Mar Biotechnol (2011) 13:496–504
DOI 10.1007/s10126-010-9320-0

ORIGINAL ARTICLE

Shotgun Proteomic Analysis of *Emiliania huxleyi*, a Marine Phytoplankton Species of Major Biogeochemical Importance

Bethan M. Jones · Richard J. Edwards · Paul J. Skipp · C. David O'Connor · M. Debora Iglesias-Rodriguez

Received: 25 March 2010 / Accepted: 7 September 2010 / Published online: 6 October 2010
© Springer Science+Business Media, LLC 2010

Abstract *Emiliania huxleyi* is a unicellular marine phytoplankton species known to play a significant role in global biogeochemistry. Through the dual roles of photosynthesis and production of calcium carbonate (calcification), carbon is transferred from the atmosphere to ocean sediments. Almost nothing is known about the molecular mechanisms that control calcification, a process that is tightly regulated within the cell. To initiate proteomic studies on this important and phylogenetically remote organism, we have devised efficient protein extraction protocols and developed a bioinformatics pipeline that allows the statistically robust assignment of proteins from MS/MS data using preexisting EST sequences. The bioinformatics tool, termed BUDAPEST (Bioinformatics Utility for Data Analysis of Proteomics using ESTs), is fully automated and was used to search against data generated from three strains. BUDAPEST increased the number of identifications over standard protein database searches from 37 to 99 proteins when data were amalgamated. Proteins involved in diverse cellular processes were uncovered. For example, experimental evidence was obtained for a novel type I polyketide synthase and for various photosystem components. The proteomic and bioinformatic approaches developed in this study are of wider applicability, particularly to the oceanographic community where genomic sequence data for species of interest are currently scarce.

Keywords BUDAPEST · *Emiliania huxleyi* · EST analysis · Mass spectrometry · Shotgun proteomics

Introduction

Coccolithophores are unicellular eukaryotic algae comprising over 200 extant species (Jordan and Green 1994). Since the mid-Mesozoic, coccolithophores have played an extremely important role in the global carbon cycle through the uptake of CO_2 by photosynthesis. They are members of a diverse group of marine phytoplankton which, despite being responsible for only 0.2% of global primary producer biomass, contribute to nearly 50% of global net primary productivity (Field et al. 1998). Coccolithophores also influence carbon fluxes in the ocean by calcification, the biogenic process of calcite ($CaCO_3$) production through the intracellular precipitation of $CaCO_3$ into elaborate plate-like structures termed coccoliths. These are then extruded to the cell surface and can subsequently fall to the seabed where they can be preserved for long geological periods. In the present-day ocean, coccolithophores are thought to be one of the most predominant sources of $CaCO_3$ in deep-sea sediments (Milliman 1993; Westbroek et al. 1993), thereby

Bethan M. Jones and Richard J. Edwards contributed equally to this work.

Electronic supplementary material The online version of this article (doi:10.1007/s10126-010-9320-0) contains supplementary material, which is available to authorized users.

B. M. Jones (✉) · M. D. Iglesias-Rodriguez
School of Ocean and Earth Science,
National Oceanography Centre, Southampton,
University of Southampton,
Waterfront Campus, European Way,
Southampton SO14 3ZH, UK
e-mail: bmj@noc.soton.ac.uk

R. J. Edwards · P. J. Skipp · C. D. O'Connor
School of Biological Sciences,
University of Southampton,
Southampton, UK

P. J. Skipp · C. D. O'Connor
Centre for Proteomic Research, University of Southampton,
Southampton, UK

forming extensive carbon reservoirs. Coccoliths may form between 20% and 80% of total sedimentary carbonate in marine environments, depending on the region studied (Baumann et al. 2004).

The most abundant and cosmopolitan representative of this important group of phytoplankton, *Emiliania huxleyi* (Lohmann) Hay et Mohler, has additional significance. The species has the capability to form mesoscale, largely monospecific blooms in high-latitude regions (Brown and Yoder 1994; Iglesias-Rodriguez et al. 2002) where cell concentrations can reach over a million cells per litre (Holligan et al. 1993; Tyrrell and Merico 2004). These blooms, sometimes over 250,000 km^2 in area (Holligan et al. 1993), are detectable by satellite imagery (Balch et al. 1991; Holligan et al. 1993; Brown and Yoder 1994) and influence how much CO_2 is taken up by the ocean in the bloom vicinity (Balch et al. 1991; Holligan et al. 1993; Robertson et al. 1994; Tyrrell and Merico 2004). Several thousand tonnes of $CaCO_3$ are precipitated in each bloom event (de Vrind-de Jong and de Vrind 1997).

As the ecological and biogeochemical importance of coccolithophores has become recognized, the physiology and biochemistry of *E. huxleyi* are being increasingly studied (Paasche 2001). Additionally, genetic studies are now also underway (Wahlund et al. 2004a, b; Nguyen et al. 2005; Quinn et al. 2006; Dyhrman et al. 2006). However, many basic scientific questions regarding *E. huxleyi* still remain unanswered; for example, the molecular biology underlying both the processes of calcification and bloom formation is almost completely unknown. No proteins have yet been isolated with a confirmed role in coccolith formation, and the molecular mechanisms behind this process remain elusive. Whilst a fully annotated genome for the species is not currently available, there are public sequence data for ESTs (Wahlund et al. 2004a, b). This resource, coupled with the availability of high-throughput, MS-based proteomic technologies, could allow large-scale studies on *E. huxleyi* at the protein level to characterize its biochemistry and response to changes in environmental conditions. However, this would first require methods to isolate proteins from *E. huxleyi* cells in a robust and reproducible manner. Additionally, it would be necessary to identify proteins for a species that currently shows taxonomical isolation whilst also utilizing EST data of unknown quality.

