J Mol Evol (2006) 63:231–239
DOI: 10.1007/s00239-006-0016-z

JOURNAL OF MOLECULAR EVOLUTION

Evidence for Positive Selection on a Sexual Reproduction Gene in the Diatom Genus *Thalassiosira* (Bacillariophyta)

Ulf Sorhannus,[1] Sergei L. Kosakovsky Pond[2]

[1] Department of Biology & Health Services, Edinboro University of Pennsylvania, Edinboro, Pennsylvania 16444, USA
[2] Antiviral Research Center, University of California San Diego, San Diego, California 92103, USA

Received: 23 January 2006 / Accepted: 15 April 2006 [*Reviewing Editor:* Dr. Martin Kreitman]

Abstract. Single likelihood ancestor counting (SLAC), fixed effects likelihood (FEL), and several random effects likelihood (REL) methods were utilized to identify positively and negatively selected sites in sexually induced gene 1 (Sig1) of four different *Thalassiosira* species. The SLAC analysis did not find any sites affected by positive selection but suggested 13 sites influenced by negative selection. The SLAC approach may be too conservative because of low sequence divergence. The FEL and REL analyses revealed over 60 negatively selected sites and two positively selected sites that were unique to each method. The REL method may not be able to reliably identify individual sites under selection when applied to short sequences with low divergence. Instead, we proposed a new alignment-wide test for adaptive evolution based on codon models with variation in synonymous and nonsynonymous substitution rates among sites and found evidence for diversifying evolution without relying on site-by-site testing. The performance of the FEL and REL approaches was evaluated by subjecting the tests to a type I error rate simulation analysis, using the specific characteristics of the Sig1 data set. Simulation results indicated that the FEL test had reasonable Type I errors, while REL might have been too liberal, suggesting that the two positively selected sites identified by FEL (codons 94 and 174) are not likely to be false positives. The evolution of these codon sites, one of which is located in functional domain II, appears to be associated with divergence among the three major *Thalassiosira* lineages.

Key words: Likelihood methods — Negative selection — Nonsynonymous substitution — Positive selection — Sexual reproduction gene — Synonymous substitution — *Thalassiosira*

Correspondence to: Ulf Sorhannus; *email:* usorhannus@edinboro.edu

Introduction

Despite their late appearance in the fossil record (e.g., Kooistra and Medlin 1996; Sorhannus 1997), diatoms form a species-rich group of algae which plays key roles in the global silica/carbon cycling (e.g., Mann 1999) and food chains of aquatic ecosystems. The diatom cell is surrounded by a silica shell that is structured like a ''lid'' and a ''box.'' The frustule (i.e., the silica shell) displays extensive shape variation and typically has a large number of tiny intricately shaped depressions, pores, spines, and passageways. In the traditional diatom classification, which was inferred from morphological features of the siliceous cell wall and the reproductive mode (e.g., Simonsen 1972, 1979; Round et al. 1990), two major subgroups were recognized (Simonsen 1979): Centrales (''centric diatoms'') and Pennales (''pennate diatoms''). The centrics typically exhibit radial or bi(multi)polar symmetry, whereas the pennates normally show bilateral symmetry with or without a central groove (i.e., raphe). The former subgroup

was further split into three taxa, while the latter was divided into two groups based on the presence/absence of the raphe (Simonsen 1979). Modern classification principles do not recognize the centric (including some of the centric subgroups) and araphid pennate diatoms as valid taxa because of their paraphyletic status (e.g., Sorhannus 2004). Molecular phylogenetic studies (e.g., Sorhannus et al. 1995; Ehara et al. 2000; Fox and Sorhannus 2003; Medlin and Kaczmarska 2004; Sorhannus 2004) have indicated, by and large, the same relationships among the major groups as those proposed by Simonsen (1979) based on nonmolecular data. In the optimal molecular tree(s), the radial centrics (except Thalassiosirales) constituted the earliest branch(es) after which the bi(multi)polar centrics/Thalassiosirales forms appeared. A still undetermined bi(multi)polar centric/Thalassiosirales lineage shares a common ancestor with the pennates, which differentiated into the earlier ''araphids'' and the subsequent ''raphids.'' However, one major difference between the nonmolecular and the molecular trees is the position of Thalassiosirales. Simonsen (1979) placed Thalassiosirales lineages among the early radial centrics, while the majority of molecular phylogenetic studies (e.g., Ehara et al. 2000; Fox and Sorhannus 2003; Kooistra et al. 2003; Medlin and Kaczmarska 2004; Sorhannus 2004) indicated that it emerged among the bi(multi)polar centrics. *Thalassiosira*, which mainly consists of marine species, is one of the most diverse genuses within Thalassiosirales (e.g., Kaczmarska et al. 2006). The species included in this study (e.g., Round et al. 1990), *Thalassiosira weissflogii* (Grun.) Fryxell et Hasle, *Thalassiosira oceanica* (Hustedt) Hasle et Heimdal, *Thalassiosira guillardii* Hasle, and *Thalassiosira pseudonana* (Hustedt) Hasle et Heimdal, are planktonic marine forms which constitute a monophyletic group, based on an extensive molecular phylogenetic study (Sorhannus 2004). Within the clade, *T. weissflogii* and *T.oceanica* shared the most recent common ancestor, followed by *T. guillardii* and *T. pseudonana* (Sorhannus 2004).

