BIOINFORMATICS APPLICATIONS NOTE
Vol. 25 no. 4 2009, pages 550–551
doi:10.1093/bioinformatics/btp013

Gene expression

MOTIFATOR: detection and characterization of regulatory motifs using prokaryote transcriptome data

Evert-Jan Blom[1], Jos B. T. M. Roerdink[2], Oscar P. Kuipers[1,*] and Sacha A. F. T. van Hijum[1,*,†]

[1]Molecular Genetics, Groningen Biomolecular Sciences, University of Groningen, PO Box 800, 9700 AV and [2]Institute for Mathematics and Computing Science, University of Groningen, Nijenborgh 9, 9747 AG, Groningen, The Netherlands

Received on September 11, 2008; revised on November 19, 2008; accepted on January 1, 2009

Advance Access publication January 25, 2009

Associate Editor: Joaquin Dopazo

ABSTRACT

Summary: Unraveling regulatory mechanisms (e.g. identification of motifs in *cis*-regulatory regions) remains a major challenge in the analysis of transcriptome experiments. Existing applications identify putative motifs from gene lists obtained at rather arbitrary cutoff and require additional manual processing steps. Our standalone application MOTIFATOR identifies the most optimal parameters for motif discovery and creates an interactive visualization of the results. Discovered putative motifs are functionally characterized, thereby providing valuable insight in the biological processes that could be controlled by the motif.

Availability: MOTIFATOR is freely available at http://www.motifator.nl

Contact: o.p.kuipers@rug.nl; sacha.vanhijum@nizo.nl

1 INTRODUCTION

Transcription factor DNA binding sites (TFBSs) are typically determined in the upstream regions of genes exhibiting similar expression patterns. Information concerning the genomic context, positional distribution, other TFBS occurrences and functional annotations provide important clues concerning the *modus operandi* of a TFBS. Existing applications that focus on motif discovery from DNA microarray data (e.g. Elemento *et al.*, 2007; Gordon *et al.*, 2005) base the actual motif discovery on gene lists obtained at seemingly arbitrary cutoffs. If for partitioning only one cutoff is used, relevant overrepresented motifs could be missed when the partitioning is too strict. Alternatively, when the partitioning is too relaxed either non-relevant motifs could be considered significant or motifs could again be missed due to 'pollution' of the gene fraction. Furthermore, a manual step of biological interpretation of the motif is required. Ideally, tuning the partitioning of genes, determining additional occurrences of the identified motifs in the remaining genes, comparing the putative motifs to known TFBSs and functional characterization of the putative motif should be automatically performed. To this end we have developed MOTIFATOR, an application that predicts and characterizes *cis*-regulatory motifs for differentially expressed genes in transcriptome studies for prokaryotes.

*To whom correspondence should be addressed.
†Present address: NIZO Food Research, PO Box 20, 6710 BA Ede, The Netherlands.

2 SOFTWARE OVERVIEW

2.1 The application and input data

MOTIFATOR is standalone and programmed in JAVA. It requires transcriptome data, genome annotation files (EMBL or Genbank) and optional annotation sources: (i) Gene Ontology; (ii) metabolic pathway; (iii) regulons; (iv) UniProt keywords; (v) InterPro domains; and (vi) user-defined functional categories.

2.2 Processing

2.2.1 Partitioning of gene expression data Instead of choosing an arbitrary cutoff for determining differentially expressed genes, our application tries multiple cutoffs to partition genes into up and down classes. In addition, expression data are partitioned into combined up-and downregulated fractions which can provide relevant insights for genes that are both repressed and activated by one regulator.

2.2.2 Motif discovery and characterization Each fraction of genes is examined for overrepresented motifs by the SCOPE method (Chakravarty *et al.*, 2007). Key aspects of SCOPE are high sensitivity and specificity, while requiring a minimum of parameters for motif detection. The putative motifs are characterized as follows:

- Additional occurrences: to search for additional occurrences of the putative motif in the entire genome sequence, position-specific scoring matrices are created from the putative motifs.
- Functional characterization of putative motifs: genes bearing putative motifs are analyzed using a functional enrichment analysis from the FIVA software (Blom *et al.*, 2007), which might point to a biological function of the putative motif.
- Matching putative to known TFBSs: this approach allows the user to focus on (i) a rapid identification of new putative motifs and (ii) the extension of existing transcriptional modules.

