APPLIED AND ENVIRONMENTAL MICROBIOLOGY, July 1999, p. 3121–3128 Vol. 65, No. 7
0099-2240/99/$04.00+0

Identification of a New Gene Family Expressed during the Onset of Sexual Reproduction in the Centric Diatom *Thalassiosira weissflogii*

E. VIRGINIA ARMBRUST*

Marine Molecular Biotechnology Laboratory, School of Oceanography, University of Washington, Seattle, Washington 98195

Received 22 February 1999/Accepted 4 May 1999

An intriguing feature of the diatom life cycle is that sexual reproduction and the generation of genetic diversity are coupled to the control of cell size. A PCR-based cDNA subtraction technique was used to identify genes that are expressed as small cells of the centric diatom *Thalassiosira weissflogii* initiate gametogenesis. Ten genes that are up-regulated during the early stages of sexual reproduction have been identified thus far. Three of the sexually induced genes, *Sig1*, *Sig2*, and *Sig3*, were sequenced to completion and are members of a novel gene family. The three polypeptides encoded by these genes possess different molecular masses and charges but display many features in common: they share five highly conserved domains; they each contain three or more cysteine-rich epithelial growth factor (EGF)-like repeats; and they each display homology to the EGF-like region of the vertebrate extracellular matrix glycoprotein tenascin X. Interestingly, the five conserved domains appear in the same order in each polypeptide but are separated by variable numbers of nonconserved amino acids. SIG1 and SIG2 display putative regulatory domains within the nonconserved regions. A calcium-binding, EF-hand motif is found in SIG1, and an ATP/GTP binding motif is present in SIG2. The striking similarity between the SIG polypeptides and extracellular matrix components commonly involved in cell-cell interactions suggests that the SIG polypeptides may play a role in sperm-egg recognition. The SIG polypeptides are thus important molecular targets for determining when and where sexual reproduction occurs in the field.

Diatoms are one of the most abundant eukaryotic unicellular algae known, with as many as 20,000 extant species distributed among marine and freshwater ecosystems (27, 36). A defining characteristic of diatoms is their siliceous cell wall, or frustule. The physical and developmental constraints associated with the generation of this relatively inflexible cell wall lead to one of the more intriguing aspects of the diatom life cycle: each mitotic division results in the formation of two differently sized daughter cells, one that is the same size as the parent and one that is slightly smaller. Over successive generations, therefore, the mean cell size of a population decreases and the standard deviation about this mean increases (26, 35).

The most common manner of escaping this trend of diminishing cell size is through sexual reproduction (reviewed in reference 12). It is generally believed that once a cell decreases in size to a diameter of less than about 30 to 40% of the maximum diameter for a given species (11), it can be induced by a wide variety of environmental triggers (see, e.g., references 2, 10, 18, and 41) to exit the mitotic cycle and initiate sexual reproduction. Centric diatoms are monoecious, and thus both sperm and eggs can be formed within a single clone (11). It remains unclear how the sperm and egg find one another, although some have hypothesized that the eggs produce pheromone-like compounds to attract sperm (see, e.g., reference 10). Once the sperm and egg do recognize one another, a series of signalling events must allow the sperm to gain entry past the egg frustule so that the gametes can fuse to create the diploid zygote, or auxospore. The auxospore then breaks free of its frustule and forms a postauxospore cell that is able to generate an entirely new frustule and thus a cell many times larger than either parent (30, 33). Importantly, these newly formed large cells rapidly resume asexual reproduction and are essentially "immune" to further induction triggers until an appropriately small cell size is obtained and external triggers can once again elicit the sexual cycle. Thus, those newly generated large cells whose genotypes convey selective advantages can be quickly dispersed throughout a population (see, for example, reference 1).

Sexual reproduction and the accompanying generation of genetic diversity in diatoms are therefore intimately coupled to the control of cell size. The mechanisms underlying this coupling remain mysterious, however. As a first step toward teasing apart the steps required for the transition from vegetative growth to sexual reproduction, I have begun to identify sexual reproduction-specific genes in the centric diatom *Thalassiosira weissflogii*. An extremely sensitive PCR-based cDNA subtraction technique was used to identify, for the first time, genes that are transcribed within the first few hours of entry into the sexual cycle, a period that precedes the major morphological changes associated with gamete formation (2). The identification of sexual reproduction-specific genes in diatoms will ultimately permit the generation of molecular markers specific to sexually reproducing cells and thus will allow a determination of when and where sexual reproduction occurs in the field, a question that has been difficult to address by more traditional techniques.

MATERIALS AND METHODS

Culture conditions. Clonal isolates of *T. weissflogii* Grun. clone Actin (from the Culture Collection of Marine Phytoplankton, Bigelow Laboratories for Ocean Sciences) were obtained by plating cells on f/2 enriched seawater (20) solidified with 1.5% agar (Meer Corporation) and then transferring individual colonies to liquid f/2 medium. The size distributions of the isolates were determined with a Coulter Multisizer II (Coulter Corporation). A number of isolates were chosen based on their size distributions and maintained in exponential growth in a semicontinuous batch culture at 20°C and 120 microeinsteins of

*Mailing address: University of Washington, School of Oceanography, Box 357940, Seattle, WA 98195. Phone: (206) 616-1783. Fax: (206) 543-6073. E-mail: armbrust@ocean.washington.edu.

continuous illumination · m^{-2} · s^{-1} (1). Each isolate was tested for responsiveness to a sexual induction trigger. The induction signal used for all experiments involves an interruption of exponential growth in continuous light with 12 hours of darkness (2). Once dark-exposed cultures are returned to continuous light, induced cells enter relatively synchronously into the sexual cycle (2). Approximately 24 h after a return to continuous light, aliquots of the induced cultures were examined microscopically to determine whether sexual stages were present. A culture that formed sexual stages in response to the dark induction was defined as responsive. A culture that did not form sexual stages in response to the dark induction trigger was defined as unresponsive.

RNA isolation. Samples were collected for total RNA isolation 5 h after dark-induced cultures were returned to continuous light. Ten liters of induced cultures at approximately 7×10^4 cells · ml^{-1} were filtered through 1.2-μm-pore-size Millipore filters, and the filtered cells were either frozen at −70°C before processing or else processed immediately. Total RNA was isolated essentially as described by Kirk and Kirk (24). Briefly, approximately 10 ml of lysis buffer (50 mM Tris [pH 8], 0.3 M NaCl, 2% sodium dodecyl sulfate [SDS], 15 mM EGTA, and 1.5% freshly added diethyldithiocarbamic acid) was used per 3.5×10^8 filtered cells, and the cells were incubated at 37°C for 30 min with intermittent vortexing. Cell debris was removed by centrifugation at 10,000 × *g* for 10 min, and 2 M KCl was added to the resulting supernatant to achieve a final concentration of 0.23 M KCl. The mixture was incubated on ice for 15 min and centrifuged at 10,000 × *g* for 10 min. A 1 M Tris (pH 9) solution was added to the resulting supernatant to achieve a final concentration of 34 mM, and then the supernatant was extracted twice with Tris-buffered phenol (Amresco). Nucleic acids were precipitated with ethanol at −20°C, and the pellet was resuspended in 4 ml of water. RNA was precipitated overnight on ice by the addition of an equal volume of 4 M LiCl. The RNA was pelleted at 10,000 × *g* for 10 min and resuspended in water or TE (10 mM Tris [pH 7.6]–1 mM EDTA). Total RNA was quantified with a GeneQuant RNA/DNA calculator (Pharmacia). Poly(A)-selected RNA was isolated according to the manufacturer's instructions by using the Oligotex mRNA Isolation Kit (Qiagen).