In many diatom species, including the *Thalassiosira* species analyzed here, successive generations of asexual reproduction result in a decrease in average cell size. Cell size is often restored through sexual reproduction, which involves the release of gametes into a surrounding body of water followed by sperm-egg fusion. Pheromone–like compounds, produced by the female gametes, are thought to play a role in attracting sperm cells (Drebes 1996). For species-specific gamete fusion to occur, particular recognition sites in the reproductive cells of both mating types are required. Three sexually induced genes (Sig1, Sig2, and Sig3), hypothesized to play a role in sperm-egg adhesion, have been sequenced in *T. weissflogii* (Armbrust 1999; Armbrust and Galindo 2001). The three polypeptides encoded by the genes have several characteristics in common: (1) three or more cysteine-rich epithelial growth factor (EGF)-like repeats; (2) five highly conserved regions (I–V) which show a high degree of similarity to the EGF-containing domain of the vertebrate extracellular matrix glycoprotein tenascin X (promotes cell to cell interaction); (3) a signal sequence; (4) the same sequential order of the conserved domains, separated by a variable number of nonconserved amino acid sites; and (5) a lack of transmembrane regions (Armbrust 1999). In particular, similarity between the EGF-containing domain of the Sig polypeptides and the extracellular matrix glycoprotein tenascin X of vertebrates suggests that Sig proteins play a role in mediating sperm-egg recognition in *Thalassiosira* species during the sexual reproduction phase (Armbrust 1999; Armbrust and Galindo 2001). Sig1, thought to be represented by at least 10 unique copies in the genome (Armbrust and Galindo 2001), is characterized by five functional domains, inferred to be active in cell adhesion (Armbrust 1999).

Genes that encode proteins involved in gamete adhesion/fusion have typically exhibited increased rates of evolution both within and between species (e.g., Wyckoff et al. 2000; Vacquier 1998; Swanson et al. 2001; Armbrust and Galindo 2001; Swanson and Vacqier 2002). Accelerated diversification of reproductive proteins has been associated with site-specific positive selection (e.g., Swanson and Vacquier 1995; Tsaur and Wu 1997; Wyckoff et al. 2000; Swanson et al. 2001; Swanson and Vacqier 2002). Amino acid sites affected by positive selection tend to show nonsynonymous (dN) substitution rates that are significantly higher than the synonymous (dS) rates. Sites identified as being influenced by positive selection are frequently located in domains thought to be involved in sperm-egg interaction (e.g, Swanson et al. 2001). This phenomenon is thought to have important consequences for the establishment of barriers to fertilization and speciation (e.g., Swanson and Vacqier 2002).

In a study carried out by Sorhannus (2003), between four and seven codon sites in Sig1 of *T. weissflogii* were inferred to be affected by positive selection. This result was questioned by Suzuki and Nei (2004), who claimed that none of the sites were influenced by positive selection, i.e., putatively selected sites were in fact false positives. In this study, we carried out a careful case study of applying realistic models of codon evolution and studying the statistical properties of the tests on the specific data set under investigation. Regrettably, data set-specific analysis of error rates of various statistical tests is not a standard practice in applied sequence analyses. Lacking external confirmatory evidence (e.g., directed

mutagenesis), which may be difficult or costly to obtain, such error rate analyses are one of the most rigorous computational tests for validity of inference. The outcome of the analyses carried out here, using a larger sample of Sig1 than the one employed by Sorhannus (2003) and additional *Thalassiosira* species, suggested that at least two sites were influenced by positive selection. One of them is located in functional domain II of Sig1. Inferred evolutionary changes in the positively selected sites appear to be associated with divergence among the three major *Thalassiosira* lineages.

Materials and Methods

Sequences

Sixty-three partial Sig1 sequences and one complete Sig1 gene were obtained from GenBank. The accession numbers and sampling locations of the *T. weissflogii* sequences are as follows: AF154499 (Li0), AF374490 (Li1)–AF374500 (Li11), AF374540 (Li12)–AF374552 (Li24) (Clone CCMP 1336; Long Island Sound, USA); AF374501 (Li25)–AF374505 (Li29) (Clone CCMP 1049; Long Island Sound, USA); AF374506 (No1)–AF374510 (No5) (Clone CCMP 1052; Skagerrak Sea, Norway); AF374521 (Ca1)–AF374525 (Ca5) (Clone CCMP 1050; Del Mar Slough, CA, USA); AF374526 (Po1)–AF374530 (Po5) (Clone CCMP 1053; North Atlantic Ocean, Portugal); AF374516 (Ha1)–AF374520 (Ha5) (Clone CCMP 1051; King Kalakaua's Fishpond, HI, USA); and AF374511 (In1)–AF374515(In5) (Clone CCMP 1587; Jakarta Harbor, Indonesia). The accession numbers and sampling locations of the other sequences were AF374537–AF374539 (Clone CCMP 1335; *Thalassiosira pseudonana*, Moriches Bay, NY, USA); AF374531–AF374533 (Clone CCMP 1005; *Thalassiosira oceanica*, Sargasso Sea); and AF374534–AF374536 (Clone CCMP 988; *Thalassiosira guillardii*, North Atlantic Ocean). Each sampling location constitutes a clone, where different sequences reflect intraindividual variation. Five identical *T. weissflogii* sequences (AF374499, AF374506, AF374512, AF374521, AF374522) and an intron located between the two coding regions (except in the cDNA sequences AF374540–AF374552) were removed from the data matrix before the alignment was performed, that is, the alignment was carried out on the remaining 59 coding sequences. A NEXUS file with aligned sequences may be downloaded from http://www.hyphy.org/pubs/Sig1.nex. The four functional domain coordinates in the alignment are as follows (Fig. 1): domain I (nucleotides 1–108, codons 1–36); domain II (nucleotides 234–342, codons 78–114); domain III (nucleotides 465–516, codons 155–172); and domain IV (nucleotides 525–591, codons 175–197).

