**Identification of a frustule-associated protein of the marine pennate
diatom *Fistulifera* sp. strain JPCC DA0580**

Michiko Nemoto¹, Yoshiaki Maeda¹, Masaki Muto^{1,2}, Masayoshi Tanaka^{1,2}, Tomoko Yoshino¹

Shigeki Mayama³ and Tsuyoshi Tanaka^{1,2*}

¹Division of Biotechnology and Life Science, Institute of Engineering, Tokyo University of Agriculture and
Technology, 2-24-16, Naka-cho, Koganei, Tokyo 184-8588, Japan

²JST, CREST, Sanbancho 5, Chiyoda-ku, Tokyo 102-0075, Japan

³Department of Biology, Tokyo Gakugei University, 4-1-1, Nukuikita-machi, Koganei, Tokyo 184-8501,
Japan

*Corresponding author

E-mail address: tsuyo@cc.tuat.ac.jp

Tel: +81-42-388-7401

Fax: +81-42-385-7713

Abstract

Among the proteins localized on the cell wall (frustule) of diatoms (frustule-associated proteins), several proteins tightly associated with the cell wall have been implicated in frustule formation. These proteins include diatom-specific unique serine- and lysine-rich sequences represented by silaffins. Taking advantage of available genome information, we used a recently described bioinformatics approach to screen silaffin-like proteins rich in serine and lysine from the genome of the marine pennate diatom *Fistulifera* sp. strain JPCC DA0580 and identified 7 proteins. All of the proteins shared a sequence motif called the XGXXG domain, which was also confirmed in a silaffin-like protein identified in other diatoms. *In vivo* localization analysis revealed that one of the identified proteins, G7408, occurs throughout the frustule with a slightly uneven distribution. This novel frustule-associated protein could be a useful tool to elucidate the mechanism of biosilica formation in diatoms and to functionalize this strain for future biotechnological applications.

Keywords:

Marine diatom, *Fistulifera* sp. strain JPCC DA0580, Frustule-associated protein, Whole genome, Biomineral

1. Introduction

Many living organisms tend to find efficient and environmentally benign ways to generate composite biomaterials consisting of organic and inorganic substances. These biomaterials have multiple functions and extraordinary physical properties at room temperature owing to their well-controlled microstructures [1].

1 Diatoms, which are eukaryotic unicellular microalgae, are an intriguing example of biomineral
2 forming organisms that produce a silica based cell wall called the frustule. Frustules are robust enough for cell
3 protection [2] and have species-specific, hierarchically organized three-dimensional structures with micro- to
4 nanoscale dimensions. These mechanical and structural features of diatom frustules have been applied in
5 materials research of photonic crystals [3], metal- or bio-catalyst platforms [4-6], and gas sensors [7]. To
6 understand frustule formation, organic component (i.e. proteins [8-17], polysaccharides [18] and long chain
7 polyamines [19-22]) of frustule have been isolated from the diatoms *Cylindrotheca fusiformis* and
8 *Thalassiosira pseudonana*. Particularly, a variety of frustule-associated proteins (e.g., frustulin [9, 10],
9 pleuralins [12, 14], silaffins [11, 13, 15], silacidin [17] and cingulin [16], for which characteristic features
10 have been previously reviewed [23]), have been identified and analyzed their roles in silica mineralization.
11 Although the higher order structure was determined by silica deposition vesicle (SDV) and insoluble organic
12 layer as indicated by previous studies [16, 24, 25], particular frustule-associated proteins have been shown to
13 involve in silica mineralization by moderating the activity of long chain polyamines and it was thought to
14 involve in nano-scale structural determination [11, 13, 15]. Previous studies have revealed that except for
15 frustulins that have been identified from the genome of *T. pseudonana* [26] and *P. tricornutum* [27],
16 frustule-associated proteins from different diatoms tend to exhibit little sequence homology [23]. In response
17 to this finding, Kröger and colleagues developed a bioinformatics approach that considered amino acid
18 composition instead of sequence homology and successfully identified novel frustule-associated proteins [16].
19 This strategy seems to be promising for discoveries of novel frustule-associated proteins from other diatoms.

1 However, these discoveries have been hampered by the lack of genome information for various diatoms,
2 which is necessary for bioinformatics screening. Thus far, only a few diatom species have been subjected to
3 whole-genome analysis [26, 27].

