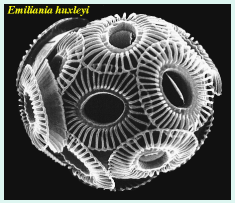


Genome-Wide Analysis of Core Promoter Regions in *Emiliania huxleyi*

[1]Justin Dailey, [1]Xiaoyu Zhang, [2]Betsy Read
[1]Department of Computer Science, California State University San Marcos, San Marcos, CA 92096
[2]Department of Computer Science, California State University San Marcos, San Marcos, CA 92096

Cal State San Marcos

1. Introduction

Emiliania huxleyi (*E. huxleyi*) is the most abundant marine coccolithophore. Because of its resistance to decomposition, it has served as a model organism for environmental studies pertaining to global carbon cycling and paleoclimatology. As one of the most important calcifying organisms on earth, *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 processes associated with certain human pathologies including osteoporosis and vascular calcification. *E. huxleyi* was selected for sequencing by the Department of Energy in 1999. Sequencing of the genome has been completed to 12x coverage and the first version was released in November 2006, which yielded 7,809 scaffolds encompassing 165 million base pairs. This particular genome was difficult to sequence, in large part due to its high GC content (~66%), relatively high rate of sequence polymorphism, and the repetitive nature of the genome. In addition to the genome sequence over 80,000 EST and cDNA sequences have been compiled under varying cellular conditions, giving us a vast amount of data available for computational mining and analysis. While the annotation of the *E. huxleyi* genome is still in progress, we extracted 3,132 potential core promoters and analyzed them for sequence motifs and physical properties in order to construct a profile of core-promoter regions for future studies.

2. Promoter Extraction

We are interested in the core-promoter regions, defined as the region extending approximately 50 to 100 base pair (bp) upstream and downstream from the transcriptional start sites (TSS). Conventional experimental methods have not yet been used to map TSSs to the *E. huxleyi* genome. Hence, we used computational methods to conservatively determine potential core-promoter regions. To locate the approximate TSSs with high confidence, and extract regions flanking this position, we aligned 190 full-length cDNA and 84,385 EST sequences to the assembled *E. huxleyi* genome and performed the following steps:

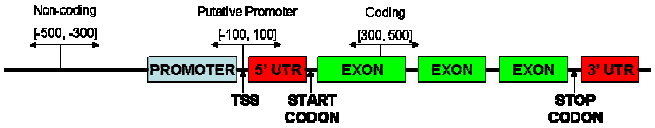
Extraction Steps

1. **Filter ESTs** – We only chose ESTs that provide information about TSS with high confidence, i.e. only those that exhibited significant homology to the amino terminus of known genes. This was done by aligning the EST sequences to the NCBI NR database using the BLAST algorithm. Only EST sequences found to have significant hits (e-value < 1xe-20) and within 5 amino acids from the subject start were used in subsequent steps.

2. **Genome Alignment** – The cDNA sequences were combined with the filtered EST sequences for a total of 11,868 sequences. These were aligned to the genome using two different algorithms, BLAT and SPLIGN, to provide cross-validation and generate the promoter regions with higher confidence. These algorithms were selected for their ability to incorporate splice signals into their alignments. Only alignments that were found to "match" between the algorithms were used for extraction. Alignments were considered a "match" if both their subject strand and the query sequence were identical, and the start and end coordinates were within 5 bp of each other.

3. **Remove Redundancies** – Alignments were sorted by genome strand and starting position. Alignments with a starting coordinate within 500 bp of the previous alignment were removed.

4. **Promoter Extraction** – Using the starting coordinate for each alignment as the putative TSS we extracted the region 100 bp upstream and 100 bp downstream as the putative core-promoter regions. Only promoter regions that had the full length of 200 bp and contained no more that ten unknown nucleotides were kept. In the end we obtained a final dataset of 3,132 putative core-promoters. In addition we extracted background datasets consisting of 200 bp sequences from both the non-coding intergenic and the coding regions of the genome. The non-coding dataset was extracted from -500 to -300 relative to the TSS, and the coding dataset from 300 to 500 relative to the TSS.



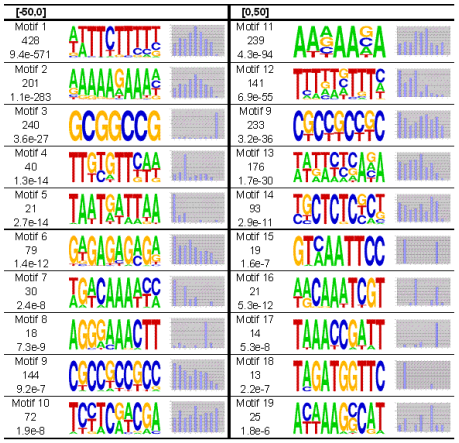
3. Motif Discovery

Core promoters are the recognition sites for the construction of the pre-initiation or the basal transcription unit complex which includes RNA polymerase II and several general 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. 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 is reported to be present in 30-40% of core promoters in many organisms.

