

Predicting Transcription Regulatory Mechanisms by Systematic Promoter Analysis

Li-Wei Chang¹, Rakesh Nagarajan², Jeffrey A. Magee³, Jeffrey Milbrandt⁴,
and Gary D. Stormo⁵

Keywords: transcription regulation, mammalian promoter analysis, regulatory network

1 Introduction.

An important part of understanding a biological pathway or a cellular function is to delineate the transcriptional regulatory mechanisms of the genes involved. Key participants in these mechanisms are special transcription factor proteins (TFs) which recognize particular sequences, called TF binding sites in promoters. Computational approaches to identify TF binding sites have been facilitated by large scale expression profiling experiments [1] and sequence analysis of multiple genomes [2]. Although these methods have been successfully applied to simple organisms such as yeast and worm [3, 4], systematic identification of bona fide transcriptional regulators in mammals still remains a challenging problem.

Two important questions which are often encountered when studying transcription regulation are: (1) Find the common transcriptional regulators of a set of correlated genes which are involved in the same biological pathway or in the same cellular process. (2) Find the genes in the genome which may be regulated by one or a set of transcriptional factors. In this study a systematic and statistical approach was taken to answer these questions by establishing an integrated model considering all the promoters and all the characterized TFs in the genome. To implement this approach, a promoter analysis pipeline (PAP) consisting of a set of analysis tools and a graphical web-based user interface was developed.

2 Methods.

PAP includes a data processing pipeline which was assembled using a series of algorithms and data manipulation tools. These applications were used to carry out genome wide promoter analysis using the following steps: 1) Promoter preparation. Promoters of multiple species were collected and repetitive elements in the promoters were masked using the program RepeatMasker. Currently in PAP, a gene's promoter was defined as 10 kilobases (kb) of sequence upstream and 5 kb downstream of the transcription start site. 2) Phylogenetic footprinting. Orthologous genes were identified according to their protein similarities and promoters of orthologous genes were aligned to identify conserved sequences in multiple organisms using the program TBA (Blanchette, Kent et al. 2004). 3) TF binding site identification. TF binding sites were predicted in all the promoters by the program

¹ Department of Biomedical Engineering, Washington University, St. Louis, Missouri 63130, USA. E-mail: lwcl@ural.wustl.edu

² Department of Pathology, Division of Laboratory Medicine, Washington University School of Medicine, 660 South Euclid Avenue, St. Louis, MO 63110, USA. E-mail: rakesh@pathbox.wustl.edu

³ Department of Pathology, Division of Laboratory Medicine, Washington University School of Medicine, 660 South Euclid Avenue, St. Louis, MO 63110, USA. E-mail: mageej@msnotes.wustl.edu

⁴ Department of Pathology, Division of Laboratory Medicine, Washington University School of Medicine, 660 South Euclid Avenue, St. Louis, MO 63110, USA. E-mail: jeff@pathbox.wustl.edu

⁵ Department of Genetics, Washington University School of Medicine, St. Louis, Missouri 63110, USA. E-mail: stormo@genetics.wustl.edu

PATSER (G. Stormo, unpublished) using weight matrix models from TRANSFAC and JASPAR databases. 4) Probabilistic model construction. Based on these predicted sites, the probability score of a particular transcription factor binding the promoter of a particular gene was calculated. For each transcription factor, the probability scores of all the promoters in the genome were then compared to each other and a “rank p-score” was computed for each promoter. Rank p-scores are normalized scores. They may be used to predict the transcription factors that are most likely to regulate a set of genes, or to predict the genes that are mostly likely to be regulated by a set of factors. 5) Statistical significance evaluation. Randomly selected promoters were used to estimate the p-value of observing a given rank p-score in PAP’s analysis by chance. After all these steps were completed, the results of calculations, including promoter alignments, TF binding sites, probability scores, and rank p-scores were stored in a relational database termed Promoter Analysis Pipeline Database (PAPdb).

3 Results and Conclusions.

Currently 43,365 human and mouse promoters and 545 weight matrices of characterized transcription factors have been included in PAP. The accuracy of PAP’s predictions was tested using genes which are known to be regulated by the same TF according to TRANSFAC and using previously identified co-regulated gene clusters collected from the literature. For test sets from TRANSFAC, PAP identified the true factor as one of the top 10% of factors in 79% of the test cases (Figure 1). When PAP was tested using five co-regulated gene sets collected from the literature, 12 out of 15 known TFs were predicted with a very high rank. PAP was also applied to analyze two novel sets of potentially correlated genes identified by previous mRNA expression profiling experiments. PAP found known TFs of these genes as well as other novel TFs. These results show that PAP was able to identify experimentally verified transcriptional regulators reliably and robustly. Therefore, by taking a systematic approach of considering all promoters and all characterized TFs in our model, we were able to make more reliable predictions about the regulation of gene expression in mammalian organisms.

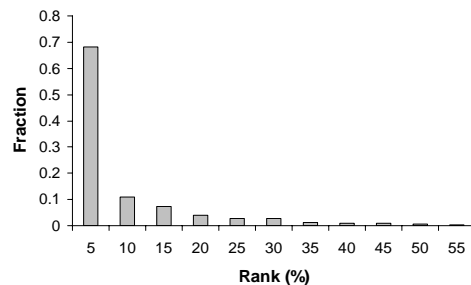


Figure 1: Test results of PAP’s performance using known co-regulated genes in TRANSFAC. PAP identified the true factor as one of the top 10% factors in 79% of the test cases.

References

- [1] Thijs, G., et al., 2002. A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes. *Journal of Computational Biology* 9:447-464.
- [2] Wasserman, W.W., et al., 2000. Human-mouse genome comparisons to locate regulatory sites. *Nature Genetics* 26:225-228.
- [3] GuhaThakurta, D., et al., 2002. Identification of a novel cis-regulatory element involved in the heat shock response in *Caenorhabditis elegans* using microarray gene expression and computational methods. *Genome Research* 12:701-712.
- [4] Hughes, J.D., et al., 2000. Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*. *Journal of Molecular Biology* 296:1205-1214.