

Prediction of Conserved Motif Networks based on Comparative Genomics

Ting Wang, Gary D. Stormo¹

Keywords: regulatory network, comparative genomics, motif discovery

1 Introduction.

Discovery of regulatory motifs in unaligned DNA sequences remains a fundamental problem in computational biology. Most motif finding algorithms perform well in finding over-represented substrings in a small collection of sequences. This requires some prior knowledge about the functional relations among genes in a study in order to select a good collection of promoters. Microarray gene profiling and Chromatin-IP are often used to provide a list of sequences that may be enriched for binding sites of some transcription factors (TFs). However, any limitation in such experiments will propagate to the motif finding step, preventing us from finding motifs for many TFs, as well as from getting a global view of motif network in the whole genome. To date, over 200 yeast TFs have been assayed, but less than 100 have confident motif predictions^[1].

2 Methods.

We have developed a new motif discovery algorithm “PhyloNet” for motif discovery at whole genome level and for *ab initio* construction of motif governed regulatory networks. Given sequences of all promoters in a genome and several related genomes as references, the algorithm combines phylogenetic information and network topology to define all the conserved sequence motifs and clusters of promoters that share the motifs and builds a regulatory motif defined network.

The algorithm contains the following components: 1) Phylogenetic footprinting: wconsensus algorithm^[2] is used to extract conserved regions of the promoters based on reference genome sequences. 2) Profile construction: multiple, suboptimal sequence alignments from phylogenetic footprinting are converted to sequence profiles. 3) Profile space partition: continuous profile space is partitioned into discrete profile clusters^[3]. Each partition is represented by a deputy profile. Distances among the space partitions are calculated by ALLR statistic^[4]. An ALLR scoring matrix is constructed for profile comparison. 4) Query hashing: the query promoter profiles are converted into a collection of formatted “seeds (or words)” of flexible length. Neighborhood words of each seed are generated via a branch and bound algorithm. A hash (or index) is built for the query promoter. 5) Motif BLAST: the entire database (defined by all promoter profiles) are searched against query hash to locate word hits then each hit is extended via local alignment to a high scoring pair (HSP). The statistic significance of these HSPs is estimated by a modified Karlin-Altschul statistic that is suitable for profile comparison^[5]. 6) HSP clustering: significant HSPs are mapped back to the query promoter, and are clustered by applying a maximum clique finding algorithm from graph theory^[6], based on the overlapping relations among HSPs. 7) Motif construction: clustered HSPs are converted to motifs using a greedy approach. Final significance of the motif is estimated based on sum of p-values. 8) Background control: the algorithm has options to shuffle either the query promoter, or the database, or both, while conserving

¹ Department of Genetics, Washington University Medical School. 4566 Scott Avenue, Campus Box 8232, St. Louis MO 63110. Email: stormo@ural.wustl.edu

the sequence identity, sequence length and length of conserved blocks. Such shuffled datasets provide background score distribution.

3 Results and Conclusions.

We applied the PhyloNet algorithm to all promoters of *S. cerevisiae* with *S. mikate*, *S. kudriazevii* and *S. bayanus* as reference genomes. Motifs were predicted using each promoter as query and ranked by their P-values and Tollr scores. Figure 1 shows the comparison between Tollr scores of the best motif predicted for each promoter and for each corresponding shuffled dataset. Highly significant motifs were only predicted based on the real promoters, indicating that a highly organized network structure exists in coordinating gene regulation via motifs. After removing redundancies, we predicted a final list of 515 conserved motifs for the yeast. We compared these motifs with a list of matrices for 100 known yeast transcription factors (TRANSFAC^[7], Habison *et al.*^[1]) using CompareTwo program, and found significant similarities for all but 1 matrices. Figure 2 shows the comparison scores between each known transcription factor and its best match in the list motifs generate by PhyloNet.

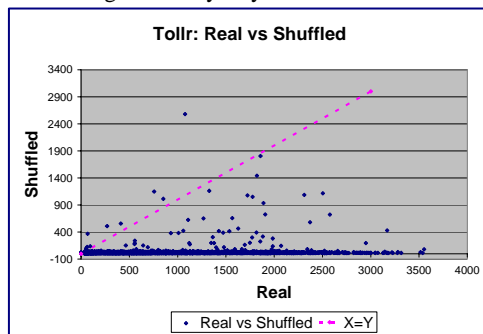


Figure 1

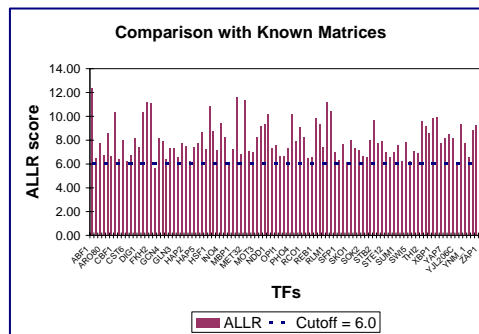


Figure 2

We conclude that PhyloNet is an efficient, effective motif finding algorithm to discover regulatory motifs at whole genome level with no prior knowledge of gene co-regulation needed. We suggest applying it to worm, fly, and ultimately, mammalian whole promoter data.

References

- [1] Harbison, C. T. et al. Transcriptional regulatory code of a eukaryotic genome. *Nature* **431**, 99-104 (2004).
- [2] Hertz, G. Z. & Stormo, G. D. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* **15**, 563-77 (1999).
- [3] Eskin, E. From Profiles to Patterns and Back Again: A Branch and Bound Algorithm for Finding Near Optimal Motif Profiles. *RECOMB 2004* (2004).
- [4] Wang, T. & Stormo, G. D. Combining phylogenetic data with co-regulated genes to identify regulatory motifs. *Bioinformatics* **19**, 2369-80 (2003).
- [5] Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J Mol Biol* **215**, 403-10 (1990).
- [6] Ji, Y., Xu, X. & Stormo, G. D. A graph theoretical approach for predicting common RNA secondary structure motifs including pseudoknots in unaligned sequences. *Bioinformatics* **20**, 1591-602 Epub 2004 Feb 12 (2004).
- [7] Matys, V. et al. TRANSFAC: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res* **31**, 374-8 (2003).