Genome analysis

PEAKS: identification of regulatory motifs by their position in DNA sequences

Nicolàs Bellora1, Domènec Farre2 and M. Mar Albà1,3,*

1Research Unit on Biomedical Informatics, Universitat Pompeu Fabra, 2Centre for Genomic Regulation and 3Catalan Institution for Research and Advanced Studies—Municipal Institute of Medical Research, Barcelona 08003, Spain

Received on July 4, 2006; revised on October 18, 2006; accepted on November 5, 2006
Advance Access publication November 10, 2006
Associate Editor: David Rocke

ABSTRACT

Summary: Many DNA functional motifs tend to accumulate or cluster at specific gene locations. These locations can be detected, in a group of gene sequences, as high frequency ‘peaks’ with respect to a reference position, such as the transcription start site (TSS). We have developed a web tool for the identification of regions containing significant motif peaks. We show, by using different yeast gene datasets, that peak regions are strongly enriched in experimentally-validated motifs and contain potentially important novel motifs.

Availability: http://genomics.imim.es/peaks
Contact: malba@imim.es
Supplementary information: Supplementary Data are available at Bioinformatics online.

The identification of regulatory motifs in DNA sequences is a challenging problem in bioinformatics. Computational predictions of known motifs, such as transcription factor binding sites (TFBS) often contain an unacceptable number of false positives, due to the short size and variability of the motifs. Focusing on motifs that are shared by several sequences can increase the specificity of motif predictions. For example, one can select sequences that have been conserved during evolution, a strategy known as phylogenetic footprinting (Lenhard et al., 2003). A different type of evolutionary constraint is related to the position of motifs along the gene sequence. There is ample evidence that many gene expression regulatory motifs show a biased location within promoter sequences (FitzGerald et al., 2004; Xie et al., 2005). That is, they are not randomly distributed but tend to accumulate or cluster in particular regions, forming high abundance ‘peaks’. This presumably reflects specific requirements of motif-binding proteins that need to interact with each other to regulate transcription. The identification of significant motif peaks can be used to increase the specificity of motif predictions, provide information on the promoter structure, and help discover regulatory motifs that are specifically involved in the regulation of genes with similar expression or function. Motivated by the lack of available computational methods to detect motif clustering we have developed a novel algorithm for this purpose, which we have termed ‘positional footprinting’ and which is implemented in the web server PEAKS.

PEAKS can be used to analyze any group of sequences that share a known reference element, such as the transcription start site (TSS), the initiation codon, a known TFBS or any other predefined site. The scope is to detect any other motifs that show a significant clustering at a particular distance from the reference element. In the first step of the procedure the sequence positions that show matches to motifs from a user-selected library are recorded. Available motif libraries are: (1) compilations of TFBS position-specific weight matrices (PSWMs), (2) all possible DNA words of a given length or (3) pre-built consensus motif collections (Zhu and Zhang, 1999; Harbison et al., 2004). Several PSWM libraries can be used: TRANSFAC (Matys et al., 2003), Jaspar (Sandelin et al., 2004) and PROMO (Messeguer et al., 2002). Using DNA words can aid in the discovery of putative new motifs in different types of DNA sequences. In the second step, the positions of predicted motifs are used to build motif frequency profiles along the sequences. A position is considered positive for a motif if the motif occurs within a sequence window surrounding that position. Increasing the window size above the default value (31) allows the detection of motifs that do no have a very precise location at the cost of decreasing the significance of motifs located at very well defined positions (see Supplementary Table S1 for a full list of program parameters). The third step is the calculation of the positional footprinting score, $S_{Spf}$, which measures the relative over-representation of a motif at a particular position (see PEAKS web server for a full mathematical description). The fourth step is the statistical evaluation of the maximum $S_{Spf}$ score obtained for each motif. To this end, we apply the same procedure described above to simulated random sequence datasets, which can be generated using an order 1 Markov model, to obtain an empirical p-value associated with the maximum $S_{Spf}$ score. If significant, we extract any other positions with a $S_{Spf}$ score above the p-value cut-off, which define the motif significant regions. The output includes a graphical representation of all the significant motifs and regions, a list of sequences containing significant motifs, motif profile pictures and a summary table.

Figure 1 shows the output produced by PEAKS in a dataset of 180 yeast genes involved in ribosome biogenesis (Mewes et al., 2002). Sequences spanned from –500 to +100 with respect to the most used TSS (Zhang and Dietrich, 2005). Motifs were detected using exact matches to a consensus motif collection containing 102 different TFBS (Harbison et al., 2004), and a sliding window of 31 nucleotides. An integrated picture (Fig. 1A) was derived from the significant regions in the profiles at $p$-value $< 1 e^{-3}$ (Fig. 1B).
Five of the seven significant motifs, Fhl1, Rap1, Sfp1, Abf1 and Reb1, are known to be involved in the regulation of ribosomal-related genes (Fig. 1C). Yox1 and Skn7, have, so far, not been associated with this function, but their distribution indicates that they are strong candidates. We calculated the ratio between the observed fraction of experimentally-validated motifs falling into a significant region and the fraction of motifs expected in this region under a random motif distribution (size of the significant region divided by the total length of the sequence). The enrichment in real motifs ranged from 2.06 for Fhl1 to 10.67 for Skn7 (Fig. 1C). New putative binding sites for these transcription factors were discovered. For example, among the 24 different Skn7 motifs in the significant region (−239 to −215) only four were previously known. A second example, using a dataset of 86 yeast genes involved in amino acid metabolism, is provided in Supplementary Figure 1S.

ACKNOWLEDGEMENTS

The authors thank Loris Mularoni, Eduardo Eyra, Robert Castelo and Oscar González for useful discussions during this work. The authors acknowledge support from Fundación Banco Bilbao Vizcaya Argentaria (FBBVA), Plan Nacional de I + D MCYT (BIO2002-04426-C02-01), EC Infobiomed NoE and Fundació ICREA.

Conflict of Interest: none declared.

REFERENCES


