BIOINFORMATICS APPLICATIONS NOTE
Vol. 23 no. 4 2007, pages 493–495
doi:10.1093/bioinformatics/btl607

Genome analysis

MotifScorer: using a compendium of microarrays to identify regulatory motifs

Matteo Brilli[1], Renato Fani[1] and Pietro Lió[2,*]

[1]Dipartimento di Biologia Animale e Genetica, via Romana 17, 50125 Firenze, Italy and [2]Computer Laboratory, University of Cambridge, Cambridge CB3 0FD, UK

Received on September 19, 2006; revised and accepted November 23, 2006
Advance Access publication November 30, 2006
Associate Editor: Martin Bishop

ABSTRACT

Summary: We describe MotifScorer, a program for systematic genome-wide identification of transcription sites. The program uses a compendium of gene expression microarrays and implements state-of-art partial least squares (PLSs) based regression and stepwise regression procedures. Candidate motifs from the upstream sequences of groups of co-regulated genes are identified and assigned a score using genomic background models and available motif finding tools. The use of a large library of expression data allows statistical comparative analysis of the specificity of motifs identified in different conditions.
Availability: MotifScorer, which is written in Java and Matlab, manual and example files are available from the authors.
Contact: pl219@cam.ac.uk

1 INTRODUCTION

The identification of the repertoire of regulatory elements in a genome is one of the major challenges in modern biology. Motifs are short and degenerated DNA sequences embedded into large regions of non-coding DNA, generally located upstream of a gene's transcription start site and, through the interaction with specific transcription factors, they modulate the expression patterns of the genes in a genome. Gene expression is usually measured on a genome-wide scale using DNA microarrays, which provide a tool for exploring the regulation of thousands of genes at once. The analysis of expression data allows the identification of co-regulated genes, likely controlled by common regulatory mechanisms. Our work is motivated by the possibility of dissecting the entire regulatory network of the genome of an organism given its genome sequence and a large library of expression datasets.

2 DESCRIPTION

MotifScorer is able to analyze a large number of genes, motif widths and hundreds of experimental conditions. The program is written in Java and regression procedures are written in Matlab. Figure 1a describes the adopted strategy:

(1) Upstream sequences of co-regulated genes are automatically retrieved using http or ftp connections, or local GenBank files.

(2) A motif finding algorithm is used to identify candidate motifs from gene upstream sequences; motifs can be imported from MDscan (Liu *et al*., 2002) and AlignACE (Roth *et al*., 1998). A genome background model based on a 3rd order Markov chain is computed with all automatically extracted intergenic sequences of the genome corresponding to the identifiers used as input. The length of intergenic sequences is user defined, and the program is able to exclude regions overlapped with coding sequences; RepeatMasker has also been implemented. The background model is used to compute the score of each motif.

(3) A score is assigned to the occurrences of each candidate motif; MotifScorer calculates the score of each upstream sequence taking into account the motif's position weight matrix (PWM), the background model and the number of motif occurrences for each sequence. The main scoring function is similar to that in Conlon *et al*. (2003) but others are implemented. Given a motif μ of length w, and occurrences $\mu_i \in \mathbf{X_{w,g}}$ in upstream sequence g; given the corresponding PWM, $\mathbf{M}_\mu := (\mathbf{P_{w,n}})_{\mathbf{w}\times\mathbf{n}}$, where $n \in \{A, C, G, T\}$, and background model $\mathbf{M_B}$, the scoring function is:

$$S_{\mu,g} = \log_2\left[\sum_{\mu_i \in X_{w,g}} \frac{P(\mu_i \mid M_\mu)}{P(\mu_i \mid M_B)}\right],$$

where $P(\mu_i \mid M_B)$ and $P(\mu_i \mid M_\mu)$ are the probabilities of each motif occurrence calculated on the motif's PWM and the background model, respectively; the sum applies over all occurrences of motif μ in sequence g.

