165. Prediction of transcriptional regulatory networks using gene expression data

Jiang Qian[1], Jimmy Lin[2], Mark Gerstein[3]

Keywords: microarray, gene expression, bioinformatics, support vector machine, transcription factor, transcriptional regulatory network, computational proteomics, gene networks

1 Introduction

With the increased prevalence of gene expression data, researchers are now afforded new opportunities to unravel the workings within the cell. In particular, due to the fact that gene expression is directly related to transcriptional control, this new abundance of data is now valuable in the establishment of transcriptional regulatory networks and prediction of transcription factor targets.

Traditional computational research in understanding transcription factors has focused more on the genomic data, such as examining sequences of promoter regions and determining binding motifs. [1] [2] We propose a method of transcription factor target prediction using gene expression profiles. Because the relationship between a transcription factor target and its expression profile is not simply correlation, traditional clustering techniques, which often assume simultaneous expression, have had less success. New efforts have been made to capture the complexity of the relationship. [3] Here, we use support vector machines (SVM) to determine the targets of specified transcription factors.

2 Method and Results

We selected *Saccharomyces cerevisiae* as the model organism due to the large amount of gene expression profiles that are available. In total, we incorporated data from 79 different DNA microarray hybridization experiments, which includes diauxic shift, the mitotic cell cycle, sporulation, and heat shock. Transcription factors were obtained from two separate databases: TRANSFAC [4] and SCPD [5].

The training data was provided using pairs of genes, which had either positive or negative transcriptional control relationships. Positive data was abstracted from the two aforementioned transcription factor databases and negative data was created with random selection of shuffled expression profiles. Assuming that the budding yeast has approximately 6000 genes, there would be 36,000,000 possible pairs – and only a small fraction of these gene pairs would have transcriptional relationships. Therefore, in order to minimize the large search space, we limited the dataset to gene expression targets whose expression is altered at least 1.5 fold when different transcription factors are knocked out. These pairs were obtained from yTAFNET. [6]

With the established training set, we used Brown's implementation of Support Vector Machine (SVM) [7], which is a supervised machine-learning algorithm designed for pattern recognition and regression. SVM establishes a hyperplane function from the labeled training data and uses it for

[1] Department of Ophthalmology, Johns Hopkins Medical School, Baltimore, MD 21287. E-mail: jqian2@jhmi.edu

[2] School of Medicine, Johns Hopkins Medical School, Baltimore, MD 21287. E-mail: jimmy.lin@jhmi.edu

[3] Department of Molecular Biochemistry and Biophysics, Yale University, New Haven, CT, USA 06520. E-mail: mark.gerstein@yale.edu

prediction. Due to the nature of the transcriptional relationship, the amount of training data, and the size of the prediction space, SVM is a well-suited algorithm for the prediction of transcription factor gene targets.

From a starting set of 36 transcription factors, we predict a total of 3419 targets. The transcription factors vary greatly in the number of targets they control. On average, transcription factors control 93 genes and the average gene is control by 1.8 transcription factors. This further highlights the complexity of the relationship between transcription factor and its target as that most relationships are not simply one-to-one.

Due to the large size and complexity of the network, simplified networks that are abstracted from the global ones show more clearly the relationships between different genes and transcription factors. We observe a hierarchical relationship between the different transcription factors. A small subset of the transcription control of transcription factors is shown in Figure 1.

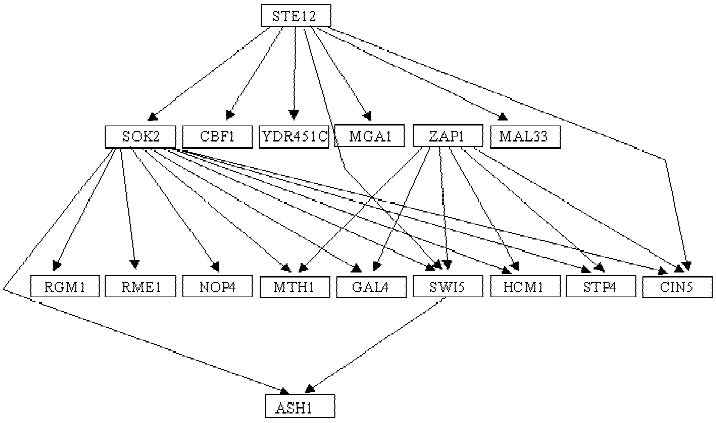

Figure 1: Transcriptional Control of Transcription Factors.

4 References and bibliography

[1] Grabe, N. 2002. AliBaba2: context specific identification of transcription factor binding sites. *In Silico Biology* 2(1): S1-15.

[2] Zhu, Z., Y. Pilpel, et al. 2002. Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (TFCC) algorithm. *Journal of Molecular Biology* 318(1): 71-81.

[3] Qian, J., M. Dolled-Filhart, et al. 2001. Beyond Synexpression Relationships: Local Clustering of Time-shifted and Inverted Gene Expression Profiles Identifies New, Biologically Relevant Interactions. *Journal of Molecular Biology* 314: 1053-66.

[4] Wingender, E., X. Chen, et al. 2001. The TRANSFAC system on gene expression regulation. *Nucleic Acids Res*earch 29(1): 281-3.

[5] Zhu, J. and M. Q. Zhang. 1999. SCPD: a promoter database of the yeast Saccharomyces cerevisiae." *Bioinformatics* 15(7-8): 607-11.

[6] F. Devaux*, P. Marc*,C. Jacq. 2001. Transcriptomes, transcription activators and microarrays. *FEBS Letters* 498(2-3):140-4.

[7] Brown, M. P., W. N. Grundy, et al. 2000. "Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences USA* 97(1): 262-7.