This paper describes the development of methods for the extraction of proteins from *E. huxleyi* for proteomic studies and an initial survey of its proteome. Protein identification was facilitated by an efficient in-house software pipeline—"Bioinformatics Utility for Data Analysis of Proteomics using ESTs" (BUDAPEST)—that allows the robust identification of proteins from ESTs using MS/MS data. The high degree of automation associated with data generation is likely to be of use for non-specialists. BUDAPEST is not species-specific and could therefore be of significant use to the oceanographic community where few genome sequences currently exist for species of biogeochemical interest. This study represents a significant increase in our knowledge of the biology and biochemistry of *E. huxleyi*, whilst the techniques developed will also assist our understanding of the biochemical and molecular processes that affect the global carbon cycle both today and in the future.

Materials and Methods

Algal Cultures and Media

E. huxleyi strain NZEH (also known as Plymouth M219 or CAWPO 6 'A') was obtained from the Plymouth Culture Collection of Marine Algae (UK). Strains CCMP1516 and CCMP371 were obtained from the Provasoli-Guillard National Centre for the Culture of Marine Phytoplankton (Maine, USA). All strains were grown in batch cultures using 0.22 μm filter-sterilized f/2 medium (Guillard and Ryther 1962; Guillard 1975). Cultures were maintained in the exponential stage of growth at 18°C and exposed to a 12:12-h light/dark light cycle with a light irradiance of 80 μmol quantam^{-2} s^{-1}. For washing of cells, an artificial seawater medium lacking nitrate, phosphate and $CaCO_3$ was used (Riegman et al. 2000).

Protein Extraction

Boiling *E. huxleyi* cells in SDS-PAGE final sample buffer or moderate sonication (<10 bursts at 10 s until all cells were lysed) gave poor quality results (Fig. 1). In contrast, both extensive sonication termed "hypersonication" (twenty-five 10-s bursts of sonication) and moderate sonication following extracellular decalcification with HCl proved to be extremely reproducible methods for the production of extracts suitable for GeLC-MS/MS analyses (Fig. 1), but only when proteins were precipitated from lysates using acetone. As the continued presence of $CaCO_3$ in samples prepared by hypersonication might interfere with some types of analyses, samples were processed by decalcification, which was successful for heavily calcified strains (NZEH and CCMP371) and for the moderately calcifying strain CCMP1516.

At the end of the exponential growth phase, 4 L of culture containing approximately 3.19×10^9–1.13×10^{10} cells was decalcified following a modified version of the coccolith removal method originally described by Linschooten et al. (1991). HCl (100 mM) was added dropwise to the culture until pH 5.0 was reached. After 2 min, rapid readjustment to the original culture pH was made using 100 mM NaOH. Following decalcification, cell integrity was checked by

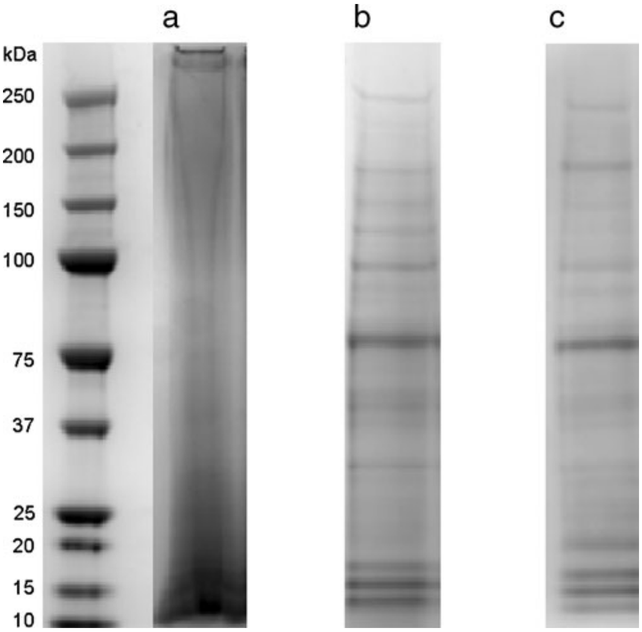

Fig. 1 SDS-PAGE of proteins of *E. huxleyi* (strain NZEH) extracted using the original moderate sonication method (**a**), the decalcification method (**b**) and by hypersonication (**c**) as described in "Materials and Methods". Gels were stained with colloidal CBB and molecular weight markers are indicated on the *left-hand side*. **a** Vertical streaking caused by interfering chemicals that are removed in the decalcification/hypersonication protocols (**b**, **c**). The proteins in **b** were analysed by GeLC-MS/MS

microscopy prior to harvesting by centrifugation (3,500×*g*, 5 min at 4°C) and four washes with an artificial seawater medium (Riegman et al. 2000). The resulting pellet of decalcified cells was resuspended in 10 mL of 100 mM triethylammonium bicarbonate (TEAB) at pH 8.0, 4°C (Sigma Aldrich, Poole, UK) and sonicated on ice for six 10-s bursts (1-min intervals between bursts) using a VC300 Vibracell sonicator (Sonics and Materials, USA) with a 20-kHz frequency, 10% duty cycle and an output of 3. Cell extracts were centrifuged at 4,630×*g* for 30 min to pellet cell debris and the supernatants were snap frozen and stored at −80°C. After gentle thawing at room temperature, supernatants were precipitated in 10× volume −20°C acetone overnight and proteins collected by centrifugation at 21,250×*g*, 4°C for 20 min. The resulting pellet was resuspended in NuPAGE 4× LDS sample buffer (Invitrogen, Paisley, UK) or 100 mM TEAB with 0.1% (*w*/*v*) SDS through 10-min incubation at 37°C and several cycles of vigorous mixing and bath ultrasonication.