cDNA subtraction. cDNAs were generated and subtracted according to the manufacturer's instructions by using the PCR-Select cDNA Subtraction Kit (Clontech). Briefly, the mRNA isolated from the responsive culture and the unresponsive culture was reverse transcribed by using avian myeloblastosis virus reverse transcriptase (Clontech) in two separate reactions to generate two populations of double-stranded cDNAs, one representative of the genes transcribed in the responsive culture and one representative of the genes transcribed in the unresponsive culture. Both sets of cDNAs were restriction digested with *Rsa*I to generate cDNA fragments. The digested cDNAs from the responsive culture were split into two aliquots. Adapter 1 (supplied with the kit) was ligated to the cDNAs from one aliquot; adapter 2R (supplied with the kit) was ligated to the cDNAs from the other aliquot. A ligation efficiency test was performed to ensure that at least 25% of the cDNAs from each aliquot possessed the appropriate adapter. For the efficiency test, two gene-specific control primers were designed to amplify the carbonic anhydrase gene recently cloned from *T. weissflogii* (34). The carbonic anhydrase-specific PCR primers are ACCTCGATATGGAGACT CTTC (forward) and CCCATTCCCATTTCTTCATCG (reverse).

Forward subtraction was designed to identify cDNAs either unique to, or up-regulated in, the responsive culture. First, an excess of cDNAs (without ligated adapters) from the unresponsive culture was mixed in two separate reactions with either adapter 1- or adapter 2R-ligated cDNAs from the responsive culture. The two tubes were separately heated to 95°C to denature the cDNAs, and then each was hybridized at 68°C for 8 h. This step promotes hybridization between the excess, unligated cDNAs from the unresponsive culture and their adapter-ligated complements from the responsive culture. To maximize the subtraction of common cDNAs, the contents of the two tubes were then mixed together without a second denaturation step, combined with an additional excess of heat-denatured, unligated cDNAs from the unresponsive culture, and hybridized overnight at 68°C. cDNA ends were then filled in by incubating the subtracted cDNAs at 75°C for 5 min in the presence of Advantage Polymerase mix (Clontech) and deoxynucleotide triphosphates. PCR primers specific to the two adapters (Clontech) were used to amplify cDNAs created during the subtraction that possessed one DNA strand ligated to adapter 1 and the other DNA strand ligated to adapter 2R. The subtracted and amplified cDNAs were cloned into pGEM-T (Promega) and used to transform One Shot TOP10 competent cells (Invitrogen). This forward-subtracted library is thus enriched for cDNAs specific to the responsive culture. As a control, a reverse subtraction was performed in a similar manner except that the excess cDNAs without ligated adapters used for subtraction had been isolated from the responsive culture and the adapter-ligated cDNAs had been isolated from the unresponsive culture. This population of reverse-subtracted cDNAs is thus enriched for genes specific to the unresponsive culture.

cDNA screening. To determine if the subtracted library contained clones up-regulated in the responsive culture, 117 colonies were randomly chosen and the cDNA insert from each was PCR amplified with the adapter-specific primers. Approximately 0.5 μg of each insert DNA was denatured by addition of an equal volume of 0.6 N NaOH and was then spotted onto duplicate maximum-strength Nytran Plus (Schleicher and Schuell) membranes. The membranes were incubated in 0.5 M Tris (pH 7.5) for 4 min and allowed to air dry, and the DNA was UV cross-linked to the membrane with a UV Stratalinker 1800 (Stratagene). The replicate membranes were then hybridized to either a forward-subtracted or a reverse-subtracted cDNA probe. These probes were generated by first sequentially digesting the populations of reverse- or forward-subtracted cDNAs with *Rsa*I, *Sma*I, and *Eag*I to ensure complete removal of the adapters. The digested cDNAs were separated from the adapters with a Qiaquick column (Qiagen), and the cDNAs were fluoroscein labeled by using the Random Prime Labeling and Signal Amplification System for the Fluorimager (Vistra). The hybridization, wash, and signal detection conditions used were those suggested in the Random Prime Labeling kit. Only those inserts that hybridized to the forward-subtracted probe but not to the reverse-subtracted probe were screened further. A subset of the inserts from these positive clones was PCR amplified and labeled with the Random Prime Labeling and Signal Amplification System. Approximately 0.5 μg of the unsubtracted cDNAs from the responsive and unresponsive cultures were denatured as before, spotted onto duplicate Nytran membranes, and separately probed with the individually labeled cDNAs.

DNA sequencing and generation of full-length cDNA clones. Plasmid DNA was prepared from positive clones with the Qiagen Mini Prep Kit and sequenced with the ABI PRISM Dye Terminator Cycle Sequencing Ready Reaction Kit with AmpliTaq DNA polymerase by using a combination of adapter-specific and gene-specific primers. DNA sequencing was performed on an Applied Biosystems 373A DNA Sequencer. Sequence data were compiled and analyzed with the Wisconsin Package, version 10.0, of the Genetics Computer Group (GCG), Madison, Wis. (9).

Full-length cDNA clones were generated by RACE (rapid amplification of cDNA ends) technology. One μg of poly(A)-selected mRNA isolated from the responsive culture was used to generate double-stranded cDNAs that were ligated to the Marathon cDNA adapter according to the manufacturer's instructions for the Marathon cDNA Amplification Kit (Clontech). To generate the 5′ end of the cDNA, a gene-specific reverse primer was designed based on the DNA sequence of the cloned cDNA fragment and was used in PCR with a forward primer specific to the Marathon cDNA adapter ligated to the 5′ end. To generate the 3′ end of the cDNA, a gene-specific forward primer was used in PCR with a reverse primer specific to the Marathon adapter ligated to the 3′ end. For clone 42, the 5′ RACE primer was TTCCAGAATCGAGATTGGGAGCAGTGC. For clone 78, the 5′ RACE primer was TCCTTTCGCATGCCTTTCCGTCGT AC. For clone 71, the 5′ RACE primer was CCGTTGGCATTCACAGATGA GTCAAC and the 3′ RACE primer was CTTGACCAACATCCGCACTTTCT CAG. Complete cDNA sequences were obtained by designing new gene-specific primers as new sequence was obtained.

Isolation of genomic clones. Total genomic DNA was isolated from *T. weissflogii* by filtering approximately $5 \cdot 10^6$ cells onto a 0.45-μm-pore-size cellulose filter. The cells were scraped from the filter and incubated at 60°C for 1 h in 300 μl of lysis buffer (10 mM TE [pH 7.5]–0.5% SDS–100 μg of proteinase K/ml); 50 μl of 5 M NaCl and 40 μl of 10% cetyltrimethylammonium bromide (CTAB) in 0.7% NaCl was then added, and the mixture was incubated at 65°C for an additional 10 min. The DNA was purified by using the column and wash reagents provided with the Qiagen DNeasy Plant Mini Kit according to the manufacturer's instructions. Gene-specific PCR primers were used to amplify genomic fragments. These fragments were cloned into pGEM-T and transformed into TOP10 *Escherichia coli* cells. Recombinant plasmid DNA was isolated and sequenced as described above.

