An algorithm for finding regulatory sequences of homologous genes

Pavesi G.(1), Stefani M.(2), Mauri G.(2), Pesole G.(3)

(1) D.I.Co., University of Milano, Milano (2) Dept. of Computer Science, University of Milano-BIcocca, Milano (3) Dept. of Biomolecular Science and Biotechnology, University of Milano, Milano

Motivation

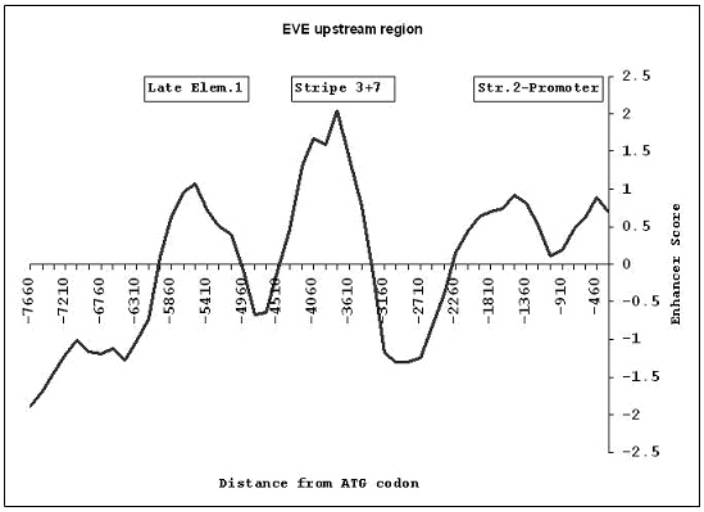
One of the greatest challenges in modern molecular biology is the identification and characterization of the functional elements regulating gene expression. Two of the most important elements are transcription factors (TFs), and the sites of the genome where they can bind (TFBSs). The TF-DNA interactions, that are responsible for the modulation of gene transcription, are at the basis of many critical cellular processes, and their malfunction often involves the onset of genetic diseases. TFBSs are located either near the transcription start site of a gene (usually within 500-1000 bps), or alternatively at very large distance (often several kilobases) from it, either upstream or downstream. When the regulation of a single gene is investigated, the idea is to increase the signal/noise ratio by comparing its flanking regions (upstream and/or downstream) with homologous genome regions of the same or other organisms at different evolutionary distances. Those parts of the regions that are more conserved throughout the different species are more likely to have been preserved by evolution for their function, and thus could be (or contain) TFBSs. Most of the methods introduced so far first build a global alignment of the sequences (some pairwise, some multiple), and report the most conserved parts of the alignment (with or without further processing, for examples by looking for known TFBSs instances in them). While this approach can produce good results, since a highly conserved region can be a good candidate for a regulatory activity, some experiments have shown that real TFBSs are often mis-aligned, and fall outside the “best regions” of the alignment (that, anyway, becomes computationally problematic for long regions, especially in the case of multiple comparisons). In this work we present an algorithm that does not require a global alignment of the sequences, nor needs to be supported by matrices or instances of known TFBSs in order to detect potential regulatory motifs.

Methods

The general idea of our method is based on three considerations: (1) single TFBSs should be conserved both in sequence and in position, that is, in the different organisms they should not have drifted too far apart in their position relative to the gene; (2) single TFBSs often are adjacent or overlapping; (3) when at long distance from a gene, TFBSs are clustered together to form enhancers, silencers, and so on; (4) finally, TFBSs should not be too frequent, since their presence in just a subset of genes determines tissue-specific expression or response to specific stimuli. The algorithm works on different levels, i.e., from the discovery of single TFBSs to the detection of full enhancers (and the conserved TFBSs therein), All in all, the algorithm, given a reference sequence (that can be several kilobases long) and one or more homologous sequences outputs: the oligos of the reference sequence with higher likelihood, given their conservation, of being instances of conserved TFBS and the regions of the reference sequence with the higher density of TFBSs, likely to represent regulatory regions (e.g. enhancers)

Results

We show how the algorithm is able to detect conserved TFBSs as well as more complex regulatory regions by examining the regions up- and downstream of genes whose regulation has been thoroughly studied and characterized. For example, the algorithm was able to discover the annotated binding sites in the human p53 core promoter (without false positives, by comparing it to the mouse and rat orthologous promoters), or the three regulatory regions upstream from the eve gene of D.melanogaster (late element 1, stripe 2, and stripe 3+7, by using D.virilis, D. erecta, and D. pseudobscura), together with the annotated TFBSs within each element. The figure shows an example of the "enhancer score" computed by the algorithm for this case: the peaks (corresponding to positive values) correspond to the three known enhancer regions, plus the core promoter.



Contact email: pavesi@dico.unimi.it