectively, rate classes across the phylogeny alternate and correspondingly ancestral reconstructions of rate class are at times uncertain (Fig. 4.1A). Asymmetric transition probabilities within particular clades are certainly possible and a closer look at lineages where the evolution of growth form is most dynamic might identify tendencies specific to particular groups.

4.5.3 Concluding remarks

I used a broadly sampled phylogeny to characterize the patterns of evolution of habitat preference and growth form across diatoms. I found support for a simple, one-parameter model for the evolution of habitat preference consistent with a view of habitat preference as a slowly-evolving trait generally conserved at the level of large clades. The evolution of growth-form, on the other hand, is more dynamic and strong support for time-heterogeneous models argues for variable pace of growth form evolution across diatoms. In some lineages, growth form evolution is essentially stagnant. In others, transitions between solitary and colonial phenotypes are frequent. In the likelihood framework, I found no support for interaction between habitat and growth form. The probability of habitat traversal does not change with growth form nor do shifts in growth form depend on habitat occupancy. Under parsimony, however, asymmetry in the number of habitat transitions exists when the lineages are colonial suggesting that coordinated evolution is plausible and statistical power of the dataset might be an issue.

Several approaches can be used to refine the inferences made here. Improved taxon sampling at the all-diatom scale or focusing attention to lineages where trait shifts are common offer the opportunity of testing specific hypotheses about the evolution or particular phenotypes (e.g. the trajectories leading to and from the planktonic colonial state). Fine-grain coding of traits that captures species' microhabitat, colony construction, and colony morphology also depends on the availability of well sampled species–level phylogenies that are currently scarce in diatoms. Understanding the evolution of these phenotypes would likely require incorporating additional traits, like species size and cell morphology, that may affect the distributions of growth form and habitat occupancy across the phylogeny. A comprehensive grasp of the evolutionary histories of these and other diatom functional traits will not be achieved without investigation of the interplay between trait evolution and species diversification.

4.6 Acknowledgements

This work was partially funded by NSF EF 0629410 and Blumberg Fellowship to ECT. Computations were performed at the Texas Advanced Computing Center (http://www.tacc.utexas.edu).

Bibliography

Altekar G, Dwarkadas S, Huelsenbeck JP, & Ronquist F, 2004. Parallel Metropolis coupled Markov chain Monte Carlo for Bayesian phylogenetic inference. *Bioinformatics*, 20:407–415. PMID: 14960467.

Alverson AJ, 2013. Timing marine-freshwater transitions in the diatom order Thalassiosirales. *Paleobiology*, pp. 91–101.

Alverson AJ, Jansen RK, & Theriot EC, 2007. Bridging the Rubicon: Phylogenetic analysis reveals repeated colonizations of marine and fresh waters by thalassiosiroid diatoms. *Molecular Phylogenetics and Evolution*, 45:193–210.

Armbrust EV, 2009. The life of diatoms in the world's oceans. *Nature*, 459:185–92.

Ashworth MP, Nakov 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.

Beaulieu JM, Jhwueng DC, Boettiger C, & O'Meara BC, 2012. Modeling stabilizing selection: Expanding the Ornstein-Uhlenbeck model of adaptive evolution. *Evolution*, 66:2369–2383. PMID: 22834738.

Beaulieu JM, O'Meara BC, & Donoghue MJ, 2013. Identifying hidden rate changes in the evolution of a binary morphological character: The evolution of plant habit in campanulid angiosperms. *Systematic Biology*, 62:725–737. PMID: 23676760.

Bergsten J, Nilsson AN, & Ronquist F, 2013. Bayesian tests of topology hypotheses with an example from diving beetles. *Systematic Biology*, 62:660–673. PMID: 23628960.

Blomberg SP, Garland T, & Ives AR, 2003. Testing for phylogenetic signal in comparatve data: behavioral traits are more labile. *Evolution*, 57:717–745.

Brown JM, Hedtke SM, Lemmon AR, & Lemmon EM, 2010. When trees grow too long: investigating the causes of highly inaccurate Bayesian branch-length estimates. *Systematic biology*, 59:145–161.