Reverse transcriptase PCR (RT-PCR). Two hundred nanograms of poly(A)-selected RNA isolated from the responsive and the unresponsive cultures were reverse transcribed in two separate reactions by using Moloney murine leukemia virus reverse transcriptase (Gibco BRL), along with the primers and conditions for first-strand synthesis provided with the SMART cDNA Synthesis kit (Clontech). Gene-specific forward and reverse PCR primers that had been designed to cross an intron were used to PCR amplify first-strand cDNAs. PCR products were analyzed by agarose gel electrophoresis.

Detection of homology and sequence alignment. Homology between the predicted amino acid sequences and those present in the GenBank database was detected by using BLAST 2.0, provided by the National Center for Biotechnology Information (32a). Multiple amino acid alignments were performed by using the ClustalW 1.7 programs (2a). Corresponding DNA sequences were aligned by hand.

Nucleotide sequence accession numbers. The GenBank accession numbers for the *Sig* sequences are AF154499, AF154500, and AF154501.

RESULTS

Identification of responsive and unresponsive isolates. Clonal cultures of *T. weissflogii* representing a range of size distributions were tested for responsiveness to a dark induction trigger (2). Two clones that displayed different size distributions and widely different responses to the induction cue were chosen (Fig. 1). Since the focus of this study was to identify genes expressed during the early stages of sexual reproduction, mRNA was isolated from the two induced cultures 5 h after their return to continuous light, a time that precedes the ob-

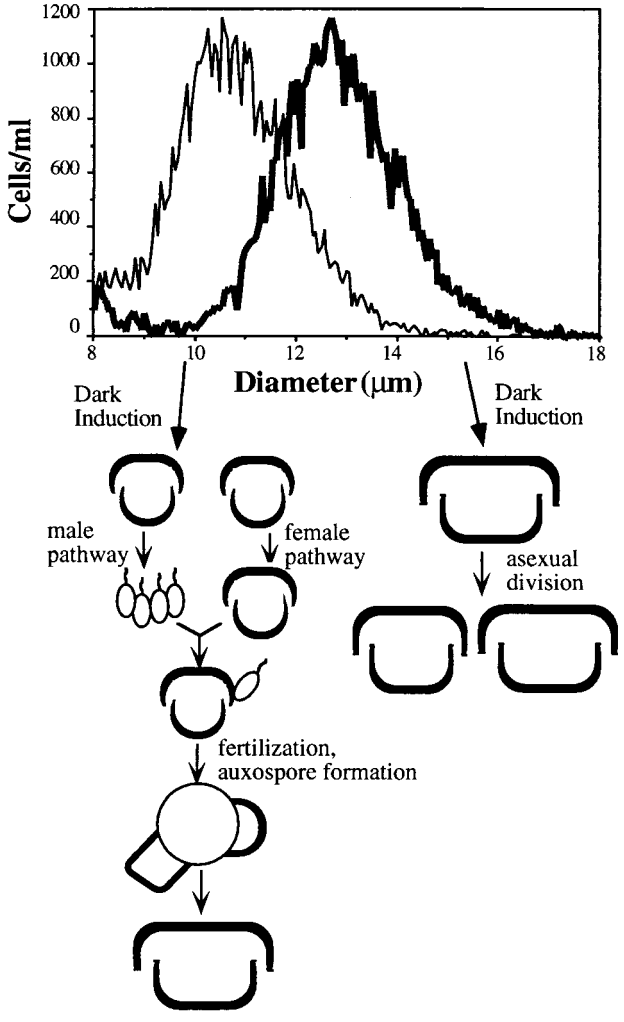

FIG. 1. Coulter size distributions of the responsive (thin line) and unresponsive (thick line) isolates used for sexual induction and a simplified schematic of the resulting life cycle features of the responsive (left side) and unresponsive (right side) cells. The mean cell diameter of the responsive culture was 10.8 μm, and the mean cell diameter of the unresponsive culture was 12.9 μm.

vious morphological changes associated with sperm formation (2). Twenty-four hours after the return to continuous light, approximately 40% of the cells observed in a remaining aliquot of the responsive culture were readily identifiable male gametes, whereas about 1% of the cells in the unresponsive culture were male gametes. Rare auxospores (data not shown) were observed only in the responsive culture, suggesting that even more cells within this culture initiated the sexual cycle, since some of the "vegetative" cells must have been eggs.

Multiple genes are up-regulated during the onset of sexual reproduction. One hundred seventeen individual clones from the sexual gene-enriched library (forward-subtracted) were screened with probes made from either the sexual (forward-subtracted) cDNAs or the vegetative (reverse-subtracted) cDNAs to determine the percentage of the library that contained cDNAs up-regulated during the onset of sexual reproduction. Of the 117 cDNAs tested, 25 hybridized specifically to the forward-subtracted probe and 15 hybridized to both the forward- and reverse-subtracted probes, indicating that these cDNAs had somehow "slipped through" the subtraction process. The vast majority of the cDNAs, however, hybridized to neither probe at a detectable level, suggesting that cDNAs of this class are expressed at relatively low levels (data not shown). Nineteen of the 25 strongly hybridizing cDNAs were used as individual probes against the total (unsubtracted) unresponsive and responsive cDNA populations to confirm that these particular cDNAs were truly up-regulated in the sexual culture. Eleven of these cDNAs appeared to be expressed at higher levels in the responsive cultures, although two of the clones were later shown to correspond to the same gene. Thus, a total of 10 of the 19 cDNAs tested appeared to be up-regulated in the responsive culture (Fig. 2), although some cDNA clones, such as clones 56 and 52, appeared to be only slightly up-regulated, while others, such as clones 29 and 71, appeared to be strongly up-regulated.

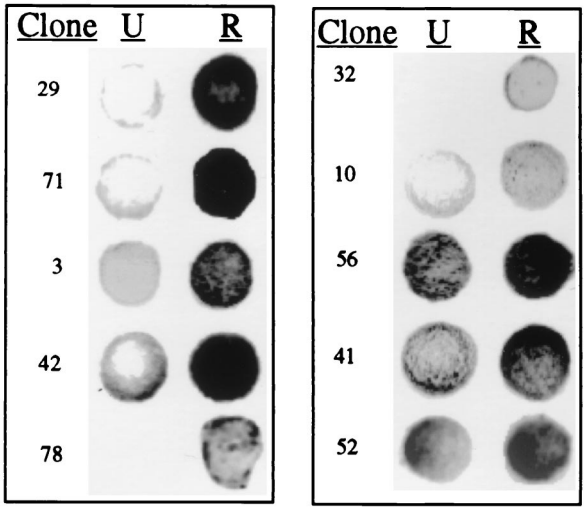

FIG. 2. Comparison of the steady-state levels of transcription of 10 differentially expressed cDNA clones. Spots of approximately 0.5 μg of the total cDNAs isolated from the unresponsive (U) and responsive (R) cultures 5 h after the dark-induced cultures were returned to continuous light are shown.

