Poster L-5

Power CoClustering – detecting multiple regulatory influences in gene expression data

Authors:
Shay Zakov (*Computer Science Department, Ben Gurion University*)
Avraham A. Melkma (*Computer Science Department, Ben Gurion University*)

Short Abstract: Thermodynamic models of transcription regulation suggest that gene expression levels may obey a power law dependence on the activity level of the regulatory mechanism. We present a new method of CoClustering micro-array data, which aims to detect such power law relations, as well as encouraging preliminary results.

Long Abstract:
The usual methods of coClustering, or bi-clustering, for the analysis of microarray expression data are ad-hoc in the sense that they attempt to identify groups of co-expressed genes without referring to a model of transcription regulation. Often it is implicitly assumed that the expression level of a gene depends linearly on the concentration of the transcription factors that regulate its expression, even though thermodynamic models indicate the existence of complex relationships between the expression level and the concentrations of the factors. We report here on encouraging preliminary results of the analysis of gene expression data, using a novel algorithm that allows for the possibility of a power law relation between a gene expression level and the activity level of the regulating mechanism. Such a relation may be expected, for example, in case the activation of transcription of a gene is driven by binding of the same transcription factor to several upstream binding si!
tes. If transcription initiation takes place only when all regulatory sites are occupied by the transcription factor, the rate of transcription initiation is proportional to a power of the concentration of the transcription factor.

In order to accommodate such power law behavior we define a new measure of distance between genes, which is smaller the closer one gene comes to being a constant power of the other in the set of experiments under consideration. The measure can also be generalized to sets of genes, in which case its value is smaller the closer each and every one of the genes comes to being modeled as a constant power of a single underlying activity variable. Although, as we prove, the problem of finding the exact value of this measure is computationally intractable, we present an iterative computational method that efficiently finds near-optimal solutions for most instances, and provably finds the optimal value for the case of two genes.

Because genes usually serve in more than one capacity it is to be expected that they are co-regulated only in certain experiments, whereas in others there is no similarity in their behavior. In order to deal with this possibility we have developed a Monte-Carlo algorithm, based on our distance measure, which identifies subsets of genes and simultaneously subsets of experiments, such that the measure of the subset of genes in the subset of experiments does not exceed some pre-specified bounding value. Preliminary results over synthetic data have shown that this algorithm has a high success rate in identifying planted

coClusters in a large data set, suggesting that the algorithm has the potential to become a powerful tool in the field of automatic biologic data analysis, pointing over some possibly previously unknown regulation mechanisms.