

GENOME-WIDE ANALYSIS OF CORE PROMOTER REGIONS IN EMILIANIA HUXLEYI

Justin Dailey and Xiaoyu Zhang*

Department of Computer Science, California State University San Marcos
San Marcos, CA 92096

Email: daile005@csusm.edu, xiaoyu@csusm.edu*

Betsy Read

Department of Biological Science, California State University San Marcos
San Marcos, CA 92096

Email: bread@csusm.edu

Emiliana huxleyi (*E. huxleyi*), the most abundant coccolithophorid, has served as a model system for environmental studies pertaining to global carbon cycling, oceanic sulfur bio-transformations, and paleoclimatology. It has also potential applications in material sciences and biomedical studies. *E. huxleyi* was selected by the Department of Energy in 1999 for genome sequencing, which was recently completed in 2006. To date, little is known about the transcription regulatory elements of *E. huxleyi* and other coccolithophorids. Because of its exceptionally high GC content (~66%), the *E. huxleyi* genome may contain unique and novel regulatory elements that have not yet been identified in other organisms. Genome-wide analysis of core promoter regions in *E. huxleyi* represents an initial and important step towards understanding the transcriptional regulation mechanism in this important organism. Using a collection of 190 full-length complementary DNAs (cDNAs) and 84,385 expressed sequence tags (ESTs), we first extracted a robust dataset of 3,132 potential core promoter regions of from the first assembly the *E. huxleyi* genome (released November 3, 2006). These promoter regions were then scanned for both known and novel sequence motifs. The *E. huxleyi* core-promoters lack the TATA motif but contain over-represented novel motifs. In addition we computed for these regions distinctive physical and structural properties that may be related to the transcriptional regulation process. Twelve out of fourteen properties analyzed showed significant distinction between the promoter regions and other background coding and non-coding regions. The core-promoter profile consisting of regulatory motifs, sequence characteristics, and structural and physical features will be used to develop computational models to predict core promoters across the entire *E. huxleyi* genome.

1. INTRODUCTION

Emiliana huxleyi (*E. huxleyi*) is the most abundant coccolithophore on a global basis. Because of its resistance to decomposition, it has served as a model organism for environmental studies pertaining to global carbon cycling, oceanic sulfur bio-transformations, and paleoclimatology^{13, 14}. As one of the most important calcifying organisms on earth¹⁷, *E. huxleyi* has also attracted the attention of material scientists who are interested in duplicating the nano-scale architecture of the coccoliths for novel materials, as well as members of the biomedical community who are interested in unraveling the genetic and molecular mechanisms of biomineralization as a means of providing insight into biomineralization processes associated with certain human pathologies including osteoporosis and vascular calcification.

Because of the widespread interest across diverse scientific fields, *E. huxleyi* was selected by the

Department of Energy in 1999 for genome sequencing. Sequencing of the genome has been completed to 12x coverage and the first version of its assembly was released recently, November 3, 2006, which yielded 7,809 scaffolds encompassing 165 million base pairs (MB). The *E. huxleyi* genome proved to be a difficult genome to sequence and assemble owing largely to its high G+C content, the relatively high rate of sequence polymorphisms and the repetitive nature of the genome. *E. huxleyi* transcripts have an unusually high GC content about 67% while the genome scaffolds have an average GC content of 66%. More than 80,000 EST/cDNA sequences have been collected for *E. huxleyi* under both calcifying and non-calcifying conditions. Hence, a vast amount of sequence data is available for computational mining and analysis.

While the genome of *E. huxleyi* has been sequenced and a large number of ESTs are available, a systematic study of the functional genomics of *E. huxleyi* promoters has yet to be undertaken. The goal of this

*Corresponding author.

project is to identify and analyze putative core promoter regions of *E. huxleyi* to discover distinctive regulatory motifs and structural features. It is an initial step in the full-scale study of the transcriptional regulation and core promoter regions of *E. huxleyi*. Previous studies on other model organisms have shown that core promoter regions contain distinctive characteristics for individual organisms, while also sharing some common cis-regulatory motifs^{3, 9, 10, 16}. However, even within an organism these regulatory motifs are not strictly preserved for every promoter region. In this paper we attempt to identify *E. huxleyi* core promoter regions and analyze them for both known and novel sequence motifs. In addition we computed distinctive physical and structural properties for these regions that may be related to the transcriptional regulation process, in order to construct a profile of core-promoter regions, which includes regulatory motifs, sequence characteristics, and structural and physical features. The feature profile will be used to develop prediction models for core promoters across the entire *E. huxleyi* genome.