To determine if *E. huxleyi* shares any known motifs with other organisms we selected position-specific scoring matrices from studies done on *Drosophila* and *Arabidopsis* promoters. Searches were performed using MAST and *E. huxleyi* promoter regions with nucleotide frequencies of the genome servering as the background parameter.

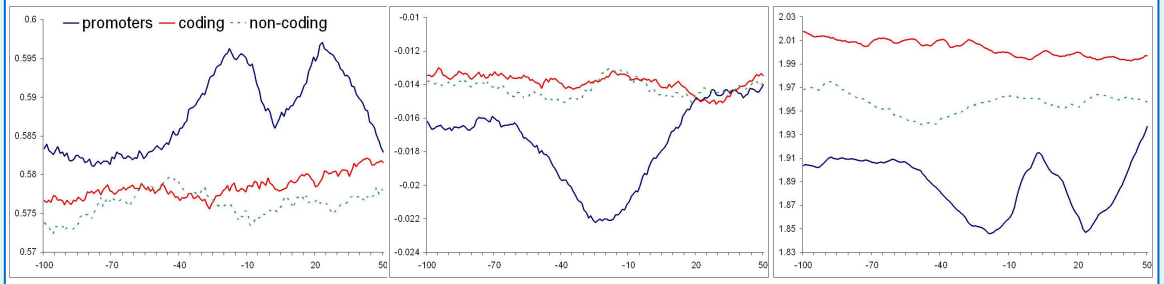
No significant matches to the *Drosophila* motifs were found. Although the *Drosophila* TATA box occurred at a frequency of 6.64% in the *E. huxleyi* promoter regions, this was not significantly different from the frequency in the background datasets. Several significant matches, however, were found to *Arabidopsis* motifs. This is not surprising since *E. huxleyi* is evolutionarily closer to *Arabidopsis* than to *Drosophila*. The motif most closely resembling the TATA box in *Arabidopsis* was found in 6.70% of the *E. huxleyi* promoters, but again this was not significant in comparison to the background models.

A search for novel motifs over-represented in the *E. huxleyi* promoter regions was performed using MEME, an expectation maximization algorithm. The table here shows the top ten motifs with lengths between 5 and 10 bp, using the zero or one occurrence per sequence model, in both the [-50, 0] and [0, 50] regions flanking the putative TSS. The numbers under the motif names represent the number of times the motif was found in the 3,132 promoter regions and it's expectation value with respect to the background nucleotide frequencies. The histograms indicate the position where the motifs were found within the region searched. While the MEME algorithm did not detect anything resembling a TATA box, motifs were detected that are very similar to motifs found in Arabidopsis. These motifs, 1, 2, and 6, are found in the [-50, 0] region. Novel motif 3 in particular is very interesting. It is a perfect pattern with a strong positional preference that is found in 7.7% of the promoters. It is, however, also found in the background datasets and could be due to the high GC content of *E. huxleyi* and will need to be investigated further.



4. Physical Property Profiles

Physical and structural properties of DNA are known to significantly influence gene regulation. Previous studies have also shown core promoter regions exhibit distinct and structural properties. For this study property profiles were constructed for *E. huxleyi* promoter regions by first aligning the putative TSSs. DNA sequences were then converted to numeric profiles using physical DNA property coefficients, and then smoothing with a mean value filter using a window size of 21. Average profiles for each property were created by computing the average at every position for all sequences. Although the putative *E. huxleyi* TSSs are based on prediction only, out of 14 properties analyzed, all but two showed well-defined patterns when compared to the background coding and non-coding datasets. Example patterns are shown below for three properties.



A-philicity – 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 transistion.

Bendability – based on Dnase I cutting frequencies, regions with a high peak correspond to regions that are more flexible that regions with a low peak.

Tri-nucleotide GC content – ratio of GC content per codon.

5. Conclusions and Future Work

This study provides a first genome-wide analysis of the composition and structure of the *E. huxleyi* core promoters. The comparison to known motifs showed strong correlation to several motifs found in *Arabidopsis*. While these particular motifs in *Arabidopsis* resemble microsatellite sequences with identifiable repeat units, the motifs in *E. huxleyi* do not. Additionally there are several other novel motifs significantly over represented in the *E. huxleyi* core promoters, suggesting the algae may contain novel transcriptional control sequences.

The results of the motif search and the physical property profiles are currently being used to train a support vector machine (SVM) for classification purposes. Results from the SVM will be used to develop a classification program that will use the property profile models together with the motifs, to more accurately predict core-promoter regions in *E. huxleyi*.