Data Analyses

The data were managed using the software package DAMBE (version 4.0.98 [Xia 2000]). The amino acid sequences of the Sig1 were aligned using CLUSTALW (version 1.7 [Thompson et al. 1994]), which is included in the DAMBE program package. Aligned amino acid sequences were mapped to corresponding codons using DAMBE. We used PAUP* 4.0 (Swofford 2002) to reconstruct a neighbor joining (Saitou and Nei 1987) gene tree based on the Tamura–Nei (1993) distance. Exact program settings for this step and all subsequent steps are given in the Supplementary Information. The resulting tree was displayed in TreeView (Page 1996). It should be noted that internal nodes with negative branch lengths were treated as unresolved polytomies. We applied a hierarchical and information theoretic model selection procedure (Kosakovsky Pond and Frost 2005) to choose a model of nucleotide substitution. HKY85 (matrix 010010) was selected as the optimal time-reversible nucleotide substitution model using the implementation in the HyPhy package (Kosakovsky Pond et al. 2005). Recent results of Kosakovsky Pond and Muse (2005) suggest that site-to-site variation in synonymous rates is widespread and can contribute to false-positive selection signal using methods which fail to account for this variation. We investigated the pattern of rate variation in the Sig1 gene as a whole, by fitting several models of site-to-site rate variation (Table 1). In order to test for evidence of positive selection in the context of models which allow for synonymous rate variation, we modified the Dual model (Kosakovsky Pond and Muse 2005) with S synonymous rate classes and N nonsynonymous rate classes to use a general discrete distribution which does not allow nonsynonymous rates to exceed synonymous rates. If this model, which we denote as Dual(–), fits the data significantly worse than the unconstrained Dual model, there is evidence of diversifying selection acting on the protein. To assess the significance of the likelihood ratio test (LRT), we noticed that the constrained model can be derived from the unconstrained model by applying N one-sided constraints (all nonsynonymous rates must be no higher than the lowest synonymous rate). An appropriate asymptotic distribution of the test statistic would be a mixture of χ^2 distributions with 0 through N degrees of freedom (Self and Liang 1987). However, for phylogenetic likelihood and $N > 1$ it is not possible to obtain the mixing proportions analytically. Instead, we use the distribution of the test statistic under the null hypothesis, i.e., the Dual(–) model, based on 100 parametric data replicates.

In order to identify codons affected by positive and negative selection, fast single-likelihood ancestor (SLAC) counting, fixed effects likelihood (FEL), and random effects likelihood (REL) methods available in the HyPhy package (Kosakovsky Pond et al. 2005) and in a free public web implementation (Kosakovsky Pond and Frost 2005) were applied to Sig1 data. Codon evolution at sites 37, 94, 150, and 174 were mapped on the tree (Fig. 2) using a maximum likelihood reconstruction of ancestral states based on a fitted model of codon substitution, as implemented in http://www.datamonkey.org (Kosakovsky Pond and Frost 2005).

A detailed account of the SLAC, FEL, and REL methods is given by Kosakovsky Pond and Frost (2005). Briefly, SLAC uses a maximum likelihood reconstruction of ancestral codon states to compare the observed ratio of nonsynonymous and synonymous substitutions with the approximate estimate of the expected ratio assuming neutral evolution. FEL uses the entire alignment to infer model parameters shared by all sites (e.g., branch lengths) and then fits dS and dN rates individually at every site. Neutrality of an individual site is tested using the likelihood ratio test. REL extends the popular methods of Nielsen and Yang (1998) to allow both synonymous and nonsynonymous substitution rates to vary among sites. All three methods used here incorporate synonymous rate variation explicitly.

To establish statistical properties of the FEL test given low divergence in Sig1 sequences and multiple polytomies, we fitted 100 data sets simulated under neutrality ($dN = dS = 1$) using the Sig1 tree with branch lengths and nucleotide substitution biases inferred by maximum likelihood. We then applied the FEL procedure to each simulated data set and tabulated the number of times a site was identified as selected (positively or negatively) by the test, as well as the p value associated with the inference. The estimated Type I error rate of the FEL test $R(p)$, treated as a function of the significance level of the test (p), can then be computed as the proportion of sites identified as selected at or below level p. For an ideal test, one would expect to find $R(p) = p$, while a conservative test would yield $R(p) \leq p$. As shown in Fig. 3, FEL is not expected to have an elevated rate of false positives in the range of $0 < p < 0.08$,which includes

