

Expressed sequence tag profiles from calcifying and non-calcifying cultures of *Emiliana huxleyi*

Thomas M. Wahlund¹, Xiaoyu Zhang², and Betsy A. Read¹

¹Department of Biological Sciences, California State University San Marcos,
333 S. Twin Oaks Valley Road, San Marcos, California 92096-0001

²Department of Computer Science, California State University San Marcos,
333 S. Twin Oaks Valley Road, San Marcos, California 92096-0001

email: bread@csusm.edu

ABSTRACT: Expressed Sequence Tag (EST) analysis is a powerful means for evaluating gene expression and discovering novel genes. The method is relatively simple and particularly important for species where knowledge of the genome is unavailable or limited. In this study, we compare EST profiles of *Emiliana huxleyi* cultures grown under conditions that promote biomineralization and coccolithogenesis (F/50 media) and conditions that reduce these processes (F/2 media). A total of 3,527 clones from the F/50 and 4,116 clones from F/2 library were randomly selected and sequenced, resulting in a unigene set of 4,057 unique sequences comprised of 540 contigs and 3517 singletons. More than 40% of the expressed genes showed similarity to known genes from other organisms deposited in GenBank ($e > 10^{-2}$) and have been functionally categorized. The comparative data provide cues for further studies pertaining to genes and proteins involved in biomineralization and coccolithogenesis, and taken together serve as a valuable resource for the coccolithophorid research community.

INTRODUCTION

Coccolithophorids, the most prominent haptophyta, emerged late in the Triassic era and radiated extensively throughout the Jurassic and Cretaceous periods according to their exceptional fossil record (Young et al. 1999). At the Cretaceous-Tertiary boundary, the coccolithophorids suffered a massive extinction, with the disappearance of two-thirds of the 50 genera. Many new coccolithophorid groups, however, appeared in the Paleocene. In contemporary oceans coccolithophorids remain as one of the major eukaryotic phytoplankton, constituting 15% of the phytoplankton biomass (Berger 1982). The distinctive calcite scales (coccoliths) that blanket the cell and form the elaborate coccosphere reflect visible light from the surface water and enable coccolithophorid blooms of today to be detected by satellite imagery. Massive blooms of 100,000 km² have been detected in this manner (Brown and Yoder 1994; Holligan et al. 1993).

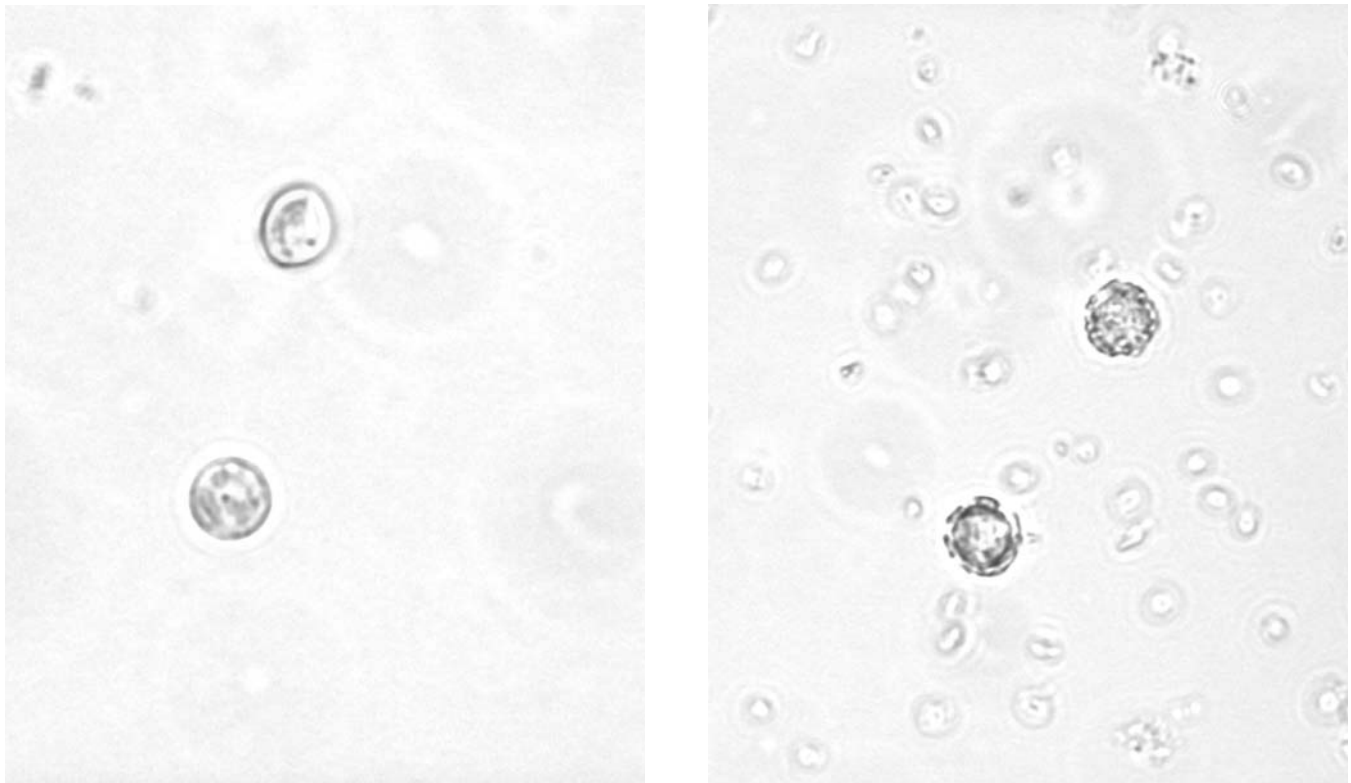
Because of their widespread distribution and bloom densities of up to 1×10^6 cells/ml, coccolithophorids are thought to play an important role in the biogeochemical cycling of carbon and sulfur. They may also influence global climate change (Westbroek et al. 1993). The reflectance capabilities of the coccoliths and the ability of the alga to produce dimethylsulfide (DMS) influences local albedo and may impact regional weather patterns (Bates et al. 1987; Charlson et al. 1987). Atmospheric CO₂ is fixed by coccolithophorids into both photosynthetic and biomineralized product, and for this reason biogeochemists are interested in the haptophytes potential to as serve as a carbon sink. It is estimated that of the total pelagic calcite deposited in the ocean sediments, as much as 60% is contributed by the coccolithophorids (Honjo 1996).

While coccoliths provide a highly characteristic sedimentary archive of the coccolithophorid evolution, the long-chain alke-

nones and alkyl alkenoates that some of these algae synthesize provide important information regarding paleoclimatology. The unsaturation ratio of the C₃₇ alkenones has been shown to be indicative of organismal growth temperatures, suggesting that the ratio measured in the deep-sea sediments reflects the overlying sea surface temperature (Brassel et al. 1986b; Brassel et al. 1986a). Since diagenetic processes do not appear to affect the original unsaturation ratio of the C₃₇ alkenones, these long-chain polyunsaturated ketones can be used as paleotemperature proxies (Conte et al. 1992; Prah et al. 1988; Sikes et al. 1991).

Coccolithophorids have long since captured the attention of oceanographers, biogeochemists, ecophysicologists and paleontologists, but it is only recently that they have attracted the attention of molecular biologists. As a consequence our knowledge of the molecular genetics of this important alga is extremely limited. We know little about the genome size, structure, and/or gene content across the different species of coccolithophorids. This information could be extremely useful in helping researchers to understand the historical origins and environmental conditions that led to the selection and radiation of the coccolithophorids, as well as the ecological and genetic processes that contribute to their continued success in contemporary oceans.

Emiliana huxleyi is an attractive model species for studying the biology of marine coccolithophorids because of its cosmopolitan distribution and ease of laboratory culturing in the laboratory. In addition, *E. huxleyi* strains are genotypically and phenotypically diverse with respect to coccolith morphology, yet no genes involved in coccolith formation have been identified and no studies comparing gene expression patterns have been reported to date. The genome of *E. huxleyi* strain 1516 is presently being sequenced in a collaborative effort between our laboratory and the U.S. Department of Energy. The sequence



TEXT-FIGURE 1

Light microscopy images of *E. huxleyi* CCMP 1516 grown in (A) phosphate-replete (f/2) and (B) phosphate-limited (f/50) media. Note the relative absence of coccoliths on f/2 versus f/50-grown cells. Magnifications, ca. $\times 1000$.

data will provide researchers with an invaluable resource that can be used for rapid analysis of gene expression patterns using microarrays and proteomics technologies, and thus address unresolved questions regarding the ecology and biology of this fascinating marine phytoplankton.