Using the datasets of *E. huxleyi* EST/cDNA sequences, we extracted a robust set of 3,132 core promoter regions. We scanned the core promoter regions for known and novel sequence motifs and found *E. huxleyi* has no hits to the motifs of *Drosophila* but shares significant motifs with *Arabidopsis*. *E. huxleyi* lacks an apparent “TATA” box motif but has other novel motifs, e.g. a perfect GCGGCCG, which shows a significant positional preference in the core-promoter regions. Additionally 12 out of the 14 analyzed physical properties showed significantly distinctive patterns in the [-50,0] range upstream of the approximate transcription start site (TSS) compared to other background coding and non-coding regions.

2. METHODS

In this section, we first describe our approach of extracting core-promoter regions from the *E. huxleyi* genome, and then the methods used to search for motifs and analyze and physical properties of the DNA.

2.1. Core-promoter Region Extraction

The core promoter region, a key component in the regulation of gene transcription, is usually considered to be a DNA sequence region extending approximately 50-

100 bp upstream and downstream of the TSS. Molecular genetics of *E. huxleyi* is still in its infancy, and thus it is not surprising that to date not a single transcriptional start site has been mapped by conventional experimental methods. Hence, to identify promoter regions, we aligned cDNA and EST sequences to the *E. huxleyi* genome to locate the approximate TSSs and extracted upstream and downstream sequence.

ESTs were initially filtered and only those exhibiting significant homology to the amino terminus of known genes were selected for further analysis. A total of 84,385 EST sequences were aligned to the NCBI NR database using the BLASTX algorithm. Only EST sequences that were found to have significant hits (e-value $< 1 \times e^{-20}$) and within 5 amino acids from the subject start were included in subsequent steps. The 190 full-length cDNA sequences are also included to provide a combined total of 11,868 sequences for the alignment step.

Next, sequences selected for further analysis were aligned to the genome using two different algorithms, BLAT⁷ and Splign⁶, to provide cross-validation and generate the core-promoter regions with higher confidence. Both algorithms were selected for their ability to incorporate splice signals into their alignments. Only “matched” alignments revealed by both algorithms were used to extract the promoter regions. An alignment is considered a “match” only if the query sequence, the genome sequence hit, and the genome sequence strand are identical for the two algorithms. In addition the start and end positions of the alignments must be within 5 base pairs of each other. These alignments were sorted by genome strand and starting position and then filtered to remove redundant or overlap sequences. If the starting position of an alignment was found to be within 500 bp downstream of the starting position of the previous alignment, it was removed from the dataset.

Finally we extracted the region 100 bp upstream and 100 bp downstream from the start positions of the alignments as the putative core-promoter regions. We took a conservative approach to extract relatively large core-promoter regions because of the putative nature of the TSSs. Additional filtering was performed such that only high-quality core-promoter regions that were 200 bp and contained no more than ten unknown

nucleotides were taken. In the end we obtained a robust set of 3,132 putative core-promoter regions.

For comparative purposes, background models consisting of the same number of coding and non-coding regions were also constructed. The background coding dataset is represented by 200 bp segments extracted from the exons in the genome using the alignments of the EST/cDNA sequences while the non-coding dataset is represented by 200 bp segments extracted 500 bp upstream from the putative TSSs.

2.2. Motif Search

Core promoters are the recognition sites for the construction of the pre-initiation complex (PIC), including RNA polymerase II and several transcription factors, which make up the basal transcription unit. Specific interactions between promoters and transcription factors are defined by highly conserved short DNA sequence motifs^{4, 15}. Although there are no universal motifs shared by all core promoters, a number of motifs have been identified that are over-represented in core-promoter regions of other model organisms. One of the most common motifs is the TATA box, which has been reported to be present in 30-40% of core promoters in many organisms.

One goal of this project was to determine if *E. huxleyi* shared any known sequence motifs with other organisms. For this purpose we chose to use the position-specific scoring matrices (PSM) from studies done on the *Drosophila*¹¹, and *Arabidopsis*⁹ promoters. Searches were performed using MAST² on *E. huxleyi* promoter sequences, with the nucleotide frequencies of the *E. huxleyi* genome as a parameter.