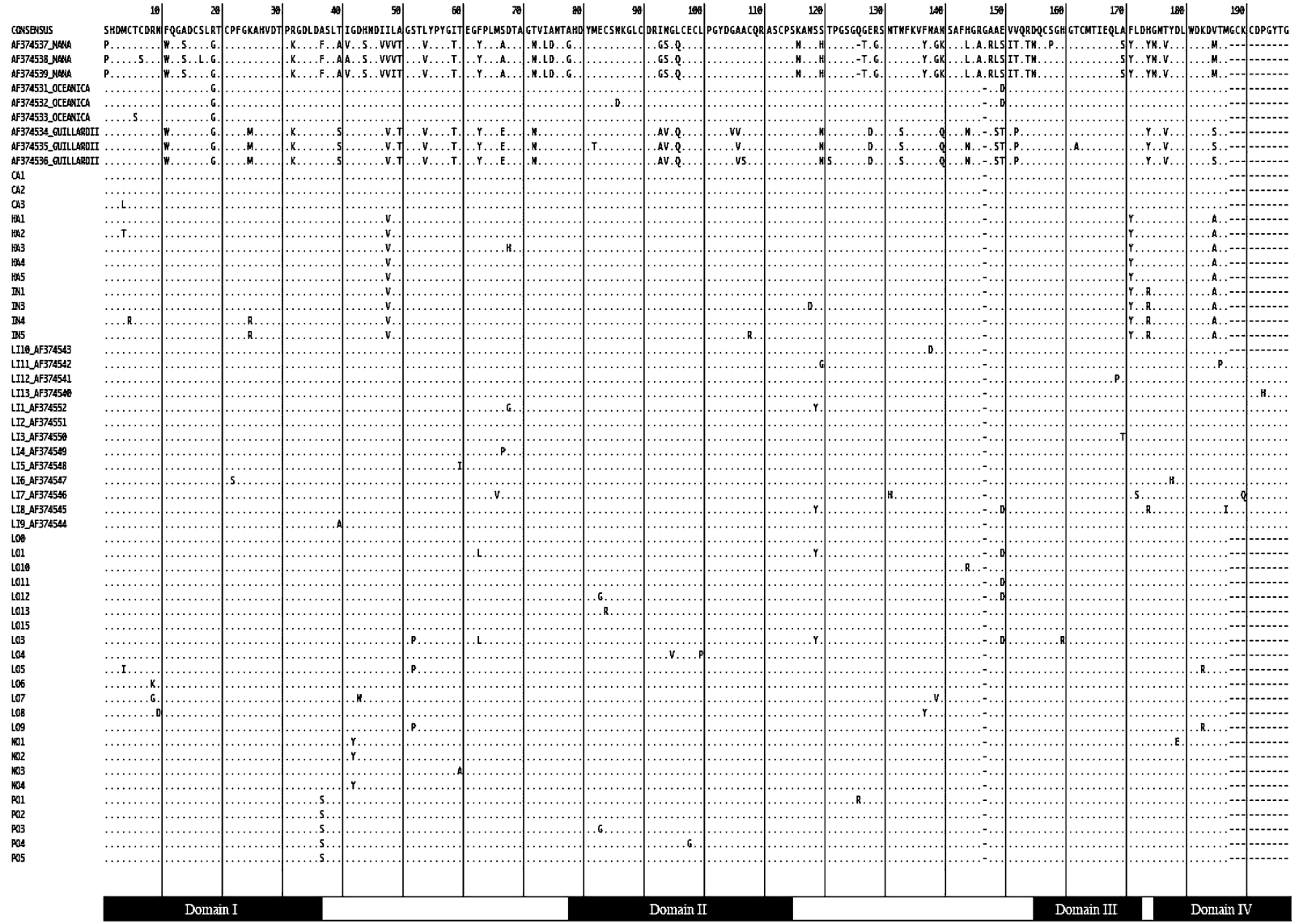

Fig. 1. Sig1 sequences used in the study (protein translated). Sites 37, 94, 150, and 174, were identified as being under diversifying selection by FEL or REL.

Table 1. Results of codon analysis of rate variation patterns in Sig1

Model	logL	No. par	dN/dS	Test, p value
Constant	−2582.48	78	0.197	N/A
Proportional	−2557.48	82	0.190	Null: constant LRT, $p \approx 10^{-9}$
Nonsynonymous	−2556.15	82	0.236	Null: constant LRT, $p \approx 10^{-10}$
Dual	−2538.03	84	0.100	Null: nonsynonymous LRT, $p \approx 10^{-8}$
Dual (–)	−2539.58	84	0.225	Alternative: dual empirical LRT, $p < 0.05$

Note. Proportional and nonsynonymous models utilized the general discrete distribution (GDD) with three bins, while Dual and Dual(–) models used GDD with two bins for synonymous rates and GDD with three bins for nonsynonymous rates.

the values inferred for codons 94 ($p = 0.07$) and 174 ($p = 0.06$). A similar analysis for REL (Fig. 4) suggests, instead, that a measurable proportion of neutrally evolving sites may be misidentified even for large cutoff Bayes factors. This may be partly due to lack of convergence (flat likelihood surface) and partly due to large errors in rate parameter estimates.

Faulty inferences about positive selection could result from recombination events (e.g., Sorhannus 2003). We ran GENECONV (version 1.81 [Sawyer 1999]) to detect possible recombination events among the sequences. The default settings were used in the analysis. This method searched for unusually long identical fragments within pairs of aligned sequences or pairwise segments with in the alignment characterized by uncommonly high matching scores (Sawyer 1999) with p values for each pair of sequences and multiple-comparison adjusted global p-value levels derived by simulation. The significance levels for global and pairwise comparisons were set at 0.05.

Finally, we modified the FEL method (Kosakovsky Pond and Frost 2005) to test for evidence of nonneutral evolution in Sig1 along a subset of three preselected branches: those which are involved in speciation events and all internal branches.

Results

The distance tree showed largely unresolved relationships among and within the *T. weissflogii* clones obtained from the North Atlantic Ocean and the California coast. Those collected from the waters around Hawaii and Indonesia formed a distinct group with regard to the Atlantic and California cell lines (Fig. 2). *T. oceanica*, unexpectedly, clustered within the Atlantic/California group of *T. weissflogii*. The sister lineage of the *T. weissflogii*/*T. oceanica* complex was *T. guillardii*. *T. pseudonana* formed a distantly related lineage with respect to *T. weissflogii*/*T.oceanica*/*T. guillardii*. With the exception of the position of *T. oceanica*, the gene/species tree is in general agreement with relationships obtained by phylogenetic studies of the diatoms (e.g., Sorhannus 2004).