Expressed Sequence Tags (ESTs) are generated by sequencing randomly selected clones from a cDNA library. Clones are sequenced from either the 5' or 3'-end and result in a partial cDNA sequence that is not monitored for sequencing errors or artifacts. ESTs are useful for functional comparisons of expressed gene populations and are a powerful tool for gene discovery, particularly when knowledge about the genome under investigation is limited. Catalogs of ESTs are also a valuable resource for the development of molecular tools and analysis methods. Sequences can be used for microarray analysis, molecular systematics, vector construction or primer design. In addition, new and ancient metabolic pathways may be traced by means of gene sequences (Bhattacharya et al. 1992; Des Marais 2000; Ferris et al. 1997; Kumada et al. 1993; So et al. 2004). Our initial EST analysis involved the partial sequencing of 3527 cDNA clones from *E. huxleyi* cultures grown in phosphate limiting media, conditions known to promote biomineralization and coccolithogenesis (Wahlund et al. 2004). In an effort to expand our EST resource for coccolithophorid gene discovery and to better understand processes related to the molecular events involved in biomineralization and coccolithogenesis, we have generated 4116 additional ESTs from cells grown in nutrient rich media, conditions that dramatically inhibit biomineralization and coccolithogenesis in *E. huxleyi* CCMP 1516. Herein, we describe the results of our analysis and compare the

gene expression profiles of *E. huxleyi* cells under calcifying and non-calcifying conditions. Compositional properties of the *E. huxleyi* genome are also discussed.

MATERIALS AND METHODS

Strains and Growth Conditions

E. huxleyi strain 1516 was obtained from the Provasoli-Guillard National Center for Culture of Marine Phytoplankton and cultured in f/2 (phosphate-replete; ca. 33 mM) or f/50 (phosphate-limited; ca. 1.3 mM) artificial sea water as described previously (Guillard 1975; Laguna et al. 2001). Batch cultures of cells were obtained by inoculating cells into 1 L of media in 4 L flasks and incubated at 17 to 18°C under cool white fluorescent light ($660 \mu\text{mol} \cdot \text{m}^{-2} \cdot \text{s}^{-2}$) under a 12-h dark, 12-h light cycle. Incubation of *E. huxleyi* CCMP 1516 in phosphate-limited (f/50) media served to promote calcification, whereas cells grown in phosphate-replete (f/2) were dramatically inhibited in their ability to calcify (text-fig. 1). Cultures of *E. huxleyi* CCMP1516 grown in f/2 media reached final cell densities of ca. 5×10^6 cells/ml, which is 2- to 3-fold higher than those obtained in f/50 media.

Library Construction

The *E. huxleyi* 1516 F/50 library was constructed as previously described (Wahlund et al. 2004). For the *E. huxleyi* 1516 f/2 library total RNA was extracted from cultures in late log phase, after 8 days of growth in phosphate replete F/2 medium, using a standard guanidinium isothiocyanate procedure (Strommer et al. 1993). The library was constructed by ResGen (Invitrogen Corp, Carlsbad, CA). First strand synthesis was performed us-

TABLE 1
Comparative analysis of *E. huxleyi* CCMP 1516 f/2 and f/50 cDNA libraries.

Library Type	Library Titer ^c	Insert Size ^d	Clones Sequenced	Contigs	Singletons	Unigenes
Phosphate-limited ^a	3 x 10 ⁷	2300	3527	566	1048	1685
Phosphate-replete ^b	6 x 10 ⁵	1.4	4116	745	1859	2841
Unigene				540	3517	4057

^a represents *E. huxleyi* 1516 cells grown in f/50 media (ca. 1.3 μM phosphate).

^b represents *E. huxleyi* 1516 cells grown in f/2 media (ca. 33 μM phosphate).

^c represents estimated number of cDNA clones in each library,

^d average length of cDNA inserts (base pairs).

ing a *Not I* primer-adaptor [GAC TAG TTC TAG ATC GCG AGC GGC CGC CC(T)₁₅] and Superscript II reverse transcriptase. Following second strand synthesis using *E. coli* DNA polymerase, *Not I*/blunt end products were directionally cloned into the *Not I*/*EcoRV* sites of the Gateway cloning vector, pCMV-SPORT 6.1. Plasmids were used to transform ElectroMax DH10B-TON A cells via electroporation, and random clones were picked for quality control analysis.

Sequencing

Sequencing templates were prepared from randomly picked cDNA clones using the Phi29 DNA Polymerase based rolling circle amplification method. Single pass sequencing was performed by Windsor Pond Associates (Chicago, IL) using the dideoxy chain termination, and dye termination chemistry (Applied Biosystems). Analysis was completed on an ABI 3100 fluorescence automated sequencer. Preparation and sequencing of clones from the f/50 library were performed as previously described (Wahlund et al. 2004).

Data handling and analysis

Expressed sequence tags from both libraries were edited and assembled using the programs PHRED, CROSS_MATCH, and CONSED from PHRAP (P. Green, <http://bozeman.mbt.washington.edu/phrp.html>). Contaminating vector sequences were removed and ESTs having a PHRED quality score of at least 20 and a minimum size of 300 nucleotides were taken for further analysis. ESTs from both libraries were grouped into classes of identical overlapping sequences (contigs) using PHRAP and were viewed to assess the accuracy of assembly using CONSED. For the assembly of contigs PHRAP parameters were set at a minimum match of 20 with a default score of 30. CROSS_MATCH parameters included a minimum match of 10 and score of 20. A set of non-redundant ESTs was created by collecting all of the ESTs from both libraries in a single set, and performing the contig assembly again.

To identify potential homologues to the *E. huxleyi* genes, homology searches were performed against the non-redundant protein database in GenBank (National Center for Biotechnology Information) using BLASTX with the BLOSUM62 matrix. For all other parameters such as filtering and word size, BLASTX default settings were employed. Similarities to known proteins were considered significant when database searches returned matches with *e*-values less than 1.0 x 10⁻². On the basis of these results, ESTs were functionally categorized according to putative cellular function as described previously (Adams et al. 1995).

RESULTS

A directional cDNA library was constructed from RNA extracted from *E. huxleyi* cells grown in phosphate-replete artificial seawater. The primary titer of the f/2 (phosphate replete) and f/50 (phosphate-limited) libraries, and their respective average cDNA insert sizes, are shown in table 1. A total of 4,116 clones from the library based on *E. huxleyi* cells grown in phosphate-replete artificial seawater were randomly chosen for PCR amplification of plasmid inserts and partial DNA sequencing. Robust DNA sequences averaging 959 bases were obtained, and when assembled, represented 2841 unique genes, 745 contigs and 1859 singletons (table 1). A significant similarity to the amino acid sequence of known proteins deposited in GenBank was obtained for 41% of the unigene sequences. General characteristics of the *E. huxleyi* 1516 f/2 cDNA library are presented in tables 1 and 2 and are compared to that of our previously-described *E. huxleyi* 1516 f/50 (phosphate-limited) cDNA library (Wahlund et al. 2004).

