Published in final edited form as:
Aquat Microb Ecol. 2009 June 3; 55(3): 241–253. doi:10.3354/ame01284.

Deciphering diatom biochemical pathways via whole-cell proteomics

Brook L. Nunn[1,*], **Jocelyn R. Aker**[1], **Scott A. Shaffer**[1], **Shannon Tsai**[1], **Robert F. Strzepek**[2], **Philip W. Boyd**[3], **Theodore Larson Freeman**[1,4], **Mitchell Brittnacher**[4], **Lars Malmström**[1], and **David R. Goodlett**[1]

[1] Medicinal Chemistry Department, University of Washington, Box 335351, Seattle, Washington 98195, USA [2] Chemistry Department, University of Otago, Box 56, Dunedin, New Zealand [3] NIWA Centre for Chemical and Physical Oceanography, Department of Chemistry, University of Otago, Dunedin, New Zealand [4] Department of Genomic Sciences, University of Washington, Box 355065, Seattle, Washington 98195, USA

Abstract

Diatoms play a critical role in the oceans' carbon and silicon cycles; however, a mechanistic understanding of the biochemical processes that contribute to their ecological success remains elusive. Completion of the *Thalassiosira pseudonana* genome provided 'blueprints' for the potential biochemical machinery of diatoms, but offers only a limited insight into their biology under various environmental conditions. Using high-throughput shotgun proteomics, we identified a total of 1928 proteins expressed by *T. pseudonana* cultured under optimal growth conditions, enabling us to analyze this diatom's primary metabolic and biosynthetic pathways. Of the proteins identified, 70% are involved in cellular metabolism, while 11% are involved in the transport of molecules. We identified all of the enzymes involved in the urea cycle, thereby describing the complete pathway to convert ammonia to urea, along with urea transporters, and the urea-degrading enzyme urease. Although metabolic exchange between these pathways remains ambiguous, their constitutive presence suggests complex intracellular nitrogen recycling. In addition, all C_4 related enzymes for carbon fixation have been identified to be in abundance, with high protein sequence coverage. Quantification of mass spectra acquisitions demonstrated that the 20 most abundant proteins included an unexpectedly high expression of clathrin, which is the primary structural protein involved in endocytic transport. This result highlights a previously overlooked mechanism for the inter- and intra-cellular transport of nutrients and macromolecules in diatoms, potentially providing a missing link to organelle communication and metabolite exchange. Our results demonstrate the power of proteomics, and lay the groundwork for future comparative proteomic studies and directed analyses of specifically expressed proteins and biochemical pathways of oceanic diatoms.

Keywords

Nitrogen cycle; Carbon cycle; Fatty acids; Protein; Carbon fixation; Clathrin-coated vesicles

*brookh@u.washington.edu.

INTRODUCTION

Proteins are essential for cellular signal transduction, structural integrity, and catalysis of most biochemical reactions. For these reasons, knowing the proteins that are expressed by an organism is central to understanding its biochemical pathways. Thus far, only a few diatom proteins have been identified from cultures (Davis et al. 2005, Hildebrand 2005, Frigeri et al. 2006), these being mostly specialized silicon-recruiting proteins involved in making the diatoms' intricate silica frustule for which there are industrial applications (Hildebrand 2005). Recent advances in genomic profiling allow cataloguing of all possible proteins that can be synthesized by an organism; for example, the genomic analysis of *Thalassiosira pseudonana* identified several potential iron uptake mechanisms and a putative urea cycle (Armbrust et al. 2004). However, there are inherent limitations in pure genomic studies, since organisms only express those proteins necessary for life under a given set of conditions. Through the analysis of expressed proteins, we can link static genomic information to dynamic cellular expression under selected environmental conditions.

In addition to proteomics providing a snapshot of cellular mechanics, recent development of label-free methods in protein quantification now allows unbiased observations of relative protein abundances in a cell at the time of harvest (Chen et al. 2008, Ryu et al. 2008). Relative quantities of proteins produced within a cell can be a direct measure of the biochemical pathways used by the organism. In previous reports examining diatom biochemistry, transcriptomics and tiling arrays have been used to compare cells under different growth conditions and examine relative gene expression (Mock et al. 2008). However, recent studies focused on human RNA transcription have demonstrated that, although the genome is extensively transcribed, much of that RNA is never translated (Birney et al. 2007). Prior genomic doctrine painted a more simplistic picture: DNA makes RNA makes protein. These new findings on human, yeast, and plant RNA show that DNA generates far more RNA than previously thought (e.g. Bertone et al. 2004, Stolc et al. 2005). Much of the transcribed RNA spans across non-protein-coding regions and is currently under investigation as to its purpose in transcription if it is not to be translated (Willingham et al. 2005). This suggests that pure genomic analysis or transcriptomic tiling arrays may be overestimating actual cellular processes, and highlights the importance of a direct measurement of protein in a cell under the growth condition of interest.

Rather than focusing on subsets of expressed proteins, our Lyse-N-Go shotgun proteomics approach (Foss et al. 2007) provides an instantaneous snapshot of all cellular pathways being used at the time of harvest. In the case of organisms whose biochemical pathways are poorly characterized, such as the diatoms, it is a logical second step after genome sequencing because it can rapidly distinguish actual from theoretical pathways and highlight previously overlooked biochemical mechanisms.

We used shotgun proteomic profiling (McDonald & Yates 2002, Nunn & Timperman 2007) to set the foundation of cellular biochemistry of marine diatoms. These single-celled photosynthetic eukaryotes are ubiquitous in the world's oceans, are key contributors to the global carbon cycle, and are responsible for as much as 40% of the organic carbon produced in the ocean (Nelson et al. 1995). Despite their environmental (Ragueneau et al. 2006) and evolutionary (Falkowski et al. 2004) importance, they remain poorly characterized at the biochemical level. Recent examinations of the *Thalassiosira pseudonana* genome (Armbrust et al. 2004) are drawing attention to many pathways that are potentially important to their success, including those responsible for carbon metabolism (Kroth et al. 2008), iron and silicic acid uptake, and silica deposition (Mock et al. 2008). Our proteomic profile, on the other hand, summarizes which pathways are actually functioning and where cellular energy is consumed.

This first report of the proteome of *T. pseudonana* provides a biochemical framework in which to view its genome and where to direct future investigations.

MATERIALS AND METHODS

Cell cultures

Axenic cultures of *Thalassiosira pseudonana* clone 3H, CCMP1335, were grown in Aquil medium under trace metal-clean conditions. Cultures were grown in triplicate 10 l polycarbonate carboys under continuous light (133.5 μE m^{-2} s^{-1} at the center of the carboy). Growth rates were calculated from least squares regressions of the natural logarithm of *in vivo* fluorescence against time during the exponential growth phase of acclimated cultures. Cells were harvested during the mid-exponential phase using centrifugation. Various cellular preparations were completed to increase the number of proteins identified and the final protein sequence coverage (see Results and discussion).

Cellular preparations

Cells were lysed using a 100 W, 20 KHz sonicator (MSE) with a titanium micro-tip (20 × 10 s, 4°C). Unlysed cells and debris were removed by centrifugation (4700 × *g*, 10 min), resulting in whole-cell lysates. To isolate the insoluble fraction from lysed cells, the whole lysates were centrifuged (17,000 × *g*, 30 min) to pellet membranes and other insoluble components, while the supernatants were aspirated and isolated for analyses. Digests of the outside of cells (or membrane components) were completed to look at cell-surface protein expression. Initial experiments were conducted to achieve organelle separations. As a result of organelle fractions being highly cross-contaminated from membrane rupture, fractionations to improve protein sequence coverage and total identifications were completed in the gas phase on a mass spectrometer (Nunn et al. 2006, Scherl et al. 2008).