A novel gene family is expressed during sexual reproduction. The 10 positive cDNA clones were sequenced, and three of these partial cDNAs—clones 42, 71, and 78—displayed similar features. The genes corresponding to these three clones are now referred to as *Sig*, for sexually induced gene. *Sig1* corresponds to cDNA clone 42, *Sig2* corresponds to cDNA clone 71, and *Sig3* corresponds to cDNA clone 78. Gene-specific RACE PCR primers were designed to amplify the 5′ and 3′ ends of these cDNAs and thus allow the full-length amino acid sequences to be predicted. In each instance, the initiator methionine was assumed to be the first methionine of the longest open reading frame. The full-length *Sig1* cDNA is 2,432 bp, with a 5′ untranslated region (UTR) of 23 bp and a 3′ UTR of 56 bp; the full-length *Sig2* cDNA is 1,395 bp, with a 5′ UTR of 20 bp and a much longer 3′ UTR of 320 bp; the full-length *Sig3* cDNA is 906 bp, with a 5′ UTR of 47 bp and a 3′ UTR of 198 bp. The *Sig1* genomic sequence contains four introns at cDNA positions (from the 5′-most end of the cDNA) 186, 709, 1198, and 1985. The *Sig2* and *Sig3* genomic sequences each contain a single intron at cDNA positions 123 and 144, respectively (Fig. 3A). The sizes of these introns are remarkably homogeneous, ranging only from 77 to 86 bp. Moreover, the introns are more A+T rich (G+C content of 26 to 39%) than the coding regions (G+C content of 47 to 49%).

Preliminary analysis of *Sig1*, *Sig2*, and *Sig3* expression by dot

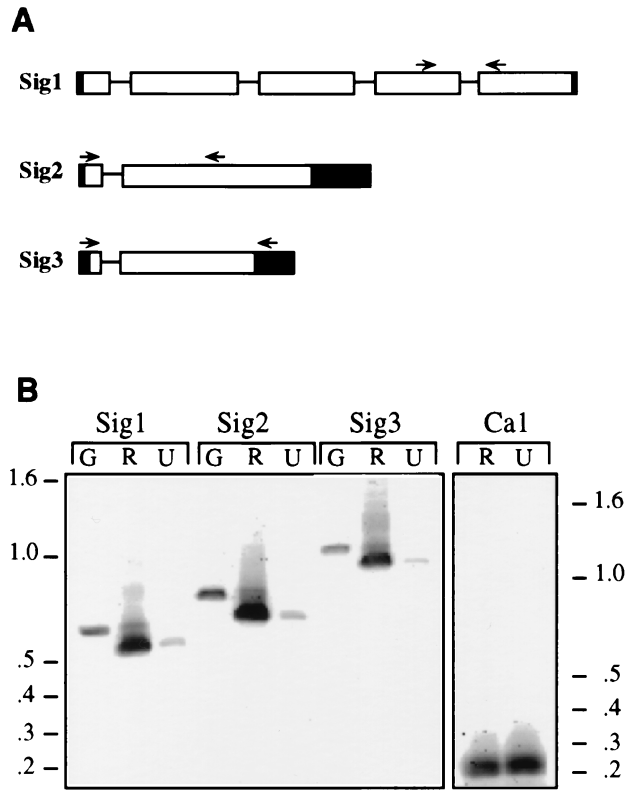

FIG. 3. Structure and expression of *Sig1*, *Sig2*, and *Sig3*. (A) Schematic of the genomic structures of the *Sig1*, *Sig2*, and *Sig3* transcription units. Rectangles, exons; lines, introns; solid rectangles, untranslated RNA. Arrows indicate approximate locations of the PCR primers used for RT-PCR. (B) Two hundred nanograms of poly (A)-selected RNA isolated from the responsive (R) and unresponsive (U) cultures 5 h after the dark-induced cultures were returned to continuous light were reverse transcribed. Total genomic DNA (G) or equal amounts of the first strand-cDNAs (R or U) were PCR amplified with the primers specific to *Sig1*, *Sig2*, and *Sig3* shown in panel A. Molecular weight markers (in kilobases) are shown on the left. PCR primers specific to carbonic anhydrase (Cal) were used to PCR amplify equal amounts of the first-strand cDNAs, and PCR products were analyzed on a second gel; molecular weight markers for this gel are shown on the right.

blots (Fig. 2) suggested that these three genes are up-regulated during the early stages of sexual reproduction. RT-PCR was used to more specifically examine the expression of these genes. Gene-specific PCR primer pairs were designed such that each pair spanned an intron, thus ensuring that the resulting PCR product originated from reverse-transcribed mRNA rather than any contaminating genomic DNA. Steady-state levels of mRNAs of *Sig1*, *Sig2*, and *Sig3* in the responsive and unresponsive cultures (at 5 h after the induced cultures were returned to continuous light) were compared to the steady-state levels of mRNA associated with carbonic anhydrase, a gene involved in inorganic carbon acquisition whose expression is not expected to be affected during the very early stages of sexual reproduction. The products observed for *Sig1*, *Sig2*, and *Sig3* (Fig. 3B, lanes U and R) were the sizes predicted for amplification of the reverse-transcribed cDNAs and are smaller than those predicted for the genomic DNA (Fig. 3B, lanes G). The amount of RT-PCR product specific to *Sig1*, *Sig2*, and *Sig3* mRNA was much greater in the responsive cultures than in the unresponsive cultures (Fig. 3B). In contrast, the amounts of RT-PCR product specific to carbonic anhydrase mRNA were approximately equal in the two cultures (Fig. 3B). No product was observed with the no-template controls (data not shown). These results imply that the steady-state levels of mRNA resulting from expression of *Sig1*, *Sig2*, and *Sig3* are greatly increased during the onset of sexual reproduction. The small amount of RT-PCR product specific to these three messages in the unresponsive culture is likely due to the low percentage of cells in this culture that were later observed to form male gametes.

The predicted amino acid sequences of the *Sig* cDNAs display a series of common features (Fig. 4A). Each possesses a putative signal sequence, characterized by a stretch of 12 to 14 hydrophobic amino acids preceded at the amino terminus by 1 or 2 basic residues (44). Each also lacks obvious hydrophobic stretches characteristic of transmembrane domains, suggesting that these three polypeptides may be secreted. Each of the predicted amino acid sequences also displays a cysteine-rich motif originally identified in human epithelial growth factor (EGF) (reviewed in reference 8). The diatom polypeptides each display a series of the consensus motif $CXCX_5GX_2C$ or $CXCX_2GaX_4C$ (where "X" refers to any amino acid and "a" denotes aromatic amino acids) that are characteristic of EGF-like motifs (Fig. 4A). A more stringent version of the EGF-like module has been defined by Campbell and Bork (5) as $X_4CX_{2-7}CX_{1-4}(G/A)XCX_{1-13}t_2aXCXCX_2GaX_2CX$ (where "t" denotes nonhydrophobic amino acids), although these authors note that large deviations are commonly observed with the motif. The predicted amino acid sequences of the diatom polypeptides display a variation of this motif, particularly in the number of amino acids that can separate cysteines 1 and 2 and cysteines 3 and 4. The diatom motif is $CX_{3-22}CX_3GXCX_{5-6}CXCX_5GX_2C$ or $CX_{3-109}CX_3GXCX_{5-25}CXCX_2GaX_4C$.