Cultures of *E. huxleyi* CCMP 1516 from which the libraries were constructed were determined to be unialgal, but not free from prokaryotic contamination. Routine checks for significant bacterial and eukaryotic contaminants was performed by plating on agar media as described previously (Laguna et al. 2001), and by rDNA PCR amplification using "universal" bacterial 16S primers (which would also amplify *E. huxleyi* chloroplast rDNA), and eukaryotic-specific 18S primers (Moon-van der Staay et al. 2001), which would not amplify bacterial controls. Each of these primer sets amplified only *E. huxleyi* rDNA sequences from liquid-grown cultures (as determined using BLASTn analyses). Cultures grown in f/50 media yielded yellow pigmented colonies when plated on solid media containing an organic carbon source (e.g. f/50 plus peptone, yeast extract, casamino acids) after 14 to 21 days incubation when incubated in the light or the dark, and 16S rDNA analysis identified these cells as *Erythrobacter/Roseobacter* species. These marine, aerobic anoxygenic phototrophic bacteria can be isolated by enrichment under photoheterotrophic conditions but grow very poorly photoautotrophically (CO₂ as sole carbon source) (Allgaier et al. 2003; Francis et al. 2001). In addition, cultures of *E. huxleyi* CCMP 1516 used for library constructions were grown in f/2 or f/50 media supplemented with low levels of kanamycin (25 mg/ml), which inhibited growth of these prokaryotic contaminants. Taken together with the data presented herein on the cDNA libraries of *E. huxleyi*, the authors feel that the impact of prokaryotic contributions to the EST data described in this study were negligible.

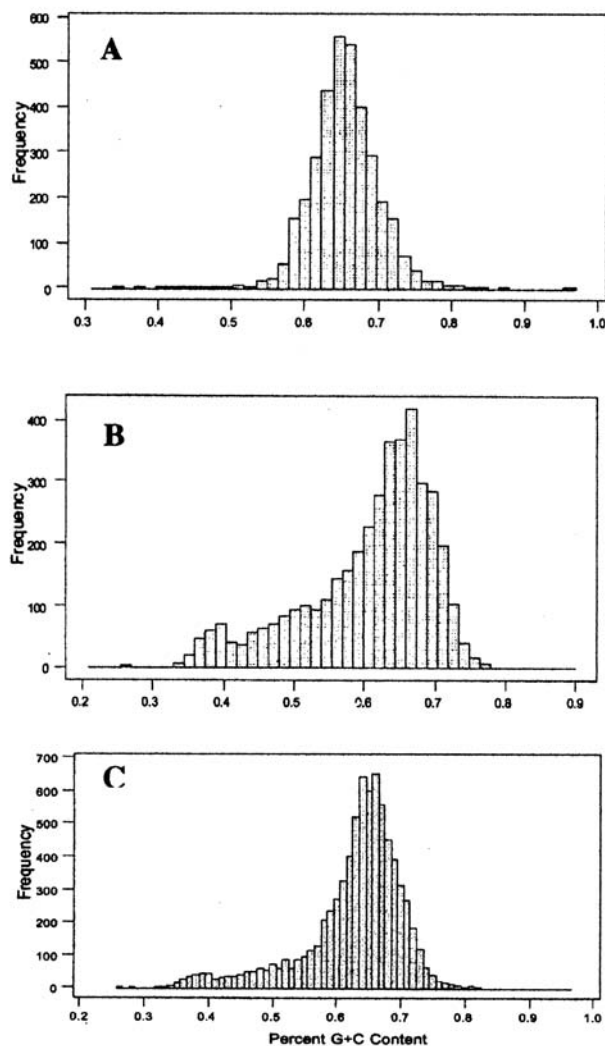
TABLE 2
BLAST search results from *E. huxleyi* cDNA libraries.

Descriptive Category	No. of ESTs
<i>P-limiting (E. huxleyi 1516, f/50)^a</i>	
ESTs matches w/ e -value $\geq 1 \times 10^{-2}$	947 (56%)
ESTs matches w/ e -value $< 1 \times 10^{-2}$	670 (40%)
ESTs w/ no GenBank match	68 (4%)
<i>Nutrient-replete (E. huxleyi 1516, f/2)^b</i>	
ESTs matches w/ e -value $\geq 1 \times 10^{-2}$	1382 (49%)
ESTs matches w/ e -value $< 1 \times 10^{-2}$	1171 (41%)
ESTs w/ no GenBank match	288 (10%)

The average G+C content of the ESTs from the *E. huxleyi* 1516 f/2 library was 60%, while the ESTs from the f/50 library clustered around a mean of 65% G+C (text-fig. 2). The histograms of G+C percentages in both libraries display distinct, but overlapping leptokurtic distributions. The *E. huxleyi* f/2 library is thought to contain a higher number of mitochondrial and chloroplast encoded transcripts, known to have a higher A+T content. The mitochondrial genome sequence for *E. huxleyi* has recently been completed and has an overall G+C content of 28% (pers.comm.). Also contained within the library are a number of viral transcripts that were assumed to be from putative culture contaminants. Fifteen to twenty of these ESTs representing 5 different gene sequences exhibited an average G+C content of 40% and very strong BLASTX homologies (e -values ranging from 0 to 1.79×10^{-119}) to Bacteriophage A511. The contaminating viral sequences together with the large number of chloroplast and mitochondrial-encoded sequences present in the f/2 library and not in the f/50 library, may in part explain the preponderance of ESTs in the with 35-50% G+C content. Although not described herein, the G+C content of the ESTs in the *E. huxleyi* f/2 library (excluding the mitochondrial, chloroplast and viral sequences) is reflective of a distinct codon bias previously described (Wahlund et al. 2004)

Next, we classified unique genes from each library by functional categories: (1) ribosomal proteins, (2) cell division, (3) DNA replication/RNA transcription, (4) cell signaling, (5) cell structure, (6) cell defense, (7) metabolism, (8) other matches, and (9) hypothetical proteins. The distribution of putatively-identified unigenes across functional categories is shown in text-figure 3A. The total number of unigenes for the f/2 and f/50 libraries from *E. huxleyi* 1516 that could be assigned a cellular role was 1459 and 728, respectively. The overall pattern of gene expression is very similar for the two different libraries with most of the ESTs ranked in the functional class "metabolism". A large number of unigenes in both libraries fell into the unclassified or hypothetical categories.

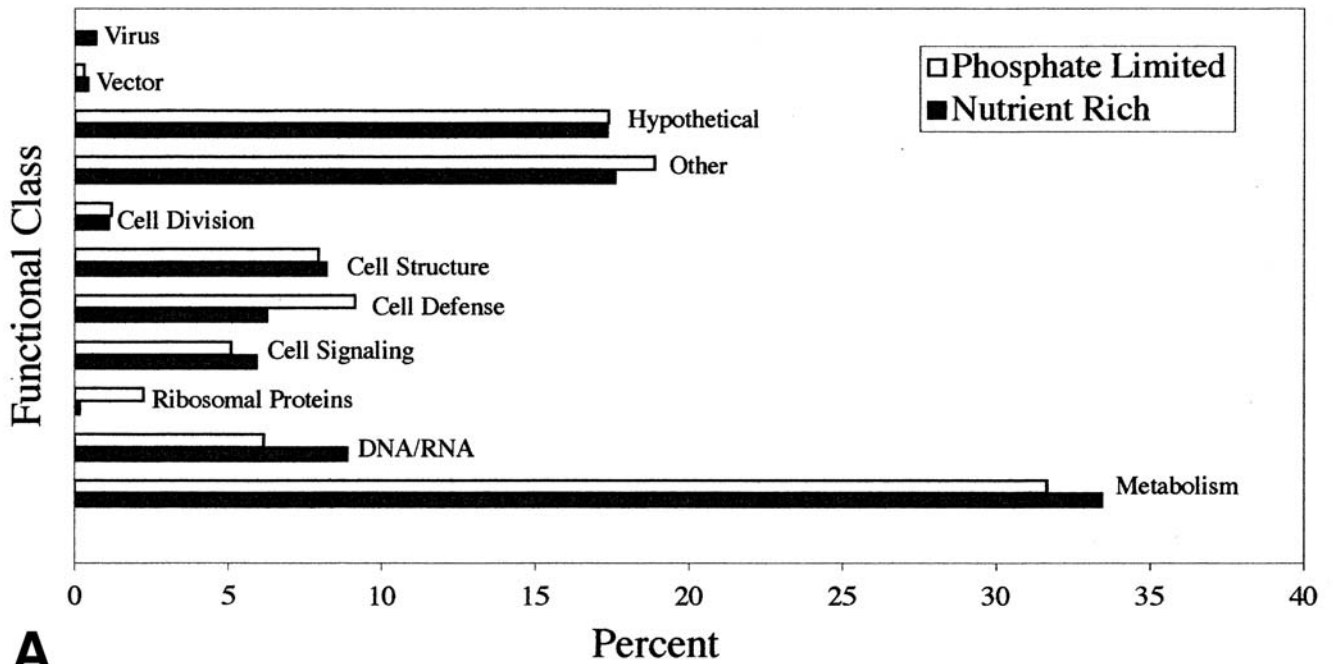
In order to reduce the number of transcripts needed to be analyzed and to provide a means for assessing the level of gene expression, ESTs from each of the two libraries were grouped into classes of identical overlapping sequences or "contigs". Contigs, formed from two or more sequences, tend to provide more definitive gene identification because of the longer reads. Analysis of what are, presumably, the more abundantly expressed genes among contigs also tend to reflect the principle metabolic pathways that were active in the cells at the time of RNA extraction. Text-figure 4 shows the distribution of gene sequences in contigs by functional class and includes a "no