4 In addition to frustule structure, biofuel production is an emerging topic in diatom research. We have
5 previously discovered the marine pennate diatom *Fistulifera* sp. strain JPCC DA0580, which accumulates oil
6 in its cells at levels up to 60% w/w [28]. Because this strain has excellent potential for biofuel production,
7 whole-genome and transcriptome analysis have been launched [29]. Based on such genome information, a
8 transformation technique for this strain has also been established [30]. Additionally, proteome analysis with
9 regard to the oil accumulation mechanism has also been attempted [31]. Establishment of these omics
10 platforms encouraged us to return to the frustule study. Structural analysis of the frustules of *Fistulifera* genera
11 by electron microscopy revealed that it has the nanoscale structure that differs greatly from that of other
12 diatoms studied with regard to frustule-associated proteins (submitted for publication) [32]. Therefore, novel
13 proteins associated with biosilica are expected to be found in this diatom.

14 In this study, we attempted to identify frustule-associated proteins from the whole-genome data of
15 *Fistulifera* sp. strain JPCC DA0580 with reference to amino acid composition-based method described
16 previously [16]. The putatively determined frustule-associated proteins were compared with those of other
17 diatom species. Furthermore, the localization of one protein was investigated through the expression of a
18 green fluorescent tagged protein in *Fistulifera* sp. strain JPCC DA0580.

2. Materials and Methods

2.1. Bioinformatics analysis

The known frustule-associated proteins in *C. fusiformis* and *T. pseudonana* [8-12, 15, 33] were used as query sequences for a Basic Local Alignment Search Tool protein (BlastP) search for the proteome predicted from draft genome assemblies of *Fistulifera* sp. strain JPCC DA0580. Gene regions in the genome have been predicted using AUGUSTUS version 2.5.5. For better prediction, it is suggested that the sequence information of approximately hundred genes is prepared for training species-specific parameters on AUGUSTUS according to the manufactures' manual. We constructed a cDNA library from *Fistulifera* sp. cells, and sequenced. Among the sequenced information, 99 of distinct full-length genes were found. These sequences were used to train the AUGUSTUS program to predict *Fistulifera* sp. gene regions, so that the prediction accuracy could be significantly improved. With regard to frustulins, sequence identity and similarity were calculated with ClustalW [34].

The amino acid compositions of 20,455 proteins deduced from the draft genome assemblies of strain JPCC DA0580 were analyzed, and serine- and lysine-rich proteins were identified as described previously with some modification [16]. Briefly, the proteins which are (1) rich in serine ($\geq 16\%$) and lysine ($\geq 10\%$), and (2) predicted to have an endoplasmic reticulum (ER)-targeting signal peptide were selected. The amino acid compositions of the protein sets deduced from these genome data were analyzed in the same manner. For signal peptide prediction, SignalP [35], Predotar [36], and HECTAR [37] software were used. Pairwise sequence alignment of amino acid sequences was performed with ClustalW [34]. The phylogenetic tree was

constructed using the neighbor-joining method. Bootstrap trials were replicated 1000 times. Gaps were treated as missing data.

2.2. Plasmid construction and transformation

A previously developed protein expression vector for strain JPCC DA0580 [30] was used for the plasmid construction. A sequence consisting of a 500-bp *fcpB* (*F-fcpB*) promoter (DDBJ/EMBL/GenBank database accession number AB642240) of strain JPCC DA0580 and a 227-bp transcription terminator sequence of the fucoxanthin chlorophyll a/c-binding protein A (*fcpA*) gene in *P. tricornutum* [38] connected by the *EcoRI* and *PstI* recognition sequences was ligated to *SalI*- and *SphI*-digested pSP-NPT/H4 [30]. The *g7408-gfp* fusion gene and the *g5605-gfp* fusion gene were synthesized with *EcoRI* and *PstI* recognition sequences at their 3' and 5' ends and cloned into the plasmid described above. The constructed plasmids were transferred into wild-type *Fistulifera* sp. strain JPCC DA0580 via microparticle bombardment as described previously [30].

2.3. Fluorescent microscopic analysis

The transformants exhibiting GFP fluorescence were observed using a fluorescence microscope (BX51; Olympus, Tokyo, Japan) equipped with a cooled digital camera (DP-70; Olympus) with a NIBA filter set for GFP and a WIG filter set for chlorophyll fluorescence. For observation of detergent-washed frustules [16], the cells were extracted twice in 2% SDS, 0.1 M ethylenediaminetetraacetic acid (EDTA) (pH8) at 55°C

for 60 min. As prepared material was washed with acetone and water until the acetone supernatant was colorless. After the last acetone wash, the frustules were washed three times with water, and subjected to the fluorescence imaging.