SIG1 and SIG2 are predicted to be negatively charged polypeptides with isoelectric points (pI) of 4.6 and 4.3, respectively (Fig. 4A). The predicted molecular mass of SIG1 is 83.5 kDa, about twice that of SIG2 (41.5 kDa). SIG3 is the smallest of the polypeptides, with a predicted molecular mass of 23.8 kDa, and is predicted to be substantially less acidic, with a pI of 6.0 (Fig. 4A). Both SIG1 and SIG2 display additional sequence motifs besides the EGF-like domains. SIG1 possesses the tri-amino acid sequence RGD (Fig. 4A). In many animal proteins and some plant proteins (39), the RGD domain has been shown to be the recognition site for a class of proteins known as integrins, which span the plasma membrane and essentially connect the inside cytoskeleton of the cell to the extracellular matrix (23, 37). SIG1 also contains the sequence DWDPENNVEVSDW, which is similar to DXDaENbVEX XDW (where "X" stands for any amino acid; "a" stands for I, L, V, F, Y, or W; and "b" stands for G or P), a version of a calcium binding motif known as an EF hand (28, 32). The SIG1 EF hand-like motif possesses the appropriate, highly conserved amino acids at positions 1, 3, and 12. SIG1 possesses an EF hand-like motif at a single location within the polypeptide, like SPARC, a protein involved in bone morphogenesis (14). In this respect, however, it is unlike most other proteins, which commonly display EF hands at multiple positions (32), including a 75-kDa protein that is a component of the cell wall of the diatom *Cylindrotheca fusiformis* (25). The second acidic SIG protein, SIG2, possesses the glycine-rich sequence GLGAGGKT (Fig. 4A), which corresponds to the ATP/GTP binding consensus sequence A (also referred to as the P loop) (38).

Although SIG1, SIG2, and SIG3 vary greatly in size and pI, there are five domains (I through V) that display strong identity among all the polypeptides (Fig. 4), suggesting that these polypeptides are encoded by a single gene family. For example, in domain I, 21 of 54 amino acids are identical in all three polypeptides and 36 of 54 are identical in two of the three polypeptides. Moreover, each domain, except domain III, has

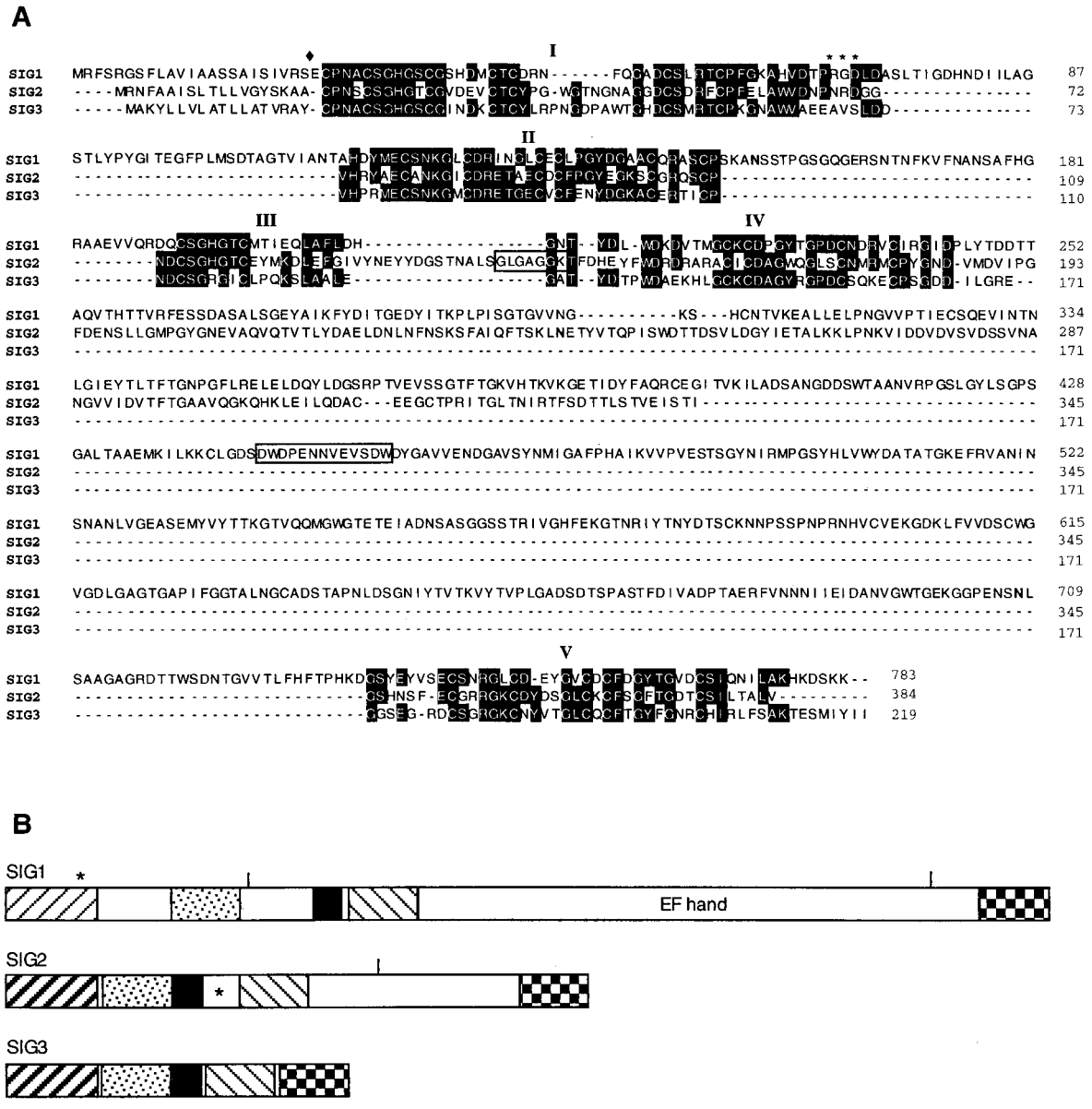

FIG. 4. Alignment and structure of the predicted amino acid sequences of SIG1, SIG2, and SIG3. (A) Alignment of the predicted amino acid sequences; amino acid numbering, beginning with the initiator methionine, is shown to the right of each line. Dashes, alignment gaps; solid diamond, potential cleavage site of the signal sequences; boldfaced N's, potential N-linked glycosylation sites; areas highlighted in black, amino acid identity domains I through V. The location of RGD in SIG1 is indicated by asterisks. The EF hand in SIG1 and the ATP/GTP binding site in SIG2 are boxed. (B) Schematic of the structure and orientation of domains I through V within the SIG polypeptides. The five different patterns represent the five different domains, arranged from left to right with domain I leftmost. Only domain I of SIG2 and domain III in each polypeptide do not contain an EGF-like motif. Open rectangles, nonconserved regions. In SIG1, the asterisk above domain I indicates the location of the RGD motif. In SIG2, the asterisk indicates the location of the ATP/GTP binding motif. The short vertical lines above SIG1 and SIG2 indicate the locations of potential N-linked glycosylation sites.

an EGF-like motif in at least two of the three polypeptides (Fig. 4). What is particularly striking is that the five domains occur in the same order within each polypeptide but are separated by different numbers of nonconserved amino acids (Fig. 4B). For example, SIG3 is essentially composed of the five domains only, whereas both SIG1 and SIG2 have variable numbers of amino acids separating each domain. It is within these nondomain sequences of SIG1 and SIG2 that potential sites for N-linked glycosylation are found (Fig. 4). Furthermore, the putative EF hand found in SIG1 and the putative ATP/GTP binding domain found in SIG2 are both located within the unique, nondomain regions (Fig. 4), suggesting that the three proteins may perform similar functions but are regulated through different mechanisms. Perhaps not surprisingly, the DNA sequences in the five domains are around 35% identical (data not shown), suggesting that these three genes may have diverged from one another in the evolutionarily distant past.