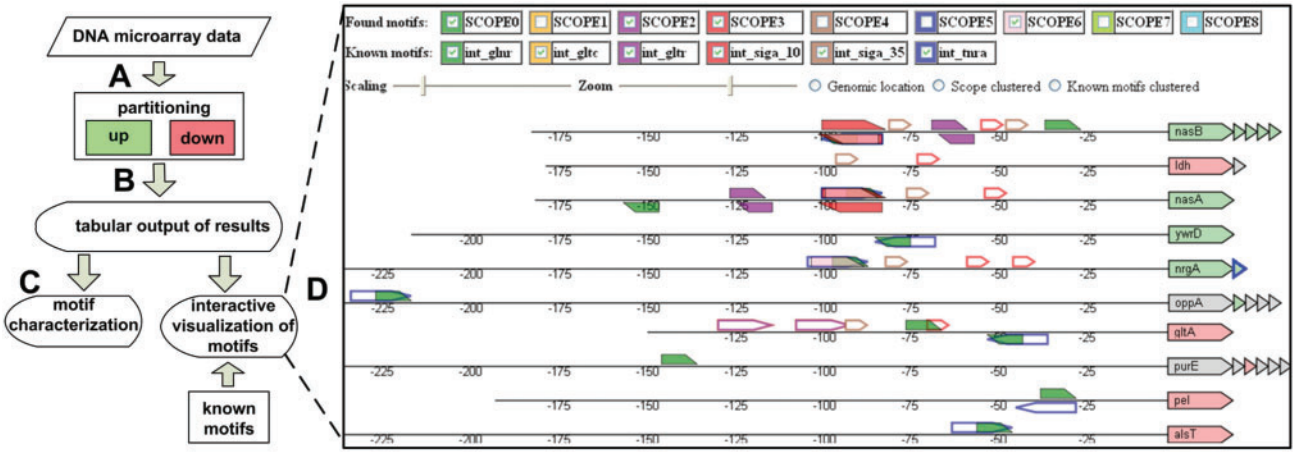

Fig. 1. Flow diagram of MOTIFATOR and an example of the visualization. (**A**) Partitioning of genes from the DNA microarray dataset. (**B**) Tabulated results of the partitioning. (**C**) More detailed analysis from the tabular display: (i) functional enrichment of putative motifs, (ii) identification of genes whose *cis*-regulatory region contain motifs similar to the putative motifs, (iii) matching of the putative motifs with known motifs and (iv) an interactive visualization. (**D**) An interactive visualization of the genomic context of putative and known motifs in the upstream sequences of transcriptional units. It is based on a dataset from Yoshida *et al.* (2003) who used DNA microarrays to identify a TFBS for a global regulator. Yoshida *et al*. reported many negatively regulated TnrA targets as well as positively regulated targets. Consequently, the most relevant results from MOTIFATOR are in the combined up and down fraction. Filled polygons: putative motifs. Open polygons: known binding sites. Genes coding for a putative regulator are colored blue.

2.3 Output

Results from each partitioning analysis are summarized into a table. Several filtering options allow user-defined prioritization of the most optimal analysis based on these filtering criteria. Identified motifs are described in depth in HTML files. An interactive SVG visualization (Fig. 1) serves multiple purposes: (i) representation of the genomic context of the putative and known motifs; (ii) complementing the sequence matching of putative motifs (see above). New motifs similar to known motifs will most likely overlap in the visualization; and (iii) facilitate determining, specific orientations and/or spacings of (combinations of) motifs pointing to more complex regulation of gene expression. More information concerning the visualization capabilities of our application can be found on the supplementary website.

3 CONCLUSIONS

MOTIFATOR offers the following advantages over existing tools:

- ready-to-use databases for over 600 prokaryotic organisms;
- sensitive *de novo* detection of motifs;
- functional enrichment analysis of putative motifs;
- matching of putative motifs with known motifs;
- standalone application allows mining of sensitive data; and
- interactive visualization includes genomic context.

The visualization facilitates determining patterns of combinations and/or spacing of motifs that might point to more complex gene regulation events. The supplementary website lists analysis results of MOTIFATOR applied to several transcriptomics datasets from the KEGG expression database.

Funding Netherlands Organisation for Scientific Research (to E.J.B.); NWO-BMI project 050.50.206 (to E.J.B.); IOP Genomics (IGE03002B to E.J.B.). FW6, Bacell Health project (LSHC-CT-2004-503468 to O.P.K.).

Conflict of Interest: none declared.

REFERENCES

Blom,E.-J. *et al*. (2007) Fiva: Functional information viewer and analyzer extracting biological knowledge from transcriptome data of prokaryotes. *Bioinformatics*, **23**, 1161–1163.

Chakravarty,A. *et al*. (2007) A novel ensemble learning method for de novo computational identification of DNA binding sites. *BMC Bioinformatics*, **8**, 249.

Elemento,O. *et al*. (2007) A universal framework for regulatory element discovery across all genomes and data types. *Mol. Cell*, **28**, 337–350.

Gordon,D.B. *et al*. (2005) TAMO: a flexible, object-oriented framework for analyzing transcriptional regulation using DNA-sequence motifs. *Bioinformatics*, **21**, 3164–3165.

Yoshida,K. *et al*. (2003) Identification of additional TnrA-regulated genes of *Bacillus Subtilis* associated with a TnrA box. *Mol. Microbiol.*, **49**, 157–165.