We then searched for novel motifs that are over-represented in the *E. huxleyi* core-promoter regions. The search was performed using MEME¹, an expectation maximization algorithm. The top ten motifs with length between 5 and 10 bp were requested using the zero or one occurrence per sequence model. To increase the sensitivity for motifs that may only appear in a small number of sequences, we limited our search to the 50 bp upstream and downstream regions flanking the TSS. We then increased the size of the flanking regions to 100 for verification, and to ascertain whether other motifs may reside outside of the immediate vicinity of the TSS.

2.3. Physical Properties

Physical and structural properties of DNA are known to significantly influence gene regulation. Previous studies have shown there are physical and structural properties unique to the promoter regions of the genome^{5, 12}. As part of this study we will incorporate the properties of our promoters into our prediction model.

To create the physical profiles the core-promoter sequences were again aligned to the putative TSSs. Sequences were converted to a numeric profile using the parameter sets of physical DNA properties⁸. Because the physical parameter coefficients refer to di- or tri-nucleotides, the profiles tend to be very noisy. We used a mean value filter to smooth the data and create profiles by computing the average at every position with a window size of 21 bp.

3. RESULTS

From a combined total of 84,575 EST/cDNA sequences we extracted 3,132 putative core-promoter regions. Figure 1 shows the accumulated distributions of those regions on the *E. huxleyi* genome.

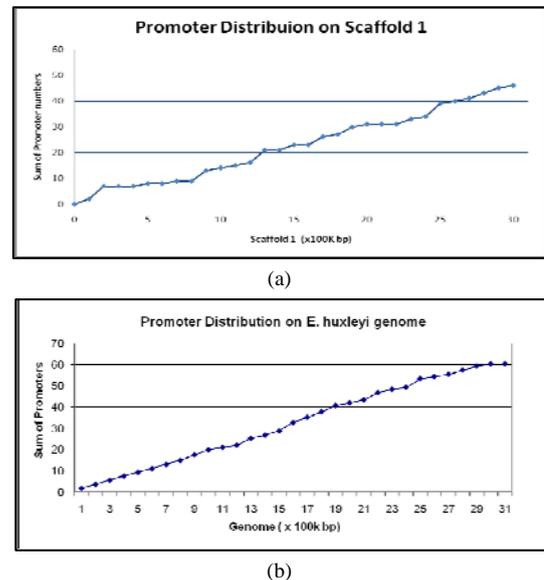


Figure 1. The accumulative distribution of the extracted regions on (a) scaffold 1 and (b) the entire genome.

As evidenced in Figure 1, the extracted core-promoter regions are distributed approximately evenly across the *E. huxleyi* genome, i.e. the data set is representative of the entire genome and is not biased.

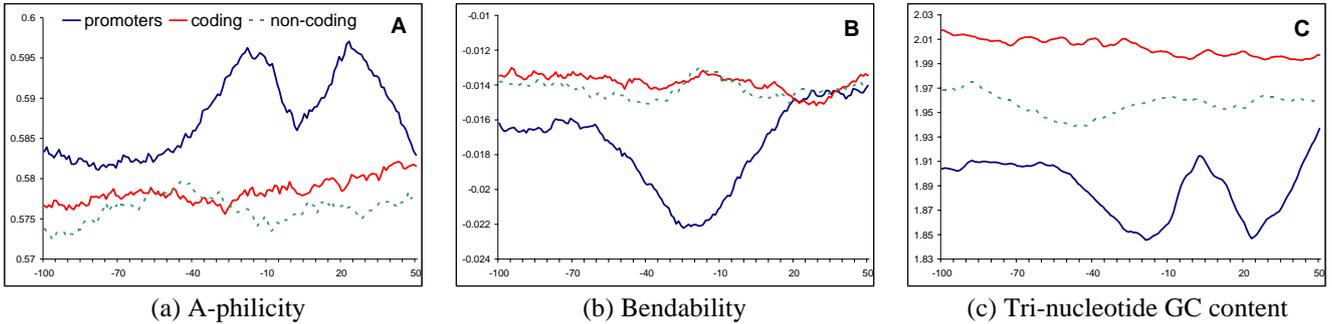


Figure 2. DNA mean physical property profiles. Transcription start sites are aligned at position 0. Profiles are smoothed with a mean filter window of 21 and averaged over all sequences. A. A-philicity; B. Bendability; C. Tri-nucleotide GC content.