Trypsin digestions were performed following Nunn et al. (2006). Briefly, protein pellets were solubilized in 300 μl of 6 M urea, followed by the addition of 20 μl of 1.5 mM Tris buffer (pH 8.8), and brought to a final concentration of 5 mM TCEP (37°C, 1 h). Disulfide bonds were reduced with dithiothreitol (DTT), alkylated with 60 μl of 200 mM iodoacetamide (IAM), and vortexed and stored in the dark for 1 h (25°C). Excess IAM was neutralized with 60 μl of 200 mM DTT for 1 h (25°C). A volume of 150 μl of each sample was aliquoted into 3 tubes, and 800 μl of NH_4HCO_3 was added to dilute the urea prior to the addition of 200 μl of MeOH and sequence-grade trypsin (Promega) at 50:1 substrate:enzyme (w/w). Trypsin digestions were vortexed and incubated at 37°C overnight. Samples were then taken to near dryness in a speedvac and same samples were pooled. To reduce the NH_4HCO_3, 200 μl of Milli-Q H_2O was added to each tube and evaporated; the process was repeated 3 times. Samples were stored at −80°C until analysis using mass spectrometry (MS).

To increase protein sequence coverage, endoproteinase Glu-C from *Staphylococcus aureus* V8 was also used. Endo Glu-C is a serine proteinase that provides users with lessspecific protein cleavage sites, offering an additional means of increasing protein coverage. This enzyme primarily targets cleavage C-terminal to glutamic (E) and aspartic acid (D) residues. Lyophilized endoproteinase Glu-C was resuspended in 25 mM NH_4HCO_3 and proteins were subjected to digestion by Glu-C at a ratio of 20:1 substrate:enzyme (w/w) for 16 h at 25°C (pH 8.0). Just prior to all MS analyses, enzymatic digestions were desalted using a micro-spin C18 column (NestGroup) following the manufacturers guidelines.

Mass spectrometry

Samples were separated and introduced into the mass spectrometer by reverse-phase chromatography using an 11 cm long, 75 μm i.d. fused silica capillary column packed with

C18 particles (Magic C18AQ, 100 A, 5 μ; Michrom, Bioresources) fitted with a 2 cm long, 100 μm i.d. precolumn (Magic C18AQ, 200 A, 5 μ; Michrom). Peptides were eluted using an acidified (formic acid, 0.1% v/v) water-acetonitrile gradient (5 to 35% acetonitrile in 60 min). Mass spectrometry was performed on either an LTQ-FT or LTQ-Orbitrap hybrid mass spectrometer (Thermo Fisher) (Table 1, Supplement 1; www.int-res.com/articles/suppl/a055p241_app.xls). Data-dependent scans were completed by precursor ion selection in the FT-based analyzer (FT or Orbitrap), followed by collision induced dissociation (CID) in the linear ion trap (LTQ). Peptide identifications were optimized by gas-phase fractionation (GPF), which was accomplished by performing repeat analyses of the sample across several narrow, but overlapping mass/charge (*m/z*) ranges (e.g. 500–600), rather than one wide *m/z* range (e.g. 400–2000) (Nunn et al. 2006).

Database search and data interpretation

All mass spectral results for this manuscript were interpreted and searched with an in-house copy of SEQUEST (PVM v.27 20070905) (Eng et al. 2008); SEQUEST is a correlative data-interpretation software that matches observed spectra to theoretical spectra generated from the predicted peptide sequences. The protein database we assembled for searching spectra included the latest release version 3.0 of the nuclear *Thalassiosira pseudonana* predicted protein database (11,390 proteins) from the Joint Genome Institute (JGI), plus 302 proteins from the PubMed Entrez Protein database, which consists of 144 proteins predicted from the chloroplast genome, 35 proteins predicted from the mitochondrial genome, 73 proteins from other miscellaneous publications, and 50 common contaminants. In order to determine the probability of false matches to proteins, the forward dataset was combined with the reverse sequences (11,692 DECOY proteins), providing each peptide with an equal probability to match the correct forward protein sequence or the reverse protein sequence (total proteins searched: 23,384). Data searches were completed with no enzyme specificity, while modifications of cysteine residues by 57 Da (resulting from the iodoacetamide modification) and methionine residues by 15.999 Da (oxidation) were allowed. Minimum protein and peptide thresholds were set at 90% on ProteinProphet and PeptideProphet (Keller et al. 2002). The SEQUEST criteria for a doubly charged peptide used a correlation factor (Xcorr) >2.5, a cross-correlation factor ΔCorr >0.1 and an Xcorr minimum of 3.5 for triply charged peptides. Protein identifications by ProteinProphet were accepted if: (1) the above mentioned thresholds were passed, (2) 2 or more peptides were identified (PeptideProphet), and (3) at least one termini was tryptic (on all trypsin digested samples) or induced by endoproteinase Glu-C (on all samples treated with Glu-C).

Using concatenated target-decoy database searches, false-positive rates were calculated according to Elias & Gygi (2007); these were all <1% and correlated well with the false-positive rate estimated using the PeptideProphet tool. In addition, proteins identified to have one peptide (Prophet scores > 90%) were scored using a single hit verification test, similar to the method published by Higdon & Kolker (2007). This included observation of the single peptide from a given protein in >1 of the 130 LC-MS/MS experiments, and/or in >1 of the biological replicates, in addition to visual inspection of at least one spectrum from the peptide to confirm the presence of a strong Y- or B-ion series that could be matched to the assigned sequence.

In addition to correlating peptide spectra with predicted proteins from the genomic model of *Thalassiosira pseudonana*, a second search was performed to correlate peptide mass spectra with open reading frames (ORF) within the 6-frame translation of the nucleotide sequence. SEQUEST treats each potential sequence of nucleotides like a protein, so this can be scored in the same manner as stated above (False Discovery Rate FDR < 1%). The minimum length for an ORF to be accepted was 30 amino acids in. A total of 1,325,597 ORFs were searched

in this database. Peptides identified from ORFs were compared to those from the JGI protein database to identify novel exon regions in the nucleotide sequence (Fig. 1).

Protein quantitation

Protein abundance was assessed at the peptide level for sets of quadruplicate data using a semi-quantitative method: peptide spectral counting (Chen et al. 2008, Ryu et al. 2008). Peptide spectral counts were determined by the number of times a peptide that matched a given protein was selected for CID, including all repeated selections of the same peptide. Thus, a protein's spectral count value was the sum of all identified peptide tandem mass spectra acquired for that protein. A protein's expression level was indicated by the sum of peptide spectral counts in quadruplicate analyses. The final result produces a 'rank order' or an estimate of the amount of a given protein compared to other observed proteins.

Proteome annotation

We assembled a database of 11,644 protein sequences, consisting of 11,390 sequences from the JGI annotation of *Thalassiosira pseudonana* chromosomal DNA (Table S2, Supplement 1), and an additional 254 proteins from the PubMed Entrez website (www.ncbi.nlm.nih.gov/pubmed/), which include chloroplast and mitochondrial proteins that are not in the JGI annotation (Table S3, Supplement 1). These sequences were analyzed by running BLAST (basic local alignment search tool, http://blast.ncbi.nlm.nih.gov/blast/Blast.cgi) against the non-redundant and transport classification databases and databases for organisms of specific interest, such as *Cyanidioschyzon merolae* and *Arabidopsis thaliana*. Bioinformatics predictions of Pfam protein domains, Prosite motifs and Gene Ontology terms from the protein sequences were obtained to suggest possible molecular functions, and subcellular localizations were predicted with the TMHMM, (http://www.cbs.dtu.dk/services/TMHMMI), TMpred (Hofmann & Stoffel 1993;www.ch.embnet.org/software/TMPRED_form.html) and TargetP (Emanuelsson et al. 2000; http://www.cbs.dtu.dk/services/TargetP/) software applications. Molecular weights and isoelectric points were calculated for all protein sequences.