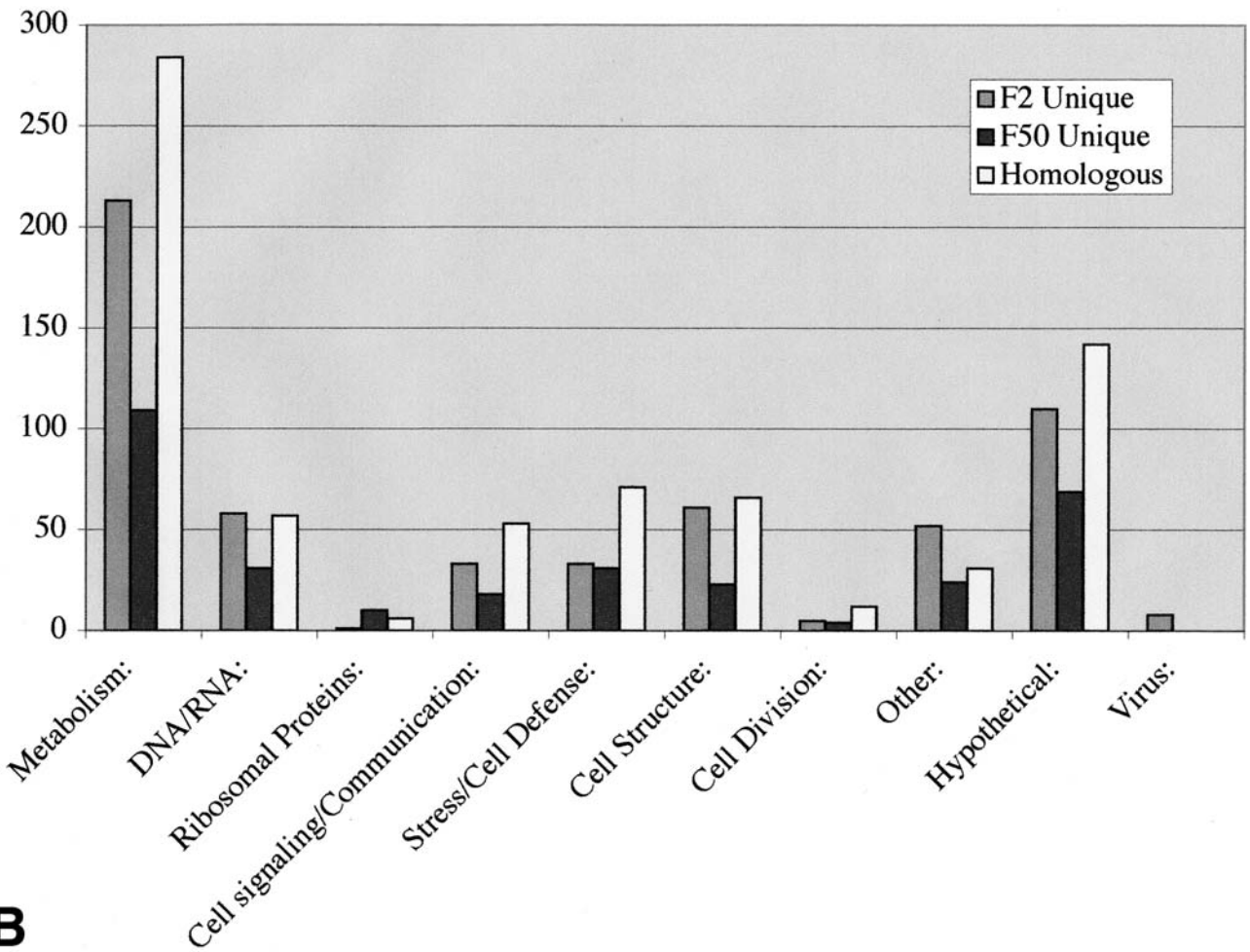
TEXT-FIGURE 2
Comparison of G+C content of ESTs sequenced from: (A) the f/50 library; (B) the f/2 library, and (C) a cumulative G+C content obtained from combining unigene sets from both libraries.

match" category representing sequences returning no significant GenBank matches (e -value $> 10^{-2}$). The 745 contigs derived from the f/2 library and 566 contigs from the f/50 library were used in the analysis. The most striking observation is the number of contigs in the f/50 library, when cells are grown under phosphate limiting conditions that promote biomineralization and coccolithogenesis, that did not return significant GenBank matches. Although these contigs may represent gene sequences that are expressed under phosphate stress, it is tempting to speculate on their potential involvement in processes of calcification and/or coccolithogenesis. The composition of the unigene set with respect to the number of homologous and unique transcripts contribute by each library across functional classes, is presented in text-fig. 3B.

Tables 3 and 4 list the most abundantly expressed transcripts, or the largest 50 contigs, in each of the two libraries. Of the top 50 most abundant ESTs in the f/2 library (table 4), 34 of the transcripts showed similarity to known proteins deposited in

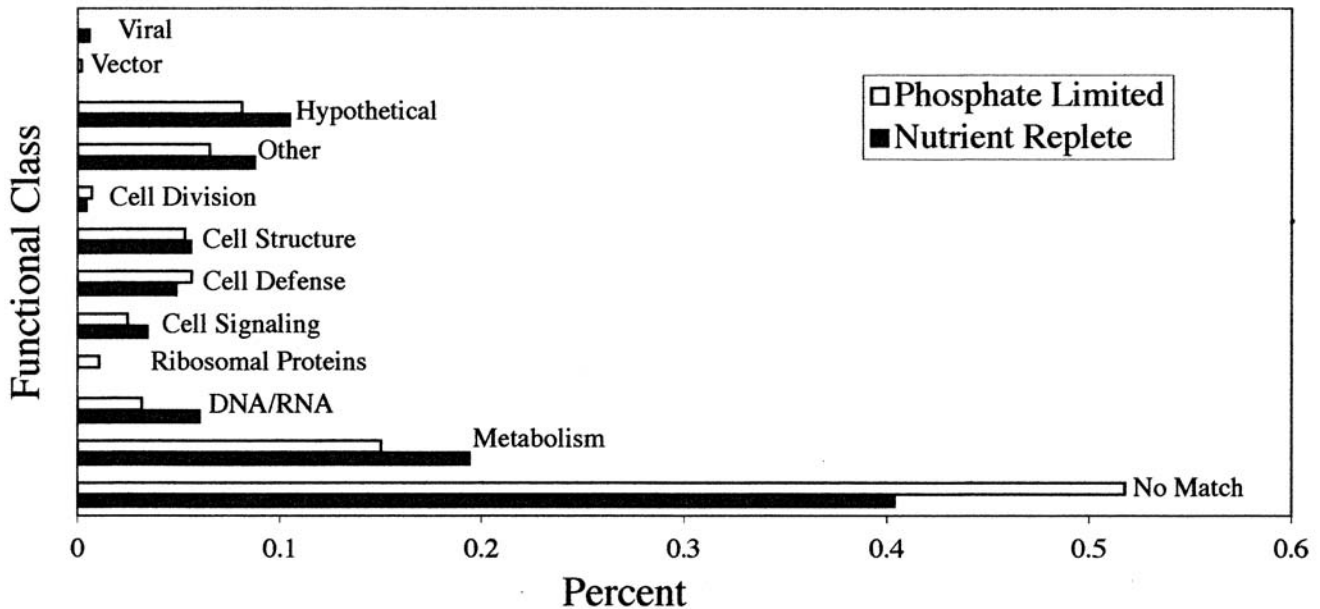


A



B

TEXT-FIGURE 3
 Distribution of *E. huxleyi* 1516 ESTs (A) and unigenes (B) from f/2 and f/50 libraries by functional categories. Putative functions were assigned for those ESTs identified by BLASTX with an e -value $\leq 10^{-2}$.