Brown JM & Lemmon AR, 2007. The importance of data partitioning and the utility of Bayes factors in Bayesian phylogenetics. *Systematic Biology*, 56:643–655.

Brown JW & Sorhannus U, 2010. A molecular genetic timescale for the diversification of autotrophic stramenopiles (Ochrophyta): Substantive underestimation of putative fossil ages. *PLoS ONE*, 5:e12 759.

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.

Butler MA & King AA, 2004. Phylogenetic comparative analysis: a modeling approach for adaptive evolution. *The American Naturalist*, 164:683–695.

Cannone JJ, Subramanian S, Schnare MN, Collett JR, D'Souza LM, Du Y, Feng B, Lin N, Madabusi LV, Müller KM, Pande N, Shang Z, Yu N, & Gutell RR, 2002. The comparative RNA web (CRW) site: An online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. *BMC bioinformatics*, 3:2. PMID: 11869452.

Cohn SA, Farrell JF, Munro JD, Ragland RL, Weitzell RE, & Wibisono BL, 2003. The effect of temperature and mixed species composition on diatom motility and adhesion. *Diatom Research*, 18:225–243.

Connolly JA, Oliver MJ, Beaulieu JM, Knight CA, Tomanek L, & Moline MA, 2008. Correlated evolution of genome size and cell volume in diatoms (Bacillariophyceae). *Journal of Phycology*, 44:124–131.

Cox EJ, 1987. Placoneis Mereschkowsky: The re-evaluation of a diatom genus originally characterized by its chloroplast type. *Diatom Research*, 2:145–157.

Cox EJ & Williams DM, 2006. Systematics of naviculoid diatoms (Bacillariophyta): A preliminary analysis of protoplast and frustule characters for family and order level classification. *Systematics and Biodiversity*, 4:385–399.

Cummins CA & McInerney JO, 2011. A method for inferring the rate of evolution of homologous characters that can potentially improve phylogenetic inference, resolve deep divergence and correct systematic biases. *Systematic Biology*, 60:833–844.

Czarnecki DB, 1987. Freshwater diatom culture collection at Loras College, Dubuque, Iowa: Notulae Naturae of The Academy of Natural Sciences of Philadelphia, No. 465. *Notulae Naturae*, 465:1–16.

Daugbjerg N & Andersen RA, 1997. A molecular phylogeny of the Heterokont algae based on analyses of chloroplast-encoded *rbcL* sequence data. *Journal of Phycology*, 33:1031–1041.

Edwards KF, Thomas MK, Klausmeier CA, & Litchman E, 2012. Allometric scaling and taxonomic variation in nutrient utilization traits and maximum growth rate of phytoplankton. *Limnology and Oceanography*, 57:554–566.

Elwood HJ, Olsen GJ, & Sogin ML, 1985. The small-subunit ribosomal RNA gene sequences from the hypotrichous ciliates Oxytricha nova and *Stylonychia pustulata*. *Molecular Biology and Evolution*, 2:399–410. PMID: 3939705.

Falkowski PG, Katz ME, Knoll AH, Quigg A, Raven JA, Schofield O, & Taylor FJR, 2004. The evolution of modern eukaryotic phytoplankton. *Science*, 305:354–360.

Fan Y, Wu R, Chen MH, Kuo L, & Lewis PO, 2011. Choosing among partition models in Bayesian phylogenetics. *Molecular Biology and Evolution*, 28:523–532.

Felsenstein J, 1985. Phylogenies and the comparative method. *American Naturalist*, 125:1–15.

Finkel ZV, Beardall J, Flynn KJ, Quigg A, Rees TAV, & Raven JA, 2010. Phytoplankton in a changing world: cell size and elemental stoichiometry. *Journal of Plankton Research*, 32:119–137.

Finkel ZV, Katz ME, Wright JD, Schofield OME, & Falkowski PG, 2005. Climatically driven macroevolutionary patterns in the size of marine diatoms over the Cenozoic. *Proceedings of the National Academy of Sciences*, 102:8927–8932.