The recombination analysis of the data matrix identified two Long Island sequences (AF374491 and AF374504) of *T. weissflogii*, belonging to two separate clones (CCMP 1336 and CCMP 1049), as having undergone a significant recombination event (global p value = 0.024). The length of the fragment was 437 nucleotides long. After both Li2 (AF374491) and Li28 (AF374504) were eliminated from the data set, the analysis showed no significant recombination events between the remaining sequences. Thus, the remaining 57 Sig1 sequences were subjected to analyses for positive/negative selection.

Codon analysis of rate variation patterns suggested that Sig1 is subject to both synonymous and nonsynonymous site-to-site rate variation (Table 1). By varying the number of rate classes in the Dual model, we determined that the best (small sample AIC) fit was achieved with two synonymous and three nonsynonymous rate classes. Furthermore, the comparison between the Dual and the Dual(–) models demonstrated that some sites in the alignment were undergoing adaptive change (parametric bootstrap $p < 0.05$ based on 100 replicates).

The SLAC analysis revealed no positively selected but 13 negatively selected sites (Table 2). FEL identified sites 94 ($p = 0.07$) and 174 ($p = 0.06$) as being influenced by positive selection and 65 negatively selected sites. One (site 94) of the two positively selected codons was located in functional domain II (Fig. 1; also see Armbrust 1999). This site was inferred to be under selection along the three branches involved in speciation. The other site (site 174) influenced by positive selection was located in a region between functional domains III and IV. The REL analysis suggested that codon sites 37 and 150 were affected by positive selection and that 64 sites were influenced by negative selection (Table 2 and Supplementary Information). The two positively selected sites were different from those discovered by FEL. Site 150 (between functional domains II and III; Fig. 1) was inferred to be under positive selection by FEL but with a p value of 0.18. Maximum likelihood ancestral state reconstruction suggested eight nonsynonymous and one synonymous substitution at that site.

Discussion

In an integrative approach to detecting positive selection, like this one, the ideal result is that every method supports the same positively selected site(s) (Kosakovsky Pond and Frost 2005). The fact that the SLAC analysis did not identify any positively selected sites is not an unexpected result since this approach