3. Results

3.1. Identification of putative frustule-associated proteins

The protein coding genes used in this study were predicted by the AUGUSTUS program [39]. During this study period, the significant improvement of gene prediction accuracy was achieved by the training of AUGUSTUS program using full-length cDNA sequence sets of *Fistulifera* sp. strain JPCC DA0580 (manuscript in preparation). Homology searches for the protein coding gene sets of *Fistulifera* sp. strain JPCC DA0580 with conventional frustule-associated proteins retrieved 7 significant hits for frustulins. From previous reports, frustulin isolated from different diatom species have been shown to share homology with each other [9, 10, 26]. The identification of frustulin-like proteins in this study indicated that those sequences were also conserved in *Fistulifera* sp. These frustulin-like proteins—G9855, G20406, G20410, G16934, G3248, G20405, and G16935—showing 34% (47%), 41% (55%), 32% (40%), 38% (53%), 30% (43%), 39% (49%), and 43% (52%) sequence identity (and similarity), respectively, to the α 1 frustulin of *C. fusiformis*. All of the frustulin-like proteins of *Fistulifera* sp. strain JPCC DA0580 conserved characteristic features of conventional frustulins, which have acidic and cysteine-rich regions and proline-rich stretches (see also supporting information)[9, 10]. By contrast, no significant hits occurred on other known

frustule-associated proteins (e.g., silaffin, silacidin and pleuralin), confirming that diatom species poorly conserve the amino acid sequences of frustule-associated proteins with the exception of frustulins. For further identification, we attempted to apply amino acid composition-based method [16] with some modification in which the silaffin-like proteins rich in serine ($\geq 16\%$) and lysine ($\geq 10\%$) throughout the full protein sequences were selected without considering the primary sequence. Among the 20,455 proteins encoded in the genome of strain JPCC DA0580, 15 serine- and lysine-rich proteins were identified. The proteins were further narrowed by searching endoplasmic reticulum signal sequences contained in all known frustule-associated proteins, and 7 proteins were eventually determined as silaffin-like proteins in *Fistulifera* sp. strain JPCC DA0580 (Fig. 1 and Fig. 2). The 7 proteins showed no homologies to known sequences. The proteins included 3 paralogous protein sets (G19907 and G7406, G5605 and G12405, G5639 and G12444) showing 77–94% identity between pairs. We also found that G7408 had the paralog G19905. However, G19905 does not have a signal peptide and thus was removed from the list of silaffin-like proteins in this study.

3.2. Sequence characterization of silaffin-like proteins

The amino acid compositions of the silaffin-like proteins of strain JPCC DA0580 are summarized in Figure 1, with a comparison to those of proteins of other diatom species (Fig. S2). It is reasonable that the identified proteins contain high numbers of serine and lysine residues because of the screening criteria used. In addition, high numbers of glycine residues were also included in the proteins. This richness in serine, lysine, and glycine resembles silaffin 1 of *C. fusiformis*, protein 41599 of *P. tricornutum* (Ptr41599), and cingulins of

1 *T. pseudonana* (Fig. S2).

2 The primary structures of each protein are depicted in Figure 2. The proteins identified from strain
3 JPCC DA0580 contain highly repetitive regions called XGXG domains because they have XGXG motifs (X =
4 S or K—e.g., SGKG, SGSG and KGKG; Fig. 2 and see also supporting information). Intensive repetition of
5 this motif was confirmed in the JPCC DA0580 proteins identified in this study as well as in protein 41599
6 from *P. tricornutum* (Ptr41599) [40] and cingulins [16]. Alternating with the XGXG domains, acidic domains
7 emerged in the proteins G7408, G19907, G7406, G5639, and G12444. The proteins G5605 and G12405 have
8 a unique domain between their signal peptides and XGXG domains that starts from the basic region, extends
9 through the zwitter region, and ends in the acidic region and is thus designated the basic-acidic gradient
10 domain. All proteins contain “RXL motif (X=H, R, L, G)”, a putative proteolytic cleavage site, that have been
11 found from the frustule associated proteins in *T. pseudonana* and *C. fusiformis* (see also supporting
12 information) [9-11, 15-17].

13 Next, to confirm the phylogenetic relationship between newly identified JPCC DA0580 proteins and
14 conventional frustule-associated proteins, we carried out clustering analysis based on the amino acid
15 sequences and created a phylogenetic tree (Fig. 3). As a result, the silaffin-like proteins of JPCC DA0580
16 were grouped into 2 new families separated from the best-known frustule-associated proteins. The proteins
17 G5639 and G12444 made a distinct clade, whereas other JPCC DA0580 proteins made another clade with
18 Ptr41599.