Finkel ZV, Sebbo J, Feist-Burkhardt S, Irwin AJ, Katz ME, Schofield OME, Young JR, & Falkowski PG, 2007. A universal driver of macroevolutionary change in the size of marine phytoplankton over the Cenozoic. *Proceedings of the National Academy of Sciences*, 104:20 416–20 420.

Guiry MD, 2012. How many species of algae are there? *Journal of Phycology*, 48:1057–1063.

Hansen TF, 1997. Stabilizing selection and the comparative analysis of adaptation. *Evolution*, 51:1341–1351.

Harmon LJ, Weir JT, Brock CD, Glor RE, & Challenger W, 2008. GEIGER: investigating evolutionary radiations. *Bioinformatics*, 24:129–131.

Hoagland KD, 1983. Short-term standing crop and diversity of periphytic diatoms in a eutrophic reservoir. *Journal of Phycology*, 19:30–38.

Hoagland KD, Roemer SC, & Rosowski JR, 1982. Colonization and community structure of two periphyton assemblages, with emphasis on the diatoms (Bacillariophyceae). *American Journal of Botany*, 69:188–213.

Hopkinson BM, Dupont CL, Allen AE, & Morel FMM, 2011. Efficiency of the CO2-concentrating mechanism of diatoms. *Proceedings of the National Academy of Sciences*, 108:3830–3837. PMID: 21321195 PMCID: PMC3054024.

Huelsenbeck JP, 1997. Is the Felsenstein zone a fly trap? *Systematic Biology*, 46:69–74. PMID: 11975354.

Huelsenbeck JP, Larget B, & Alfaro ME, 2004. Bayesian phylogenetic model selection using reversible jump Markov chain Monte Carlo. *Molecular Biology and Evolution*, 21:1123–1133.

Huelsenbeck JP, Nielsen R, & Bollback JP, 2003. Stochastic mapping of morphological characters. *Systematic Biology*, 52:131–158. PMID: 12746144.

Jiang L, Schofield OME, & Falkowski PG, 2005. Adaptive evolution of phytoplankton cell size. *The American naturalist*, 166:496–505. PMID: 16224705.

Johnson RE, Tuchman NC, & Peterson CG, 1997. Changes in the vertical microdistribution of diatoms within a developing periphyton mat. *Journal of the North American Benthological Society*, 16:503–519.

Jones HM, Simpson GE, Stickle AJ, & Mann DG, 2005. Life history and systematics of *Petroneis* (Bacillariophyta), with special reference to British waters. *European Journal of Phycology*, 40:61–87.

Kass RE & Raftery AE, 1995. Bayes factors. *Journal of the American Statistical Association*, 90:773–795.

Kermarrec L, Ector L, Bouchez A, Rimet F, & Hoffmann L, 2011. A preliminary phylogenetic analysis of the Cymbellales based on 18S rDNA gene sequencing. *Diatom Research*, 26:305–315.

Key T, McCarthy A, Campbell DA, Six C, Roy S, & Finkel ZV, 2010. Cell size trade-offs govern light exploitation strategies in marine phytoplankton. *Environmental Microbiology*, 12:95–104.

Kilham S, Kreeger D, & Lynn S, 1998. COMBO: a defined freshwater culture medium for algae and zooplankton. *Hydrobiologia*, pp. 147–159.

Kociolek JP & Stoermer EF, 1986. Phylogenetic relationships and classification of monoraphid diatoms based on phenetic and cladistic methodologies. *Phycologia*, 25:297–303.

Kociolek JP & Stoermer EF, 1988. A preliminary investigation of the phylogenetic relationships among the freshwater, apical pore field-bearing cymbelloid and gomphonemoid diatoms (Bacillariophyceae). *Journal of Phycology*, 24:377–385.

Kociolek JP & Stoermer EF, 1989. Phylogenetic relationships and evolutionary history of the diatom genus *Gomphoneis*. *Phycologia*, 28:438–454.