RESULTS AND DISCUSSION

Proteomic profile

A total of 130 HPLC-tandem mass spectrometry (MS/MS) experiments on *Thalassiosira pseudonana* were completed to explore and enhance our understanding of diatom biochemistry, define the expressed proteome, and verify gene models under optimal growth conditions. Genomic models of *T. pseudonana* include over 11,500 protein-coding genes, suggesting that this organism can adapt quickly to surrounding environments. Shotgun mass spectrometry is ideal for getting an instantaneous snapshot of the cellular biochemistry at the time of harvest with minimal sample preparation. To circumvent losses in identifications possibly due to lack of pre-fractionation, our laboratory typically performs peptide separations in the mass spectrometer by performing narrow mass/charge (*m/z*) gas-phase fractionation of ions (Nunn et al. 2006, Scherl et al. 2008). Typically, as the number of analyses increases, confidence in protein identifications also increases (Fig. 2); eventually, the number of identifications reaches a plateau. To ensure that our data is precise and accurate, the present study included 3 biological replicates analyzed in duplicate or quadruplicate on 2 different mass spectrometry platforms (Table A1, Supplement 1) with and without gas-phase fractionations. From our total of 130 tandem mass spectrometry experiments on biological and analytical replicates of whole-cell and cellular fractions of *T. pseudonana*, we verified the translation of 17,590 unique peptides correlated with 1928 expressed proteins (16.5% of the genomic model) (Table A4, Supplement 1). Through quadruplicate analyses of whole-cell tryptic digests, we were able to achieve relative quantities of the most abundant proteins expressed (Chen et al. 2008, Ryu et al.

2008). Numerous proteins in the genome have the same function and homologous amino acid sequences; there are 1762 unique protein identities, many of which (166) are homologues. Because the actual number of these homologues observed cannot be deciphered, they are grouped by the proteomic mass spectra searches and a single protein (from the 1762) is used to represent the group in the remaining discussion and presentation of the results. We verified identities using a false discovery rate of <1%. To date, this is the greatest number of diatom proteins observed.

Analysis of proteins encoded by the genomes of the nucleus (Armbrust et al. 2004), mitochondria and chloroplast (Oudot-Le Secq et al. 2007) suggest a higher expression of total chloroplast and mitochondrial genes (55 and 40% of the predicted proteins observed, respectively). From the 23 nuclear chromosomes with predicted protein sequences (no protein-coding genes predicted on chromosome 21), ~15% were identified using tandem mass spectrometry.

Unpredicted peptides and proteins

In order to search for novel transcripts, we searched for matches using a database of all ORFs in a 6-frame translation of the *Thalassiosira pseudonana* (v.3, www.jgi.doe.gov) nucleotide sequence. Each 6-frame translation ORF was treated like a protein when searching mass spectra. By excluding all peptide mass spectra that matched annotated protein sequences from the gene model (17,590 peptides), we identified over 1433 additional ‘unpredicted’ peptide units (PUs) that were previously not predicted in the genome model (Fig. 1; Table A5, Supplement 1). Of these 1433 peptides, 1039 are fully tryptic peptides, 253 have one tryptic terminus and 141 are non-tryptic from Endo-Glu-C digestions. A recent transcriptomic study observed an equivalent number of unpredicted transcriptional units (TUs), demonstrating a discovery on the same order of magnitude (1132 TUs) (Mock et al. 2008). We also observed 185 PUs that overlap the nucleotide sequence reported by Mock et al. (2008) either in a different read-direction or when initialized from a different codon. Mapping their unpredicted TUs on the nucleotide sequence with our unpredicted PUs revealed 96 common expressed novel sequences between the 2 studies, and we achieved 1152 units novel to only the peptide analyses. Discrepancy between the novel sequences and proteins observed may result from harvesting the diatom under different growing conditions (24 vs. 12 h light cycles). Many of these unpredicted PUs are most likely from alternative intron and exon splice sites relative to the current genomic model (v.3); 14 of these unpredicted peptides are within 30 nucleic acids in the same reading direction, but do not overlap with previously modeled TUs. These newly discovered exon units are being more thoroughly analyzed to determine if there are specific nucleotide sequences that simplify exon predictions in diatoms.

Only 4% of the 1433 peptides were found to be significantly homologous to previously identified proteins in the publicly available database; this likely results from the inherently short nature of peptides. Analysis of the BLAST results identified TUs that correlate with NAD (P)H nitrate reductases (cytochrome b5 reductase), additional photosystem reaction centers, serine/threonine protein phosphatases and several others. Some of these new units may correlate with proteins required for silica deposition, but because so little is known about silicic acid uptake and biochemistry, function may not be assigned. Further analyses of these novel TUs are currently being performed, including folding models to determine functional domains and structures. These findings, in conjunction with Mock et al. (2008), suggest that genomics modeling of diatom nucleotides is more complicated than previously anticipated (Armbrust et al. 2004).

Biochemical pathways of diatoms

To examine the energetic demands of *Thalassiosira pseudonana*, we examined the number of proteins expressed from each category in a gene ontology (GO) analysis (Ashburner et al. 2000). The Gene Ontology database provides a consistent description of each of the proteins based on homologous sequences to model organisms and has different levels of specification, which are referred to as tiers. Analysis of the biological process ontology (3rd tier) revealed that most expressed proteins are involved in cellular metabolism, such as light harvesting and energy transfer (Fig. 3). Specifically, we observed all of the major proteins predicted from the genome (and their subunits) to be involved in enzymatic carbon fixation, light harvesting (photosynthetic antennae) and electron transport (photosynthetic reaction centers). Proteins responsible for inter- and intracellular transport were also highly expressed. For example, 21% of the proteins involved in the directed transfer of macromolecules and ions into, out of, within (between organelles) or between cells were observed. Sixty expressed proteins were related to cellular component organization and biogenesis, which are processes that involve the formation, deformation, or destruction of cellular components. This GO category includes the primary components of clathrin-mediated endocytosis (discussed below).

Analysis of enzymes observed in individual biochemical pathways, as defined by the Kyoto Encyclopedia of Genes and Genomes (KEGG), also demonstrates that the majority of enzymes required for nitrogen, carbon, and fatty acid (FA) synthesis and metabolism were observed (Table 1). Interestingly, examination of the enzymes involved in biosynthesis of unsaturated FAs reveals that under optimal growth conditions, diatoms primarily increase FA chain length, although these remain saturated. This is evidenced by the identification and high sequence coverage of the enzymes involved in FA elongation and the lack of desaturation enzymes identified. This suggests that the majority of fats in phytoplankton grown under replete light and nutrients are saturated FAs, where each carbon in the FA contains 2 hydrogens, making it easy for phytoplankton to stack the molecules for storage. FA profiles from *Thalassiosira pseudonana* (Tonon et al. 2005) harvested at late-exponential growth revealed that palmitic acid, a C_{16} saturated fatty acid, is the most abundant FA in the diatom, confirming our findings. Understanding nutrient conditions that control FA metabolism will help aquaculturists optimize growth media for phytoplankton currently under investigation as biofuels and food supplement sources.

Although many of these pathways are well established in other organisms, such as in *Saccharomyces cerevisiae* and *Arabidopsis thaliana*, several enzymes involved in major pathways have not yet been modeled from the *Thalassiosira pseudonana* genome. Over one-quarter of the identified protein sequences from *T. pseudonana* (504) yield BLAST homologues to ‘hypothetical’ proteins with no known function, underscoring the need for further studies on diatom protein function. This proteomic profile provides a group of proteins on which these initial studies on unknowns should be focused. We are currently folding all hypothetical proteins *in silico* to better determine tertiary structures and predict their function (Malmstrom et al. 2007).