(4) regression methods allow to select those motifs acting together to affect the expression of genes from the scores and the expression levels. We have implemented different algorithms for PLS regression, i.e.: PLS-nipals (Abdi, 2003), multilinear PLS (Andersson and Bro, 2000, http://www.models.kvl.dk/source/nwaytoolbox/) and robust PLS (Verboven and Hubert, 2005). The analysis outputs regression coefficients.

(5) Motifs identified by PLS or stepwise regression procedures are compared to identify those motifs acting specifically in different conditions.

The program can also output files containing digraph representations which are readable by open source network visualization programs, such as Visone (www.visone.info). Using a compendium of expression data MotifScorer can produce a motifs—conditions digraph, while using single expression data a network describing the relationships among genes, motifs, expression level and regression results can be produced.

*To whom correspondence should be addressed.

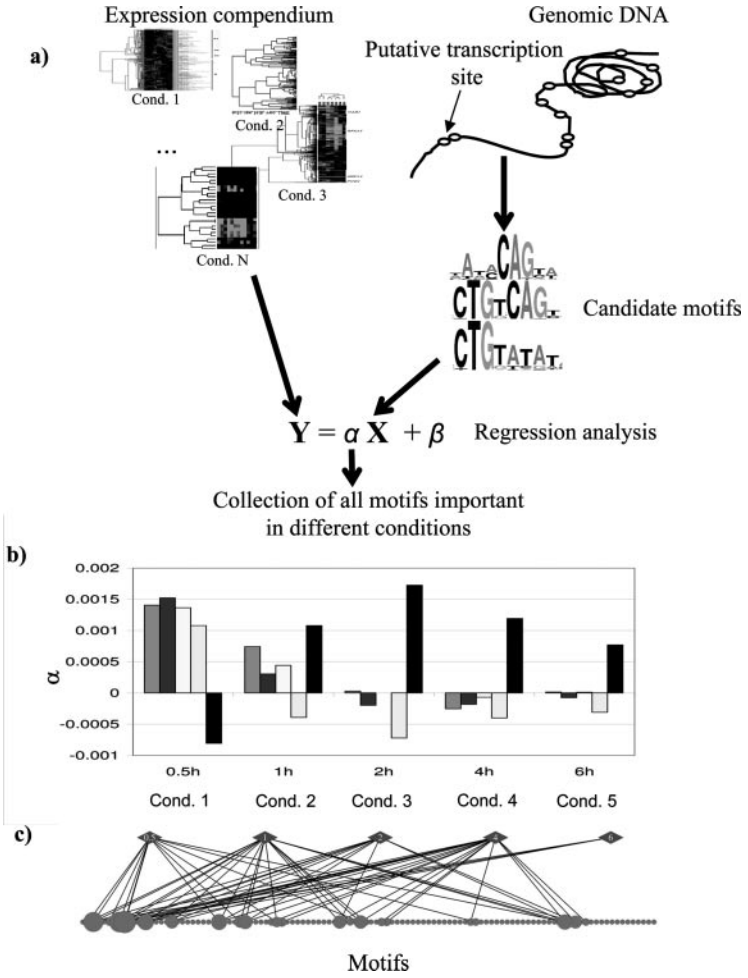

Fig. 1. (**a**) Schematic pipeline for motif searching using MotifScorer; see text for details; (**b**) Histogram showing the changes in PLS regression coefficients of several motifs in a time-series. *y*-axis is the regression coefficient, *x*-axis is the time of amino acid starvation. Motifs are, from left to right (in parenthesis we indicate the retrieved binding site using the consensus in SCPD (Zhu and Zhang, 1999); the number of asterisks indicates the number of mismatches): TCACGTGCACATCA (that includes a MET4p/PHO4p binding site); TGACTC (GCN4p*); ATGAGTC (GCN4p*); GAGTCATTCCGA (BAS1p**); CACCGGTAGAGTCA (unknown); (**c**) Network graph of motifs—conditions: each motif (circles) is linked to condition(s) (diamonds) if its regression coefficient is significantly different from zero.

Compendium of expression data may be composed by data from nearby species such as *Escherichia coli* and *Salmonella typhimurium* or *Schizosaccharomyces japonicus* and *Schizosaccharomyces pombe* [see for instance Gu (2004); Felsenstein (1988)].