Kociolek JP & Stoermer EF, 1993. Freshwater gomphonemoid diatom phylogeny: preliminary results. In: Hv Dam, ed., *Twelfth International Diatom Symposium*, no. 90 in Developments in Hydrobiology, pp. 31–38. Springer Netherlands.

Koester JA, Swanson WJ, & Armbrust EV, 2013. Positive selection within a diatom species acts on putative protein interactions and transcriptional regulation. *Molecular Biology and Evolution*, 30:422–434.

Kolaczkowski B & Thornton JW, 2009. Long-branch attraction bias and inconsistency in Bayesian phylogenetics. *PLoS ONE*, 4:e7891.

Kooistra WH, Gersonde R, K Medlin L, & G Mann D, 2007. Chapter 11 - The origin and evolution of the diatoms: Their adaptation to a planktonic existence. In: PG Falkowski & AH Knoll, eds., *Evolution of Primary Producers in the Sea*, pp. 207–249. Academic Press, Burlington.

Kooistra WHCF, Forlani G, & De Stefano M, 2009. Adaptations of araphid pennate diatoms to a planktonic existence. *Marine Ecology*, 30:1–15.

Krammer K, 1982. *Valve morphology in the genus* Cymbella *C.A. Agardh*, vol. XI of *Micromorphology of Diatom Valves*. J. Cramer.

Krammer K, 2003. *Diatoms of Europe: Diatoms of the European inland waters and comparable habitats. Volume 4:* Cymbopleura, Delicata, Navicymbula, Gomphocymbellopsis, Afrocymbella *supplements to cymbelloid taxa.*

Krammer K & Lange-Bertalot H, 1986. *Bacillariophyceae. 1. Teil: Naviculaceae.* No. Band 2/1 in Süsswasser flora von Mitteleuropa. Gustaf Fischer Verlag, Stuttgart, New York.

Lanfear R, Calcott B, Ho SYW, & Guindon S, 2012. PartitionFinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses. *Molecular biology and evolution*, 29:1695–1701.

Lange-Bertalot H & Metzeltin D, 1996. *Indicators of oligotrophy: 800 taxa representative of three ecologically distinct lake types: Carbonate buffered, oligodystrophic, weakly buffered soft water.* Koeltz Scientific Books.

Lartillot N, Brinkmann H, & Philippe H, 2007. Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model. *BMC Evolutionary Biology*, 7:S4. PMID: 17288577 PMCID: PMC1796613.

Lartillot N & Philippe H, 2004. A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution*, 21:1095–1109. PMID: 15014145.

Lartillot N, Rodrigue N, Stubbs D, & Richer J, 2013. PhyloBayes MPI: Phylogenetic Reconstruction with infinite mixtures of profiles in a parallel environment. *Systematic Biology*, 62:611–615. PMID: 23564032.

Leblanc K, Aristegui J, Armand L, Assmy P, Beker B, Bode A, Breton E, Cornet V, Gibson J, Gosselin MP, Kopczynska E, Marshall H, Peloquin J, Piontkovski S, Poulton AJ, Quéguiner B, Schiebel R, Shipe R, Stefels J, van Leeuwe MA, Varela M, Widdicombe C, & Yallop M, 2012. A global diatom database – abundance, biovolume and biomass in the world ocean. *Earth System Science Data*, 4:149–165.

Litchman E, Klausmeier CA, & Yoshiyama K, 2009. Contrasting size evolution in marine and freshwater diatoms. *Proceedings of the National Academy of Sciences*, 106:2665–2670.

Litchman E, de Tezanos Pinto P, Klausmeier CA, Thomas MK, & Yoshiyama K, 2010. Linking traits to species diversity and community structure in phytoplankton. *Hydrobiologia*, 653:15–28.

Mackay AW, Edlund MB, & Khursevich G, 2010. Diatoms in ancient lakes. In: JP Smol & EF Stoermer, eds., *The Diatoms: Applications for the environmental and earth sciences*, pp. 209–228. Cambridge University Press, Cambridge, 2 edn.