Urea cycle

Several theoretical biochemical pathways were inferred from the *Thalassiosira pseudonana* genome (Armbrust et al. 2004); although gene expression via tiling arrays provides additional evidence suggesting that a biochemical pathway is being utilized, direct observations of translated proteins provides indisputable evidence of a functional biochemical mechanism in the organism. For example, detection of a putative urea cycle in diatoms was an unanticipated finding in the genome, and this proteomic profile is the first evidence of expression of all enzymes involved in this pathway. Most organisms use the urea cycle to manage the excretion of nitrogenous waste that results from the catabolism of proteins and amino acids; however,

diatoms have adapted to be able to utilize urea as a primary nitrogen source (Peers et al. 2000). The urea cycle is also of interest because some of its intermediate metabolites are involved in other biochemical pathways. Ornithine, for example, is a precursor in the manufacture of long-chain polyamines, which are some of the primary components of diatom biosilica (Kröger et al. 2000). Examination of the organelle-targeting protein pre-sequences suggests that the urea cycle is carried out in the mitochondria, as in most organisms with an active urea cycle, while urease is located in the cytosol (Armbrust et al. 2004). Although the constitutive activity of urease has been demonstrated under numerous nutrient conditions (Peers et al. 2000), constitutive activities of urea cycle enzymes under optimal growth conditions have not been examined. Our whole-cell proteomic analysis confirms the presence of all proteins involved in the urea cycle as well as the enzyme urease (Fig. 4). In addition, one urea transporter (ID: 24250) was also identified. Unfortunately, since we were unable to successfully isolate clean organelle fractions, this proteomic analysis is neither able to determine the cellular location of these enzymes, nor whether these 2 pathways are in metabolic communication. The identification of these enzymes under nitrogen-replete conditions may provide diatoms with an evolutionarily advantageous mechanism to control internal nitrogen storage and transfer if the 2 pathways are able to exchange metabolites across organelles. Below, we discuss the discovery of a highly expressed protein involved in the transfer of macromolecules across membranes. This may be a missing link that can help explain the transfer of proteins that lack targeting pre-sequences and/or metabolites across the diatoms' complex system of organelles.

Carbon fixation

The carbon fixation pathway used by diatoms is contentious (Reinfelder et al. 2004, Kroth et al. 2008, McGinn & Morel 2008). The 2 primary pathways, C_3 and C_4, differ in their initial steps of CO_2 incorporation and, as a result, fractionate stable isotopes differently. This fractionation of carbon isotopes allows the study of trophic relationships in marine systems and can be used to reconstruct past CO_2 levels or phytoplankton growth. Traditionally, the C_4 pathway was thought to be a more efficient means of carbon fixation used by evolutionarily advanced plants adapted to dry or otherwise harsh environments; however, the analysis of the diatom genome has revealed the presence of all genes for C_4 carbon fixation. Results from ^{14}C-tracer experiments following malate production in *Thalassiosira pseudonana* led Roberts et al. (2007) to conclude that the C_3 fixation pathway was primarily being used. More recently, McGinn & Morel (2008) demonstrated that direct inhibition of 2 key enzymes in the C_4 pathways of 3 model diatoms, including *T. pseudonana*, 'cripple carbon uptake' and O_2 evolution, and therefore suggest that the C_4 pathway is being followed. They further suggest that the study following malate production (Roberts et al. 2007) possibly missed the metabolite as a result of rapid degradation. Here, we present direct physical evidence from MS-based detection of all enzymes involved in C_4 carbon fixation when the diatom is neither nutrient-, CO_2-, or light-stressed (Fig. 5). Over 22% of the sequence of the phosphoenolpyruvate carboxykinase (PEPCKase) enzyme was observed, and spectral counting on quadruplicate analyses of the diatom ranked this enzyme as the 165th most abundant protein (out of 1762) observed under optimal growth conditions. Two PEPCase enzymes were observed with high protein sequence coverage (63 and 50%), and ranked in the top 100 most abundant proteins, based on spectral counting. These enzymes are responsible for catalyzing the addition of HCO_3^- to phosphoenolpyruvate (PEP) to form oxaloacetate, the 4-carbon molecule after which the C_4 pathway is named. The finale of the C_4 pathway involves concentration of CO_2 by RuBisCO through the conversion of oxaloacetate to malate (performed by a mitochondrial malate dehydrogenase, MDH; 67% sequence coverage), and malate to pyruvate + CO_2 via malic enzyme (ME1; 5% sequence coverage). Although Kroth et al. (2008) proposed a model for carbohydrate metabolism based on genomic analysis of signal peptides from *Phaeodactylum tricornutum*, we believe that, in order to fully decipher the carbon pathway of

this diatom, organellar proteomics must be completed. We made several attempts to isolate clean organelles from diatom cultures, but were unsuccessful.

Clathrin-coated vesicles

An often overlooked advantage of proteomic analyses of whole cells is that it provides an unbiased method to discover and explore important biological pathways. Using a quantitative method that counts the number of times peptide components from each protein are observed in an HPLC-tandem mass spectrometry experiment, we determined the relative abundance of expressed proteins in *Thalassiosira pseudonana* (Fig. 6) (Chen et al. 2008). Surprisingly, this analysis revealed that clathrin was the 6th most abundant protein. Clathrin is the primary structural protein that helps deform membranes to facilitate the invagination of molecular cargo into vesicles. Clathrin-mediated endocytosis (CME) is a process by which virtually all eukaryotic cells internalize nutrients, antigens, growth factors, and pathogens (Takei & Haucke 2001,Conner & Schmid 2003). This protein is not typically highly expressed (Foss et al. 2007,Ryu et al. 2008,Goo et al. 2009, D. R. Goodlett unpubl. data on *Salmonella typhimurium*, *Saccharomyces cerevisae*, humans), making its rank as the 6th most abundant protein in *T. pseudonana* an unanticipated discovery. Only recently has CME been demonstrated to be an important means for internalization in *Arabidopsis thaliana* (Dhonukshe et al. 2007). In *T. pseudonana*, where genomic predictions note that proteins are deficient in N-terminal pre-sequences for sorting and directing molecular traffic through up to 4 membranes (Montsant et al. 2007,Kroth et al. 2008), the discovery of such a prominent mechanism is noteworthy. Vesicle-mediated protein sorting can play an important role in the segregation of intracellular molecules into distinct organelles, as clathrin-coated vesicles (CCV) can be rapidly uncoated and the cargo released, or the uncoated vesicle can be transported to internal organelles such as endosomes and the Golgi complex (Kaksonen et al. 2006,Donohoe et al. 2007,Limbach et al. 2008). Proteins involved in clathrin coats (heavy and light chains), vesicle budding (adaptor proteins), scission (yeast homologue Vps9, Vps16, Vps54) and uncoating enzymes (DnaJ-like auxilin, HSC70) were also well represented (i.e. >50% sequence coverage) in the whole-cell proteomic profile. Although this analysis neither allows the identification of the molecules transported using CCVs, nor reveals the membranes that are invaginated, their high abundance draws attention to these coated vesicles for future research. Three intriguing possibilities are that (1) this mechanism provides a means for transporting proteins and metabolites across the unique organelle arrangement of diatoms, and the ability to recycle receptors and proteins for further use (e.g. Stoorvogel et al. 1996,Donohoe et al. 2007), (2) clathrin-coated vesicles are important in polarized growth, such as in algal rhizoids initiating cell division (Limbach et al. 2008), or (3) clathrin is involved in silica deposition vesicles, since large amounts of membrane are required for cell division under exponential growth.

Most abundant proteins

Fucoxanthin chlorophyll a/c proteins (FCP; 18 kDa) were the most highly expressed proteins in *Thalassiosira pseudonana* whole-cell lysates. A total of 14 different FCP homologues were identified, 5 of which were among the 20 most abundant proteins (rank orders 1, 2, 4, 7, 18; Fig. 6). In contrast, the expression of the carbon fixation enzyme RuBisCo (large subunit; 54 kDa) was lower than anticipated (rank 13). Prior to the present study, it had been assumed that RuBisCo would be the most abundantly expressed protein under nutrient-replete conditions, as it constitutes as much as 40% of higher plant protein mass (Dhingra et al. 2004) and has been projected to be the most abundant protein on earth (Cooper 2000). Our results indicate that most protein in diatoms is associated with the light harvesting antennae, which contain FCPs, while the RuBisCo enzyme is either recycled for multiple uses or has a short half-life. Variations in growth conditions followed by protein profiling will provide greater insight into phytoplankton adaptations and physiological responses.