One-dimensional SDS-PAGE and GeLC-MS/MS

Resuspensions were centrifuged at 21,250×*g* for 30 min to pellet insoluble material and proteins in the supernatant were reduced with 0.5 M DTT. Approximately 30 μg of protein was fractionated on NuPAGE 4–12% gradient SDS polyacrylamide gels (Invitrogen; dimensions: 8.0×8.0 cm). After visualisation with 0.2% colloidal CBB, each gel lane was excised, cut in 28 equal-sized pieces (each approximately 2.5×7.0 mm) and subjected to in situ trypsin digestion following the method of Shevchenko et al. (1996). Tandem mass spectrometry was performed as previously described (Skipp et al. 2005). Briefly, the resulting peptides were separated by nanocapillary RP-LC using a Waters C18 RP, 3 μm, 100 Å (150 mm×75 μm, i.d.) column (Waters, Elstree, UK) and were subsequently electrosprayed into a quadrupole time-of-flight tandem mass spectrometer (Q-Tof Global Ultima, Waters) fitted with a nanoLockSpray™ source (Waters). A survey scan was acquired from *m*/*z* 300–1,700 and the switching criteria for MS to MS/MS included ion intensity and charge state. The collision energy used to perform MS/MS was varied according to the mass and charge state of the eluting peptide.

Protein Identifications from Protein Databases

Peak lists were generated using Mascot Distiller v.2.2 (Matrix Science, London, UK) and MS/MS data were searched against the NCBI-nr database. Data were also searched against a taxonomically restricted database, which contained 196,774 proteins (93,921,998 residues), and was constructed from all Alveolata, Cryptophyta, Haptophyceae, Rhodophyta and Stramenopile sequences from UniProtKB Release 15.1 (UniProt 2008) in addition to the proteome from the closest available full genome sequence, that of *Thalassiosira pseudonana* (Armbrust et al. 2004). This database was constructed in an attempt to maximize the coverage of closely related taxa but minimize the redundancy generated by excessively large databases. Both searches were performed using the Mascot search engine (Matrix Science) with parent mass tolerance set at 150 ppm and fragment mass tolerance at 0.25 Da. Carbamidomethylation was also set as a fixed modification, oxidation of methionine as a variable modification, and a maximum of one missed cleavage was allowed. The significance threshold for search results was set at $p<0.05$, and only proteins with 2+ supporting peptides were accepted. To remove redundancy issues due to the presence of proteins from multiple species, non-*E. huxleyi* proteins were only accepted if they had 2+ supporting peptides that were not found in identified *E. huxleyi* proteins. These were then annotated using homology to known proteins and phylogenetic context as will be described below (Electronic Supplementary Materials (ESM) Data 1).

Protein Identifications from ESTs Using the BUDAPEST Pipeline

Data were searched with Mascot against 90,100 *E. huxleyi* ESTs downloaded from NCBI-dbEST using the search term: "*Emiliania huxleyi*"[orgn:_txid2903] (20/10/2008). Results were used to assign peptides to proteins using an in-house software pipeline known as BUDAPEST (http://www.personal.soton.ac.uk/re1u06/software/budapest/; Fig. 2).

Conversion of ESTs to RF Translations

Mascot identifies ESTs using peptides from all six possible reading frame (RF) translations, and so the first stage of BUDAPEST identified and filtered out incorrect RFs and erroneous Mascot matches of peptides to these RFs. Each initially identified EST was first translated in all six RFs, and any RFs containing open reading frames (ORFs) of 10+ contiguous amino acid residues (i.e. without interrupting stop codons) were considered as candidates for the correct (i.e. coding) RF for that EST. If the sense strand could be predicted by the presence of a 3′ poly-A repeat or 5′ poly-T repeat (≥10 nucleotides), only the three sense RFs were considered and these were truncated at the 3′-proximal stop codon to exclude obvious 3′ UTR sequences. In silico translations (including internal stop codons) of these candidate RFs were then searched against the taxonomically restricted protein database (described above) using the BLASTP algorithm ($e<10^{-4}$; Altschul et al. 1997). To minimize the inclusion of probable UTRs and/or sequencing frameshifts, any translations with BLAST-detectable homology ($e<10^{-4}$) to a protein in the database were truncated if necessary so that only translated ORFs overlapping the BLAST local alignments were retained (i.e. both ends were removed up to, and including, any translated stop codons that flanked the region involved in the BLAST alignment; if no such flanking stop codon was present, the translated ORF was retained up to the end of the sequence).

Translations with BLAST hits were assumed to represent the true coding RF, and therefore, any translations of different RFs from the same EST that did not themselves have BLAST-detectable ($e<10^{-4}$) homologues in the database were removed from the analysis. If none of the RF translations for a given EST had a BLAST-detectable homologue, all of the candidate translations were kept without additional truncation. As a final step, translations were compared against the original Mascot output derived from the MS/MS data. Any peptides that did not match the processed RF translations were removed. Conversely, processed RF translations that did not match any peptides were also excluded.

Consensus Sequence Generation

Because ESTs represent redundant and partial sequences, it was possible that several ESTs encoded different stretches of the same protein and hence could be combined for identification purposes. Translations that were wholly or partially overlapping by ≥20 amino acids (at ≥95% sequence identity) were therefore combined into a single "consensus" (contig) sequence using another in-house programme, FIESTA (Fasta Input EST Analysis; http://www.personal.soton.ac.uk/re1u06/software/fiesta/). Any consensus sequences that were only supported by a single peptide derived from Mascot searches of MS/MS data (hereafter termed "supporting peptides") were excluded from the analysis at this stage. The remaining consensus sequences were then clustered using BLASTP, with all sequence sharing BLAST-detectable homology ($e<10^{-4}$) assigned to the same consensus cluster. Supporting peptides were then designated as either unique to a consensus sequence or shared by two or more sequences. Details can be found in the BUDAPEST manual (available online).