19

3.3. Localization analysis

To estimate the *in vivo* localization of the silaffin-like proteins, we tried to construct expression vectors for the GFP-fused proteins G7408 and G5605 (G7408-GFP and G5605-GFP, respectively). The plasmid for G5605-GFP was successfully created, whereas that for G7408-GFP was obtained in only a truncated form lacking the residues G199–G245 in the second XGXG domain despite several trials (the lacking amino acid sequence is described in supporting information). The possible reason of the failure of full-length g7408 cloning could be the toxicity of full-length G7408 expression in *E. coli*. Therefore, we used the plasmid expressing the truncated G7048 with GFP fusion in this study.

Although antibiotic-resistant clones were obtained with both of the constructed expression vectors, the transformants emitting significant GFP fluorescence were confirmed only with G7408-GFP (Fig. 4 and Fig. S3). Even in the same clone of transformants with *g7408-gfp*, some cells did not emit fluorescence (Fig. 4a and Fig. S4). The reason why nonfluorescent transformants emerged remains unclear, although the same phenomena have occurred in previous GFP fusion experiments [31]. Fluorescent microscopic analysis of the transformants with *g7408-gfp* revealed that GFP fluorescence was distributed throughout the frustule surface (Fig. 4), whereas the valve regions (top faces of the frustules) displayed slightly stronger fluorescence than that of the girdle bands (Fig. 4a). In particular, intensive GFP fluorescence was observed in a linear arrangement around the center of the valve face (Fig. 4b). The liner structure is likely to be a raphe sternum (a slit through the valve face, Fig. S1), which is a silica-rich structure [41]. Further fluorescent microscopic analysis of detergent-treated cell wall of transformants revealed that G7408-GFP tightly associated with silica

and could not be extracted by SDS (Fig. S4).

4. Discussion

In this study, genome-wide screening of the genes encoding the frustule-associated proteins in *Fistulifera* sp. strain JPCC DA0580 was performed. The morphometric analysis for the frustule of strain JPCC DA0580 revealed the lightly silicified frustules, a distinct median costa (raphe sternum), and a wart-like central pore (fistula), as with other species in the *Fistulifera* genus [32] (our report on the morphometric analysis of *Fistulifera* sp. strain JPCC DA0580 was submitted for publication). In addition to overall morphology, nano-scale structure of *Fistulifera* sp. is also dissimilar to that of *C. fusiformis*, *P. tricornutum*, and *T. pseudonana*. Therefore, we assumed that the proteins responsible for controlling the nanostructure of the JPCC DA0580 frustules would be different from those in previously studied diatoms. Furthermore, comparison of the frustule-associated proteins identified from phylogenetically distant diatom species might enable us to determine the component essential for silica mineralization.

Taking advantage of available genome information, we identified 7 putative silaffin-like proteins of JPCC DA0580 based on amino acid composition and endoplasmic reticulum-directing signal peptides with reference to the previous report [16]. In contrast to the previous report [16], that search for the protein with ≥ 100 amino acid domain containing $\geq 18\%$ serine and $\geq 10\%$ lysine residues, this study searched for the proteins containing $\geq 16\%$ serine and $\geq 10\%$ lysine residues throughout the full protein sequences. CingulinY proteins of *T. pseudonana* that fulfill the search criteria of the previous report [16] do not fulfill the search

1 criteria used in this study. Therefore, a subset of silaffin-like proteins could be overlooked in this study. A
2 lysine rich “KXXX motif” has been previously reported in the silaffins and cingulins from *T. pseudonana* [15,
3 16, 42] and silaffin from *C. fusiformis* [11]. The “KXXX motif” was thought to mediate silica precipitation
4 activity of those proteins [16]. The silaffin-like proteins identified in this study do not have “KXXX motif” in
5 their sequences. Alternatively, those proteins possess “XGXG motif (X=S or K)”. Owing to these “XGXG
6 motif”, overall amino acid compositions of silaffin-like proteins in *Fistulifera* sp. are similar to that of silaffins
7 and cingulins in *T. pseudonana* and *C. fusiformis*. Although the specific function of each motif is unknown,
8 silica mineralization activities of those proteins are likely to be independent on the specific motif but
9 dependent on overall amino acid composition. Recently, Poulsen et al. reported that a lysine-rich 12-14-amino
10 acid peptide motif conserved in the silaffins from *T. pseudonana* and *C. fusiformis* was important to promote
11 silica targeting [43]. Based on that report, the proteins identified in this study that also contain lysine-rich
12 peptide motifs could also have silica targeting activity. Despite this common tendency in amino acid
13 composition, the sequences of the frustule-associated proteins were hypothesized to be species specific [16].
14 This hypothesis was further supported by our phylogenetic analysis in which the proteins of JPCC DA0580
15 were categorized into new families separated from almost all conventional frustule proteins of other diatoms.
16 Furthermore, previous studies demonstrated that the serine residues of silaffin in *C. fusiformis* were highly
17 post-translationally modified with phosphate groups and its importance for silica precipitation activity was
18 reported [13]. In addition, the lysine residue of silaffins in *C. fusiformis* was shown to be post-translationally
19 modified with methyl groups and polyamines [11, 42]. Those post-translational modifications of serine and