CONCLUSIONS

On average, whole-cell analysis using a Lyse-N-Go shotgun proteomic technique on a high-end nano-flow HPLC LTQ-FTMS costs around $250 per analysis and can yield ~200 to 500 proteins, depending on preparation techniques, chromatography and instrumentation methods. Our results demonstrate how shotgun proteomics can be employed as a low-investment (both time and money) screening technique for verifying putative physiological pathways (e.g. urea cycle and C_4 pathway) and discovering the importance of overlooked biochemical mechanisms (e.g. clathrin-mediated endocytosis). Although our study included an exhaustive number of mass spectrometry experiments to investigate the effects of cellular and gas-phase fractionations on total number of proteins identified, current evidence shows that gas-phase fractionation could be optimized to further decrease the number of experiments (Scherl et al. 2008). We note that proteomic methods continue to be reformed, to provide more information at a lower cost, as with genomic methods. Because current field techniques are not capable of isolating enough of a single microorganism from the ocean to allow holistic proteomic profiling, we propose that future studies should be conducted on phytoplankton cultures grown under controlled laboratory conditions (e.g. with variations in light, nitrogen, or iron concentrations). Knowledge of which proteins are expressed under particular environmental perturbations will allow the development of inexpensive protein-specific assays to rapidly assess the physiological status of phytoplankton communities. In addition to the development of environmental markers that monitor the current status of phytoplankton blooms, a mechanistic understanding of biochemical pathways through protein expression studies may elucidate how environmental perturbations affect phytoplankton physiology and their role in global carbon, nitrogen and silicon cycles (MacIntyre & Cullen 2005).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the University of Otago for access to facilities and administrative support, and B. Gallis and Y. Goo for input on the manuscript. B.L.N. thanks the National Science Foundation Office of Polar Programs for Postdoctoral funding. This work was supported by NSF grants OCE0453737 to B.L.N. and OCE-0825790 to B.L.N. and D.R.G.; and NIH grants NIEHS 5P30ES007033-10, NCRR 1S10RR02344-01, and NCRR 1S10RR17262-01 to D.R.G.

References

Armbrust EV, Berges JA, Bowler C, Green BR, et al. The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. Science 2004;306:79–86.10.1126/science.1101156 [PubMed: 15459382]

Ashburner M, Ball CA, Blake JA, Botstein D, et al. Gene ontology: tool for the unification of biology. Nat Genet 2000;25:25–29.10.1038/75556 [PubMed: 10802651]

Bertone P, Stolc V, Royce TE, Rozowsky JS, et al. Global identification of human transcribed sequences with genome tiling arrays. Science 2004;306:2242–2246.10.1126/science.1103388 [PubMed: 15539566]

Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature 2007;447:799–816.10.1038/nature05874 [PubMed: 17571346]

Chen J, Ryu S, Gharib SA, Goodlett DR, Schnapp LM. Exploration of the normal human bronchoalveolar lavage fluid proteome. Proteomics Clin Appl 2008;2:585–595.

Conner SD, Schmid SL. Regulated portals of entry into the cell. Nature 2003;422:37–44.10.1038/nature01451 [PubMed: 12621426]

Cooper, GM. The cell: a molecular approach. Vol. 2. Sinauer Associates; Sunderland, MA: 2000.

Davis AK, Hildebrand M, Palenik B. A stress-induced protein associated with the girdle band region of the diatom Thalassiosira pseudonana (Bacillariophyta). J Phycol 2005;41:577–589.10.1111/j.1529-8817.2005.00076.x

Dhingra A, Portis AR, Daniell H. Enhanced translation of a chloroplast-expressed RbcS gene restores small subunit levels and photosynthesis in nuclear RbcS antisense plants. Proc Natl Acad Sci USA 2004;101:6315–6320.10.1073/pnas.0400981101 [PubMed: 15067115]

Dhonukshe P, Aniento F, Hwang I, Robinson DG, Mravec J, Stierhof YD, Friml J. Clathrin-mediated constitutive endocytosis of PIN auxin efflux carriers in Arabidopsis. Curr Biol 2007;17:520–527.10.1016/j.cub.2007.01.052 [PubMed: 17306539]

Donohoe BS, Kang BH, Staehelin LA. Identification and characterization of COPIa- and COPIb-type vesicle classes associated with plant and algal Golgi. Proc Natl Acad Sci USA 2007;104:163–168.10.1073/pnas.0609818104 [PubMed: 17185411]

Elias JE, Gygi SP. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. Nat Methods 2007;4:207–214.10.1038/nmeth1019 [PubMed: 17327847]

Eng JK, Fischer B, Grossmann J, Maccoss MJ. A fast SEQUEST cross correlation algorithm. J Proteome Res 2008;7:4598–4602.10.1021/pr800420s [PubMed: 18774840]

Falkowski PG, Katz ME, Knoll AH, Quigg A, Raven JA, Schofield O, Taylor FJR. The evolution of modern eukaryotic phytoplankton. Science 2004;305:354–360.10.1126/science.1095964 [PubMed: 15256663]

Foss EJ, Radulovic D, Shaffer SA, Ruderfer DM, Bedalov A, Goodlett DR, Kruglyak L. Genetic basis of proteome variation in yeast. Nat Genet 2007;39:1369–1375.10.1038/ng.2007.22 [PubMed: 17952072]

Frigeri LG, Radabaugh TR, Haynes PA, Hildebrand M. Identification of proteins from a cell wall fraction of the diatom Thalassiosira pseudonana. Mol Cell Proteomics 2006;5:182–193. [PubMed: 16207702]

Goo YA, Liu AY, Ryu S, Shaffer S, et al. Identification of secreted glycoproteins of human prostate and bladder stromal cells by comparative quantitative proteomics. Prostate 2009;69:49–61.10.1002/pros.20853 [PubMed: 18792917]

Higdon R, Kolker E. A predictive model for identifying proteins by a single peptide match. Bioinformatics 2007;23:277–280.10.1093/bioinformatics/btl595 [PubMed: 17121779]

Hildebrand M. Prospects of manipulating diatom silica nanostructure. J Nanosci Nanotechnol 2005;5:146–157.10.1166/jnn.2005.013 [PubMed: 15762173]

Kaksonen M, Toret CP, Drubin DG. Harnessing actin dynamics for clathrin-mediated endocytosis. Nat Rev Mol Cell Biol 2006;7:404–414.10.1038/nrm1940 [PubMed: 16723976]

Keller A, Nesvizhskii AI, Kolker E, Aebersold R. Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. Anal Chem 2002;74:5383–5392.10.1021/ac025747h [PubMed: 12403597]

Kröger N, Deutzmann R, Bergsdorf C, Sumper M. Species-specific polyamines from diatoms control silica morphology. Proc Natl Acad Sci USA 2000;97:14133–14138.10.1073/pnas.260496497 [PubMed: 11106386]

Kroth PG, Chiovitti A, Gruber A, Martin-Jezequel V, et al. A model for carbohydrate metabolism in the diatom Phaeodactylum tricornutum deduced from comparative whole genome analysis. PLoS One 2008;3:e1426.10.1371/journal.pone.0001426 [PubMed: 18183306]

Limbach C, Staehelin LA, Sievers A, Braun M. Electron tomographic characterization of a vacuolar reticulum and of six vesicle types that occupy different cytoplasmic domains in the apex of tip-growing Chara rhizoids. Planta 2008;227:1101–1114.10.1007/s00425-007-0684-y [PubMed: 18193275]

MacIntyre, HL.; Cullen, JJ. Using cultures to investigate physiological ecology of microalgae. In: Anderson, RA., editor. Algal culturing techniques. Elsevier; Amsterdam: 2005. p. 287-326.