Maddison WP & Maddison DR, 2011. Mesquite: a modular system for evolutionary analysis.

Mann DG, 1982a. Structure, life history and systematics of *Rhoicosphenia* (Bacillariophyta). I. The vegetative cell of Rh. Curvata. *Journal of Phycology*, 18:162–176.

Mann DG, 1982b. Structure, life history and systematics of *Rhoicosphenia* (Bacillariophyta). II. Auxospore formation and perizonium structure of *Rh. Curvata*. *Journal of Phycology*, 18:264–274.

Mann DG, 1984. Structure, life history and systematics of *Rhoicosphenia* (Bacillariophyta). V. Initial cell and size reduction in *Rh. Curvata* and a description of the Rhoicospheniaceae Fam. Nov. *Journal of Phycology*, 20:544–555.

Mann DG & Stickle AJ, 1995. Sexual reproduction and systematics of *Placoneis* (Bacillariophyta). *Phycologia*, 34:74–86.

Mann DG & Vanormelingen P, 2013. An inordinate fondness? The number, distributions, and origins of diatom species. *Journal of Eukaryotic Microbiology*, pp. 414–420.

Marshall DC, 2010. Cryptic failure of partitioned Bayesian phylogenetic analyses: Lost in the land of long trees. *Systematic Biology*, 59:108–117.

McCormick PV & Stevenson RJ, 1991. Mechanisms of benthic algal succession in lotic environments. *Ecology*, 72:1835–1848.

Medlin L, Elwood HJ, Stickel S, & Sogin ML, 1988. The characterization of enzymatically amplified eukaryotic 16S-like rRNA-coding regions. *Gene*, 71:491–499.

Naselli-Flores L, Padisák J, & Albay M, 2007. Shape and size in phytoplankton ecology: Do they matter? *Hydrobiologia*, 578:157–161.

Nawrocki EP, 2009. Structural RNA homology search and alignment using covariance models. *Electronic Theses and Dissertations*.

Nawrocki EP, Kolbe DL, & Eddy SR, 2009. Infernal 1.0: inference of RNA alignments. *Bioinformatics*, 25:1335–1337. PMID: 19307242.

Nelson DM, TrÃĺguer P, Brzezinski MA, Leynaert A, & Quéguiner B, 1995. Production and dissolution of biogenic silica in the ocean: Revised global estimates, comparison with regional data and relationship to biogenic sedimentation. *Global Biogeochemical Cycles*, 9:359–372.

Niklas KJ & Newman SA, 2013. The origins of multicellular organisms. *Evolution & Development*, 15:41–52.

Nylander JAA, 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. PMID: 17766271.

O'Meara B, Ané C, Sanderson M, & Wainwright P, 2006. Testing for different rates of continuous trait evolution using likelihood. *Evolution*, 60:922–933.

Padisák J, Soréczki-Pintér É, & Rezner Z, 2003. Sinking properties of some phytoplankton shapes and the relation of form resistance to morphological diversity of plankton – an experimental study. *Hydrobiologia*, 500:243–257.

Pagel M, 1994. Detecting correlated evolution on phylogenies: A general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society B: Biological Sciences*, 255:37–45.

Pagel M, 1999. Inferring the historical patterns of biological evolution. *Nature*, 401:877–884.

Pagel M & Meade A, 2006. Bayesian analysis of correlated evolution of discrete characters by reversible-jump Markov chain Monte Carlo. *American Naturalist*, 167:808–825.

Paradis E, Claude J, & Strimmer K, 2004. APE: Analyses of phylogenetics and evolution in R language. *Bioinformatics*, 20:289–290.

Patrick R & Reimer C, 1966. *The Diatoms of the United States: Exclusive of Alaska and Hawaii: Vol. 1: Fragilariaceae, Eunotiaceae, Achnanthaceae, Naviculaceae.*, vol. 1 of *Monographs of the Academy of Natural Sciences of Philadelphia*. Philadelphia.