TEXT-FIGURE 4

Distribution of ESTs from *E. huxleyi* 1516 f/2 and f/50 libraries grouped into contigs. Putative functions were assigned for those ESTs identified by BLASTX with an e -value $\leq 10^{-2}$. A “no match” category represents sequences with no significant GenBank matches (e -value $> 10^{-2}$).

GenBank. In contrast, only 24 of the top 50 most abundantly expressed transcripts in the f/50 library showed similarity to known proteins in GenBank (table 3). Three putatively identified sequences that are among the most abundantly expressed transcripts in both libraries include phosphoenolpyruvate carboxykinase, acetyl-CoA synthetase, and actin. Comparisons using BLASTN (a nucleotide-nucleotide pairwise alignment program) revealed two other transcripts returning no significant GenBank matches as being listed in both libraries as one of the largest contigs. Contig 526 in the f/50 library corresponds to contig 735 in the f/2 library and in both libraries similar numbers of transcripts are present. Contig 562 in the f/50 library corresponds to contig 717 in the f/2 library and again similar numbers of transcripts are present in each library.

Several interesting differences in the transcript populations between the two libraries can be readily observed, one of which is the prevalence of genes involved in central and intermediate metabolism in f/2-grown cells versus f/50-grown cells (e.g., glyceraldehyde-3-phosphate DH, fructose-1,6-bisphosphate aldolase, phosphoglycerate kinase, acyl-CoA synthetase, and S-adenosyl methionine synthetase; table 4). The relative absence of these types of transcripts in f/50-grown cells suggests a down-regulation of these central and intermediate metabolic pathways under phosphate-limitation conditions. Several genes involved in DNA replication and RNA transcription are also among the most prevalent contigs in the f/2 library but not in the f/50 library, reinforcing the notion that cells grown in the nutrient replete media are growing and actively reproducing as compared to those grown in the phosphate limiting media. Also of note is the presence of the bacteriophage tail sheath protein amongst the most abundantly expressed transcripts in the f/2 library that, as discussed above, is presumed to represent a viral contaminant in the nutrient replete culture that is not present in the phosphate-limited culture.

In sharp contrast, among the most prevalent putatively identified transcripts in the *E. huxleyi* 1516 f/50 library are those associated with stress and cell defense. The sixth most abundantly sequenced clone from the f/50 library is a transcript that shows high sequence similarity to the ubiquitin B protein. Other transcripts that are present in multiple copies and known to be linked to stress responses include homologs to HSP70, cathepsin, 14-3-3 like protein, and a cysteine proteinase.

DISCUSSION

In spite of the marked morphological differences between calcifying and noncalcifying *E. huxleyi* cells, no major differences are apparent in the quantitative distribution of ESTs by functional class when comparing the two cDNA libraries (text-fig. 3A). The composition of the unigene set, however, suggests non-calcifying *E. huxleyi* cells express a more complex set of genes (text-fig. 3B). While calcifying and non-calcifying cells express many of the same presumably housekeeping genes, the non-calcifying cells express nearly twice as many library specific genes. This is true across nearly all functional categories with the exception of ribosomal, cell division, and cell defense related genes where the number of unique transcripts originating from each of the two libraries is approximately the same.

Trends indicative of differential gene expression between the two culture media conditions employed in this study are apparent. The abundance of various metabolic enzymes and proteins involved in DNA replication and transcription, amongst the most highly expressed genes under phosphate replete growth conditions, is one such example. In sharp contrast is the number and diversity of stress related proteins that are amongst the most highly expressed genes under the phosphate limited growth conditions. While a number of heat shock proteins were present in both libraries, HSP70 is one of the most abundant transcripts in the f/50 library, along with cathepsin, ubiquitin B, and a

TABLE 3
Contigs of most prevalent ESTs from f/50 library.

ID	Putative ID	e-value	Reads	Length
526	Unknown	>E-2	9	1032
527	CLAVATA1 receptor kinase-like protein [<i>O. sativa</i>]	3.01E-2	9	649
528	pep carboxykinase [<i>Toxoplasma gondii</i>]	9.77E-93	9	1090
529	Unknown	>E-2	10	610
530	Unknown	>E-2	10	673
531	Unknown	>E-2	10	327
532	Unknown	>E-2	10	601
533	cytoplasmic actin [<i>Lethenteron japonicum</i>]	3.5E-102	10	582
534	Unknown	>E-2	10	529
535	cysteine proteinase Cysp1 [<i>C. salmositica</i>]	1.12E-36	11	890
536	hypothetical protein [<i>Nostoc punctiforme</i>]	8.55E-06	11	1128
537	ENSANGP00000001657 [<i>Anopheles gambiae</i>]	1.83E-2	11	1823
538	hypothetical protein Rv1147 [<i>Mycobacterium tuberculosis H37Rv</i>]	2.65E-5	11	550
539	inositol 2-dehydrogenase-like protein [<i>A. thaliana</i>]	4.75E-13	11	430
540	actin [<i>Podocoryne carnea</i>]	3.3E-132	11	769
541	thiolase family protein [<i>C. crescentus</i>]	2.05E-46	11	585
542	Unknown	>E-2	11	1261
543	CALRETICULIN PRECURSOR (CRP55)	3.42E-42	11	814
544	Putative endoglucanase type K precursor	1.52E-08	12	559
545	acyl-CoA synthetase family [<i>A. thaliana</i>]	3.13E-38	12	1179
546	cathepsin Y [<i>Rattus norvegicus</i>]	5.65E-47	12	1102
547	Unknown	>E-2	12	638
548	ENSANGP00000019055 [<i>A. gambiae</i>]	3.19E-91	12	564
549	14-3-3-LIKE PROTEIN A	3.65E-40	12	577
550	Unknown	>E-2	12	501
551	ENSANGP00000017006 [<i>A. gambiae</i>]	4.77E-4	13	546
552	Unknown	>E-2	13	1010
553	alpha adrenergic receptor 2B [<i>Phoca vitulina</i>]	3.63E-2	13	507
554	putative minor tail protein - phage associated	1.79E-7	13	1527
555	ACTIN 3	7.1E-174	14	1026
556	rRNA intron-encoded homing endonuclease [<i>Oryza sativa</i>]	4.29E-10	14	1344
557	sulfite oxidase, putative [<i>Deinococcus radiodurans</i>]	6.21E-39	14	982
558	Unknown	>E-2	14	583
559	NAM phosphate synthase [<i>M. acetivorans</i>]	6.00E-50	14	564
560	putative HSP70	0	15	1759
561	hypothetical protein MG06499.4 [<i>M. grisea</i>]	6.92E-2	17	663
562	Unknown	>E-2	17	509
563	Unknown	>E-2	17	1160
564	OSJNBb0085H11.25 [<i>Oryza sativa</i>]	7.00E-5	17	680
565	Unknown	>E-2	19	796
566	hypothetical protein [<i>Pyrococcus abyssi</i>]	5.62E-2	22	1368
567	Unknown	>E-2	25	595
568	Unknown	>E-2	25	709
569	hypothetical protein Chr3_0400 [<i>Leishmania major</i>]	5.28E-41	30	1005
570	UBB protein [<i>Homo sapiens</i>]	5.33E-95	31	628
571	Unknown	>E-2	38	546
572	Unknown	>E-2	49	621
573	Unknown	>E-2	62	610
574	hypothetical protein [<i>B. thetaiotaomicron</i> VPI-5482]	1.88E-2	65	627
575	Unknown	>E-2	106	794

cysteine proteinase. HSP70 assumes a prominent role in stress resistance in eukaryotes, and in other algae exhibits elevated levels of expression in response to different types of stress including pH, light, temperature, high CO₂ levels, nitrates, and phosphates (Beuf et al. 1999; Biel et al. 2002). "Stress genes" such as the HSP70 and some of these other genes may be important in determining whether there is a link between calcification and stress. Some of the apoptotic genes also identified herein including cathepsin, metacaspase, and a hypersensitive response element may shed light on phytoplankton mortality and the evolutionary development and functionality of autocatalytic cell death in unicellular plankton.