Malmstrom L, Riffle M, Strauss CE, Chivian D, Davis TN, Bonneau R, Baker D. Superfamily assignments for the yeast proteome through integration of structure prediction with the gene ontology. PLoS Biol 2007;5:e76.10.1371/journal.pbio.0050076 [PubMed: 17373854]

McDonald WH, Yates JR III. Shotgun proteomics and biomarker discovery. Dis Markers 2002;18:99–105. [PubMed: 12364816]

McGinn PJ, Morel FM. Expression and inhibition of the carboxylating and decarboxylating enzymes in the photosynthetic C4 pathway of marine diatoms. Plant Physiol 2008;146:300–309.10.1104/pp.107.110569 [PubMed: 17993542]

Mock T, Samanta MP, Iverson V, Berthiaume C, et al. Whole-genome expression profiling of the marine diatom Thalassiosira pseudonana identifies genes involved in silicon bioprocesses. Proc Natl Acad Sci USA 2008;105:1579–1584.10.1073/pnas.0707946105 [PubMed: 18212125]

Montsant A, Allen AE, Coesel S, De Martino A, et al. Identification and comparative genomic analysis of signaling and regulatory components in the diatom Thalassiosira pseudonana. J Phycol 2007;43:585–604.10.1111/j.1529-8817.2007.00342.x

Nelson DM, Treguer P, Brzezinski MA, Leynaert A, Queguiner B. Production and dissolution of biogenic silica in the ocean: revised global estimates, comparison with regional data and relationship to biogenic sedimentation. Global Biogeochem Cycles 1995;9:359–372.10.1029/95GB01070

Nunn BL, Timperman AT. Marine proteomics. Mar Ecol Prog Ser 2007;332:281–289.10.3354/meps332281

Nunn BL, Shaffer SA, Scherl A, Gallis B, Wu M, Miller SI, Goodlett DR. Comparison of a Salmonella typhimurium proteome defined by shotgun proteomics directly on an LTQ-FT and by proteome pre-fractionation on an LCQ-DUO. Brief Funct Genomics Proteomics 2006;5:154–168.10.1093/bfgp/ell024

Oudot-Le Secq MP, Grimwood J, Shapiro H, Armbrust EV, Bowler C, Green BR. Chloroplast genomes of the diatoms Phaeodactylum tricornutum and Thalassiosira pseudonana: comparison with other plastid genomes of the red lineage. Mol Genet Genomics 2007;277:427–439.10.1007/s00438-006-0199-4 [PubMed: 17252281]

Peers GS, Milligan AJ, Harrison PJ. Assay optimization and regulation of urease activity in two marine diatoms. J Phycol 2000;36:523–528.10.1046/j.1529-8817.2000.99037.x

Ragueneau O, Schultes S, Bidle K, Claquin P, La Moriceau B. Si and C interactions in the world ocean: importance of ecological processes and implications for the role of diatoms in the biological pump. Global Biogeochem Cycles 2006;20:GB4S02.10.1029/2006GB002688

Reinfelder JR, Milligan AJ, Morel FM. The role of the C4 pathway in carbon accumulation and fixation in a marine diatom. Plant Physiol 2004;135:2106–2111.10.1104/pp.104.041319 [PubMed: 15286292]

Roberts K, Granum E, Leegood RC, Raven JA. C3 and C4 pathways of photosynthetic carbon assimilation in marine diatoms are under genetic, not environmental, control. Plant Physiol 2007;145:230–235.10.1104/pp.107.102616 [PubMed: 17644625]

Ryu S, Gallis B, Goo YA, Shaffer SA, Radulovic D, Goodlett DR. Comparison of a label- free quantitative proteomic method based on peptide ion current area to the isotope coded affinity tag method. Cancer Inform 2008;4:243–255. [PubMed: 19259412]

Scherl A, Shaffer SA, Taylor GK, Kulasekara HD, Miller SI, Goodlett DR. Genome-specific gas-phase fractionation strategy for improved shotgun proteomic profiling of proteotypic peptides. Anal Chem 2008;80:1182–1191.10.1021/ac701680f [PubMed: 18211032]

Stolc V, Samanta MP, Tongprasit W, Sethi H, et al. Identification of transcribed sequences in Arabidopsis thaliana by using high-resolution genome tiling arrays. Proc Natl Acad Sci USA 2005;102:4453–4458.10.1073/pnas.0408203102 [PubMed: 15755812]

Stoorvogel W, Oorschot V, Geuze HJ. A novel class of clathrin-coated vesicles budding from endosomes. J Cell Biol 1996;132:21–33.10.1083/jcb.132.1.21 [PubMed: 8567724]

Takei K, Haucke V. Clathrin-mediated endocytosis: membrane factors pull the trigger. Trends Cell Biol 2001;11:385–391.10.1016/S0962-8924(01)02082-7 [PubMed: 11514193]

Tonon T, Qing R, Harvey D, Li Y, Larson TR, Graham IA. Identification of a long-chain polyunsaturated fatty acid acyl-coenzyme A synthetase from the diatom Thalassiosira pseudonana. Plant Physiol 2005;138:402–408.10.1104/pp.104.054528 [PubMed: 15821149]

Willingham AT, Orth AP, Batalov S, Peters EC, et al. A strategy for probing the function of noncoding RNAs finds a repressor of NFAT. Science 2005;309:1570–1573.10.1126/science.1115901 [PubMed: 16141075]

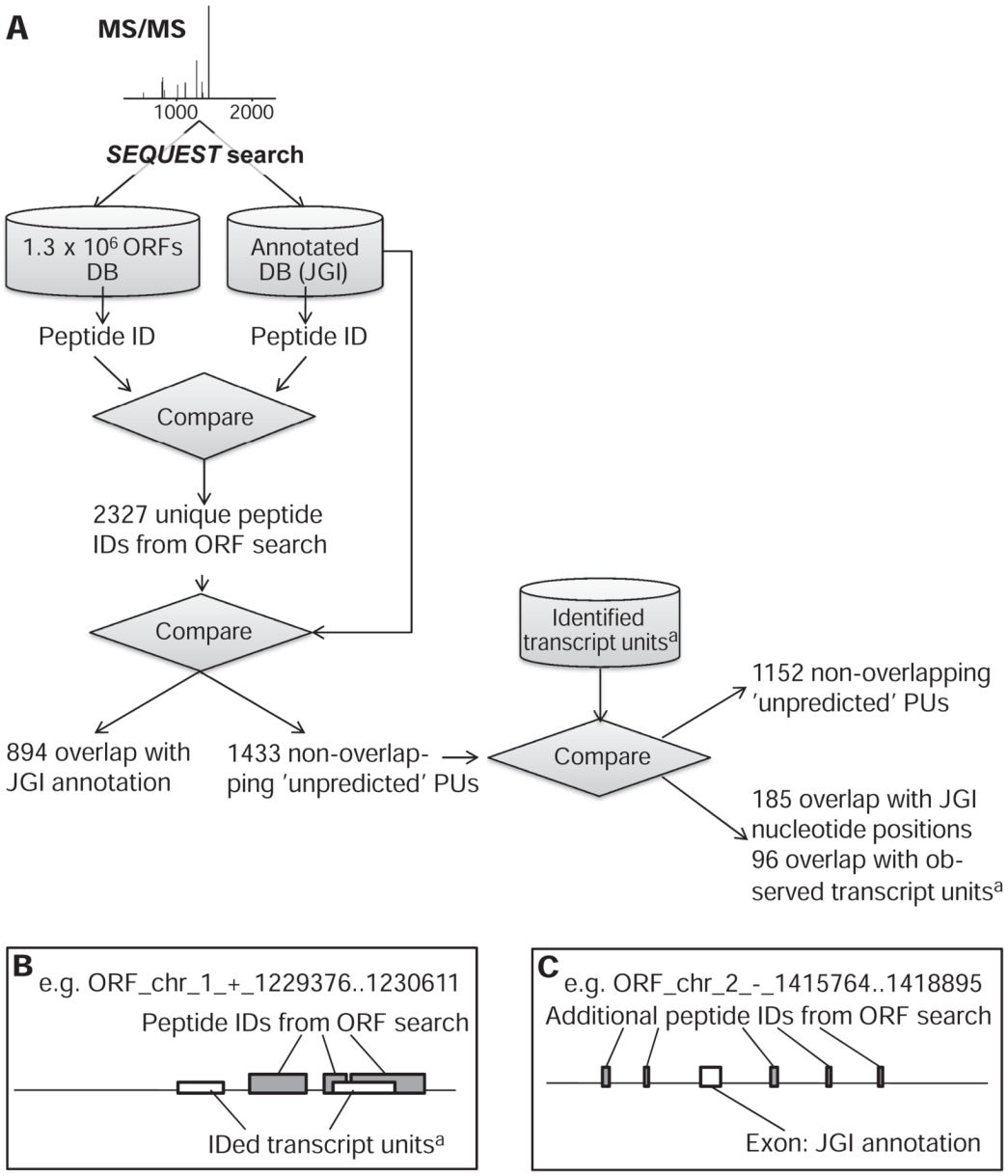

Fig. 1.