1 lysine residues were also identified from tpsil3p in *T. pseudonana* [15]. Based on those reports, the serine and
2 lysine residues which are highly modified have been considered to be absolutely imperative for the function of
3 silaffins. By utilizing those characteristic features of silaffins, Scheffel et al successfully screened the
4 cingulins involving in silica morphogenesis [16]. In this study, by reference to the bioinformatics approach
5 developed by Scheffel et al, serine and lysine rich proteins were screened from the genome of *Fistulifera* sp.
6 Since the structural analyses by mass spectrometry of those proteins have not been conducted, it is unclear if
7 those proteins undergo post-translational modification. Besides the XGXG domains, several characteristic
8 domains were found in the identified proteins, including acidic and basic-acidic gradient domains. The acidic
9 domains were also found in serine-, lysine-rich proteins from other diatoms such as cingulin W2 of *T.*
10 *pseudonana* and silaffin 1 precursor of *C. fusiformis* [11, 16]. In the case of silaffin 1 precursor of *C.*
11 *fusiformis*, the acidic domain is removed during maturation. The basic-acidic gradient domains found in the
12 proteins G5605 and G12405 are not present in the silaffins from other diatoms, and therefore unique to
13 *Fistulifera* sp. At this stage, functional assignments for these regions remain unclear.

14 *In vivo* localization analysis for the identified protein was carried out using a GFP fusion method.

15 Although the amino acid sequence of G7408 was truncated, the localization of the truncated G7408-GFP
16 fusion protein was confirmed to be on the frustules. This result indicates that protein G7408 can at least
17 interact with the frustule. In order to strictly confirm whether G7408 distributed in other subcellular
18 localization, confocal imaging as well as immunoelectron microscopic analysis would be required. Some
19 frustule-associated proteins have been demonstrated to distribute unevenly on the frustules; for instance,

tps3p of *T. pseudonana* localizes predominantly in the valve region [6], and cingulins localize in the girdle bands [16]. The G7408-GFP fusion protein did not show a strong tendency for uneven localization but was slightly accumulated in the valve region rather than in the girdle bands (Fig. 4b). TEM analysis of cross-section of *Fistulifera* cells revealed that the valve region contain more silica than the girdle bands (Fig. S1). Furthermore, intense GFP fluorescence was also observed around the raphe sternum where a thick silica structure exists (Fig. 4a), suggesting that G7408 has a tendency to accumulate in silica-rich regions and might be involved in silica deposition. Previous studies revealed that silaffins and cingulins in *C. fusiformis* and *T. pseudonana* were tightly associated with silica structure even after the SDS treatment and thought to play a role in cell wall formation. The other frustule-associated proteins, such as frustulins, that could be extracted by EDTA are believed to be involved in cell wall protection, but may not be involved in cell wall formation [44]. In this study, tight association of G7408-GFP with silica was also confirmed with fluorescent microscopic analysis of SDS treated cell wall of *g7408-gfp* transformants (Fig. S4). Since frustule associated proteins have previously been used to immobilize functional molecules on the silica cell wall [6], frustule-associated protein G7408 may be a useful new tool for further functionalization of this strain for future biotechnological applications.

Acknowledgement

Financial support for this research was provided by JST, CREST.

References

- [1] R. de la Rica, H. Matsui, Applications of peptide and protein-based materials in bionanotechnology, *Chem Soc Rev*, 39 (2010) 3499-3509.
- [2] C.E. Hamm, R. Merkel, O. Springer, P. Jurkojc, C. Maier, K. Prechtel, V. Smetacek, Architecture and material properties of diatom shells provide effective mechanical protection, *Nature*, 421 (2003) 841-843.
- [3] T. Fuhrmann, S. Landwehr, M. El rharbi-kucki, M. Sumper, Diatoms as living photonic crystals, *Appl Phys B*, 78 (2004) 257-260.
- [4] Y. Yu, J. Addai-Mensah, D. Losic, Synthesis of self-supporting gold microstructures with three-dimensional morphologies by direct replication of diatom templates, *Langmuir*, 26 (2010) 14068-14072.
- [5] K. Umemura, Y. Gao, T. Nishikawa, Preparation of photocatalyst using diatom frustules by liquid phase deposition method, *J Nanosci Nanotechnol*, 10 (2010) 4883-4888.
- [6] N. Poulsen, C. Berne, J. Spain, N. Kröger, Silica immobilization of an enzyme through genetic engineering of the diatom *Thalassiosira pseudonana*, *Angew Chem Int Ed Engl*, 46 (2007) 1843-1846.
- [7] Z. Bao, M.R. Weatherspoon, S. Shian, Y. Cai, P.D. Graham, S.M. Allan, G. Ahmad, M.B. Dickerson, B.C. Church, Z. Kang, H.W. Abernathy, 3rd, C.J. Summers, M. Liu, K.H. Sandhage, Chemical reduction of three-dimensional silica micro-assemblies into microporous silicon replicas, *Nature*, 446 (2007) 172-175.
- [8] A.K. Davis, M. Hildebrand, B. Palenik, A stress-induced protein associated with the girdle band region of the diatom *Thalassiosira pseudonana* (Bacillariophyta), *Journal of phycology*, 41 (2005) 577-589.
- [9] N. Kröger, C. Bergsdorf, M. Sumper, A new calcium binding glycoprotein family constitutes a major