Identification of Proteins from Consensus Sequences

The final stage of the BUDAPEST pipeline assigned protein identifications to consensus sequences in a phylogenetic context using sequence homology to known proteins. Each consensus sequence was used as a query for HAQESAC (Homologue Alignment Quality, Establishment of Subfamilies and Ancestor Construction; Edwards et al. 2007) in an automated analysis pipeline to generate multiple sequence alignments and phylogenetic trees of protein families. Each query was searched using BLAST against the taxonomically restricted protein database (described above) and a representative subset of queries (≤50 closest homologues) was aligned with MAFFT (Katoh and Toh 2008). These were subsequently used to construct a phylogenetic tree with 1,000 bootstraps using FastTree (Price et al. 2009). Phylogenetic information was then used to identify the consensus sequences for protein identification. In the case of multi-gene families, discrete identities were given to different consensus sequences from the same cluster wherever possible. All trees and alignments generated are provided as ESM Data 2. All programmes, datasets and results files are available from http://www.personal.soton.ac.uk/re1u06/research/ehux/. Data are available in the PRIDE database (Martens et al. 2005; www.ebi.ac.uk/pride) under accession no. 9197.

Decoy Search Databases

We performed decoy database searches against the NCBI-nr database, which provided false discovery rates of 0.81% for strain NZEH, 0.9% for strain CCMP371 and 4.22% for strain CCMP1516. As an additional control for the incorrect matching of *E. huxleyi* peptides to randomly translated RFs in BUDAPEST, two additional sequence databases were created for Mascot searches in all RFs. In each case, 90,100 decoy ESTs of the same length to those in the input database were generated using a Monte Carlo

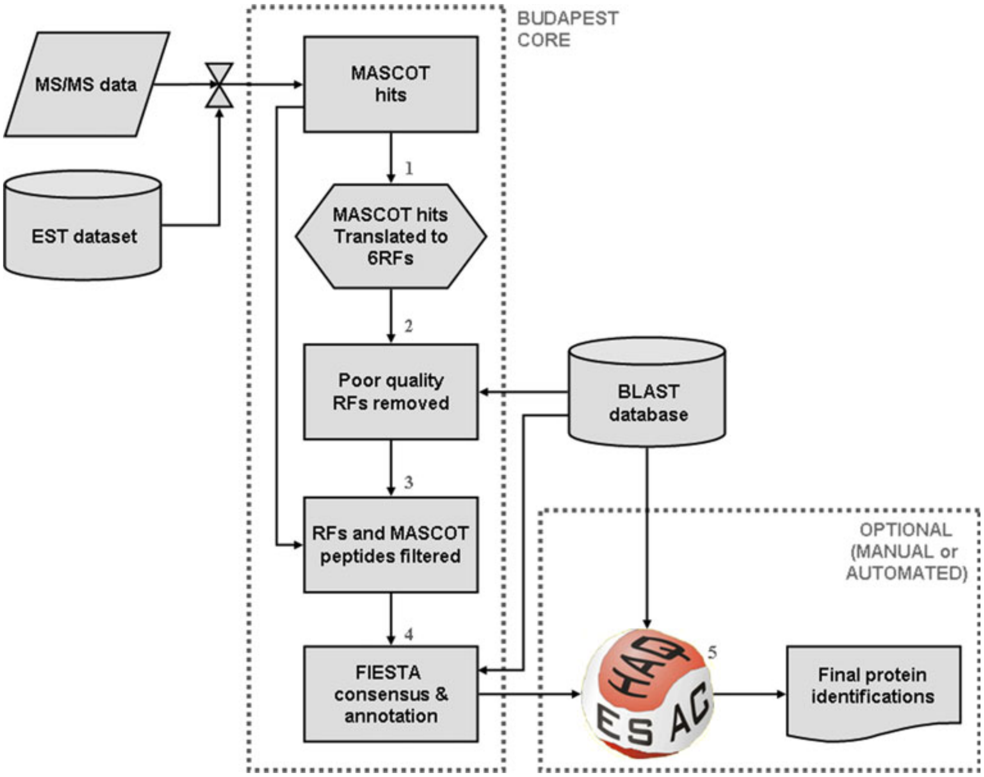
BUDAPEST CORE
MS/MS data
EST dataset
MASCOT hits
1
MASCOT hits Translated to 6RFs
2
Poor quality RFs removed
BLAST database
3
RFs and MASCOT peptides filtered
4
FIESTA consensus & annotation
OPTIONAL (MANUAL or AUTOMATED)
HAQ ESAC
5
Final protein identifications

Fig. 2 Summary of the BUDAPEST workflow. Stages executed by BUDAPEST workflow are indicated by the *dashed box*. *1* EST sequences (as identified by the initial Mascot search) were translated in all possible reading frames (*RFs*) using terminal poly-A or poly-T repeats to identify strand orientation where possible. *2* Poor quality translations, predicted to be from an incorrect reading frame on the basis of a BLASTP homology search ($e<10^{-4}$) against a protein database (e.g. UniProt), are removed. *3* Any peptides derived from Mascot searches of MS/MS data (supporting peptides) that only matched excluded RF translations are removed, as are RF translations that lack any supporting peptides. *4* The remaining RF translations are combined into consensus sequences based on overlapping regions of 20+ amino acids at 95%+ sequence identity and annotated using similarity (if any) to known proteins as established by a second BLASTP search. *5* Optionally, phylogenetic analysis of consensus sequences using HAQESAC can be used to improve sequence annotation and construct the final list of protein identifications

Markov chain method as implemented in RJE_SEQGEN (http://www.southampton.ac.uk/~re1u06/software/rje_seq/). For each EST, poly-[AT] repeats were removed, a new sequence was randomly generated, and then the poly-[AT] repeat (if present) was reattached. The first EST decoy database (decoy 1) was generated using nucleotide frequencies alone where the probability of a given nucleotide being added to the random sequence was determined by its frequency in the original sequence. The other database (decoy 2) took trinucleotide frequencies into account, where the probability of a given nucleotide being added to the random sequence was determined by the preceding two nucleotides. MS/MS data obtained from all three *E. huxleyi* strains (NZEH, CCMP1516 and CCMP371) were searched against decoy databases 1 and 2 using Mascot and the results processed using BUDAPEST.