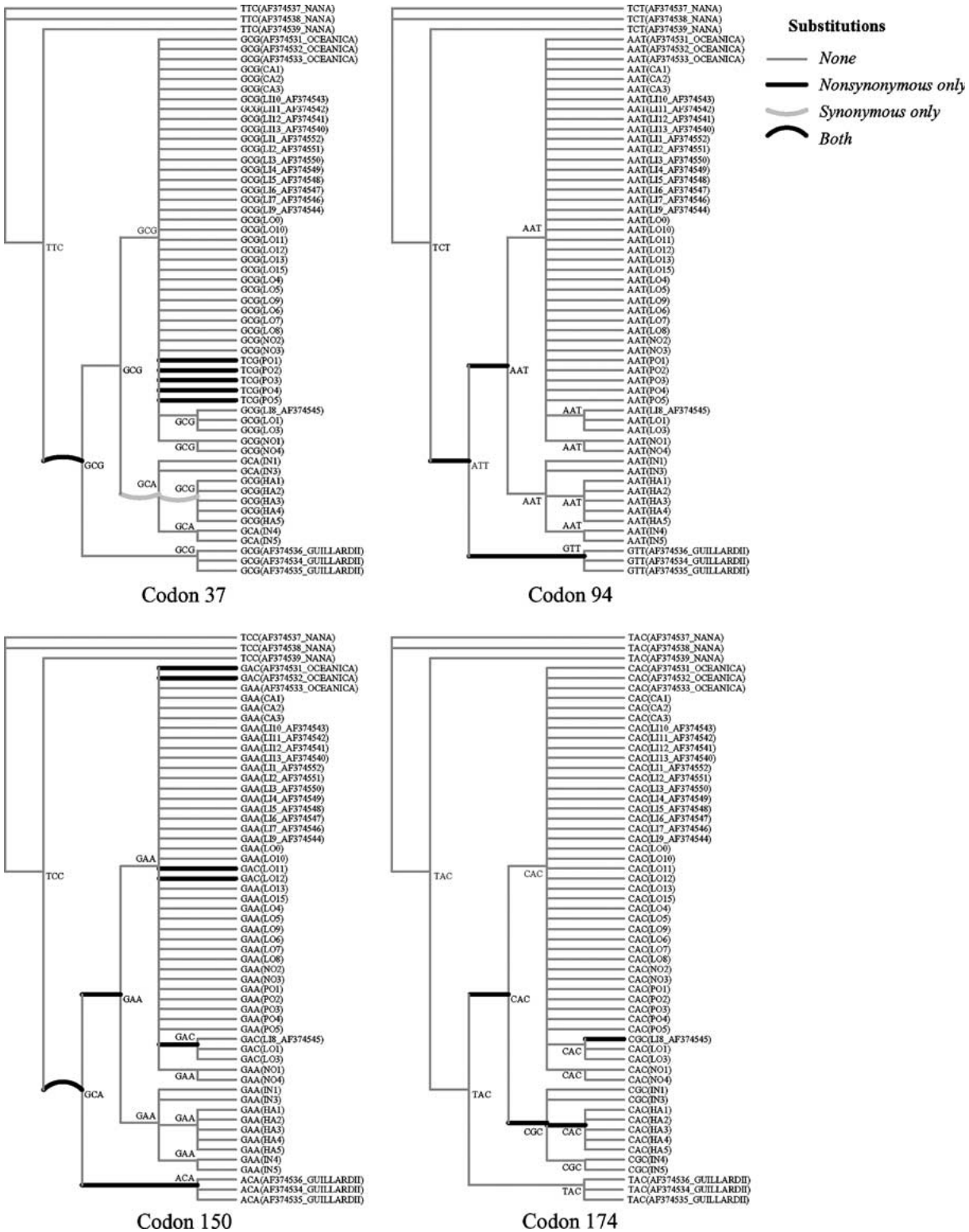

Fig. 2. Codon evolution in sites 37, 94, 150, and 174 mapped on the distance tree used in the SLAC, FEL, and REL analyses. Ancestral codons, shown at each node, were inferred using SLAC. All terminal codons are shown. Li0–Li29 represent *T. weissflogii* sequences from the Long Island isolates; No1–No4 represent *T. weissflogii* sequences from the Norwegian isolate; Po1–Po5 represent *T. weissflogii* sequences from the Portuguese isolate; Ca1–Ca3 represent *T. weissflogii* sequences from the California isolate; Ha1–Ha5 represent *T. weissflogii* sequences from the Hawaiian isolate; In1–In4 represent *T. weissflogii* sequences from the Indonesian isolate. Nana1–Nana3 represent *T. pseudonana* sequences; Guillardii1–Guillardii3 represent *T. guillardii* sequences; Oceanica1–Oceanica3 represent *T. oceanica* sequences.

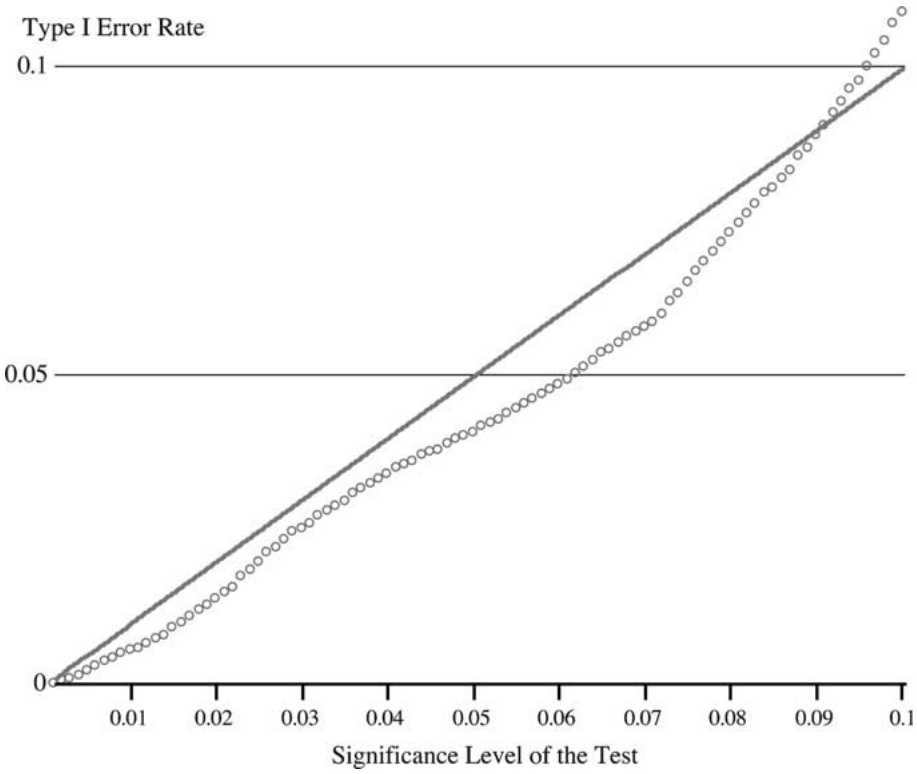

Fig. 3. Type I error rate as a function of the significance level of the FEL test. The circles represent the actual false-positive errors made by FEL on 100 data sets simulated using the tree and model from Sig1 but assuming neutral evolution. For portions of the graph which lie below the expected error rate (solid line), FEL is behaving conservatively, but since both lines are reasonably close, the test is not unduly conservative.