The SIG polypeptides display homology to the EGF-like domain of the vertebrate extracellular matrix protein tenascin. The SIG polypeptides display strong homology to the EGF-like domain of a family of extracellular matrix glycoproteins known as tenascin. Each member of the tenascin family (tenascin X, R, and C) is composed of four functional domains that

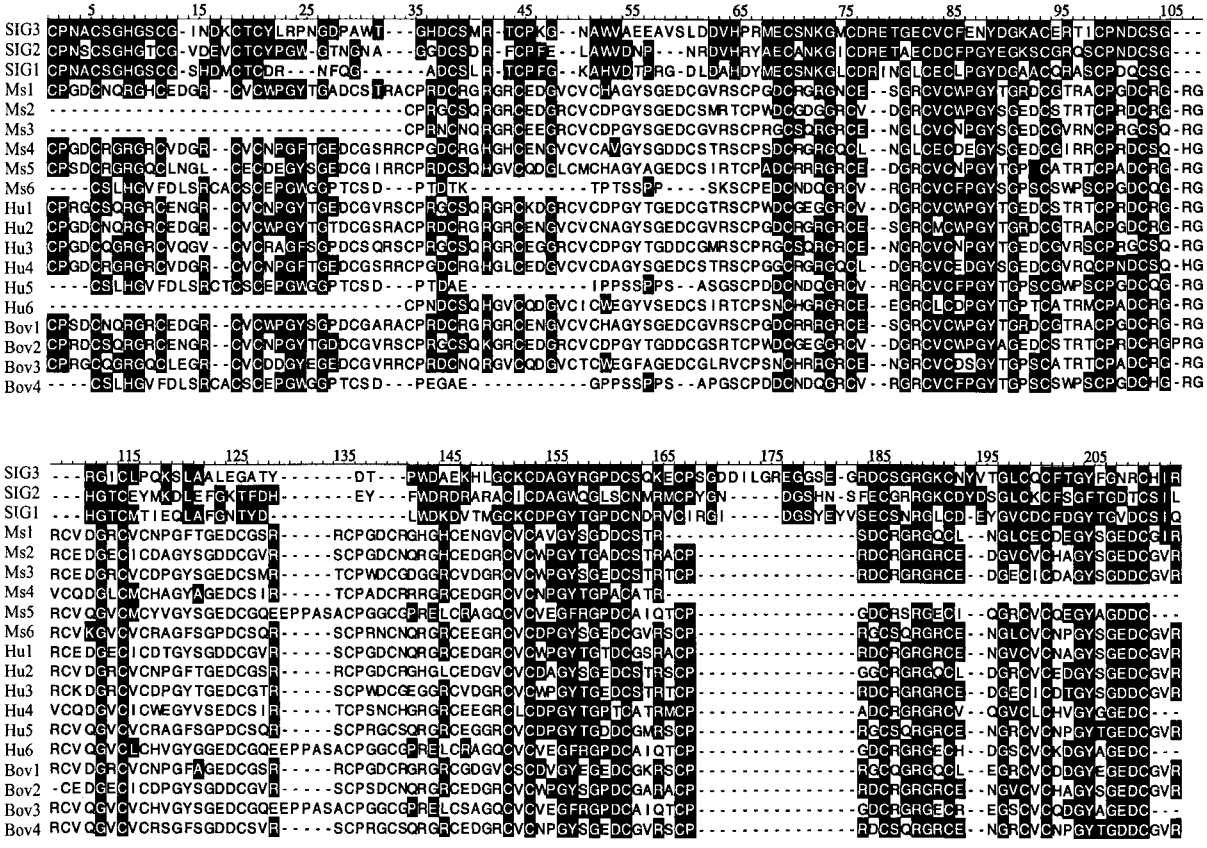

FIG. 5. Alignments of composites of the predicted amino acid sequence from domains I through V for SIG1, SIG2, and SIG3 with the EGF-like domains of tenascin X from mice (GenBank accession no. 2564958), humans (4), or cows (13). Alignment gaps are indicated by dashes. Identical and similar amino acids (similarly charged acidic [D and E] or basic [R, K, and H] residues or uncharged [M, L, V and A] residues) are highlighted in black. Amino acids 19 to 206 of SIG3 are included. Amino acids 391 to 575 (Ms1), 298 to 451 (Ms2), 236 to 389 (Ms3), 484 to 637 (Ms4), 546 to 732 (Ms5), and 132 to 296 (Ms6) of the mouse sequence, amino acids 283 to 467 (Hu1), 407 to 591 (Hu2), 221 to 405 (Hu3), 500 to 681 (Hu4), 146 to 312 (Hu5), and 593 to 748 (Hu6) of the human sequence, and amino acids 403 to 619 (Bov1), 279 to 463 (Bov2), 558 to 744 (Bov3), and 142 to 308 (Bov4) of the bovine sequence are included.

together promote cell-cell interactions during different stages of development (7, 15). The strongest homology is observed between SIG3 and EGF-like domains of tenascin X from mice (GenBank accession no. 2564958), humans (4), and cows (13). Since SIG3 is composed primarily of domains I through V only (Fig. 4), and because these five domains occur in the same order in each SIG polypeptide (Fig. 4), composites of the amino acid sequences corresponding to the five domains for each polypeptide were constructed and compared to tenascin X. Sequence homologous to the entire length of the five domains is found in six overlapping repeats throughout the EGF-like domains of mouse and human tenascin X and in four overlapping repeats throughout the EGF-like domain of bovine tenascin X (Fig. 5). This strongly conserved homology between a diatom and a vertebrate protein again suggests that the five domains within the SIGs may reflect functional domains.

DISCUSSION

The ability of centric diatoms to form anisogamous sperm and eggs was first suspected in the mid-1930s (see, e.g., references 3 and 19) and eventually confirmed some 15 years later (45). Since then, the life cycles of numerous centric diatoms have been described based on both laboratory and field observations. A common feature of many of these studies, however, is a certain amount of luck, because easily observable sexual events are rarely found. Thus, most of our knowledge of the impact and extent of diatom sexual reproduction that occurs in the field is based on inference (see, e.g., reference 6). Consequently, our understanding of how the sexual cycle is induced, how gametes find one another, and how sexual events influence the genetic diversity of field populations remains quite limited. The approach taken here has been to identify the genes expressed during the early stages of sexual reproduction in the sexually manipulable diatom *T. weissflogii* (1, 2, 43) with the goals of both understanding the onset of sexual events in more detail and developing molecular markers for monitoring these events in the field.

Sexual reproduction-specific gene expression. Sexual reproduction in diatoms is often triggered when vegetative cells are exposed to unfavorable growth conditions (12), an event that likely induces expression of a wide variety of genes, some of which will be integral to the onset of sexual reproduction and some of which will be part of a more general response to stress. The key step, then, in the identification of genes specifically required for sexual reproduction was to eliminate the stress response genes from analysis. The fact that the ability to undergo sexual reproduction in diatoms is coupled to the obtainment of an appropriate cell size provided an ideal means for distinguishing between the two response types. Within a given species of diatom, large cells tend to "ignore" sexual induction triggers and continue to divide asexually, while small cells tend

to respond to the same induction triggers by forming gametes. This implies that a similar suite of stress response genes is expressed whether small or large cells are exposed to an induction trigger. In contrast, sexual reproduction genes should be expressed only when appropriately sized small cells are exposed to the induction trigger. Furthermore, because the small cells of *T. weissflogii* produce gametes in a relatively synchronous manner once a dark-induced culture is returned to continuous light (2), a snapshot of the pattern of gene expression occurring during the transition to sexual reproduction could be obtained.