Interestingly, the *E. huxleyi* promoter regions have no significant matches to any of the *Drosophila* promoter motifs, but contain several significant hits to motifs found in the *Arabidopsis*⁹. This is not completely surprising since *E. huxleyi* is evolutionarily closer to *Arabidopsis* than *Drosophila*. While TATA boxes are found in promoters for approximately 30% of *Drosophila* genes¹⁰ and 28.8% of *Arabidopsis* genes, the motif most closely resembling the TATA box (Motif 3) in *Arabidopsis*⁹ is found in only 4.5% of the *E. huxleyi* promoter regions.

TATA boxes were also not detected in the *E. huxleyi* promoter regions when the MEME algorithm was applied to search for novel motifs. The 10 most significant motifs in the 50 bp upstream ([-50, 0]) and 50 bp downstream ([1, 50]) are listed in Table 1. Only motif 5 in the [-50, 0] region remotely resembles a TATA box. Interestingly, motif 1 and 2 found in the [-50,0] region are very similar to the first two motifs found in *Arabidopsis* promoters⁹. These motifs in *Arabidopsis*, however, are considered to be microsatellites. The motifs in *E. huxleyi* on the other hand do not resemble microsatellites. The similarity nonetheless in sequence and prevalence between those two genomes is intriguing. With the exception of these two motifs, the remaining motifs identified in the *E. huxleyi* promoter regions are not similar to promoter motifs found in any other organisms, suggesting that *E. huxleyi* may contain novel transcriptional control sequences. Motif 3 in Table 1 is particularly interesting. It is a perfect pattern that appears 240 times in the 3,132 *E. huxleyi* core-promoter regions used in this study. Moreover, this particular motif shows significant positional preference in the promoter regions centered near the -5 position relative to the TSS, suggesting it

may be important to transcriptional regulation of *E. huxleyi* genes.

The analysis of the chemical and physical properties for the promoter regions was also intriguing. Although the putative *E. huxleyi* TSSs are based on prediction only, out of the 14 properties analyzed, all but two showed well-defined patterns when mapped across the promoter regions as compared to background coding and non-coding regions. The majority of the properties show a distinct configuration centered around position -20 relative to the TSS. Figure 2 shows the profiles of three physical properties with a smoothing window of size 21 bp: A-philicity, DNA bendability, and Tri-nucleotide GC content. While the *E. huxleyi* genome has high GC content, Figure 2 (c) shows that the core-promoter regions have relatively lower Tri-nucleotide GC content compared to coding and non-coding regions and two dips at approximately 20 bp upstream and 20 bp downstream. Florquin et al.⁵ discussed a common structural profile for bendability as being low in a region upstream of the TSS and significantly higher downstream of the TSS. It is indeed observed in the trace pattern in Figure 2 (b). A region in the DNA with a high A-philicity value is more easily converted to the A-form than a low value region, which is more resistant to transition. Two regions exhibiting the highest potential for A-philicity can be observed 20 bp upstream and downstream in the core-promoter profile in Figure 2 (a). This position may be shifted slightly on account of the predicted TSS position. It is clear, nonetheless, that promoters in *E. huxleyi* are dependent upon some of these observed chemical and physical properties. These signals which may serve as a promoter blueprint will be incorporated into the promoter prediction model.

[-50,0]		[0,50]	
Motif 1 428 9.4e-571		Motif 11 239 4.3e-94	
Motif 2 201 1.1e-283		Motif 12 141 6.9e-55	
Motif 3 240 3.6e-27		Motif 9 233 3.2e-36	
Motif 4 40 1.3e-14		Motif 13 176 1.7e-30	
Motif 5 21 2.7e-14		Motif 14 93 2.9e-11	
Motif 6 79 1.4e-12		Motif 15 19 1.6e-7	
Motif 7 30 2.4e-8		Motif 16 21 5.3e-12	
Motif 8 18 7.3e-9		Motif 17 14 5.3e-8	
Motif 9 144 9.2e-7		Motif 18 13 2.2e-7	
Motif 10 72 1.9e-8		Motif 19 25 1.8e-6	

Table 1. Analysis of motifs present in the [-50, 0] and [0, 50] regions of *E. huxleyi*. Motifs are ordered by the strength of their expected value. The numbers under the motif name correspond to the number of sequences the motif was found in, and the expected value of this motif occurring by chance with the background nucleotide frequencies of the genome. Sequence frequency logos were graphed using WebLogo, where the size of the characters represent the frequencies of occurrence. Positional preference graphs were created by grouping the motif starting position for each sequence into 5 bp bins within the selected range.