Gene Ontology Analysis

The functional diversity and subcellular distributions of the proteins identified by BUDAPEST analysis were broadly assessed by gene ontology (GO) terms using QuickGO at GOA (Barrell et al. 2009; accessed through UniProtKB). The GO function and location were noted for the closest hit to each consensus sequence. If there was no annotated GO term for a given protein or if the protein was not present in UniProtKB (i.e. *T. pseudonana* proteome hits), then the term “not available” was given. Proteins identified by the taxonomically restricted database search were also annotated as described.

Results

Shotgun Proteomic Analysis Using GeLC-MS/MS and Protein Identification Using BUDAPEST

Following the development of methods to extract proteins from *E. huxleyi* (see “Materials and Methods”), proteins from cellular lysates of strains NZEH, CCMP1516 and CCMP371 were fractionated by SDS-PAGE prior to the excision of gel slices and trypsin treatment. Recovered peptides were characterized using nano-LC and MS/MS and the data searched against a taxonomically restricted database of proteins from Alveolata, Cryptophyta, Haptophyceae, Rhodophyta and the Stramenopiles using Mascot. In all, these yielded 37 proteins with significant matches (ESM Table 1), but only 19 of these were of *E. huxleyi* origin, highlighting the lack of genomic sequence information for this organism. Searches were also performed against NCBI-nr, but yielded no additional protein identifications (data not shown).

To increase the number of proteins identified, the MS/MS data were searched against an EST database of *E. huxleyi* cDNA sequences, yielding a total of 401 hits (ESM Data 3). However, it was plausible that some peptides were incorrectly matched to translated RFs derived from ESTs, thereby causing protein misidentification (Cooper et al. 2007). Therefore, a bioinformatics pipeline (BUDAPEST) was developed to allow a more rigorous assignment of MS/MS peptide data to ESTs. This constructed consensus sequences from corrected RFs to enable the identification of proteins in a phylogenetic context. It also clustered these consensus sequences by similarity.

Processing of the data using BUDAPEST yielded 83 consensus sequences in 41 clusters for NZEH, 68 sequences in 33 clusters for CCMP1516 and 63 sequences in 35 clusters for CCMP371. Conservative annotation using sequence homology and phylogenetics produced at least 80 unique protein identifications when data for all three strains were combined (ESM Table 2 and Data 3).

Combining data from the EST and protein database searches (by nomenclature) yielded 99 proteins in total, representing a substantial increase in protein identifications compared to the use of protein databases alone and highlighting the complementary nature of these two approaches. Protein descriptions were assigned as specifically as possible based on phylogenetic analysis (ESM Data 1 and 2). As with most genome annotations, the assigned protein descriptions were inferred from annotations of homologous proteins in other organisms. Thus, the precise roles of the proteins in *E. huxleyi* may require experimental validation even though their assignment to consensus clusters is unambiguous. Additionally, multiple consensus sequences in a cluster may represent several protein variants (e.g. splice isoforms) or different (albeit closely related) members of a protein family. It was not possible to distinguish between these two scenarios, so we have adopted a conservative identification strategy and it is likely that the number of identifications is a lower estimate.

In contrast to the results from the *E. huxleyi* EST library, decoy databases based on mononucleotide frequencies (decoy 1) and trinucleotide frequencies (decoy 2) each yielded in total only a single RF with 2+ supporting peptides from all three strains. We therefore conclude that protein identifications using the BUDAPEST pipeline are likely to be genuine rather than stochastic hits to translations of incorrect RFs.

Diversity of the Identified Proteins

It was important to determine if the protein extraction/assignment protocols developed in this study yielded a representative range of proteins from *E. huxleyi*. Accordingly, we grouped proteins identified by both BUDAPEST and Mascot searching according to their GO terms to determine their functional classes and subcellular compartments. Proteins involved in a range of molecular processes within numerous cellular compartments were detected (ESM Supplementary Results). For strain NZEH, 30% of all proteins were membrane-associated proteins, which are typically difficult to discover using 2-DE. Since the same protein can have a different function and cellular location in a different organism, it was important to only assign the GO terms of the closest hit to prevent misidentification and misclassification. This was difficult for most proteins in our study since all the closest hits in the UniProtKB database were electronically annotated and often lacked GO information. Consequently, 69% of all proteins identified for CCMP1516 could not be assigned a subcellular location and 45% did not have a designated process. Relative proportions of membrane-related proteins are therefore likely to be higher than reported for each sample, as are the proportions relating to associated processes.

Discussion

Protein Identification Using the BUDAPEST Pipeline

E. huxleyi plays an important biogeochemical role in the oceans. However, its underlying biochemistry and physiology are poorly understood. The problems of protein extraction from macroalgae species are well documented (Wong et al. 2006; Contreras et al. 2008; Nagai et al. 2008; Wang et al. 2009) and versatile protein extraction protocols did not exist for *E. huxleyi* prior to this investigation. Accordingly, in this study, we have developed highly adaptable methods for extracting proteins from this organism as well as an efficient strategy for identifying *E. huxleyi* proteins using available EST data. These methods remove interfering substances that have previously compromised the fractionation of *E. huxleyi* proteins by SDS-PAGE gels and rendered protein quantification assays ineffective. Of the alternative methods developed, decalcification and hypersonication are likely to be more versatile than protein extraction by boiling, which was not always reproducible and is incompatible with enzymatic assays.

Using the BUDAPEST strategy, 83 different consensus sequences in 41 clusters were identified for NZEH, 68 in 33 clusters for CCMP1516 and 63 in 35 clusters for CCMP371. This resulted in 80 positive protein assignments overall, of which 74 had BLAST-detectable homology ($e<10^{-4}$) with known proteins. Cross-species Mascot searches against protein data yielded an additional 19 proteins, bringing the total number of distinct proteins to 99 when overlapping identifications (by nomenclature) were taken into account.

