

Inferring transcription factor activities using empirical reference distributions

Daniel E. Zak¹, Ronald K. Pearson², Rajanikanth Vadigepalli²,
Gregory E. Gonye², James S. Schwaber², & Babatunde A. Ogunnaike¹ *

¹Department of Chemical Engineering, University of Delaware, Newark, DE 19716

²Department of Pathology, Cell Biology and Anatomy, Thomas Jefferson University, Philadelphia, PA 19107

ABSTRACT

Genome-wide expression profiles are valuable because they provide system-wide views of cellular responses to environmental stimuli. There is often a need, however, to reduce the dimensionality of the data to make it more interpretable. One approach is to combine it with promoter information and bioinformatic tools to infer the transcriptional regulatory elements (TREs) and transcription factors (TFs) that largely govern the variations in gene expression. This approach has been used successfully in yeast and other systems [1]. The present work is concerned with the problem of assigning statistical significance to TF activity predictions made from combined expression profile/promoter analyses.

The basic steps in predicting TF activity from promoters and expression profiles are to (1) generate clusters of genes whose expression profiles are similar by some metric, and (2) identify TREs in the promoter regions of the genes that clustered together. This may be viewed as a process in which a *library* of DNA sequences is passed through an experimental selection process with the objective of creating a subset of sequences that is *enriched* for the TREs regulating the system. *Enriched* in the present context means appearance at a higher frequency than if there had been no selection. This selection is indirect in that subsets of promoters are chosen from a total promoter population using gene expression data clustering. Putative TREs in the subsets may be identified by using bioinformatics tools (for example, MatInspector [6]) that query databases of known TREs (for example, Transfac [5]). Alternatively, the sequences themselves may be searched for enriched patterns using a pattern discovery algorithm [7, 8]. Given that it makes use of the accumulated biological knowledge, we are presently employing the database-driven approach.

Most TREs are short (8 - 25 bp) sequences, and thus the probability of random appearance can be high, necessitating tests for statistical significance of those found in the clustered gene promoters. Individual TREs will often appear repeatedly in a single promoter sequence (as many as 40 times, depending on sequence length), complicating analytical tests for significance. Since multiple appearances of a TRE are

less likely to be caused by random variation, it is nevertheless important to factor this into the test for significance. One group has employed extended hypergeometric distributions [4], but this approach becomes computationally intractable for realistic numbers of TRE/promoter sequence. An alternative approach, employed in the present work, is to construct empirical reference distributions (ERDs) through repeated random re-sampling from the original DNA sequence library in groups of equivalent size as the cluster. The ERDs can then be used to test the null hypothesis that the TREs found in the experimentally-selected group of DNA sequences can be explained by random occurrence. A similar analysis was ultimately performed in [4].

To validate the ERD method, we applied it to the 30 clusters of yeast genes from [8] obtained from yeast cell cycle expression profiles [2]. We used the *Saccharomyces cerevisiae* promoter database (SCPD, [10]) to predict putative TREs 500bp upstream of the start codon in the 3000 ORFs used in [8]. Of the 45 searchable TREs, 20 were significant ($p \leq 0.01$) in at least one of the 30 clusters, compared to 18 for [8] using a method that directly discovers TREs from the promoter sequences [7]. Seventeen clusters had at least one significant TRE, compared to 12 for [8]. With one exception (out of 6), all previously known TREs that were present in SCPD that were identified in [8] were significant ($p = 0$) in their appropriate clusters when ERD was used. Additionally, ERD identified physiologically relevant co-occurrence of TREs that [8] did not. For example, both RAP1 and GCR1, which form a complex to regulate ribosomal genes [3], were significant in the ribosomal gene cluster using ERD, while [8] only identified RAP1. These results demonstrate the utility of the ERD approach. It must be kept in mind, however, that the ERD approach is ultimately complementary to the approach used by [8] and other motif discovery algorithms. While it is unable to identify novel TREs directly from promoter sequences, once a pattern discovery algorithm has defined a potential TRE, it is straightforward to use ERD to evaluate the enrichment of the motif relative to a random collection of sequences.

Our current efforts involve the application of ERD to discover TFs involved in mammalian neuromodulatory processes. Our approach combines the collection of gene expression responses of neuronal cell lines to neuromodulators and promoter/TRE information obtained using the bioinformatic

* Author to whom correspondence should be addressed: Babatunde A. Ogunnaike, ogunnaik@che.udel.edu, (302) 831-4504, FAX: (302) 831-1048.

matic tool PAINT [9]. PAINT takes as input a list of genes (presently restricted to mouse) and compiles upstream regulatory regions from genomic sequence and then identifies TREs using MatInspector. Our preliminary results have identified TREs known to be involved in neuromodulation from the literature, as well as unexpected TFs for the neuronal response to angiotensin that we are confirming experimentally.

Acknowledgments

We thank Praveen Chakravarthula for expert technical support with PAINT, Hester Liu, Dan Miller and Grace Straszewski for obtaining microarray data. We thank NIH/NHLBI, NIH/NIAAA IRPG, NIH/NIGMS BISTI, and DARPA BioCOMP for funding. DEZ additionally thanks the University of Delaware Department of Chemical Engineering for funding.

REFERENCES

- [1] R. B. Altman and S. Raychaudhuri. Whole-genome expression analysis: challenges beyond clustering. *Curr. Opin. Struct. Biol.*, 11(3):340–7, 2001.
- [2] R. J. Cho, M. J. Campbell, E. A. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. G. Wolfsberg, A. E. Gabrielian, D. Landsman, D. J. Lockhart, and R. W. Davis. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell*, 2(1):65–75, 1998.
- [3] S. J. Deminoff and G. M. Santangelo. Rap1p requires Gcr1p and Gcr2p homodimers to activate ribosomal protein and glycolytic genes, respectively. *Genetics*, 158(1):133–43, 2001.
- [4] R. Elkon, C. Linhart, R. Sharan, R. Shamir, and Y. Shiloh. Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. *Genome Res.*, 13(5):773–80, 2003.
- [5] V. Matys, E. Fricke, R. Geffers, E. Gossling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, D. U. Kloos, S. L. S, B. Lewicki-Potapov, H. Michael, R. Munch, I. Reuter, S. Rotert, H. Saxel, M. Scheer, S. Thiele, and E. Wingender. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Rec.*, 31(1):374–8, 2003.
- [6] K. Quandt, K. Frech, H. Karas, E. Wingender, and T. Werner. Matind and matinspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Rec.*, 23(23):4878–84, 1995.
- [7] F. P. Roth, J. D. Hughes, P. W. Estep, and G. M. Church. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, 16(10):939–45, 1998.
- [8] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. *Nat. Genet.*, 22:281–285, 1999.
- [9] R. Vadigepalli, P. Chakravarthula, D. E. Zak, J. S. Schwaber, and G. E. Gonye. PAINT: A promoter analysis and interaction network generation tool for genetic regulatory network identification. *Omic*s, In press.
- [10] J. Zhu and M. Q. Zhang. SCPD: a promoter database of the yeast *Saccharomyces cerevisiae*. *Bioinformatics*, 15(7–8):607–11, 1999.