Thalassiosira pseudonana. (A) Outline of the methodology for finding expressed peptides that were not previously annotated by JGI (2005) or observed in tiling array transcript units. A total of 1433 peptides that do not correlate with the current *T. pseudonana* gene model were expressed; 185 of these overlap with nucleotide positions from the JGI annotation, but the sequence is in a different direction or different codon, while 96 peptides overlap with previous discoveries of novel transcripts. (B) Example of novel peptide units (dark grey) observed on a single open reading frame (ORF) (black line). Some of these peptides overlap with transcript units observed in tiling arrays (Mock et al. 2008). (C) Example of novel peptide units (dark grey) observed from a single ORF (black line) near an exon, as modeled by JGI. MS/MS: tandem mass spectrometry, DB: database, PU: peptide units. a Transcription unit observations made by Mock et al. (2008)

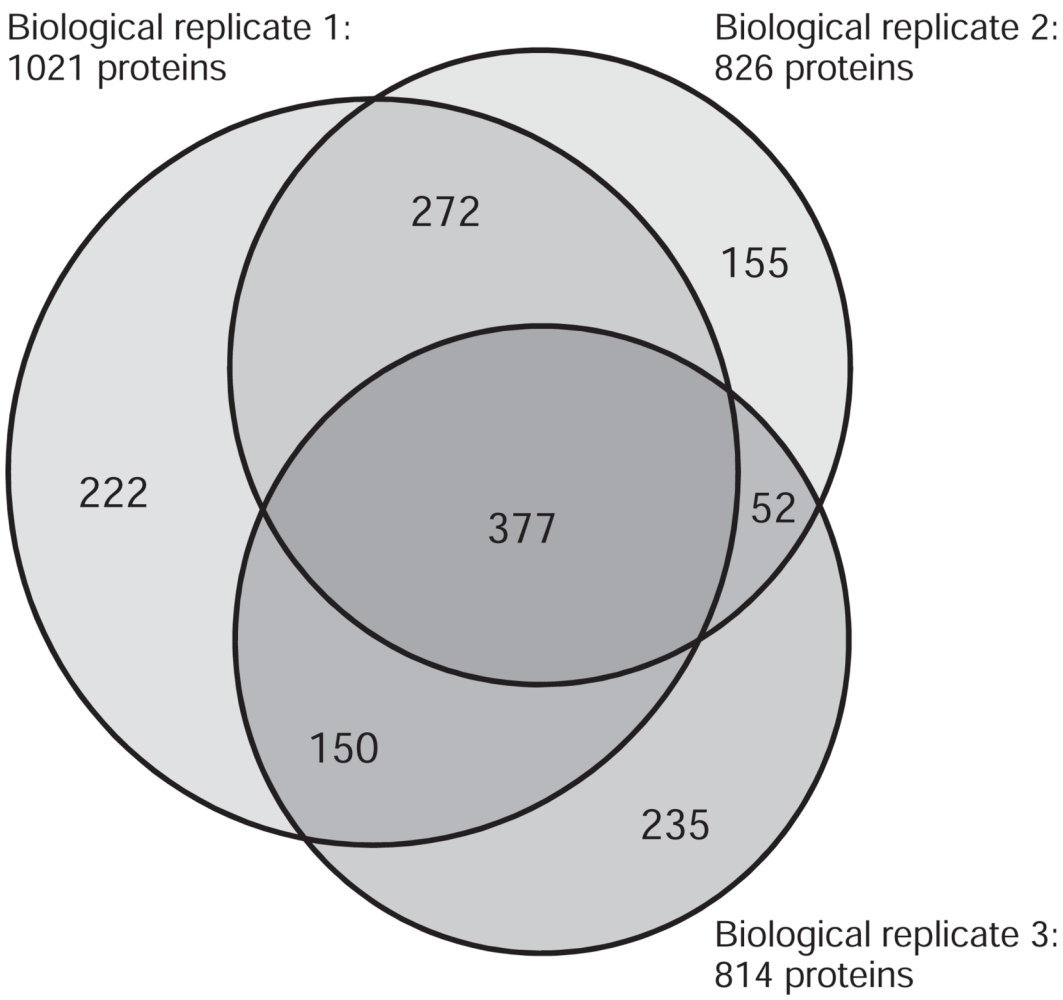

Fig. 2.

Thalassiosira pseudonana. Overlap of proteins identified from 3 biological replicates analyzed using the same sample preparation and mass spectrometry (MS) protocol as in Fig. 1. Biological replicates were each grown in separate containers and harvested in the same manner. To reduce sample handling, 6 gas-phase fractionations were performed in the mass spectrometer for each biological replicate. This method increases the total number of peptide ions to isolate for MS2 analyses, and ultimate sequencing. With each additional MS experiment performed on these biological replicates, greater overlap occurred in protein identifications. Numbers in the circles indicate the number of identified proteins

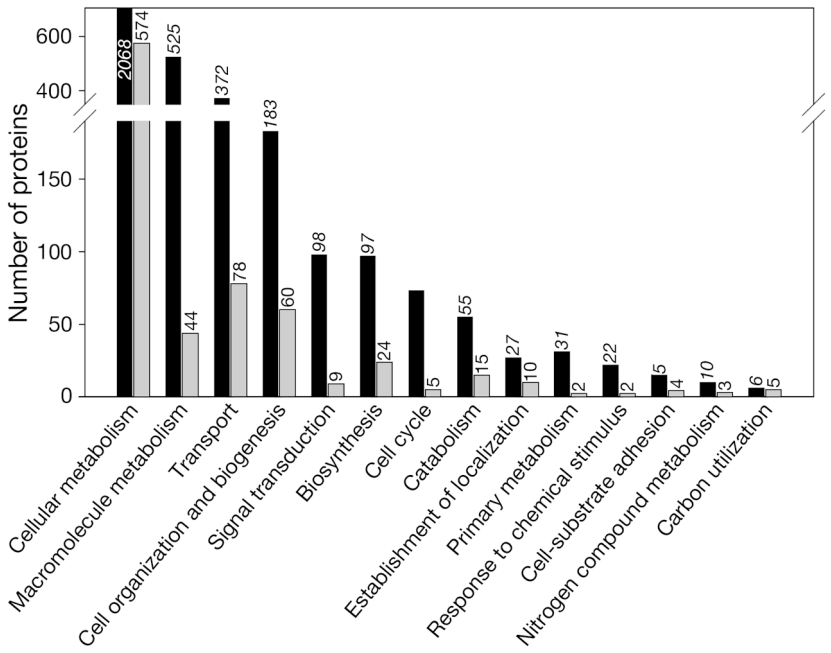

Fig. 3.

