

Assigning Significance to the Appearance of Regulatory Elements in Experimentally Defined Collections of DNA Sequences Using Empirical Reference Distributions

Daniel E. Zak¹, Ronald K. Pearson², Rajanikanth Vadigepalli², Gregory E. Gonye²
James S. Schwaber², and Babatunde A. Ogunnaik¹

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

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

Key words: Gene networks, transcriptional regulation
regulatory elements, empirical distributions

Prepared for Presentation at the 2003 Annual Meeting, San Francisco, CA, Nov. 16–21
Copyright ©2003, D.E. Zak and B.A. Ogunnaik, University of Delaware
R.K. Pearson, R. Vadigepalli, G.E. Gonye, and J.S. Schwaber, Thomas Jefferson University

November 2003

Unpublished

AIChE shall not be responsible for statements or opinions contained in papers or in its
publications.

¹Author to whom correspondence should be addressed: ogunnaik@che.udel.edu

The present work is concerned with assigning statistical significance to the frequency of appearance of pre-defined patterns in any collection of DNA sequences. This problem is closely tied to several key challenges in functional genomics, including the system-wide localization of transcriptional regulatory elements (TREs) in gene promoters, and the quantification of the activity of the transcription factors (TF's) that bind to the TREs under various cellular conditions. Progress on these challenges lays a foundation for system-wide dynamic modeling of gene expression. Through genome-wide location analysis [7] and other biochemical methods, several groups have laid such foundations in *E. coli* [9], yeast [4], and sea urchin development [2]. While the problem of TRE significance did not emerge in the approaches taken by the above groups, it does emerge in alternative approaches that are widely in use. Examples include promoter analysis of co-regulated genes [10], and TRE activity profiling in nuclear extracts (Regulatory Element Activity Profile, [REAP] data: Genpathway, Inc.). The present work includes a description of the general problem of TRE significance and the empirical reference distribution (ERD) – based approach we have taken to resolve it. Results from applying the ERD approach to promoters of co-regulated genes and REAP data are also presented. The ERD approach appears promising in that it has led to the identification of TREs that could be verified in the literature, as well as a prioritized list of TREs to be verified by alternative methods.

Several functional genomic methods involve passing a library of DNA sequences through an experimental or analytical selection process with the objective of creating a subset of sequences that is enriched for the TREs regulating the system. We use the term enriched to mean appearance at a higher frequency than if there had been no selection. The selection may be direct, as in the case of REAP data (Genpathway, Inc.), or indirect, as in the case where subsets of promoters are selected from a total promoter population using gene expression data clustering. Given a population of DNA sequences, putative TREs may be identified by using bioinformatics tools (for example, MatInspector [6]) that will query databases of TREs (for example, TransFac [5]). Alternatively, the sequences themselves may be searched for enriched patterns using a pattern discovery algorithm [10]. Given that it makes use of the accumulated biological knowledge, we are presently employing the database-driven approach. Since most TREs are short (8 – 25 bp) sequences, the probability of random appearance can be high, thereby necessitating tests for statistical significance.

It is straightforward to characterize the probability of random occurrence of a given TRE in theoretical random DNA (i.e., randomly generated sequences of the four possible bases, each occurring with equal probability). This probability is more difficult to characterize in highly structured random genomic DNA sequences (e.g., in sequences randomly drawn from the set of all potential promoter sequences from a real genome), however, because

the underlying biases may not be known. The ERD method overcomes this difficulty by empirically constructing reference distributions through random re-sampling from the original DNA sequence library. These reference distributions are then used to test the null hypothesis that the TREs found in the experimentally-selected group of DNA sequences can be explained by random occurrence. The TREs may then be prioritized for further study by $(1 - p)$, where p is the hypothesis test p -value. Given that the ERD method empirically constructs reference distributions, no distributional assumptions are made. Subtlety does exist, however, in the definition of the DNA sequence library, since it must be as faithful to the experimental process as possible. In the case of the gene expression cluster analysis study, the DNA library is the population of promoters of all genes on the microarrays. For REAP data, the library consists of species-matched genomic DNA with an identical length distribution as that observed in the experimental data. Thus the ERD algorithm can be generalized to analyze the results of any sequence sampling technology.