The genes either expressed specifically or up-regulated during the early stages of sexual reproduction were identified by an extremely sensitive PCR-based cDNA subtraction protocol. One of the more powerful aspects of this approach was that differentially expressed genes could be identified even though only a portion of the induced population initiated the sexual cycle; the key was simply that a subset of genes expressed under the experimental conditions differed from those expressed under the control conditions.

Thus far, 10 sexually induced genes have been identified, although it seems likely that these genes represent only a minor subset of the possible suite of sexual reproduction-specific genes expressed in *T. weissflogii*. The transition from asexual to sexual reproduction in diatoms undoubtedly requires the induction of numerous genes in order to allow the diploid vegetative cells to exit the mitotic cycle and irreversibly undergo meiosis to form functional haploid gametes. The flagellated sperm that are formed lack a frustule and look and behave nothing like a vegetative cell. In fact, some of the early controversy over the centric diatom life cycle centered around the difficulty in distinguishing sperm from potentially contaminating flagellates (19). The changes that accompany the formation of egg cells are morphologically more subtle but nonetheless result in the creation of cells that the sperm can readily distinguish from vegetative cells of its own and other species. Lastly, once the sperm and egg cells do find and recognize one another, a signalling event must allow the sperm entry past the frustule of the egg cell so that plasma membrane fusion and subsequent nuclear fusion can occur to create the zygote. Thus, genes necessary for recognition and signalling events, in addition to genes necessary for morphological changes, are likely to be differentially expressed during the early stages of the sexual cycle.

Potential involvement of the extracellular matrix in gamete recognition. Despite the relatively small number of genes analyzed thus far, a new gene family, known as *Sig*, consisting of at least three members strongly up-regulated during the onset of sexual reproduction, has been identified. The three polypeptides predicted to be encoded by these genes display a number of common features: they each contain three or more EGF-like motifs; they each have five highly conserved domains that display strong homology to the EGF-containing domain of the vertebrate extracellular matrix glycoprotein tenascin X; they each possess a signal sequence; and they each appear to lack transmembrane domains. Hundreds of (predominantly animal) proteins that contain the EGF-like, cysteine-rich motif (5, 8) have been identified. These EGF-containing proteins fall into four general categories: growth factors, transmembrane receptors or adhesion molecules, soluble secreted proteins, and extracellular matrix proteins. Because of the similarity of the SIG polypeptides to the extracellular matrix glycoprotein tenascin X and the absence of transmembrane domains, the SIG polypeptides are hypothesized to be components of the extracellular matrix.

The functional theme shared by proteins of the extracellular matrix is cell adhesion, which is commonly accomplished through protein-protein interactions mediated by the cysteine-rich EGF-like domains. Because of the timing of expression of the *Sig* gene family, the SIG polypeptides are hypothesized to be involved in sperm-egg recognition events. A number of proteins, such as SPE-9 from *Caenorhabditis elegans* (42) and a PKD1-like protein from the sea urchin (31), also contain EGF-like motifs and also have been hypothesized to be required for sperm-egg interactions. The SPE-9 protein, for example, is composed essentially of 10 EGF-like repeats (42). What distinguishes these invertebrate proteins from the SIG polypeptides, however, is that the invertebrate proteins either possess a transmembrane domain with a short cytoplasmic tail or are attached to the plasma membrane through a glycosylphosphatidylinositol (GPI) anchor (29). Presumably, this membrane anchoring allows adhesion events that occur outside the cell to be communicated to the inside of the cell.

The SIG polypeptides display neither transmembrane domains nor the putative recognition sequence for interaction with a GPI anchor (17). Instead, only SIG1 displays an apparent means for communicating with the inside of the cell. SIG1 possesses an RGD domain which is known to act as an attachment site for integrins. Commonly, extracellular adhesion events are communicated to the inside of the cell by a "relay" from the extracellular matrix to the membrane-spanning integrins to the intracellular cytoskeleton (23). It is unclear whether SIG2 and SIG3, both of which lack the RGD domain, also interact with integrins, since only about 25% of extracellular matrix proteins known to bind to integrins actually possess the RGD domain (23).

It is not yet known whether the SIG polypeptides are produced by the sperm or eggs or both, since both gamete types were present in the responsive culture. In general, sperm cells do not appear to possess a substantial extracellular matrix. However, sperm commonly possess an acrosome vesicle that contains matrix material released during the initial recognition event (21). Polyclonal antibodies are currently being generated against recombinant versions of these polypeptides to examine this question in more detail.

The observed homology between the EGF-like motif of diatom and vertebrate proteins is particularly intriguing. The EGF-like motifs of SIG polypeptides differ slightly from the more commonly observed motif (5)—for example, the first C of one of the motifs can be up to 102 amino acids from the second C—but the conservation of the arrangement of the remaining C's is striking. Furthermore, the diatom versions of the EGF-like motif have been conserved among the three SIG polypeptides despite the fact that the corresponding DNA sequences have diverged from one another. The EGF-like motif has rarely been observed in plant or unicellular organisms: of the more than 2,400 described proteins with EGF or EGF-like motifs, only 4 are found in fungal proteins and 35 are found in plant proteins. None have been previously documented in unicellular algae (see, for example, the SMART database [16, 40]). Interestingly, the only EGF-containing protein from plants or fungi that appears to be part of the extracellular matrix and thus potentially involved in adhesion was found in a pathogenic fungus and presumably allows the fungus to attach to its host before invasion (22). The SIGs may therefore be one of the more ancient examples of polypeptides containing an EGF-like motif involved in gamete recognition.

ACKNOWLEDGMENTS

I thank Ann Murkowski for help with the culture work; Lila Koumandou and Tatiana Rynearson for helpful comments on drafts of the manuscript; and Pam Jensen for help with the sequencing gels.

This work was supported by National Science Foundation grant OCE 9702158.

REFERENCES

1. **Armbrust, E. V., and S. W. Chisholm.** 1992. Patterns of cell size change in a marine diatom: variability evolving from clonal isolates. J. Phycol. **28:**146–156.
2. **Armbrust, E. V., R. J. Olson, and S. W. Chisholm.** 1990. Role of light and the cell cycle on the induction of spermatogenesis in a centric diatom. J. Phycol. **26:**470–478.

2a. **Baylor College of Medicine.** 3 May 1999, revision date. [Online.] Clustal W. Human Genome Sequencing Center, Baylor College of Medicine, Houston, Tex. http://dot.imgen.bcm.tmc.edu:9331/multi-align/multi-align.html. [20 February 1999, last date accessed.]