Both the f/2 and f/50 libraries contained multiple copies of phosphoenolpyruvate carboxykinase (PEPCK), although the f/2 library contained a higher number of PEPCK transcripts than f/50 (16 versus 9, respectively). PEPCK is an enzyme that,

in most organisms, catalyzes the decarboxylation and phosphorylation of oxaloacetate (OAA) to form phosphoenolpyruvate (PEP). This is the first step in gluconeogenesis, wherein citric acid cycle intermediates are converted to hexoses. However, PEPCK also functions in the anapleuotic pathway in some organisms, converting PEP to OAA to replenish TCA cycle intermediates, and also has been shown to play a key role during photosynthesis in many C₄ plants and algae (Raven 1997). Therefore, depending upon the organism (or an organism's growth conditions), PEPCK may play different physiological roles. To date, there is little known about the regulation of genes and proteins involved carbon metabolism in *E. huxleyi*. We are investigating the role(s) of PEPCK in the biology of *E. huxleyi*, including the possibility that it may be involved as an alternative carbon concentrating mechanism for photosynthetic CO₂ fixation, as recently described in the marine diatom, *Thalassiosira weissflogii* (Reinfelder et al. 2000). These studies

may shed light on the debate surrounding the evolution of C₄ photosynthesis, a process that may have originally evolved in microalgae, thus predating its appearance in terrestrial C₄ and crassulacean acid metabolism plants.

Whether or not protein-coding genes play a role in regulating the growth, shape, and physical properties of the coccoliths, as they do in biomineralization processes in the diatom, oyster, sea urchin, and mollusk (Gotliv et al. 2003; Hazelaar et al. 2003; Illies et al. 2002; Miyamoto et al. 1996; Zhang et al. 2000) is not known. Nineteen of the most abundant ESTs in the f/50 library derived from lith forming cells show no significant match to any sequence present in the f/2 library derived from non-calcifying cells, nor do they show any significant match to any of our GenBank sequences, including biomineralization proteins. Our laboratory intends to obtain full length sequences for these clones that may represent marker sequences for calcification, and characterize their biochemical and biophysical properties with respect to what is known regarding silaffins (Kroger et al. 2000; Kroger et al. 2001; Kroger et al. 2002), lustrin (Zhang et al. 2002), pearlins (Miyashita et al. 2000), nacrein (Miyamoto et al. 1996) and other biomineralization proteins (Michenfelder et al. 2003; Wustman et al. 2002; Zhang et al. 2000). Many biomineral-associated proteins tend to have no distinct secondary or tertiary structure but rather feature extended, repeating β -turn, loop, or random coil conformations (Michenfelder et al. 2003; Wustman et al. 2002; Xu and Evans 1999; Zhang et al. 2002; Zhang et al. 2000) and contain specific amino acid sequence motifs. They also tend to be acidic in nature and feature distinct amino acid sequence repeats (Sollner et al. 2003).

N-acetylneuraminic acid phosphate synthase may also be a marker for calcification in *E. huxleyi* as the enzyme is highly expressed in the library derived from lith forming cells and is absent from the library derived from nonlith forming cells. N-acetylneuraminic acid phosphate synthase is involved in the synthesis of sialic acids, a group of carboxylated amino sugars that have a number of important cellular functions. Sialic acids are known to bind Ca²⁺ ions (Jacques et al. 1977) and could play a role in biomineralization in *E. huxleyi* by providing interactions between the organic matrix and the calcite platelets. Because sialic acids have been implicated in so many other mineralization processes (Ameye et al. 2001; Butler et al. 2003; Fujisawa et al. 1995) the abundance of N-acetylneuraminic acid phosphate synthase in calcifying *E. huxleyi* cells is intriguing.

Functional gene sequences identified through this EST project represent a powerful tool for reconstructing a phylogenetic history of coccolithophorids that can be compared with fossil records. Analysis of gene sequences across species can be used to trace the timing and the tempo of coccolithophorid evolution to the extent that the degree of sequence divergence represents elapsed time since a common ancestor. cDNA sequences can also be employed to identify and compare introns in corresponding genomic sequences to resolve more detailed relationships. Some of the nuclear encoded functional genes, such as nitrate reductase, light harvesting complex proteins, elongation factor Tu, glyceraldehydes-3-phosphate dehydrogenase, and FtsH that have been identified in our libraries, are ideally suited for these purposes and can be used to complement work being done with rRNA sequences (Saez et al. 2003). Functional gene sequences and their products can define physiological capabilities and as such can provide independent information about paleoenvironmental features to which coccolithophorids have

adapted. Geochemical changes in the oceans causing alterations in environmental metal concentrations may trigger selection of species and metalloenzymes with different metal requirements (Kirschvink et al. 2000).

In conclusion, the results presented here are a first step in the description of the expressed part of the *E. huxleyi* genome under calcifying and non-calcifying growth conditions. We have partially sequenced 3527 and 4,166 cDNAs from cultures of calcifying and non-calcifying cells, respectively; contributing 2841 new sequences and bringing the total unigene set for *E. huxleyi* to 4057. The proportion of ESTs that match known proteins in GenBank is fairly high, approximately 40%, and represents nearly a 100-fold increase in the number of protein sequences that had been previously reported for *E. huxleyi*. Although still in its preliminary stages, the establishment of this high-quality *E. huxleyi* unigene set will greatly facilitate the investigation of gene function, structure, and regulation and will help to establish the evolutionary relationship between *E. huxleyi* and other phytoplankton. Furthermore, several genes revealed are relevant to the development of new molecular tools and techniques that will increase our understanding of the biochemical and physiological processes specific to this important alga. And finally, the cDNA sequence information will also be a valuable resource for the Joint Genome Institute (U.S. Department of Energy), which is currently sequencing the entire *E. huxleyi* genome in collaboration with our laboratory. The genome sequence is expected to be completed in the fall of 2004.

ACKNOWLEDGMENTS

The authors would like to thank Drs. Colombar de Vargas and Linda Medlin for the critical review and helpful suggestions for our manuscript.

REFERENCES

- ADAMS, M. D., KERLAVAGE, A. R., FLEISCHMANN, R. D., FULDNER, R. A., BULT, C. J., LEE, N. H., KIRKNESS, E. F., WEINSTOCK, K. G., GOCAYNE, J.D and WHITE, O., 1995. Initial assessment of human gene diversity and expression patterns based upon 83 million nucleotides of cDNA sequence. *Nature*, 377(6547 Suppl.): 3-174.
- ALLGAIER, M., UPHOFF, H., FELSKE, A. and WAGNER-DOBLER, I., 2003. Aerobic Anoxygenic Photosynthesis in *Roseobacter* Clade Bacteria from Diverse Marine Habitats. *Applied and Environmental Microbiology*, 69(9): 5051-5059.
- AMEYE, L., DE BECKER, G., KILLIAN, C., WILT, F., KEMPS, R., KUYPERS, S. and DUBOIS, D. P., 2001. Proteins and saccharides of the sea urchin organic matrix of mineralization: characterization and localization in the spine skeleton. *Journal of Structural Biology*, 134(1): 56-66.
- BATES, T. S., CHARLSON, R. J. and GAMMON, R. H., 1987. Evidence for the climatic role of marine biogenic sulfur. *Nature*, 329: 319-321.
- BERGER, W. H., 1982. Deglacial CO₂ buildup: Constraints on the coral reef model. *Paleogeography, Paleoclimatology, Paleoecology*, 40: 235-253.
- BEUF, L., KURANO, N. and MIYACHI, S., 1999. Rubisco activase transcript (*rca*) abundance increases when the marine unicellular green alga *Chlorococcum littorale* is grown under high-CO₂ stress. *Plant Molecular Biology*, 41(5): 27-635.
- BHATTACHARYA, D., MEDLIN, L., WAINRIGHT, P. O., ARIZTIA, E., BIBEAU, C., STICKEL, S. K. and SOGIN, M. L.,

TABLE 4
Contigs of most prevalent ESTs from f/2 library.

