

Integrated Clustering and Motif Detection For Genome-Wide Sequence and Expression Data

Daniel N. Hill
Theoretical Biology and Biophysics
Theoretical Division
Los Alamos National Laboratory
Los Alamos, NM 87545
danhill@lanl.gov

William S. Hlavacek
Theoretical Biology and Biophysics
Theoretical Division
Los Alamos National Laboratory
Los Alamos, NM 87545
wish@lanl.gov

ABSTRACT

Regulation of gene expression is largely responsible for translating genotype into phenotype, as it seems that genes are expressed only when their products are needed. After the genome sequence of an organism is in hand, the next logical step in understanding and characterizing the organism is to probe the function of its genes and their regulation. As genome sequences rapidly accumulate and fill electronic databases, we are challenged to understand the function and regulation of genes at the same pace.

An important step in elucidating genetic regulatory networks is identification of the DNA binding sites of transcription factors. To identify these binding sites, we are evaluating automated methods of detection that take advantage of genomic sequence data and large-scale measurements of gene expression, enabled by microarray technology. The premise of such approaches is that genes regulated by a common transcription factor are likely to have similar expression profiles. It is assumed that motif detection can be focused on a relatively small set of sequences containing related transcription factor binding sites by identifying genes with similar expression profiles. A promising algorithm based on this premise, developed by Church and co-workers [1], involves a two-step approach. In the first step, the k-means algorithm is used to assign genes to clusters based on measurements of expression. In the second step, a Gibbs sampling algorithm is used to identify statistically significant ungapped motifs in the non-coding sequences upstream of putative translation start sites. Tavazoie et al. [1] applied this method, using the genome sequence of yeast [2] and microarray measurements of gene expression for synchronized cell cultures [3], to identify transcription factor binding sites known to play a role in cell cycle control.

Recently, Holmes and Bruno [4] have proposed an alternative to the approach described above, dubbed the kimono algorithm. In this approach, clustering and motif detection are integrated. This integration is accomplished by allowing the output of motif detection to feed into the clustering algorithm and vice versa at each iteration of the two processes. The clusters produced thus represent genes that have similar expression patterns and a common DNA

subsequence. These clusters are more likely to represent sets of co-regulated genes. Likewise, the motifs produced are more likely to be real transcription factor binding sites as they are based on more relevant clusters. Because this algorithm is computationally intensive and is not feasible on a single processor, it has yet to be tested on real biological data.

We have developed an efficient parallel implementation of the kimono algorithm and have begun to test it using publicly available sequence and expression data. We have also begun to compare its performance to the algorithm of Tavazoie et al [1]. Results are preliminary at present.

ADDITIONAL AUTHORS

The authors are members of the project team at Los Alamos National Laboratory focused on Systems Research in Genetic Regulatory Networks. Other members of the team include J. Ambrosiano (ambro@lanl.gov), M.E. Wall (mewall@lanl.gov), D.H. Sharp (dhs@lanl.gov), T. Cleland (cleland@lanl.gov), S.M. Mniszewski (smm@lanl.gov), and A. Evangelisti (amevang@lanl.gov).

REFERENCES

- [1] Tavazoie, S., Hughes, J., Campbell, M., Cho, R., and Church, G. (1999) Systematic Determination of Genetic Network Architecture. *Nat. Genet.* **22**, 281–285.
- [2] Saccharomyces Genome Database. <http://genome-www.stanford.edu/Saccharomyces>
- [3] Cho, R.J. et al. (1998) A genome wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* **2**, 65–73
- [4] Holmes, I. and Bruno, W.J. (2000) Finding regulatory elements using joint likelihoods for sequence and expression profile data. *Proc. ISMB* **8**, 202–210. Also see the kimono website (<http://whitefly.lbl.gov/~ihh/kimono>).