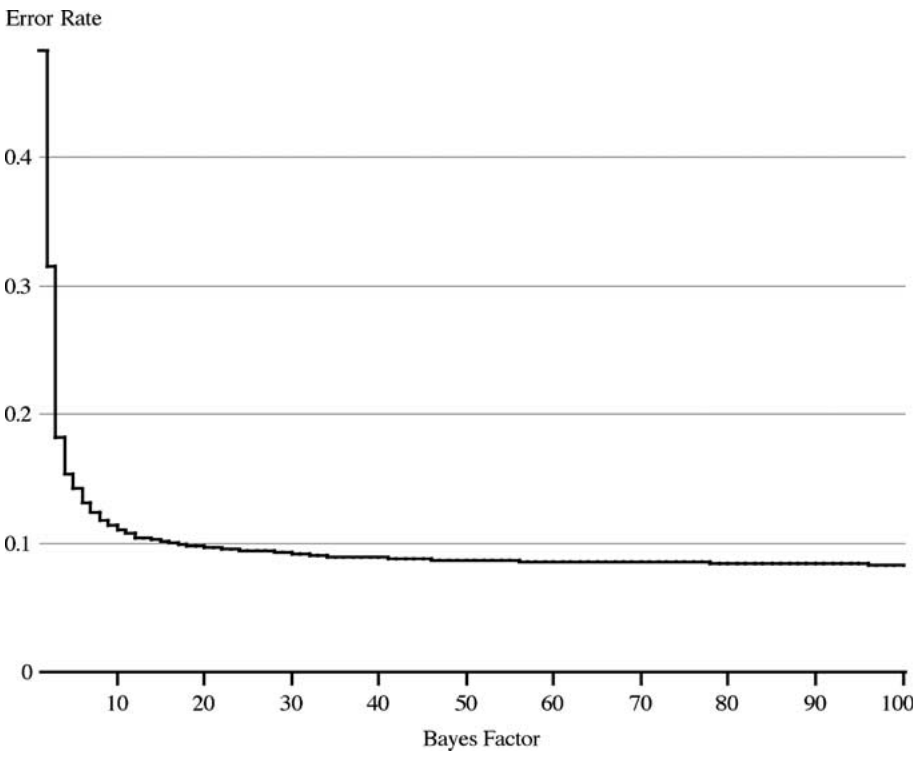

Fig. 4. Type I error rate as a function of the significance level of the REL test. The line represents the actual false-positive errors (only for identifying sites as positively selected) made by REL on 100 data sets simulated using the tree and model from Sig1 but assuming neutral evolution.

may lack power for data sets consisting of sequences with low divergence (Kosakovsky Pond and Frost 2005). The FEL and REL methods each discovered two positively selected sites that were different between methods. Since FEL has been found, in general, to be a more conservative method (i.e., less likely to identify false-positive results) than REL, more confidence can be given to the results of the former approach (Kosakovsky Pond and Frost 2005). The results of the Type I error rate analysis carried out for

Table 2. Positively selected sites identified by at least one method

	SLAC		FEL		REL		
Codon	dN − dS	*p* value	dN − dS	*p* value	dN − dS	Posterior probability	Bayes factor
37	−0.12	0.60	2.51	0.71/0.47/0.47	**3.10**	0.98	4269.55
94	1.99	0.30	**2.81**	0.07/0.02/0.04	−0.60	0.012	0.99
150	2.97	0.21	8.59	0.19/0.18/0.21	**3.10**	0.99	9446.52
174	2.06	0.35	**3.24**	0.06/0.37/0.18	−0.60	0.03	2.6

Note. FEL tests list three *p* values: whole tree, internal branches only, and the three internal branches separating the three phylogenetically resolved species. dN − dS values are normalized to be directly comparable between methods. When a test is significant for a given method, normalized dN − dS values are shown in boldface.

the Sig1 data set/tree supported the conservative nature of the FEL test (Fig. 3), suggesting that the positively selected sites identified by this method are unlikely to be false positives. Moreover, the REL approach is expected to be unstable on this sequence alignment data due to poor parameter estimating properties induced by short sequences of low divergence and polytomies in the tree.

In light of the capacity to disperse over great distances and the lack of apparent barriers to gene flow (e.g., Darling et al. 2000), a low degree of genetic variation among unicellular planktonic organisms is expected (e.g., Palumbi 1994; Norris 2000). However, many molecular evolutionary studies, including that by Armbrust and Galindo (2001), have revealed a high degree of genetic differentiation in many pelagic organisms (e.g., De Vargas et al. 1999; Darling et al. 2000; Darling et al. 2004; Rynearson and Armbrust 2004). To explain the unexpected high degree of biological diversity in the open ocean, both allopatric and sympatric mechanisms have been proposed (e.g., De Vargas et al. 1999; Darling et al. 2004; Rynearson and Armbrust 2004). In this study, amino acid evolution due to positive selection, especially in site 94 of functional domain II and site 174 (Fig. 1), was associated with divergence among the major *Thalassiosira* lineages (Fig. 2). Based on the available information, it is difficult to infer whether codon changes in the two sites generated gamete isolation in sympatry or allopatry among the diverging lineages. However, the findings here suggest that positive selection for species-specific amino acids in a functional domain of a reproductive protein can be one of several mechanisms contributing to the fast diversification of unicellular pelagic organisms (also see discussion by Swanson and Vacquier 2002).

Acknowledgments. The authors would like to thank Simon D. W. Frost and Craig Van Bell for helpful discussion and insightful comments. Suggestions made by Martin Kreitman, Allen Rodrigo, and an anonymous reviewer improved an earlier version of this article. This research was supported in part by a University of California, San Diego Center for AIDS Research Developmental Award to S.L.K.P (AI36214).

References

Armbrust EV (1999) Identification of a new gene family expressed during the onset of sexual reproduction in the centric diatom Thalassiosira weissflogii. Appl Environ Microbiol 65:3121–3128

Armbrust EV, Galindo HM (2001) Rapid evolution of a sexual reproduction gene in centric diatoms of the genus *Thalassiosira*. Appl Environ Microbiol 67:3501–3513

Darling KF, Wade CM, Stewart IA, Kroon D, Dingle R, Brown AJL (2000) Molecular evidence for genetic mixing of Arctic and Antarctic subpolar population of planktonic forminifers. Nature 405:43–47

Darling KF, Kucera M, Pudsey CJ, Wade CM (2004) Molecular evidence links cryptic diversification in polar planktonic protests to Quaternary climate dynamics. Proc Natl Acad Sci USA 101:7657–7662

De Vargas C, Norris R, Zaninetti L, Gibb SW, Pawlowski J (1999) Molecular evidence of cryptic speciation in planktonic foraminifers and their relation to oceanic provinces. Proc Natl Acad Sci USA 96:2864–2868