4. CONCLUSIONS AND DISCUSSIONS

This study provides a first genome-wide analysis of the composition and structure of *E. huxleyi* core promoters. When compared to *Drosophila* we found no significant matches to known motifs, and the comparison to *Arabidopsis* showed strong correlation with two of the motifs found. Additionally there are several other novel motifs that are significantly over represented in the core promoter regions. These motifs combined with the physical property profiles of the core promoters will be

used in developing the prediction model for the entire genome.

The next step of the research will focus on more accurately defining the TSSs and using these to determine positional preference of the motifs discovered, and expanding the search to look for potentially weaker but statistically significant motifs. These, in combination with microarray data will be used to develop promoter models that focus on subsets of genes whose expression is controlled thru *cis*-regulatory elements

REFERENCES

1. Bailey, T.L. and C. Elkan. *Fitting a mixture model by expectation maximization to discover motifs in biopolymers.* in *Second International Conference on Intelligent Systems for Molecular Biology.* 1994.
2. Bailey, T.L. and M. Gribskov, *Combining evidence using p-values: Application to sequence homology searches.* *Bioinformatics*, 1998. 14: p. 48-54.
3. Bajic, V.B., V. Choudhary, and C.K. Hock, *Content analysis of the core promoter region of human genes.* In *Silico Biol*, 2004. 4(2): p. 109-125.
4. Deng, W. and S.G.E. Roberts, *Core promoter elements recognized by transcription factor IIB.* *Biochemical Society Transactions*, 2006. 34: p. 1051-1053.
5. Florquin, K., et al., *Large-scale structural analysis of the core promoter in mammalian and plant genomes.* *Nucleic Acids Res*, 2005. 33(13): p. 4255-64.
6. Kapustin, Y., A. Souvorov, and T. Tatusova. *Splign: a Hybrid Approach to Spliced Sequence Alignments.* in *RECOMB 2004.* 2004. San Diego, CA.
7. Kent, W.J., *BLAT - The BLAST-Like Alignment Tool.* *Genome Research*, 2002. 12: p. 656-664.
8. Liao, G.-c., E.J. Rehm, and G.M. Rubin. *Insertion site preferences of the P transposable element in Drosophila melanogaster.* in *Proceedings of the National Academy of Sciences of the United States of America.* 2000.
9. Molina, C. and E. Grotewold, *Genome wide analysis of Arabidopsis core promoters.* *BMC Genomics*, 2005. 6(1): p. 25.
10. Ohler, U., et al., *Computational analysis of core promoters in Drosophila genome.* *Genome Biology*, 2002. 3(12).
11. Ohler, U., et al., *Computational analysis of core promoters in the Drosophila genome.* *Genome Biol*, 2002. 3(12): p. 87.
12. Ohler, U., et al., *Joint modeling of DNA sequence and physical properties to improve eukaryotic promoter recognition.* *Bioinformatics*, 2001. 17 Suppl 1: p. 199-206.
13. Paasche, E., *A review of the coccolithophorid Emiliana huxleyi (prymnesiophyceae), with particular reference to growth, coccolith formation, and calcification-photosynthesis interactions.* *Phycological Reviews*, 2002. 40: p. 503-529.
14. Stolte, W., et al., *Genetic and Physiological Variation in Pigment Composition of Emiliana Huxleyi (Prymnesiophyceae) and the Potential Use of Its Pigment Ratios As a Quantitative Physiological Marker.* *Journal of Phycology*, 2000. 36(3): p. 529-539.
15. Tsai, F.T.F. and P.B. Sigler, *Structural basis of preinitiation complex assembly on human Pol II promoters.* *The EMBO Journal*, 2000. 19(1): p. 25-36.
16. Wasserman, W.W., et al., *Human-mouse genome comparisons to locate regulatory sites.* *Nat Genet*, 2000. 26(2): p. 225-228.
17. Westbroek, P., et al. *Mechanism of calcification in the marine alga Emiliana huxleyi.* in *Royal Society of London Philosophical Transactions Series B.* 1984.