Pol D & Siddall ME, 2001. Biases in maximum likelihood and parsimony: A simulation approach to a 10-Taxon Case. *Cladistics*, 17:266–281.

R Development Core Team, 2013. R: A language and environment for statistical computing.

Rajan V, 2013. A Method of Alignment Masking for Refining the Phylogenetic Signal of Multiple Sequence Alignments. *Molecular Biology and Evolution*, 30:689–712.

Rambaut & Drummond, 2007. Tracer - BEAST software.

Raven JA, 1998. The twelfth Tansley Lecture. Small is beautiful: the picophytoplankton. *Functional Ecology*, 12:503–513.

Raven JA, Finkel ZV, & Irwin AJ, 2005. Picophytoplankton: bottom-up and top-down controls on ecology and evolution. *Small*, 55:209–215.

Revell LJ, 2010. Phylogenetic signal and linear regression on species data. *Methods in Ecology and Evolution*, 1:319–329.

Revell LJ, 2012. phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, 3:217–223.

Reynolds CS, 2006. *The Ecology of Phytoplankton.* Cambridge University Press, Cambridge.

Richlen ML & Barber PH, 2005. A technique for the rapid extraction of microalgal DNA from single live and preserved cells. *Molecular Ecology Notes*, 5:688–691.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, & Huelsenbeck JP, 2012. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Systematic biology*, 61:539–542. PMID: 22357727 PMCID: PMC3329765.

Round FE, Crawford RM, & Mann DG, 1990. *Diatoms: Biology and morphology of the genera.* Cambridge University Press.

Ruck EC & Theriot EC, 2011. Origin and evolution of the canal raphe system in diatoms. *Protist*, 162:723–737. PMID: 21440497.

Sanderson MJ, 2002. Estimating absolute rates of molecular evolution and divergence times: A penalized likelihood approach. *Molecular biology and evolution*, 19:101–109. PMID: 11752195.

Sarno D, Kooistra WHCF, Balzano S, Hargraves PE, & Zingone A, 2007. Diversity in the genus *Skeletonema* (Bacillariophyceae): III. Phylogenetic position and morphological variability of *Skeletonema costatum* and *Skeletonema grevillei*, with the description of *Skeletonema ardens* sp. nov. *Journal of Phycology*, 43:156–170.

Sarno D, Kooistra WHCF, Medlin LK, Percopo I, & Zingone A, 2005. Diversity in the genus *Skeletonema* (Bacillariophyceae). II. An assessment of the taxonomy of *S. costatum* -like species with the description of four new species. *Journal of Phycology*, 41:151–176.

Shimodaira H, 2002. An approximately unbiased test of phylogenetic tree selection. *Systematic Biology*, 51:492–508. PMID: 12079646.

Shimodaira H & Hasegawa M, 2001. CONSEL: For assessing the confidence of phylogenetic tree selection. *Bioinformatics*, 17:1246–1247. PMID: 11751242.

Siddall ME & Whiting MF, 1999. Long-branch abstractions. *Cladistics*, 15:9–24.

Spaulding SA & Kociolek JP, 2000. Freshwater diatom biogeography. *Nova Hedwigia*, 71:223–242.

Sperling EA, Peterson KJ, & Pisani D, 2009. Phylogenetic-signal dissection of nuclear housekeeping genes supports the paraphyly of sponges and the monophyly of Eumetazoa. *Molecular Biology and Evolution*, 26:2261–2274.

Stabelli OR, Lartillot N, Philippe H, & Pisani D, 2012. Serine codon usage bias in deep phylogenomics: Pancrustacean relationships as a case study. *Systematic Biology*, pp. 121–133.

Stamatakis A, 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, 22:2688–2690.

Stamatakis A, Hoover P, & Rougemont J, 2008. A rapid bootstrap algorithm for the RAxML web servers. *Systematic biology*, 57:758–771. PMID: 18853362.

Stevenson RJ, 1997. An introduction to algal ecology in freshwater benthic habitats. In: RJ Stevenson, ML Bothwell, & R Lowe, eds., *Algal ecology: Freshwater benthic ecosystems*, Aquatic ecology series. Academic Press: San Diego.