1 diatom cell wall component, EMBO J, 13 (1994) 4676-4683.

2 [10] N. Kröger, C. Bergsdorf, M. Sumper, Frustulins: domain conservation in a protein family associated with

3 diatom cell walls, Eur J Biochem, 239 (1996) 259-264.

4 [11] N. Kröger, R. Deutzmann, M. Sumper, Polycationic peptides from diatom biosilica that direct silica

5 nanosphere formation, Science, 286 (1999) 1129-1132.

6 [12] N. Kröger, G. Lehmann, R. Rachel, M. Sumper, Characterization of a 200-kDa diatom protein that is

7 specifically associated with a silica-based substructure of the cell wall, Eur J Biochem, 250 (1997) 99-105.

8 [13] N. Kröger, S. Lorenz, E. Brunner, M. Sumper, Self-assembly of highly phosphorylated silaffins and their

9 function in biosilica morphogenesis, Science, 298 (2002) 584-586.

10 [14] N. Kröger, R. Wetherbee, Pleuralins are involved in theca differentiation in the diatom *Cylindrotheca*

11 *fusiformis*, Protist, 151 (2000) 263-273.

12 [15] N. Poulsen, N. Kröger, Silica morphogenesis by alternative processing of silaffins in the diatom

13 *Thalassiosira pseudonana*, J Biol Chem, 279 (2004) 42993-42999.

14 [16] A. Scheffel, N. Poulsen, S. Shian, N. Kröger, Nanopatterned protein microrings from a diatom that direct

15 silica morphogenesis, Proc Natl Acad Sci U S A, 108 (2011) 3175-3180.

16 [17] S. Wenzl, R. Hett, P. Richthammer, M. Sumper, Silacidins: highly acidic phosphopeptides from diatom

17 shells assist in silica precipitation in vitro, Angew Chem Int Ed Engl, 47 (2008) 1729-1732.

18 [18] E. Brunner, P. Richthammer, H. Ehrlich, S. Paasch, P. Simon, S. Ueberlein, K.H. van Pee, Chitin-based

19 organic networks: an integral part of cell wall biosilica in the diatom *Thalassiosira pseudonana*, Angew Chem

1 Int Ed Engl, 48 (2009) 9724-9727.

2 [19] M. Sumper, A phase separation model for the nanopatterning of diatom biosilica, *Science*, 295 (2002)

3 2430-2433.

4 [20] M. Sumper, E. Brunner, G. Lehmann, Biomineralization in diatoms: characterization of novel polyamines

5 associated with silica, *FEBS Lett*, 579 (2005) 3765-3769.

6 [21] M. Sumper, G. Lehmann, Silica pattern formation in diatoms: species-specific polyamine biosynthesis,

7 *Chembiochem*, 7 (2006) 1419-1427.

8 [22] N. Kröger, R. Deutzmann, C. Bergsdorf, M. Sumper, Species-specific polyamines from diatoms control

9 silica morphology, *Proc Natl Acad Sci U S A*, 97 (2000) 14133-14138.

10 [23] N. Kröger, N. Poulsen, Diatoms-from cell wall biogenesis to nanotechnology, *Annu Rev Genet*, 42

11 (2008) 83-107.

12 [24] B. Tesson, M. Hildebrand, Extensive and intimate association of the cytoskeleton with forming silica in

13 diatoms: control over patterning on the meso- and micro-scale, *PLoS One*, 5 (2010) e14300.

14 [25] B. Tesson, M. Hildebrand, Characterization and localization of insoluble organic matrices associated with

15 diatom cell walls: insight into their roles during cell wall formation, *PLoS One*, 8 (2013) e61675.