3 EXAMPLE

MotifScorer has very flexible input format i.e. reads the output of several motif finding programs, such as MDscan and AlignACE. Currently available tools for regulatory motif identification, such as REDUCE (Roven and Bussemaker, 2003) and MotifRegressor (Conlon *et al*., 2003) are very specialized, therefore several researchers suggest to use two or more algorithms (Tompa *et al*., 2005). PLS regression has good performance with collinear and numerous (comparable to observation number) predictors, while stepwise regression is not suitable in these conditions (Andersson and Bro, 2000; Abdi, 2003). We show in Figure 1b an example of the pipeline implemented in MotifScorer, used in conjunction with the motif finding program MDscan (Liu *et al*., 2002): we downloaded the list of documented GCN4 regulated genes for a total of 287 sequences and we retrieved all the corresponding upstream sequences with MotifScorer. Upstream sequences were used for multiple MDscan runs searching for the 10 top ranking motifs of 5−15 nt; the outputs were used to calculate the score of each upstream sequence using a 3rd order Markov model trained on the full-intergenic set from the yeast genome. In the next step, MotifScorer performed a robust PLS (RSIMPLS) using the entire set of scores as a predictor matrix for expression levels at all the time points from the amino acid and adenine starvation experiment of Gasch *et al*. (2000) which has five time points. In Figure 1b we report the regression coefficients of some of the motifs, and their changes in different conditions. Most motifs found have a consensus related to binding sites of transcription factors involved in amino-acid and nitrogen metabolisms i.e. GCN4p, a leucine zipper transcriptional regulator known to promote the expression of amino acid biosynthetic genes when their availability is limited. We found that

the importance of motifs related to the metabolism of aminoacids decreases in the last time points, when cells experience a general stress response, and slow down their biosynthetic activities to face the adverse conditions.

ACKNOWLEDGEMENTS

The authors thank the BioinfoGRID project which is funded by the EU within the framework of the Sixth Framework Programme for Research and Technological Development (FP6).

Conflict of Interest: none declared.

REFERENCE

Abdi,H. (2003) Partial least squares regression (PLS-regression). In Lewis-Beck,M., Bryman,A. and Futing,T. (eds), *Encyclopedia for Research Methods for the Social Sciences*, Sage, Thousand Oaks, CA, pp. 1–17.

Andersson,C.A. and Bro,R. (2000) The N-way toolbox for MATLAB. *Chemom. Intell. Lab. Syst.*, **52**, 1–4.

Conlon,E.M. *et al.* (2003) Integrating regulatory motif discovery and genome-wide expression analysis. *Proc. Natl Acad. Sci. USA*, **18**, 3339–3344.

Felsenstein,J. (1988) Phylogenies and quantitative characters. *Annu. Rev. Ecol. Syst.*, **19**, 445–471.

Gasch,A.P. *et al.* (2000) Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, **11**, 4241–4257.

Gu,X. (2004) Statistical framework for phylogenetic analysis of expression profiles. *Genetics*, **167**, 531–542.

GuhaThakurta,D. (2006) Computational identification of transcriptional regulatory elements in DNA sequence. *Nucleic Acids Res.*, **34**, 3585–3598.

Liu,X.S., Brutlag,D.L. and Liu,J.S. (2002) An algorithm for finding protein–DNA interaction sites with applications to chromatin immunoprecipitation microarray experiments. *Nat. Biotechnol.*, **20**, 835–839.

Roth,F.P. *et al.* (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, **16**, 939–945.

Roven,C. and Bussemaker,H.J. (2003) REDUCE: an online tool for inferring cis-regulatory elements and transcriptional module activities from microarray data. *Nucleic Acids Res.*, **31**, 3487–3490.

Tompa,M. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.

Verboven,S. and Hubert,M. (2005) LIBRA: a MATLAB Library for Robust Analysis. *Chemom. Intell. Lab. Syst.*, **75**, 127–136.

Zhu,J. and Zhang,M.Q. (1999) SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics*, **15**, 607–611.