To validate the ERD method, we applied it to the 30 clusters of yeast genes from [10] obtained from yeast cell cycle expression profiles [1]. We used the *Saccharomyces cerevisiae* promoter database (SCPD, [12]) to predict putative TREs 500bp upstream of the start codon in the 3000 ORFs used in [10]. Of the 45 searchable TREs, 20 were significant ($p \leq 0.01$) in at least one of the 30 clusters, compared to 18 for [10] using a method that directly discovers TREs from the promoter sequences [8]. Seventeen clusters had at least one significant TRE, compared to 12 for [10]. With one exception (out of 6), all previously known TREs that were present in SCPD that were identified in [10] were significant ($p = 0$) in their appropriate clusters when ERD was used. Additionally, ERD identified physiologically relevant co-occurrence of TREs that [10] 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 [10] 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 [10] 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.

The ERD method has also been applied to a cluster of regulated genes identified using cDNA arrays as well as REAP data for the response of a neuronal cell line to angiotensin. The analysis was greatly facilitated by the use of the bioinformatic tool PAINT [11], which, when given a list of genes, will compile upstream regulatory regions from genomic sequence and then use MatInspector to identify the TREs present in those regions. Several TREs were significant in the results that are known to be activated in response to angiotensin in several

cell types. Results from both analyses also suggested that novel TRE families may also play significant roles in the neuronal response to angiotensin, and these hypotheses are being tested experimentally. Ongoing work involves continued benchmarking of the ERD approach against existing tests for significance, exploring the robustness of the approach to the size of the DNA sequence collection, and testing for significant cases of TRE co-occurrence, an essential step in deciphering combinatorial aspects of transcriptional regulation. Given the flexibility of the ERD approach, we expect it will see broad application in the discovery of transcriptional regulatory networks.

Acknowledgments: We thank Mary Harper and Paul Labhart of GENpathway Inc. for discussions and REAP data, Praveen Chakravarthula for expert technical support with PAINT, Hester Liu, Dan Miller and Grace Straszewski for obtaining the microarray data. DEZ acknowledges the University of Delaware Department of Chemical Engineering for funding.

References

- [1] 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.
- [2] E. H. Davidson, J. P. Rast, P. Oliveri, A. Ransick, C. Caletani, C. H. Yuh, T. Minokawa, G. Amore, V. Hinman, C. Arenas-Mena, O. Otim, C. T. Brown, C. B. Livi, P. Y. Lee, R. Revilla, A. G. Rust, Z. Pan, S. M. J., P. J. Clarke, M. I. Arnone, L. Rowen, R. A. Cameron, D. R. McClay, L. Hood, and H. Bolouri. A genomic regulatory network for development. *Science*, 295(5560):1669–78, 2002.
- [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] T. I. Lee, N. J. Rinaldi, F. Robert, D. T. Odom, Z. Bar-Joseph, G. K. Gerber, N. M. Hannett, C. T. Harbison, C. M. Thompson, I. Simon, J. Zeitlinger, E. G. Jennings, H. L.

- Murray, D. B. Gordon, B. Ren, J. J. W. JJ, J. B. Tagne, T. L. Volkert, E. Fraenkel, D. K. Gifford, and R. A. Young. Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, 298(5594):799–804, 2002.
- [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] B. Ren, F. Robert, J. J. Wyrick, O. Aparicio, E. G. Jennings, I. Simon, J. Zeitlinger, J. Schreiber, N. Hannett, E. Kanin, T. L. Volkert, C. J. Wilson, S. P. Bell, and R. A. Young. Genome-wide location and function of DNA binding proteins. *Science*, 290(5500):2306–9, 2000.
- [8] 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.
- [9] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat. Genet.*, 31(1):64–8, 2002.
- [10] 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.
- [11] 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. *Omics*, In press.
- [12] J. Zhu and M. Q. Zhang. SCPD: a promoter database of the yeast *saccharomyces cerevisiae*. *Bioinformatics*, 15(7–8):607–11, 1999.