

# A Probabilistic Boolean Approach to Find Regulatory Networks from Microarray Data

Wataru Fujibuchi and David Landsman

National Center for Biotechnology Information

National Library of Medicine

8600 Rockville Pike, Bethesda, MD 20894

{wataru, landsman}@ncbi.nlm.nih.gov

The new microarray technology for assaying the expression of thousands of genes simultaneously is providing substantial data on mRNA expression levels in a variety of species under a plethora of cellular conditions (Marshall 1999). As the amount of experimental data increases, searching for biologically meaningful genetic networks in microarray data could assist in the understanding of the molecular mechanisms of gene expression. However, it is difficult to extract genetic network information from microarray data due to discrepancies between real biological systems and idealized mathematical systems. Although many hopeful models (Erb & Michaels 1999; D’Haeseleer *et al.* 1999; Chen, Filkov, & Skiena 1999) have been proposed, these models either have not described biological systems well in terms of experimental reality or are lacking in the evaluation of known biological systems. Biological systems contain noisy and unstable outputs because of a number of inevitable and variable experimental processes such as temperature, time, buffer conditions and the uncertain nature of molecules.

Here we attempt to extract the maximum information on genetic interactions from microarray experiments and evaluate the resulting networks by comparison to known gene regulatory events. We introduce a probabilistic approach to find realistic biological regulatory circuits that represent Boolean networks (Kauffman 1993; Akutsu, Miyano, & Kuhara 1999) from gene expression patterns. In this approach, stable logic circuits between master and slave genes are searched and ranked by the probability based on a binomial distribution model. All permutations of logic circuits between genes are generated. Input and output frequencies are used to calculate the probability of the stability of each circuit. The logic circuit that gives the best probability is taken as the candidate regulation between the genes. We then assemble logic circuits between input and output genes to construct a global network.

Among recently sequenced organisms, the yeast *Saccharomyces cerevisiae* is one of the best eukaryotic model systems because of its small genome size (~6000 genes) and the abundance of experimental data. To make use of published microarray data, we made an experimental database called GExDB-Yeast, containing

121 experimental values (included are 103 time-course dependent values for each yeast gene). In order to identify significant gene expression or repression, we use a three state model of normalized microarray data, represented by “activation”, “repression”, and “no change”. We reduce the number of genes by excluding these genes which do not show expression or repression above a cut-off level. Furthermore, we obtained information on the acceptable time lags between the expression of master and slave genes by prior analysis of known regulation events. By using a window-search type of variable time lags of expression, we have applied the probabilistic method to the yeast cell-cycle and glucose-response system data which follow time-course expression. Statistical evaluation shows that the reproducibility of known regulatory networks is more significant than randomly drawn networks with a probability range of  $10^{-2} \sim 10^{-10}$ , depending on the parameters. We report that this type of probabilistic method is suitable for deciphering genetic networks from noisy experimental data.

## References

- Akutsu, T.; Miyano, S.; and Kuhara, S. 1999. Identification of genetic networks from a small number of gene expression patterns under the boolean network model. In *Pac Symp Biocomput*, 17–28.
- Chen, T.; Filkov, V.; and Skiena, S. S. 1999. Identifying gene regulatory networks from experimental data. In *RECOMB ’99*.
- D’Haeseleer, P.; Wen, X.; Fuhrman, S.; and Somogyi, R. 1999. Linear modeling of mRNA expression levels during CNS development and injury. In *Pac Symp Biocomput*, 41–52.
- Erb, R. S., and Michaels, G. S. 1999. Sensitivity of biological models to errors in parameter estimates. In *Pac Symp Biocomput*, 53–64.
- Kauffman, S. A. 1993. *The origins of order*. Oxford University Press, New York.
- Marshall, E. 1999. Do-it-yourself gene watching. *Science* 286(5439):444–447.