ID	Putative ID	e-value	Reads	Length
696	Unknown	E > -2	5	976
697	Unknown	E > -2	5	1428
698	DNA-directed DNA polymerase; phage SP01	1.4E-11	5	1391
699	Unknown	E > -2	5	1279
700	BiP-isoform D [Glycine max]	5.0E-107	6	950
701	Unknown	E > -2	6	1277
702	MC107L [<i>Molluscum contagiosum</i> virus]	4.87E-08	6	1594
703	pG1 protein [<i>Homo sapiens</i>]	4.55E-20	6	1380
704	fucoxanthin chlorophyll a /c protein	1.73E-25	6	818
705	DNA repair protein RAD25 [<i>Pyrococcus abyssi</i>]	2.88E-12	6	2771
706	yzkA~similar to cell wall-binding protein [<i>Bacillus subtilis</i>]	2.85E-11	6	1605
707	intracellular carbonic anhydrase [<i>Thalassiosira weissflogii</i>]	1.28E-55	6	995
708	PFTAIRE-interacting factor 2 [<i>Drosophila melanogaster</i>]	1.68E-3	6	1994
709	Probable monooxygenase cyt P450 [<i>Mesorhizobium loti</i>]	1.31E-2	6	2350
710	COG5276: Uncharacterized conserved protein [<i>Microbulbifer degradans</i> 2-40]	2.37E-24	6	2084
711	uncoupling protein UCP-4 [<i>Rattus norvegicus</i>]	6.94E-66	6	1315
712	glyceraldehyde-3-phosphate dehydrog	7.8E-136	6	1323
713	actin, type 1 - <i>Emiliania huxleyi</i> (fragment)	0	7	1575
714	pep carboxykinase [<i>T. gondii</i>]	4.01E-71	7	885
715	Tryptophan synthase [<i>Neurospora crassa</i>]	2.5E-157	7	2228
716	transmembrane protein SBB199 [<i>Homo sapiens</i>]	2.97E-16	7	1628
717	Unknown	E > -2	7	1177
718	phosphoglycerate kinase [<i>C. crispus</i>]	3.5E-160	8	1556
719	Unknown	E > -2	8	926
720	Endolysin (N-acetylmuramoyl-L-alanine amidase)	3.75E-99	8	4208
721	Hypothetical ANK-repeat protein	2.24E-27	8	1655
722	Uncharacterized protein conserved in bacteria [<i>R. sphaeroides</i>]	1.78E-31	8	1888
723	Unknown	E > -2	9	961
724	Unknown	E > -2	9	2208
725	Unknown	E > -2	9	1185
726	S-adenosyl methionine synthetase]	2.4E-110	9	1480
727	nicotinamide phosphoribosyl transferase [<i>H. ducreyi</i>]	7.71E-11	9	2994
728	Unknown	E > -2	9	1613
729	putative helicase [<i>Salmonella enterica</i>]	3.13E-20	9	3836
730	pep carboxykinase [<i>Emericella nidulans</i>]	5.06E-89	9	949
731	adenosylhomocysteinase-like protein [<i>Oryza sativa</i>]	0	10	1773
732	Unknown	E > -2	10	1000
733	Unknown	E > -2	10	970
734	AT5g56010/MDA7_5 [<i>A. thaliana</i>]	2.24E-96	10	1632
735	Unknown	E > -2	10	2180
736	conserved hypothetical protein [<i>Shewanella oneidensis</i>]	2.49E-62	10	3725
737	acyl-CoA synthetase family [<i>A. thaliana</i>]	5.51E-94	11	2120
738	EsV-1-216 [<i>Ectocarpus siliculosus</i> virus]	7.85E-4	11	3326
739	major tail sheath protein [Bacteriophage A511]	0	12	4700
740	Unknown	E > -2	14	1506
741	putative membrane protein [<i>E. coli</i> K12]	1.70E-20	15	1030
742	fructose-1,6-bisphosphate aldolase precursor [<i>Odontella sinensis</i>]	3.8E-130	17	1558
743	putative elongation factor 3 [<i>Schizosaccharomyces pombe</i>]	3.0E-166	20	3576
744	blr0930 [<i>Bradyrhizobium japonicum</i>]	2.0E-4	21	1655
745	predicted AAA family ATPase [<i>Mycoplasma penetrans</i>]	2.65E-22	85	1408