Sullivan J & Swofford DL, 1997. Are Guinea pigs rodents? The importance of adequate models in molecular phylogenetics. *Journal of Mammalian Evolution*, 4:77–86.

Theriot EC, Ashworth M, Ruck E, Nakov T, & Jansen RK, 2010. A preliminary multigene phylogeny of the diatoms (Bacillariophyta): Challenges for future research. *Plant Ecology and Evolution*, 143:278–296.

Theriot EC, Cannone JJ, Gutell RR, & Alverson AJ, 2009. The limits of nuclear-encoded SSU rDNA for resolving the diatom phylogeny. *European Journal of Phycology*, 44:277–290. PMID: 20224747.

Thingstad TF, Øvreäs L, Egge JK, Løvdal T, & Heldal M, 2005. Use of non-limiting substrates to increase size; a generic strategy to simultaneously optimize uptake and minimize predation in pelagic osmotrophs? *Ecology Letters*, 8:675–682.

Verdy A, Follows M, & Flierl G, 2009. Optimal phytoplankton cell size in an allometric model. *Marine Ecology Progress Series*, 379:1–12.

Von Dassow P, Petersen TW, Chepurnov VA, & Virginia Armbrust E, 2008. Inter- and intraspecific relationships between nuclear DNA content and cell size in selected members of the centric diatom genus *Thalassiosira* (Bacillariophyceae). *Journal of Phycology*, 44:335–349.

Vyverman W, Verleyen E, Sabbe K, Vanhoutte K, Sterken M, Hodgson DA, Mann DG, Juggins S, Vijver BVd, Jones V, Flower R, Roberts D, Chepurnov VA, Kilroy C, Vanormelingen P, & Wever AD, 2007. Historical processes constrain patterns in global diatom diversity. *Ecology*, 88:1924–1931.

Whitney KD & Garland T, 2010. Did genetic drift drive increases in genome complexity? *PLoS Genetics*, 6:e1001 080.

Will S, Reiche K, Hofacker IL, Stadler PF, & Backofen R, 2007. Inferring noncoding RNA families and classes by means of genome-scale structure-based clustering. *PLoS computational biology*, 3:e65. PMID: 17432929.

Xie W, Lewis PO, Fan Y, Kuo L, & Chen MH, 2011. Improving marginal likelihood estimation for Bayesian phylogenetic model selection. *Systematic Biology*, 60:150–160.

Yang Z, 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends in Ecology & Evolution*, 11:367–372.

Yokota K & Sterner RW, 2010. Trade-offs limiting the evolution of coloniality: Ecological displacement rates used to measure small costs. *Proceedings of the Royal Society B: Biological Sciences*, pp. 458–463. PMID: 20739317.

Vita

Teofil Nakov is from Skopje, Macedonia. He spent his childhood playing street soccer in the Hrom neighborhood of Skopje, collecting tea (*Sideritis scardica*), berries, rose hips, or just climbing mountains in the carst of Bistra National Park in Macedonia. Deliberations with his brother, Novica Nakov, during his high school years at Rade Jovčevski Korčagin and beyond molded much of his interests and character. In 2005, he received a Bachelor of Science degree in Biochemistry and Physiology from the Institute of Biology, Faculty of Natural Sciences and Mathematics at Sts. Cyril and Methodius University, Skopje, Macedonia. Along the way, he learned about diatoms and met a plethora of enthusiastic biologists at the Biology Students' Research Society. After his bacalaureat, he worked as a teaching and research assistant in the laboratory of Systematics of Lower Plants at the Institute of Biology, Skopje. In August 2007, he joined the Plant Biology Graduate Program at the University of Texas at Austin. He remains a Manchester United fan despite having never visited Old Trafford.

Permanent address: teofiln@gmail.com

This dissertation was typeset with LaTeX[†] by the author.

[†] LaTeX is a document preparation system developed by Leslie Lamport as a special version of Donald Knuth's TeX Program.