Lawrence Berkeley National Laboratory
Recent Work

Title
Over/Under-Representation of sigma-70 promoter-like signals in Different Genomic Regions

Permalink
https://escholarship.org/uc/item/9p03t37n

Authors
Froula, Jeff
Francino, Maria Pilar
Huerta, Araceli H.

Publication Date
2005-06-15
Over and Under-Representation of Sigma-70 Promoter-like Signals in Different Genomic Regions

Jeff Fronek, Araceli H. Huerta, M. Pilar Francino
Department of Evolutionary Genomics, Joint Genome Institute, Walnut Creek, CA.

Background

It has recently been shown [Huerta and Collado-Vides, 2003] that regulatory regions in E. coli contain high-densities of potential interaction sites for sigma 70, the principal sigma transcription factor which binds to the -10 (TATAAT) and -35 (TTGACA) promoter sequences and is essential for general transcription in exponentially growing bacterial cells [Hawley & McReynolds, 1981; Harley & Reynolds, 1987; Luise & Margalit, 1993; Graber & Gross, 2001]. In contrast, coding regions and regions located between convergently-transcribed genes do not harbor an overrepresentation of potential interaction sites for sigma 70, the principal sigma transcription factor which binds to the -10 (TATAAT) and -35 (TTGACA) promoter sequences and is essential for general transcription in exponentially growing bacterial cells [Hawley & McReynolds, 1981; Harley & Reynolds, 1987; Luise & Margalit, 1993; Graber & Gross, 2001]. These potential sigma 70 sites do not represent identical sequences, but rather form clusters containing different variations of the -10 and -35 consensus. It is currently believed that most of these sites are non-functional and are more accurately termed promoter-like sequences. Functional promoter sites identified experimentally are often found in the subregions of highest density of overlapping signals, even when individual sites have higher binding affinity for RNA polymerase than found in the subregions of highest density of overlapping signals within the region [Reznikoff, 1992; Huerta & Collado-Vides, 2003]. Using a prediction algorithm by [Huerta and Collado-Vides, 2003] we obtained sequence for 522 potential sigma 70 binding sites.

Objective

To help determine if the promoter-like sequences have a biological function, we have shown that these sites are over-represented in regulatory DNA beyond the expectations based on nucleotide composition. Additionally these sites are under-represented in other parts of the genome, possibly to prevent unwanted RNA polymerase binding. In estimating the occurrence of these sites, we used the hexamers corresponding to all the variations of the -10 box consensus present in the E. coli genome. From the 522 putative binding sites, we got 185 unique hexamers. These hexamers have been shown to be the most essential part of the sigma-70 binding site. We will compare the expected and observed binding site frequencies for each of the four categories of sequence: coding, regulatory, non-regulatory, and as a control, totally random DNA.

Methodology & Results: Statistical Evaluation of Over/Under Representation of Promoter-like Sequences

Generating Random DNA Permutations to get Expectation

To determine whether or not a set of promoter-like sequences were over/under represented in different parts of the genome, we needed to compare the observed frequency with the expected. To acquire neutral expectations for the hexamer frequencies we counted each word occurrence on sequence that had been shuffled in such a manner that it maintained the original nucleotide bias as the original sequence. We used a strategy described by Altschul et al. [1990] that would take a seed sequence and randomize it while still maintaining the mono-, di- and trinucleotide bias. We wrote an algorithm in Perl designed around the strategy outlined by Altschul et al. which essentially finds all the possible mono- to trinucleotide combinations within the original sequence and then randomly shuffles them before reassembling them into a new sequence of the same length.

Calculating Significance of Promoter Representation

All 185 promoter-like sequences were counted in each of the 1000 shuffled sequences. A total sum was obtained for each of the 1000 sequences and then sorted from lowest to highest. The values excluding only the top and bottom 2.5% of the list could be identified and compared to the sum obtained from the original, un-shuffled, sequence. If the real value was outside the 2.5% confidence limits (blue lines), it was considered significant. Confidence limits could be established by sorting the 1000 sums and finding the top and bottom 2.5% values.

How We Made the Histograms

Calculating Significance of Promoter Representation

Average Density of Promoter-like Signals in Coding and Non-Coding Regions of the E. coli Genome

This increased density of promoter-like sites 400 bp upstream of a gene start is illustrated in this figure. "Regulatory" regions are 500 bases upstream and 500 bases downstream of the gene start site (position 0). "Coding" regions contain sizes above 1kb; from these genes, the middle point was taken as the position 0 and 500 bases upstream and 500 bases downstream of this position were included. "Convergent" regions include the region between conversely transcribed genes (on opposite strands) and is considered a non-regulatory region since it is not upstream of any gene. This region is defined here as: the end of the 3' gene position 0 and 500 bases upstream and 500 bases downstream of this position were included.

The promoter-like sites were counted by PATSER [Hertz & Stormo, 1999]. This figure was pulled from [Huerta and Collado-Vides, 2003].

Showing Over/Under Representation Graphically

Each of 185 promoter-like sequences were counted in each category and summed. A summation was also made for each of the 1000 shuffled sequences for the same four categories. A histogram was made from the 1000 summation values and the real count was placed onto the histogram for comparison. If the real value was outside the 2.5% confidence limits (blue lines), it was considered significant.

Literature Cited


