Discovering Transcription Factor Binding Sites in the Yeast *Saccharomyces Cerevisiae*

Xue-wen Chen[1], Jianwen Fang[2*], Xinkun Wang[3*]

Keywords: transcription factor binding site, genome sequence, microarray, fuzzy clustering

1 Introduction.

With the availability of whole genome sequence information and the large amount of gene expression profiling data from high throughput functional genomics approaches, it becomes clear that genome wide computational tools are needed for the analysis of transcript factors (TF) and for the identification of their binding sites. Even without knowing either a set of binding sites for a particular TF or a set of co-regulated sequences, computational algorithms are capable of predicting binding sites and finding their locations in sequences. Typical binding site identification algorithms find co-regulated genes using clustering algorithms. As microarray data are typically noisy, expression profile based clustering is fuzzy by nature. Thus, it is desirable to allow each gene to associate to all clusters with some degrees of certainty. In this paper, we introduce a fuzzy clustering based method for discovering binding signals. An initialized fuzzy C-means algorithm (initFCM) [1] is used for clustering genes in terms of their expression profiles; a web-based interface is developed to search for gene upstreams; furthermore, a standard motif discovery tool MEME [2] is used to find consensus patterns. The statistical significance of consensus patterns is measured in terms of both p-value and E-value, as this helps find distinctive patterns with little chance to be the background.

2 Methods.

The proposed system consists of the following components: fuzzy clustering, upstream search, and multiple sequence analysis. Gene expression profiles are first analyzed by the initFCM algorithms to find co-regulated genes. This algorithm is an iterative partitioning method and gives well-separated initial centers while avoiding the choice of outliers. Instead of assigning each gene to one and only one cluster, we allow each gene to have a membership value associated to each cluster. This cluster membership measures the degree of believe that a gene belongs to that particular cluster. Only genes with a certain degree of believe are considered to be coregulated. A web-based interface connected to a local yeast genome database is developed to retrieve the upstreams of co-regulated genes. It takes a list of gene ID, the length of an upstream sequence, and the position of the upstream starting site as inputs. The output is upstream sequences in FASTA format, which are fed to a multiple local alignment program, MEME, to identify the motifs and thus putative binding sites. The statistical significance of these motifs are evaluated in terms of their E-values (expectation values) and *p*-values. The E-value is an estimate of the number of motifs that would have equal or higher log likelihood ratio if the training set sequences have been generated randomly, while the *p*-value is the probability of a random sequence having the same match score or higher. The consensus sequences identified are considered to be potential regulatory signals.

[1] Corresponding author. Information and Telecommunication Technology Center, Electrical Engineering and Computer Science Department, The University of Kansas, Lawrence, KS 66045. E-mail: xwchen@ku.edu.

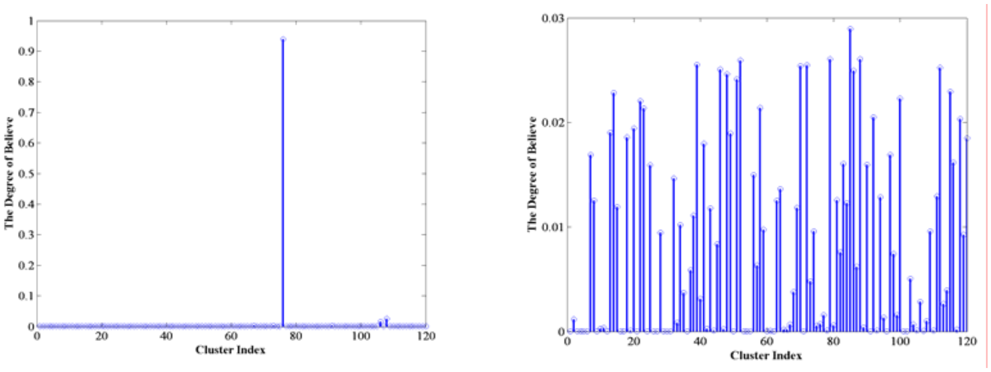
[2]Molecular Biosciences Department, The University of Kansas, Lawrence, KS 66045. E-mail: jwfang@ku.edu.

[3] Higuchi Biosciences Center, The University of Kansas, Lawrence, KS 66045. E-mail: xwang@ku.edu.

[*] Both authors contributed equally.

3 Results.

The data set analyzed here is the cDNA microarray data of *Saccharomyces Cerevisiae* in cell structures which were collected for 6221 genes under 80 different experimental conditions (e.g., cell cycle, sporulation, and diauxic shift) [3]. The initFCM algorithm is first applied for clustering yeast genes. The number of clusters is chosen to be 120. After clustering, some genes can be easily assigned to a cluster, while some can not. Figure 1 (a) and (b) show the degrees of believe versus the cluster index for two ORFs, YAL023C and YAL008W, respectively. For each gene *j*, we calculate the ratio of the maximum value of u_{kj} (k = 1, … , 120) and the second largest value of u_{kj}. If this ratio is no less than two, we assign this gene j to cluster $k = \arg\max_{k} u_{kj}$. Otherwise, we conclude that the cluster membership of gene *j* is fuzzy, and the corresponding gene will not be considered further. Clusters of sizes of 20 or more genes will remain for further analysis. For each gene in a cluster, the 600 bp upstreams are extracted using our web based interface; these upstreams are then fed to the MEME program to identify regulatory signals. Further analysis shows that most of the binding sites identified by our system are either verified by biological experiments or found in TRANSFAC database. For example, one cluster with 48 co-regulated ORFs, identified by our approach, has a consensus upstream sequence GAGGAAATTGAA (E-value = 8.1×10^{-17}), which is a substring of the experimentally verified sequence GAAGAGGAAATTGAA in SCPD database [4]. This sequence is related to the transcript factor GAL4 MCM1 UAS2CHA UASH.



(a) (b)

Figure 1: The degree of believes in clusters for (a) YAL023C and (b) YAL008W.

References

[1] Chen, X. 2002. Clustering Gene Expressing Data with Min-max-median Initialized Fuzzy C-Means Algorithms. *Workshop on Genomics Signal Processing and Statistics* (*GENSIPS 2002*), *NC: IEEE.*

[2] Bailey, T. and Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers, *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology*, pp. 28-36, AAAI Press, Menlo Park, California.

[3] Spellman, P., Sherlock, G., Zhang, M., Iyer, V., Anders, K., Eisen, M., Brown, P., Botstein, D., and Futcher, B. 1998. Comprehensive Identification of Cell Cycle-regulated Genes of the Yeast *Saccharomyces cerevisiae* by Microarray Hybridization. *Molecular Biology of the Cell* **9**, 3273-3297.

[4] Zhu, J. and Zhang, M. 1999. SCPD: a promoter database of the yeast Saccharomyces Cerevisiae. *Bioinformatics* 15 (7/8), pp. 607-611.