Since many clusters have multiple consensus sequences, it is likely that numerous variants are present for some of the proteins. For example, it is possible that "unknown protein 1" (strain NZEH with four sequences in its cluster) may represent at least two isoforms of an unknown protein since sequence 11a has distinct peptides to sequences 11b, c and d (ESM Data 3). However, in most instances, further sequence data (e.g. a fully annotated genome) would be the only way to determine the nature of these different consensus sequences with more confidence.

Roles of the Identified Proteins

Identified components are associated with a range of cellular processes, including photosynthesis, protein folding and numerous metabolic pathways such as nitrogen metabolism (carbamoyl phosphate synthetase for the first step of the urea cycle), glycolysis (phosphoglycerate kinase) and the citric acid cycle (isocitrate dehydrogenase).

Given the prodigious rate of calcification exhibited by *E. huxleyi* strains such as NZEH and CCMP371, some of the abundant proteins that we have robustly identified could also have roles in the currently uncharacterized carbon mineralisation process. For example, the regulation of carbonic anhydrase (identified for strain NZEH), which is a key enzyme in the carbon concentrating mechanism of phytoplankton and catalyzes the interconversion of CO_2 and water with bicarbonate and protons (Raven and Beardall 2003), has also been linked to biomineralisation in coccolithophores (Isenberg et al. 2007). Further examples include the ATP synthase (ATPase) subunits, in particular the vacuolar ATPase (V-ATPase) units that were detected for all three strains. Calcium-stimulated V-ATPase activity has been demonstrated on calcifying vesicle membranes of the related species *Pleurochrysis carterae* (Araki and Gonzalez 1998) and has been suggested to maintain an alkali pH consistent with $CaCO_3$ formation within these compartments (Corstjens et al. 2001). It is plausible that some of the ATPase subunits identified within this study form part of a complex within *E. huxleyi* membranes with a similar function. The presence of clathrin heavy chain may also have significance since clathrin-coated vesicles have been reported to contain V-ATPases (Forgac 2000). Clathrin is a major component of coated vesicles (Pearse 1976; Crowther and Pearse 1981) which transport proteins and lipids in eukaryote cells (Kirchhausen 2000) and participate in many membrane trafficking pathways, including those operating from the *trans*-Golgi network (Lewin and Mellman 1998; Kirchhausen 2000). Current evidence suggests that coccolith formation (coccolithogenesis) in *E. huxleyi* occurs in coccolith vesicles that derive from the Golgi (Westbroek et al. 1989; de Vrind-de Jong and de Vrind 1997; Young and Henriksen 2003). Since clathrin-coated vesicles contain V-ATPases (Forgac 2000), it is plausible that some of these components could play a role in the alkalinisation of a clathrin-coated structure for coccolithogenesis. Interestingly, a proteomics-based study on the diatom *T. pseudonana* also found that clathrin was highly abundant (Nunn et al. 2009). Since diatoms are encased by biogenically precipitated silica, it was proposed that clathrin could be a component of silica deposition vesicles (Nunn et al. 2009).

Validation of Previously Hypothetical Components

The proteomic data obtained in this study both validate the limited genomic data available for *E. huxleyi* and reveal some interesting facets of its biology. For example, the analysis confirmed the expression of a modular polyketide synthase type I (PKS) protein which was recently predicted from unannotated trace reads of the *E. huxleyi* genome sequence (John et al. 2008). Polyketides are a diverse group of natural products that are extremely important from a medicinal perspective, possessing various antibiotic, antifungal, anti-cancer and immunosuppressive properties (Staunton and Weissman 2001). Until recently, it was thought that type I PKS proteins were found exclusively within bacteria or fungi. However, it has now been reported that some protists such as dinoflagellates (Snyder et al. 2003, 2005) and the parasite *Cryptosporidium parvum* also possess type I PKS (Zhu et al. 2002), although phylogenetic analysis indicates that type I PKS presence within protist groups is sparse and patchy (John et al. 2008). Since *E. huxleyi* completely dominates the vast blooms that it forms (Holligan et al. 1983), it is possible that polyketides could be used to minimize grazing pressure or gain an advantage over other coccolithophore, diatom and phytoplankton species.

Whilst DNA sequencing technology has made enormous progress, fully sequencing and annotating genomes still requires a considerable investment of resources and expertise. This is particularly the case for a eukaryote genome and it follows that the vast majority of organisms on the planet will have completely unknown or only partially characterized genome sequences for the foreseeable future. This lack of genomic information significantly complicates proteomic analyses, which typically rely heavily on such data for the robust identification of proteins. Generating EST libraries from transcribed genes is currently considerably more cost-effective than full genome sequencing and annotation. However, the accurate assembly and annotation of ESTs from phylogenetically isolated organisms such as *E. huxleyi* presents challenges because of the lack of homologous sequence data. It is also difficult to distinguish sequencing errors from polymorphisms; when ESTs are obtained from a publically available database, sequence quality is also generally unknown.

Despite these problems associated with ESTs, we have demonstrated that it is possible to obtain robust proteomic data in the absence of extensive genomic sequence data. For this reason, the custom-designed pipeline BUDAPEST is therefore likely to be applicable to many studies involving marine organisms. Since protein extraction methods developed are versatile, future studies combining these approaches with quantitative proteomic methods will increase our understanding of responses to environmental change and the biochemistry of *E. huxleyi* with respect to calcification and other pathways of global biogeochemical significance.

Acknowledgements BMJ was supported by a Natural Environment Research Council (NERC) PhD studentship (NER/S/A2006/14211). This work is a contribution to the "European Project on Ocean Acidification" (EPOCA) which received funding from the European Community's Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 211384. We gratefully acknowledge the technical support of Dr John Gittins, Miss Shanon Pead and Mrs Therese Cavassa. We would also like to thank two anonymous reviewers for their comments on this manuscript.