16 [26] E.V. Armbrust, J.A. Berges, C. Bowler, B.R. Green, D. Martinez, N.H. Putnam, S. Zhou, A.E. Allen, K.E.

17 Apt, M. Bechner, M.A. Brzezinski, B.K. Chaal, A. Chiovitti, A.K. Davis, M.S. Demarest, J.C. Detter, T.

18 Glavina, D. Goodstein, M.Z. Hadi, U. Hellsten, M. Hildebrand, B.D. Jenkins, J. Jurka, V.V. Kapitonov, N.

19 Kröger, W.W. Lau, T.W. Lane, F.W. Larimer, J.C. Lippmeier, S. Lucas, M. Medina, A. Montsant, M. Obornik,

- 1 M.S. Parker, B. Palenik, G.J. Pazour, P.M. Richardson, T.A. Ryneerson, M.A. Saito, D.C. Schwartz, K.
- 2 Thamatrakoln, K. Valentin, A. Vardi, F.P. Wilkerson, D.S. Rokhsar, The genome of the diatom *Thalassiosira*
- 3 *pseudonana*: ecology, evolution, and metabolism, Science, 306 (2004) 79-86.
- 4 [27] C. Bowler, A.E. Allen, J.H. Badger, J. Grimwood, K. Jabbari, A. Kuo, U. Maheswari, C. Martens, F.
- 5 Maumus, R.P. Otilar, E. Rayko, A. Salamov, K. Vandepoele, B. Beszteri, A. Gruber, M. Heijde, M. Katinka, T.
- 6 Mock, K. Valentin, F. Verret, J.A. Berges, C. Brownlee, J.P. Cadoret, A. Chiovitti, C.J. Choi, S. Coesel, A. De
- 7 Martino, J.C. Detter, C. Durkin, A. Falciatore, J. Fournet, M. Haruta, M.J. Huysman, B.D. Jenkins, K.
- 8 Jiroutova, R.E. Jorgensen, Y. Joubert, A. Kaplan, N. Kröger, P.G. Kroth, J. La Roche, E. Lindquist, M.
- 9 Lommer, V. Martin-Jezequel, P.J. Lopez, S. Lucas, M. Mangogna, K. McGinnis, L.K. Medlin, A. Montsant,
- 10 M.P. Oudot-Le Secq, C. Napoli, M. Obornik, M.S. Parker, J.L. Petit, B.M. Porcel, N. Poulsen, M. Robison, L.
- 11 Rychlewski, T.A. Ryneerson, J. Schmutz, H. Shapiro, M. Siaut, M. Stanley, M.R. Sussman, A.R. Taylor, A.
- 12 Vardi, P. von Dassow, W. Vyverman, A. Willis, L.S. Wyrwicz, D.S. Rokhsar, J. Weissenbach, E.V. Armbrust,
- 13 B.R. Green, Y. Van de Peer, I.V. Grigoriev, The *Phaeodactylum* genome reveals the evolutionary history of
- 14 diatom genomes, Nature, 456 (2008) 239-244.
- 15 [28] M. Matsumoto, H. Sugiyama, Y. Maeda, R. Sato, T. Tanaka, T. Matsunaga, Marine diatom, *Navicula* sp.
- 16 strain JPCC DA0580 and marine green alga, *Chlorella* sp. strain NKG400014 as potential sources for
- 17 biodiesel production, Appl Biochem Biotechnol, 161 (2010) 483-490.
- 18 [29] T. Tanaka, Fukuda, Y., Yoshino, T., Maeda, Y., Muto, M., Matsumoto, M., Mayama, S., Matsunaga, T. ,
- 19 High-throughput pyrosequencing of the chloroplast genome of a highly neutral-lipid-producing marine

1 pennate diatom, *Fistulifera* sp. strain JPCC DA0580, *Photosynth Res*, 109 (2011) 7.

2 [30] M. Muto, Y. Fukuda, M. Nemoto, T. Yoshino, T. Matsunaga, T. Tanaka, Establishment of a genetic
3 transformation system for the marine pennate diatom *Fistulifera* sp. strain JPCC DA0580--a high triglyceride
4 producer, *Mar Biotechnol* (NY), 15 (2013) 48-55.

5 [31] D. Nojima, T. Yoshino, Y. Maeda, M. Tanaka, M. Nemoto, T. Tanaka, Proteomics Analysis of Oil
6 Body-Associated Proteins in the Oleaginous Diatom, *J Proteome Res*, (2013).

7 [32] F.F. Pniewski, T. Friedl, A. Latała, Identification of diatom isolates from the Gulf of Gdańsk: testing of
8 species identifications using morphology, 18S rDNA sequencing and DNA barcodes of strains from the
9 Culture Collection of Baltic Algae (CCBA), *Oceanological and Hydrobiological Studies*, 39 (2010) 3-20.

