Global Gene Expression Analysis with Genetic Programming

Yuh-Jyh Hu
Computer Science and Engineering Department
Tatung University
yhu@cse.ttu.edu.tw

Abstract

Computer-assisted methods have become very important in analyzing biosequence data. However, most of the current methods are limited to finding motifs only. Genes can be regulated in many ways, including combinations of regulatory elements. This research is aimed at combinatorial motif analysis and hypothesis generation. A genome-wide gene expression analysis demonstrated the value of the studies of motif combinations and classification hypotheses.

1 Introduction

The advance of the microarray and the genechip technology provides a view of changes in gene expression on a genomic scale. As more is learned about the functions of every gene in the entire genome, we have the ability to gain insight into a cell's response to a changing environment through its global gene expression patterns. Potential applications include predicting drug interaction or drug resistance, exploring immune systems, etc. We first propose using multiple objective functions to detect meaningful motifs from sequences. Our experimental results demonstrated the synergy of the information content and the multiplicity significance helps maintain the balance between the consensus quality and the over-representation of motifs. To learn beyond how genes behave in the course of time on a genomic scale, we propose a novel view of the gene regulation analysis. With the assistance of the genechip technology, genes could be grouped into families according to different temporal patterns. Our analysis of gene regulation is focused on the search for significant combinatorial motifs involved in the regulation as well as potential hypotheses of how the gene regulation is related to the motifs. This type of analyses not only complement the global study of the changes of gene expression by looking into the involvement of combinatorial motifs in regulation, but also suggest to biologists further biological tests on the genes through the inference from the hypotheses produced by the analyses.

We tested the integrated system on the regulation of thermal stress response in yeat genome. The experimental results showed that our new motif-detecting method (GPMD) outperformed several current representative motif-findinf algorithms, including CONSENSUS (Hertz *et. al.*, 1995), Gibbs (Lawrence *et. al.*, 1993) and MEME (Bailey and Elkan, 1995). The integration of GPMD, GPCI (Hu, 1998) and C4.5 (Quinlan, 1993) successfully generated useful hypotheses for biologists.

References

Bailey, T. and Elkan, C. "Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization", Machine Learning, 21, p51-80, 1995.

Hertz, G. and Stormo, G. "Identification of Consensus Patterns in Unaligned DNA and Protein Sequences: A Large-Deviation Statistical Basis for Penalizing Gaps", in Proceedings of the 3rd International Conference on Bioinformatics and Genome Research, p201-216, 1995.

Hu, Y. "A Genetic Approach to Constructive Induction", in Proceeding of the 3rd Annual Genetic Programming Conference, p146-151, 1998.

Lawrence, C., Altschul, S., Boguski, M., Liu, J., Neuwald, A. and Wootton, J. "Detecting Subtle Sequence Signals: A Gibbs Sampling Strategy for Multiple Alignments", Science, Vol 262, p208-214, 1993.

Quinlan, J. R. C4.5 : Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA., 1993.