References

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl Acid Res 25:3389–3402

Araki Y, Gonzalez EL (1998) V- and P-type Ca^{2+}-stimulated ATPases in a calcifying strain of *Pleurochrysis* sp. (Haptophyceae). J Phycol 34:79–88

Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M, Brzezinski MA, Chaal BK, Chiovitti A, Davis AK, Demarest MS, Detter JC, Glavina T, Goodstain D, Hadi MZ, Hellstein U, Hildebrand M, Jenkins BD, Jurka J, Kapitonov VV, Kroger N, Lau WWY, Lane TW, Larimer FW, Lippmeier JC, Lucas S, Medina M, Montsant A, Obornik M, Parker MS, Palenik B, Pazour GJ, Richardson PM, Rynearson TA, Saito MA, Schwartz DC, Thamatrakoln K, Valentin K, Vardi A, Wilkerson FP, Rokhsar DS (2004) The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. Science 306:79–86

Balch WM, Holligan PM, Ackleson SG, Voss KJ (1991) Biological and optical properties of mesoscale coccolithophore blooms in the Gulf of Maine. Limnol Oceanogr 36:629–643

Barrell D, Dimmer E, Huntley RP, Binns D, O'donovan C, Apweiler R (2009) The GOA database in 2009—an integrated gene ontology annotation resource. Nucl Acids Res 37:D396–D403

Baumann K-H, Böckel B, Frenz M (2004) Coccolith contribution to South Atlantic carbonate sedimentation. In: Thierstein HR, Young JR (eds) Coccolithophores: from molecular processes to global impact. Springer, Berlin

Brown CW, Yoder JA (1994) Coccolithophore blooms in the global ocean. J Geophys Res 99:7467–7482

Contreras L, Ritter A, Dennett G, Boehmwald F, Guitton N, Pineau C, Moenne A, Potin P, Correa JA (2008) Two-dimensional gel electrophoresis analysis of brown algal protein extracts. J Phycol 44:1315–1321

Cooper B, Neelam A, Campbell KB, Lee J, Liu G, Garrett WM, Scheffler B, Tucker ML (2007) Protein accumulation in the germinating *Uromyces appendiculatus* uredospore. Mol Plant–Microb Interact 20:857–866

Corstjens PLAM, Araki Y, Gonzalez EL (2001) A coccolithophorid calcifying vesicle with a vacuolar-type ATPase proton pump: cloning and immunolocalization of the V_o subunit c^1. J Phycol 37:71–78

Crowther R, Pearse B (1981) Assembly and packing of clathrin into coats. J Cell Biol 91:790–797

De Vrind-De Jong EW, De Vrind JPM (1997) Algal deposition of carbonates and silicates. Rev Mineral 35:267–307

Dyhrman ST, Haley ST, Birkeland SR, Wurch LL, Cipriano MJ, Mcarthur AG (2006) Long serial analysis of gene expression for gene discovery and transcriptome profiling in the widepread marine coccolithophore *Emiliania huxleyi*. Appl Environ Microbiol 72:252–260

Edwards RJ, Moran N, Devocelle M, Kiernan A, Meade G, Signac W, Foy M, Park SDE, Dunne E, Kenny D, Shields DC (2007) Bioinformatic discovery of novel bioactive peptides. Nat Chem Biol 3:108–112

Field CB, Behrenfield MJ, Randerson JT, Falkowski P (1998) Primary production of the biosphere: integrating terrestrial and oceanic components. Science 281:237–240

Forgac M (2000) Structure, mechanism and regulation of the clathrin-coated vesicle and yeast vacuolar H(+)-ATPases. J Exp Biol 203:71–80

Guillard RRL (1975) Culture of phytoplankton for feeding marine invertebrates. In: Smith WL, Chanley MH (eds) Culture of marine invertebrate animals. Plenum, New York

Guillard RRL, Ryther JH (1962) Studies of marine planktonic diatoms. I. *Cyclotella nana* Hustedt and *Detonula confervacea* Cleve. Can J Microbiol 8:229–239

Holligan PM, Viollier M, Harbour DS, Camus P, Champagne-Philippe M (1983) Satellite and ship studies of coccolithophore production along a continental shelf edge. Nature 304:339–342

Holligan PM, Fernandez E, Aiken J, Burkill PH, Finch M, Groom SB, Malin G, Muller K, Purdie DA, Robinson C, Trees CC, Turner SM, Van Der Wal P (1993) A biogeochemical study of the coccolithophore, *Emiliania huxleyi*, in the North Atlantic. Glob Biogeochem Cycles 7:879–900

Iglesias-Rodriguez MD, Brown CW, Doney SC, Kleypas J, Kolber D, Kolber Z, Hayes PK, Falkowski PG (2002) Representing key phytoplankton functional groups in ocean carbon cycle models: coccolithophorids. Glob Biogeochem Cycles 16:1100

Isenberg HD, Lavine LS, Weissfellner H (2007) The suppression of mineralization in a coccolithophorid by an inhibitor of carbonic anhydrase. J Eukaryot Microbiol 10:477–479

John U, Beszteri B, Derella E, Van De Peer Y, Read BR, Moreau H, Cembella A (2008) Novel insights into evolution of protistan polyketide synthases through phylogenomic analysis. Protist 159:21–30

Jordan RW, Green JC (1994) A checklist of the extant haptophyta of the world. J Mar Biol Assoc UK 74:149–174

Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. Brief Bioinform 9:286–298

Kirchhausen T (2000) Clathrin. Ann Rev Biochem 69:699–727

Lewin DA, Mellman I (1998) Sorting out adaptors. Biochim Biophys Acta 1401:129–145

Linschooten C, Bleijswijk JDL, Emburg PR, Vrind JPM, Kempers ES, Westbroek P, Jong EWV-D (1991) Role of the light–dark cycle and medium composition on the production of coccoliths by *Emiliania huxleyi* (Haptophyceae). J Phycol 27:82–86

Martens L, Hermjakob H, Jones P, Adamski M, Taylor C, States D, Gevaert K, Vandekerckhove J, Apweiler R (2005) PRIDE: the proteomics identifications database. Proteomics 5:3537–3545

Milliman JD (1993) Production and accumulation of calcium carbonate in the ocean: budget of a nonsteady state. Glob Biogeochem Cycles 7:927–957