Drebes G (1977) Sexuality. In: Werner D (ed) The biology of diatoms. University of California Press, Berkeley, pp 250–283

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

Fox M, Sorhannus U (2003) RpoA: A useful gene for phylogenetic analysis in diatoms. J Eukaryot Microbiol 50:471–475

Kaczmarska I, Beaton M, Benoit AC, Medlin LK (2006) Molecular phylogeny of selected members of the order Thalssiosirales (Bacillariophyta) and evolution of the fultoportula. J Phycol 42:121–138

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

Kooistra WHCF, De Stefano M, Mann DG, Salma N, Medlin LK (2003) The phylogenetic position of *Toxarium*, a pennate-like lineage within centric diatoms (Bacillariophyceae). J Phycol 39:185–197

Kosakovsky Pond SL, Frost SDW (2005) Not so different after all: a comparison of methods for detecting amino-acid sites under selection. Mol Biol Evol 22:1208–1222

Kosakovsky Pond SL, Muse SV (2005) Site-to-site variation in synonymous substitution rates. Mol Biol Evol 22(12):2375–2385

Kosakovsky Pond SL, Simon SDW (2005) Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics 21(10):2531–2533

Kosakovsky Pond SL, Frost SDW, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics 21:676–679

Mann DG (1999) Species concepts in diatoms. Phycologia 38:437–495

Medlin LK, Kaczmarska I (2004) Evolution of the diatoms: V. Morphological and cytological support for the major clades and a taxonomic revision. Phycologia 43:245–270

Muse SV, Gaut BS (1994) A likelihood approach for comparing synonymous and nonsynonymous nucleotide substitution rates, with application to the chloroplast genome. Mol Biol Evol 11:715–724

Norris RD (2000) Pelagic species diversity, biogeography and Evolution. In: Erwin DH, Wing SL (eds) Deep time: paleobiology's perspective. Allen Press, Lawrence, KS, pp 236–258

Page RDM (1996) TREEVIEW: An application to display phylogeneticrees on personal computers. Comp Appl Biosci 12:357–358

Palumbi SR (1994) Genetic divergence, reproductive isolation and marine speciation. Annu Rev Ecol Syst 25:547–572

Round FE, Crawford RM, Mann DG (1990) The diatoms: biology and morphology of the genera. Cambridge University Press, Cambridge

Rynearson T, Armbrust VE (2004) Genetic differentiation among populations of planktonic marine diatom *Ditylum brightwellii* (Bacillariophyceae). J Phycol 40:34–43

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

Sawyer SA (1999) GENECONV: A computer package for the statistical detection of gene conversion. Distributed by the author. Department of Mathematics, Washington University, St. Louis, MO; available at http://www.math.wustl.edu/∼sawyer

Self SG, Liang K-Y (1987) Asymptotic properties of maximum likelihood estimators and likelihood ratio test statistic under nonstandard conditions. JASA 82(398):605–610

Simonsen R (1972) Ideas for a more natural system of the centric diatoms. Nova Hedwigia 29:37–54

Simonsen R (1979) The diatom system: ideas on phylogeny. Bacillaria 2:9–71

Sorhannus U (1997) The origination time of diatoms: an analysis based on ribosomal RNA data. Micropaleontology 43:215–218

Sorhannus U (2003) The effect of positive selection on a sexual reproduction gene in *Thalassiosira weissflogii* (Bacillariophyta): results obtained from maximum likelihood and parsimony-based methods. Mol Biol Evol 20:1326–1328

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

Sorhannus U, Gasse F, Perasso R, Baroin-Tourancheau A (1995) A preliminary phylogeny of diatoms based on 28S ribosomal RNA sequence data. Phycologia 34:65–73

Suzuki Y, Nei M (2004) False-positive selection identified by ML-based methods: examples from the Sig1 gene of the diatom *Thalassiosira weissflogii* and the tax gene of a human T-cell lymphotropic virus. Mol Biol Evol 21:914–921

Swanson WJ, Vacquier VD (1995) Extraordinary divergence and positive Darwinian selection in a fusagenic protein coating the acrosomal process of abalone spermatozoa. Proc Natl Acad Sci USA 92:4957–4961

Swanson WJ, Vacquier VD (2002) Reproductive protein evolution. Annu Rev Ecol Syst 33:161–179

Swanson WJ, Yang Z, Wolfner MF, Aquadro CF (2001) Positive Darwinian selection drives the evolution of several female reproductive proteins in mammals. Proc Natl Acad Sci USA 98:2509–2514

Swofford DL (2002) PAUP*: Phylogenetic Analysis Using Parsimony (*and other methods), version 4.0b10. Sinauer Associates, Sunderland, MA

Tamura K, Nei M (1993) Estimation of the number of nucleotide substitutions in control region of mithocondrial DNA in humans and chimpanzees. Mol Biol Evol 10:512–526

Thompson JD, Higgins DG, Gibson TJ (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions–specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673–4680

Tsaur SC, Wu CI (1997) Positive selection and the molecular evolution of a gene of male reproduction, Acp26Aa of Drosophila. Mol Biol Evol 14:544–549

Vacquier VD (1998) Evolution of gamete recognition proteins. Science 281:1995–1998

Wyckoff GJ, Wang W, Wu CI (2000) Rapid evolution of male reproductive genes in the descent of man. Nature 403:304–309

Xia X (2000) Data analysis in molecular biology and evolution Kluwer Academic, Dordrecht