10 [33] A. Montsant, K. Jabbari, U. Maheswari, C. Bowler, Comparative genomics of the pennate diatom
11 *Phaeodactylum tricornutum*, *Plant Physiol*, 137 (2005) 500-513.

12 [34] J.D. Thompson, D.G. Higgins, T.J. Gibson, CLUSTAL W: improving the sensitivity of progressive
13 multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix
14 choice, *Nucleic Acids Res*, 22 (1994) 4673-4680.

15 [35] T.N. Petersen, S. Brunak, G. von Heijne, H. Nielsen, SignalP 4.0: discriminating signal peptides from
16 transmembrane regions, *Nat Methods*, 8 (2011) 785-786.

17 [36] I. Small, N. Peeters, F. Legeai, C. Lurin, Predotar: A tool for rapidly screening proteomes for N-terminal
18 targeting sequences, *Proteomics*, 4 (2004) 1581-1590.

19 [37] B. Gschloessl, Y. Guermeur, J.M. Cock, HECTAR: a method to predict subcellular targeting in

1 heterokonts, BMC Bioinformatics, 9 (2008) 393.

2 [38] K.E. Apt, P.G. Kroth-Pancic, A.R. Grossman, Stable nuclear transformation of the diatom *Phaeodactylum*

3 *tricornutum*, Molecular & general genetics : MGG, 252 (1996) 572-579.

4 [39] M. Stanke, O. Keller, I. Gunduz, A. Hayes, S. Waack, B. Morgenstern, AUGUSTUS: ab initio prediction

5 of alternative transcripts, Nucleic Acids Res, 34 (2006) W435-439.

6 [40] G. Sapriel, M. Quinet, M. Heijde, L. Jourden, V. Tanty, G. Luo, S. Le Crom, P.J. Lopez, Genome-wide

7 transcriptome analyses of silicon metabolism in *Phaeodactylum tricornutum* reveal the multilevel regulation

8 of silicic acid transporters, PLoS One, 4 (2009) e7458.

9 [41] E.J. Cox, Ontogeny, homology, and terminology-Wall morphogenesis as an aid to character recognition

10 and character state definition for pennate diatom systematics, J Phycol, 48 (2012) 1-31.

11 [42] M. Sumper, R. Hett, G. Lehmann, S. Wenzl, A code for lysine modifications of a silica biomineralizing

12 silaffin protein, Angew Chem Int Ed Engl, 46 (2007) 8405-8408.

13 [43] N. Poulsen, A. Scheffel, V.C. Sheppard, P.M. Chesley, N. Kröger, Pentyllysine clusters mediate silica

14 targeting of silaffins in *Thalassiosira pseudonana*, J Biol Chem, 288 (2013) 20100-20109.

15 [44] W.H. van de Poll, E.G. Vrieling, W.W.C. Gieskes, Location and expression of frustulins in the pennate

16 diatoms *Cylindrotheca fusiformis*, *Navicula pelliculosa*, and *Navicula salinarum* (Bacillariophyceae), Journal

17 of phycology, 35 (1999) 1044–1053.

18

19

1 **Figure legends**

2 **Figure 1** Amino acid composition of the silaffin-like proteins identified in the *Fistulifera* sp. JPCC DA0580
3 genome.

4

5 **Figure 2** Schematic primary structures of the silaffin-like proteins screened from the genome of *Fistulifera* sp.
6 JPCC DA0580. Each protein has a predicted signal peptide for cotranslational import into the endoplasmic
7 reticulum (ER). The acidic domain, basic domain, and XGXG domain are aspartic acid- and glutamic
8 acid-rich domains, lysine-, arginine-, and histidine-rich domains, and serine- and lysine-rich domains
9 including XGXG (X = serine or lysine) motifs, respectively. The proteins G5606 and G12405 have a
10 basic-acidic (BA) gradient domain that starts from the basic region, extends through the zwitter region, and
11 ends in the acidic region.

12

13 **Figure 3** Phylogenetic tree of the putative frustule-associated proteins from *Fistulifera* sp. JPCC DA0580
14 [Fso], *Phaeodactylum tricornutum* [Ptr], *Cylindrotheca fusiformis* [Cfu], and *Thalassiosira pseudonana* [Tps].

15

16 **Figure 4.** Fluorescence microscopy of *Fistulifera* sp. strain JPCC DA0580 cells transformed with
17 *g7408-green fluorescent protein* gene. Apical view (a) and valve view (b). (Scale bar = 2 μm.)

18

Figure 1

Figure 2

Figure 3

Figure 4