Nagai K, Yotsukura N, Ikegami H, Kimura H, Morimoto K (2008) Protein extraction for 2-DE from the lamina of *Ecklonia kurome* (laminariales): recalcitrant tissue containing high levels of viscous polysaccharides. Electrophoresis 29:672–681

Nguyen B, Bowers RM, Wahlund TM, Read BR (2005) Suppressive subtractive hybridization of and differences in gene expression content of calcifying and noncalcifying cultures of *Emiliania huxleyi* strain 1516. Appl Environ Microbiol 71:2564–2575

Nunn BL, Aker JR, Shaffer SA, Tsai S, Strzepek RF, Boyd PW, Freeman TL, Brittnacher M, Malmstrom L, Goodlett DR (2009) Deciphering diatom biochemical pathways via whole-cell proteomics. Aquat Microb Ecol 55:241–253

Paasche E (2001) A review of the coccolithophorid *Emiliania huxleyi* (Prymnesiophyceae), with particular reference to growth, coccolith formation and calcification–photosynthesis interactions. Phycologica 40:503–529

Pearse B (1976) Clathrin: a unique protein associated with intracellular transfer of membrane by coated vesicles. Proc Natl Acad Sci USA 73:1255–1259

Price MN, Dehal PS, Arkin AP (2009) FastTree: computing large minimum evolution trees with profiles instead of a distance matrix. Mol Biol Evol 26:1641–1650

Quinn P, Bowers RM, Zhang X, Wahlund TM, Fanelli MA, Olszova D, Read BR (2006) cDNA microarrays as a tool for identification of biomineralization proteins in the coccolithophorid *Emiliania huxleyi* (Haptophyta). Appl Environ Microbiol 72:5512–5526

Raven JA, Beardall J (2003) CO_2 acquisition mechanisms in algae: carbon dioxide diffusion and carbon dioxide concentrating mechanisms. In: Larkum A, Raven JA, Douglas S (eds) Photosynthesis in the Algae. Kluwer, Dordrecht

Riegman R, Stolte W, Noordeloos AAM, Slezak D (2000) Nutrient uptake and alkaline phosphatase (EC 3:1:3:1) activity of *Emiliania huxleyi* (Prymnesiophyeae) during growth under N and P limitation in continuous cultures. J Phycol 36:87–96

Robertson JE, Robinson C, Turner DR, Holligan P, Watson AJ, Boyd P, Fernandez E, Finch M (1994) The impact of a coccolithophore bloom on oceanic carbon uptake in the northeast Atlantic during summer 1991. Deep Sea Res I 41:297–314

Shevchenko A, Jensen ON, Podtelejnikov AV, Sagliocco F, Wilm M, Vorm O, Mortensen P, Shevchenko A, Boucherie H, Mann M (1996) Linking genome and proteome by mass spectrometry: large-scale identification of yeast proteins from two dimensional gels. Proc Natl Acad Sci USA 93:14440–14445

Skipp P, Robinson J, O'Connor CD, Clarke IN (2005) Shotgun proteomic analysis of *Chlamydia trachomatis*. Proteomics 5:1558–1573

Snyder RV, Gibbs PDL, Palacios A, Abiy L, Dickey R, Lopez JV, Rein KS (2003) Polyketide synthase genes from marine dinoflagellates. Mar Biotechnol 5:1–12

Snyder RV, Guerrero MA, Sinigalliano CD, Winshell J, Perez R, Lopez JV, Rein KS (2005) Localization of polyketide synthase encoding genes to the toxic dinoflagellate *Karenia brevis*. Phytochemistry 66:1767–1780

Staunton J, Weissman K (2001) Polyketide biosynthesis: a millennium review. Nat Prod Rep 18:380–416

Tyrrell T, Merico A (2004) *Emiliania huxleyi*: bloom observations and the conditions that induce them. In: Thierstein HR, Young JR (eds) Coccolithophores: from molecular processes to global impact. Springer, Berlin

Uniprot (2008) The universal protein resource (UniProt). Nucleic Acids Res 36:D190–D195

Wahlund TM, Hadaegh AR, Clark R, Nguyen B, Fanelli M, Read BR (2004a) Analysis of expressed sequence tags from calcifying cells of marine coccolithophorid (*Emiliania huxleyi*). Mar Biotechnol 6:278–290

Wahlund TM, Zhang X, Read BA (2004b) Expressed sequence tag profiles from calcifying and non-calcifying cultures of *Emiliania huxleyi*. Micropaleontology 50:145–155

Wang D-Z, Lin L, Hong H-S (2009) Comparative studies of four protein preparation methods for proteomic study of the dinoflagellate *Alexandrium* sp. using two-dimensional electrophoresis. HarmAlgae 8:685–691

Westbroek P, Young J, Linschooten K (1989) Coccolith production (biomineralization) in the marine alga *Emiliania huxleyi*. J Protozool 36:368–373

Westbroek P, Brown CW, Van Bleijswijk J, Brownlee C, Brummer GJ, Conte M, Egge J, Fernandez E, Jordan R, Knappertsbusch M, Stefels J, Veldhuis M, Van Der Wal P, Young J (1993) A model system approach to biological climate forcing. The example of *Emiliania huxleyi*. Global Planet Change 8:27–46

Wong P-F, Tan L-J, Nawi H, Abubakar S (2006) Proteomics of the red alga, *Gracilaria changii* (Gracilariales, Rhodophyta). J Phycol 42:113–120

Young JR, Henriksen K (2003) Biomineralization within vesicles: the calcite of coccoliths. Rev Mineral Geochem 54:189–215

Zhu G, Lagier MJ, Stejskal F, Millership JJ, Cai X, Keithly JS (2002) *Cryptosporidium parvum*: the first protist known to encode a putative polyketide synthase. Gene 298:79–89

All programmes, datasets and result files are available from: http://www.personal.soton.ac.uk/re1u06/research/ehux/.