Thalassiosira pseudonana. Proteins present at time of cell harvest (white bars) relative to the proteins predicted from the genome (black bars). There are 3 main categories within the gene ontology (GO): biological process, molecular function, and cellular component. Each of these ontologies has a hierarchical system containing more specificity. Here, the 3rd level in the biological processes ontology is presented, showing that the majority of cellular proteins is involved in metabolic and transport processes. The actual number of proteins for each process is shown above the bars

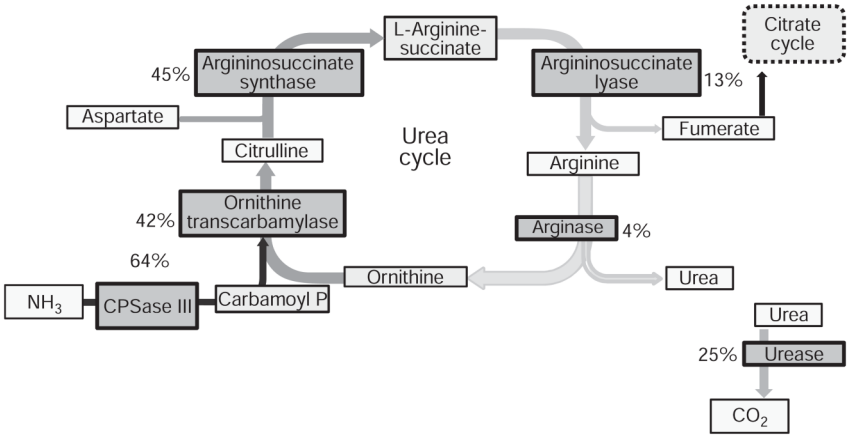

Fig. 4.

Thalassiosira pseudonana. (grey boxes) Enzymes utilized during the cycling of urea, presented with % protein sequence coverage (a relative measure of protein abundance) observed. (White boxes) Metabolites utilized or produced during the synthesis and degradation of urea (not monitored). CPSase III: carbamoyl-phosphate synthetase (40323), ornithine transcarbamylase (997), argininosuccinate synthase (42719), arginosuccinate lyase (29075), arginase (35561), urease (30193). Numbers in parentheses are protein annotation identifiers

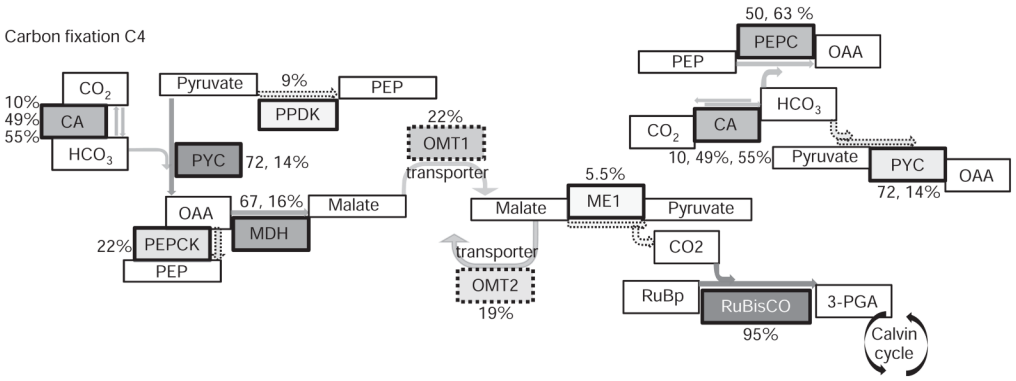

Fig. 5.

Thalassiosira pseudonana. (grey boxes) Carbon fixation enzymes observed, with % protein sequence coverage observed. (White boxes) Synthesized or degraded metabolites (not monitored). Enzymes illustrated more than once indicate the annotation, resulting proteomic profiles include multiple isoforms, and N-terminal pre-sequences suggest different cellular locations. CA: carbonic anhydrase (25840, 34125, 34094), OMT: oxogluterate/malate translocator protein (20731, 26366), ME1: malic enzyme (34030), MDH: malate dehydrogenase (20726, 41425), PEPC: phosphoenolpyruvate carboxylase (34543, 268546), PEPCK: phosphoenolpyruvate carboxykinase (5186), PYC: pyruvate carboxylase (11076, 269908, 11075), RuBisCo: ribulose-bisphosphate carboxylase (116793850, 118411103, 118411022). Numbers in parentheses are protein annotation identifiers. Metabolites include OAA: oxaloacetate, PEP: phosphoenolpyruvate, RuBp: ribulose-bisphosphate

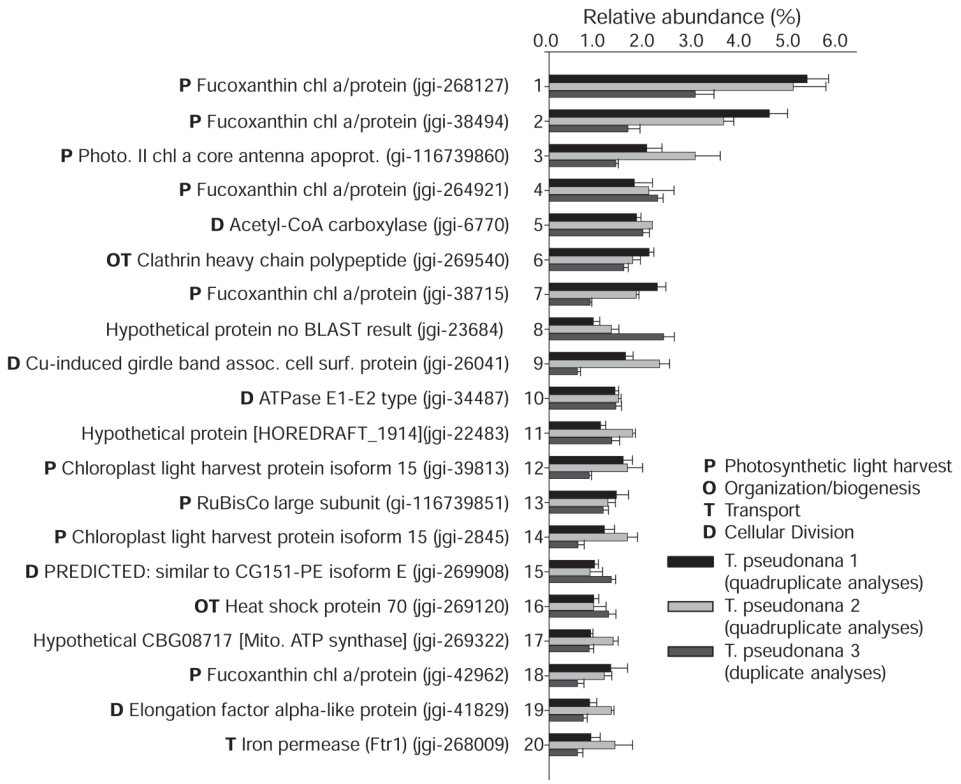

Fig. 6.

Thalassiosira pseudonana. Illustration of the 20 most abundant proteins expressed when grown under resource-replete conditions. Peptide spectral counting was used to evaluate protein abundance at the peptide level for 2 sets of quadruplicate data (using a Thermo Fisher LTQ-Orbitrap mass spectrometer) and 1 set of duplicate data (using a Thermo Fisher LTQ-FT mass spectrometer) from 3 different biological samples. Regardless of biological replicate analyzed, instrument platform or data acquisition type, the 20 most abundant proteins remained the same. The maximum relative amount of any one protein was ~6%. Error bars are SDs of the average of quadruplicate or duplicate data suites observed in replicate tandem MS analyses. P, O, T, and R are indicators for gene ontology (GO) categories represented. *D is indicative of proteins used in purine nucleotide metabolism and replication

Table 1

Thalassiosira pseudonana. Number of theoretical proteins predicted in genome annotation (v.3; www.jgi.doe.gov) and total number observed in 10 primary biochemical pathways (as defined by the Kyoto Encyclopedia of Genes and Genomes, KEGG)

	Theoretical	Observed
Carbon fixation	17	14
Citrate cycle (TCA cycle)	14	14
Glycolysis/gluconeogenesis	21	17
Pyruvate metabolism	23	19
Reductive carboxylate cycle (CO_2 fixation)	8	7
Nitrogen metabolism	17	13
Urea cycle and amino acid metabolism	29	12
Fatty acid metabolism	14	10
Fatty acid biosynthesis	9	7
Biosynthesis of unsaturated fatty acids	6	3