3. **Braarud, T.** 1939. Microspores in diatoms. Nature **143:**899.
4. **Bristow, J., M. K. Tee, S. E. Gitelman, S. H. Mellon, and W. L. Miller.** 1993. Tenascin-X: a novel extracellular matrix protein encoded by the human XB gene overlapping P450c21B. J. Cell Biol. **122:**265–278.
5. **Campbell, I. D., and P. Bork.** 1993. Epidermal growth factor-like modules. Curr. Opin. Struct. Biol. **3:**385–392.
6. **Crawford, R. M.** 1995. The role of sex in sedimentation of a marine diatom bloom. Limnol. Oceanogr. **40:**200–204.
7. **Crossin, K. L.** 1996. Tenascin: a multifunctional extracellular matrix protein with a restricted distribution in development and disease. J. Cell. Biochem. **61:**592–598.
8. **Davis, C. G.** 1990. The many faces of epidermal growth factor repeats. New Biol. **2:**410–419.
9. **Devereux, J., P. Hoeberli, and O. Smithie.** 1984. A comprehensive set of sequence analysis programs for the VAX. Nucleic Acids Res. **12:**387–395.
10. **Drebes, G.** 1996. On the life history of the marine plankton diatom *Stephanopyxis palmeriana*. Helgol. Wiss. Meeresunters. **13:**101–114.
11. **Drebes, G.** 1977. Sexuality, p. 250–283. *In* D. Werner (ed.), The biology of diatoms. University of California Press, Berkeley.
12. **Edlund, M. B., and E. F. Stoermer.** 1997. Ecological, evolutionary, and systematic significance of diatom life histories. J. Phycol. **33:**897–918.
13. **Elefteriou, F., J.-Y. Exposito, R. Garrone, and C. Lethias.** 1997. Characterization of the bovine tenascin-X. J. Biol. Chem. **272:**22866–22874.
14. **Engel, J., W. Taylor, M. Paulsson, H. Sage, and B. Hogan.** 1987. Calcium binding domains and calcium-induced conformational transition of SPARC/BM-40/osteonectin, an extracellular glycoprotein expressed in mineralized and nonmineralized tissues. Biochemistry **26:**6958–6965.
15. **Erickson, H. P.** 1993. Tenascin-C, tenascin-R and tenascin-X: a family of talented proteins in search of functions. Curr. Opin. Cell Biol. **5:**869–876.
16. **European Molecular Biology Laboratory.** 21 April 1999, revision date. SMART database. [Online.] coot.embl-heidelberg.de/SMART/. [20 February 1999, last date accessed.]
17. **Ferguson, M. A. J., and A. F. Williams.** 1988. Cell-surface anchoring of proteins via glycosyl-phosphatidylinositol structures. Annu. Rev. Biochem. **57:**285–320.
18. **Furnas, M. J.** 1985. Diel synchronization of sperm formation in the diatom *Chaetoceros curvisetum* Cleve. J. Phycol. **21:**667–671.
19. **Gross, F.** 1937. The life history of some plankton diatoms. Phil. Trans. R. Soc. Lond. B **228:**1–47.
20. **Guillard, R. R. L.** 1975. Culture of phytoplankton for feeding marine invertebrates, p. 29–60. *In* W. L. Smith and M. H. Chaney (ed.), Culture of marine invertebrate animals. Plenum Press, New York, N.Y.
21. **Hardy, D. M., M. N. Oda, D. S. Friend, and T. T. F. Huang.** 1991. A mechanism for differential release of acrosomal enzymes during the acrosome reaction. Biochem. J. **275:**759–766.
22. **Hogan, L. H., S. Josvai, and B. S. Klein.** 1995. Genomic cloning, characterization, and functional analysis of the major surface adhesin WI-1 on *Blastomyces dermatitidis* yeasts. J. Biol. Chem. **270:**30725–30732.
23. **Hynes, R. O.** 1994. The impact of molecular biology on models for cell adhesion. Bioessays **16:**663–669.
24. **Kirk, M. M., and D. L. Kirk.** 1985. Translational regulation of protein synthesis, in response to light, at a critical stage of *Volvox* development. Cell **41:**419–428.
25. **Kroger, N., C. Bergsdorf, and M. Sumper.** 1994. A new calcium binding glycoprotein family constitutes a major diatom cell wall component. EMBO J. **13:**4676–4683.
26. **MacDonald, J. D.** 1869. On the structure of the diatomaceous frustule, and its genetic cycle. Ann. Mag. Nat. Hist. **4:**1–8.
27. **Mann, D. G., and S. J. M. Droop.** 1996. Biodiversity, biogeography and conservation of diatoms. Hydrobiologia **336:**19–32.
28. **Maurer, P., and E. Hohenester.** 1997. Structural and functional aspects of calcium binding in extracellular matrix proteins. Top. Matrix Biol. **15:**569–580.
29. **Mendoza, L. M., D. N. Nishioka, and V. D. Vacquier.** 1993. A GPI-anchored sea urchin sperm membrane protein containing EGF domains is related to human uromodulin. J. Cell Biol. **121:**1291–1297.
30. **Migita, S.** 1967. Sexual reproduction of *Melosira moniliformis* Agardh. Bull. Fac. Fish. Nagasaki Univ. **23:**123–133.
31. **Moy, G. W., L. M. Mendoza, J. R. Schulz, W. J. Swanson, C. G. Glabe, and V. D. Vacquier.** 1996. The sea urchin sperm receptor for egg jelly is a modular protein with extensive homology to the human polycystic kidney disease protein, PKD1. J. Cell Biol. **133:**809–817.
32. **Nakayama, S., and R. H. Kretsinger.** 1994. Evolution of the EF-hand family of proteins. Annu. Rev. Biophys. Biomol. Struct. **23:**473–507.

32a. **National Center for Biotechnology Information.** 13 April 1999, revision date. [Online.] BLAST 2.0. http://www.ncbi.nlm.nih.gov/BLAST/. [20 February 1999, last date accessed.]

33. **Rao, V. N. R.** 1971. Studies on *Cyclotella meneghiniana* Kutz. II. Induction of axuospore formation. Phykos **10:**84–98.
34. **Roberts, S. B., T. W. Lane, and F. M. M. Morel.** 1997. Carbonic anhydrase in the marine diatom *Thalassiosira weissflogii* (Bacillariophyceae). J. Phycol. **33:**845–850.
35. **Round, F. E.** 1972. The problem of cell size during diatom cell division. Nova Hedwigia **31:**485–493.
36. **Round, F. E., R. M. Crawford, and D. G. Mann.** 1990. The diatoms, p. 747. Cambridge University Press, Cambridge, United Kingdom.
37. **Ruoslahti, E., and M. D. Pierschbacher.** 1987. New perspectives in cell adhesion: RGD and integrins. Science **238:**491–497.
38. **Saraste, M., P. R. Sibbald, and A. Wittinghofer.** 1990. The P-loop—a common motif in ATP- and GTP-binding proteins. Trends Biochem. Sci. **15:**430–434.
39. **Schindler, M., S. Meiners, and D. A. Cheresh.** 1989. RGD-dependent linkage between plant cell wall and plasma membrane: consequences for growth. J. Cell Biol. **108:**1955–1965.
40. **Schultz, J., F. Milpetz, P. Bork, and C. P. Ponting.** 1998. SMART, a simple modular architecture research tool: identification of signalling domains. Proc. Natl. Acad. Sci. USA **95:**5857–5864.
41. **Schultz, M. E., and F. R. Trainor.** 1968. Production of male gametes and auxospores in the centric diatoms *Cyclotella meneghiniana* and *C. cryptica*. J. Phycol **4:**85–88.
42. **Singson, A., K. B. Mercer, and S. W. L'Hernault.** 1998. The *C. elegans spe-9* gene encodes a sperm transmembrane protein that contains EGF-like repeats and is required for fertilization. Cell **93:**71–79.
43. **Vaulot, D., and S. W. Chisholm.** 1987. Flow cytometric analysis of spermatogenesis in the diatom *Thalassiosira weissflogii* (Bacillariophyceae). J. Phycol. **23:**132–137.
44. **von Heijne, G.** 1985. Signal sequences: the limits of variation. J. Mol. Biol. **184:**99–105.
45. **Stosch, von H. A.** 1950. Oogamy in a centric diatom. Nature **165:**531–532.