1992. Algae containing chlorophylls *a* + *c* are paraphyletic: Molecular evolutionary analysis of the chromophyta. *Evolution*, 46: 801-1817.
- BIEL, A. J., CANADA, K., HUANG, D., INDEST, K. and SULLIVAN, K., 2002. Oxygen-Mediated Regulation of Porphobilinogen Formation in *Rhodobacter capsulatus*. *Journal of Bacteriology*, 184(6): 685-1793.
- BRASSEL, S. C., BRERETON, R. G., EGLINTON, G., GRIMALT, J., LIEBEZEIT, G., MARLOWE, I. T., PLAUFMANN, U. and SARNTHEIN, M., 1986b. Paleoclimate signals recognized by chemometric treatment of molecular stratigraphic data. In: Leythaeuser, D. and Rullkotter, J., Eds., *Advances in Organic Geochemistry*, 10: 649-660.
- BRASSEL, S. C., EGLINTON, G., MARLOWE, I. T., PFLAUMANN, U. and SARNTHEIN, M., 1986a. Molecular stratigraphy: a new tool for climatic assessment. *Nature*, 320: 129-133.
- BROWN, C. W. and YODER, J. A., 1994. Coccolithophorid blooms in the global ocean. *Journal of Geophysical Research*, 99(C4): 7467-7482.
- BUTLER, W.T., BRUNN, J.C. and QIN, Q. C., 2003. Dentin extracellular matrix (ECM) proteins: comparison to bone ECM and contribution to dynamics of dentinogenesis. *Connective Tissue Research*, 44(Suppl. 1): 171-178.
- CHARLSON, R. J., LOVELOCK, J. E., ANDREAE, M. O. and WARREN, S. G. 1987. Oceanic phytoplankton, atmospheric sulfur, cloud albedo, and climate. *Nature*, 326: 665-661.
- CONTE, M. N., EGLINTON, G. and MADUREIRA, L. A. S., 1992. Long-chain alkenones and alkyl alkenoates as paleotemperature indicators: their production, flux, and early sedimentary diagenesis in the Eastern North Atlantic. In: Eckardt, C. B. and Larter, S. R., Eds., *Advances in Organic Chemistry*, 19: 287-298.
- DES MARAIS, D. J. 2000. EVOLUTION: When Did Photosynthesis Emerge on Earth? *Science*, 289(5485): 1703-1705.
- FERRIS, P. J., PAVLOVIC, C., FABRY, S. and GOODENOUGH, U. W., 1997. Rapid evolution of sex-related genes in *Chlamydomonas*. *Proceedings of the National Academy of Sciences of the United States of America*, 94: 8634-8639.
- FRANCIS, C. A., CO, E.-M. and TEBO, B. M., 2001. Enzymatic Manganese(II) Oxidation by a Marine α -Proteobacterium. *Applied and Environmental Microbiology*, 67(9): 4024-4029.
- FUJISAWA, R., NODASAKA, Y. and KUBOKI, Y., 1995. Further characterization of interaction between bone sialoprotein (BSP) and collagen. *Calcification Tissue International*, 56(2): 140-144.
- GOTLIV, B.-A., ADDADI, L. and WEINER, S., 2003. Mollusk Shell Acidic Proteins: In Search of Individual Functions. *ChemBioChem*, 4: 522-529.
- GUILLARD, R. R. L. 1975. Culture of phytoplankton for feeding marine invertebrates. In: Smith, W. L. and Chanley, M. H., Eds., *Culture of Marine Invertebrate Animals*. New York: Plenum Press, 29-60.
- HAZELAAR, S., VAN DER STRATE, H. J., GIESKES, W. W. C. and VRIELING, E. G., 2003. Possible role of ubiquitin in silica biomineralization in diatoms: identification of a homologue with high silica affinity. *Biomolecular Engineering*, 20(163-169).
- HOLLIGAN, P. M., FERNANDEZ, E., AIKEN, J., BALCH, W. M., BOYD, P., BURKILL, P. H., FINCH, M., GROOM, S. B., MALIN, G., MULLER, K., PURDIE, D. A., ROBINSON, C., TREES, C., TURNER, S. M. and WAL, P. V.D., 1993. A biogeochemical study of the coccolithophore *Emiliania huxleyi* in the North Atlantic. *Global Biogeochemical Cycles*, 7: 879-900.
- HONJO, S., 1996. Fluxes of particles to the interior of the open oceans. In: Ittekkot, P. S. V., Honjo, S. and Depetris, P.J., Eds., *Particle Flux in the Ocean*, SCOPE, New York: Wiley, 57: 91-57.
- ILLIES, M. R., PEELER, M. T., DECHTIARUK, A. M. and ETTENSOHN, C. A., 2002. Identification and developmental expression of new biomineralization proteins in the sea urchin *Strongylocentrotus purpuratus*. *Development Genes and Evolution*, 212: 419-431.
- JACQUES, L. W., BROWN, E. B., BARRET, J. M., BREY, W. S. and WELTNER, W. J., 1977. Sialic acid - A calcium-binding carbohydrate. *Journal of Biological Chemistry*, 252: 4533-4538.
- KIRSCHVINK, J. L., GAIDOS, E. J., BERTANI, L. E., BEUKES, N. J., GUTZNER, J., MAEPA, L. N. and STEINBERGER, R. E., 2000. Paleoproterozoic snowball earth: extreme climate and geochemical global change and its biological consequences. *Proceedings of the National Academy of Sciences of the United States of America*, 97(4): 1400-1405.
- KROGER, N., DEUTZMANN, R., BERGSDORF, C. and SUMPER, M., 2000. Species-specific polyamines from diatoms control silica morphology. *Proceedings of the National Academy of Sciences of the United States of America*, 97(26): 14133-14138.
- KROGER, N., DEUTZMANN, R. and SUMPER, M., 2001. Silica-precipitating peptides from diatoms. The chemical structure of silaffin-1A from *Cylindrotheca fusiformis*. *Journal of Biological Chemistry*, 276(28): 26066-26070.
- KROGER, N., LORENZ, S., BRUNNER, E. and SUMPER, M., 2002. Self-assembly of highly phosphorylated silaffins and their function in biosilica morphogenesis. *Science*, 298(5593): 584-586.
- KUMADA, Y., BENSON, D., HILLEMANN, D., HOSTED, T., ROCHEFORT, D., THOMPSON, C., WOHLLEBEN, W. and TATENO, Y., 1993. Evolution of the glutamine synthetase gene, one of the oldest existing and functioning genes. *Proceedings of the National Academy of Sciences of the United States of America*, 90: 3009-3013.
- LAGUNA, R., ROMO, J., READ, B. A. and WAHLUND, T. M., 2001. Induction of phase variation events in the life cycle of the marine coccolithophorid *Emiliania huxleyi*. *Applied and Environmental Microbiology*, 67(9): 3824-3831.
- MICHENFELDER, M., FU, G., LAWRENCE, C., WEAVER, J. C., WUSTMAN, B.A., TARANTO, L., EVANS, J. S. and MORSE, D. E., 2003. Characterization of two molluscan crystal-modulating biomineralization proteins and identification of putative mineral binding domains. *Biopolymers*, 70: 522-533.
- MIYAMOTO, H., MIYASHITA, T., OKUSHIMA, M., NAKANO, S., MORITA, T. and MATSUSHIRO, A., 1996. A carbonic anhydrase from the nacreous layer in oyster pearls. *Proceedings of the National Academy of Sciences of the United States of America*, 93(18): 9657-9660.
- MIYASHITA, T., TAKAGI, R., OKUSHIMA, M., NAKANO, S., MIYAMOTO, H., NISHIKAWA, E. and MATSUSHIRO, A., 2000. Complementary DNA cloning and characterization of pearlins, a new class of matrix protein in the nacreous layer of oyster pearls. *Marine Biotechnology*, 2: 409-418.
- MOON-VAN DER STAAY, S. Y., DE WACHTER, R. and VAULOT, D., 2001. Oceanic 18S rDNA sequences from picoplankton reveal unsuspected eukaryotic diversity. *Nature*, 409: 607-610.

- PRAHL, F. G., MUEHLHAVSEN, L. A. and ZAHNLE, D. L., 1988. Further evaluation of long-chain alkenones as indicators of paleoceanographic conditions. *Geochimica et Cosmochimica Acta*, 52: 2303-2310.
- RAVEN, J. A. 1997. Putting the C in phycology. *European Journal of Phycology*, 32:319-333.
- REINFELDER, J. R., KRAEPIEL, A. M. and MOREL, F. M. 2000. Unicellular C₄ photosynthesis in a marine diatom. *Nature*, 407: 996-999.
- SAEZ, A., G. PROBERT, I., GEISEN, M., QUINN, P., YOUNG, J. R. and MEDLIN, L. K., 2003. Pseudo-cryptic speciation in coccolithophores. *Proceedings of the National Academy of Sciences of the United States of America*, 100(12): 7163-7168.
- SIKES, E. L., FARRINGTON, J. W. Z and KEIGWIN, L. D., 1991. Use of alkenone unsaturation ratios to determine past sea surface temperatures: Core-top SST calibrations and methodology considerations. *Earth and Planetary Science Letters*, 104: 36-47.
- SO, A. K.-C., ESPIE, G. S., WILLIAMS, E. B., SHIVELY, J. M., HEINHORST, S. and CANNON, G. C., 2004. A novel evolutionary lineage of carbonic anhydrase (γ class) is a component of the carboxysome shell. *Journal of Bacteriology*, 186(3): 623-630.
- SOLLNER, C., BURGHAMMER, M., BUSCH-NENTWICH, E., BERGER, J., SCHWARZ, H., RIEKEL, C. and NICOLSON, T., 2003. Control of crystal size and lattice formation by starmaker in otolith biomineralization. *Science*, 302(5643): 282-286.
- STROMMER, J., GREGERSON, R. and VAYDA, M., 1993. Isolation and characterization of Plant mRNA. In: Glick, B. R. and Thompson, J. E., Eds., *Methods in Plant Molecular Biology and Biotechnology*. Boca Raton: CRC Press, 49-66.
- WAHLUND, T. M., HADAEGH, A. R., CLARK, R., NGUYEN, B., FANELLI, M. and READ, B. A., 2004. Analysis of expressed sequence tags from calcifying cells of the marine coccolithophorid, *Emiliana huxleyi*. *Marine Biotechnology*, 6(3): 278-290.
- WESTBROEK, P., BROWN, C. W., VAN BLEUSWIJK, J., BROWN-LEE, C., BRUMMER, G. J., CONTE, M., EGGE, J., FERNANDEZ, E., JORDAN, R., KNAPPERTSBUSCH, M., STEFELS, J., VELDHUIS, M. and WAL, P. VANDER, 1993. A model system approach to biological climate forcing. the example of *Emiliana huxleyi*. *Global and Planetary Change*, 8: 27-46.
- WUSTMAN, B. A., SANTOS, R., ZHANG, B. and EVANS, J.S., 2002. Identification of a "glycine-loop"-like coiled structure in the 34 AA Pro, Gly, met repeat domain of the biomineral-associated protein, PM27. *Biopolymers*, 65: 362-372.
- XU, G. and EVANS, J., 1999. Model peptide studies of sequence repeats derived from the intracrystalline biomineralization protein, SM50. I. GVGGR and GMGGQ repeats. *Biopolymers*, 49(4): 303-312.
- YOUNG, J. R., DAVIS, S. A., BOWN, P. R. and MANN, S., 1999. Coccolith ultrastructure and biomineralisation. *Journal of Structural Biology*, 126: 195-215.
- ZHANG, B., WUSTMAN, B. A., MORSE, D. and EVANS, J. S., 2002. Model peptide studies of sequence regions in the elastomeric biomineralization protein, Lustrin A. I. The C-Domain consensus-PG-NVNCT-motif. *Biopolymers*, 63: 358-369.
- ZHANG, B., XU, G. and EVANS, J. S., 2000. Model peptide studies of sequence repeats derived from the intracrystalline biomineralization protein, SM50. II. Pro, ASN-rich tandem repeats. *Biopolymers*, 54: 464-475.

