Discovering Biological Functional Modules from Gene Expression Data

Chen-Hsiang Yeang

Artificial Intelligence Laboratory
Massachusetts Institue Of Technology
Cambridge, Massachusetts 02139

http://www.ai.mit.edu

The Problem: The goal of this research project is to identify genes belonging to the same functional modules on the basis of relatively few DNA array (gene chip) experiments.

Motivation: Gene chip or microarray technologies measure the expression levels of the entire genome within a single substrate. It is widely believed that the genes participating in the same functional modules in biological processes often share (at least parts of) the same regulatory mechanism. This common mechanism imposes constraints on the (MRNA) transcription levels that we can measure using either HDA or microarray technology. Biological modules are therefore to a degree identifiable from expression data. However, this is made more difficult by the stochastic nature of biological processes, variation due to experiment and measurement technology, as well as due to the very few (informative) experiments that are available (in tens) in comparison to the number of genes (from 6000 to 100000). In this project we develop clustering and other algorithms for the purpose of identifying functional modules from relatively few expression measurements.

Previous Work: There are two major approaches for clustering expression data: hierarchical clustering [1] and self-organizing map [2]. The hierarchial clustering starts from individual genes as clusters, successively merging pairs of clusters with the smallest distance. The pairwise distances between clusters are modified after each merge. Self-organizing map (SOM) starts from a fixed number of seed nodes arranged in a specific geometric structure. The seed nodes selectively move toward randomly chosen points until convergence. Both methods have distinct disadvantages. In a hierarchical clustering algorithm, the distance between points become distorted after each merge. For example, neighboring points might be pulled away if they are assigned to different clusters at early stages. Self-organizing map, on the other hand, requires designating both the number and geometry of the nodes, both of which govern the clustering solution. Moreover, both methods assign a single gene to a single cluster. This is not always desirable since a single gene product can participate in multiple biological module. For other clustering and modeling approaches in the context of gene expression data, see, e.g., [3], [4] and [5].

Approach: Every clustering problem contains two subproblems: the definition of clustering metric and the clustering algorithm. If we assume that the logarithm of gene expression levels are governed by a Gaussian distribution, the standard correlation coefficient suffices initially as a similarity measure. We actually use the confidence value associated with the hypothesis that the correlation is positive (genes are co-expressed) as the pairwise similarity measure. This is a monotonic function of the correlation coefficient: $-Nlog(1 - corr_N(X_i, X_j)^2)$, where N is the number of samples. This measure can be defined over multiple genes as well. We are developing various extensions of this simple pairwise measure by constructing graph models for capturing *co-regulation* explicitly rather than just relying on observed *co-expression*.

We devise a clustering algorithm with following properties: (1) a single gene can participate in multiple clusters or may not belong to any of them; (2) all genes within the cluster are tightly coupled according to the clustering metric. To this end, we first construct a sparse graph by thresholding confidence values; only pairs of genes whose co-expression is identifiable with high confidence are linked in the graph. The cliques (complete and maximal subgraphs) of the resulting graph serve as candidates for modules. There are two major issues. First, finding the maximal cliques in a graph is an NP-complete problem. Since we are only interested in the links of very high confidence values, however, the graph is very sparse. Depending on the number of edges in this reduced graph, all cliques can be retrieved with reasonable running time (various approximation algorithms could be used in this context as well). Second, each clique requires the corresponding genes to be strongly interconnected. This means that the cliques are small and highly figures/yeang.ps hoffset=-80 voffset=-6008in1.8in

overlapped with each other. To reduce the number of redundant clusters, we merge cliques into super-cliques which are no longer complete subgraphs, in spite they still maintain high level of connectivity within the cluster.

Preliminary results: Figure 1 shows a subset of clustering result on 45 gene expression data with 491 genes in *S. Cerevisiae* (yeast) genome. Many clusters comprise genes belonging to the same biological complexes. Furthermore, certain genes participating the same biological process but are not in the same complex are grouped together (for example, in one cluster, gene YEL024W belong to cytochrome bc1 complex, genes YGL187C and YLR038C belong to cytochrome c oxidase, both are involved in electron transport of glucose metabolism).

Impacts: The problem of discovering functional modules from expression data is both biologically important and computational challenging. From biological perspective, to identify members of functional modules is the first step toward understanding the regulatory network of the cell. Although the genomes of several organisms have been sequenced, the functions of many genes remain unknown. Our algorithm can help attributing functional roles for unknown genes. From computational perspective, one of the key challenges is dealing with overfitting as the number of data samples is so small.

Future work: The correlation coefficient we have used initially in the clustering algorithm measures only co-expression, not co-regulation. We are in the process of constructing statistical graph models that better capture the underlying co-regulation and using such models as part of the clustering algorithm. The graph models are appropriate for representing various conditional independence constraints and dependencies associated with combinatorial regulation. They can be used to filter out "false positive" genes within each cluster as well to induce more realistic clustering metrics.

Research Support: This project is partially funded by a grant from MIT/NTT partnership (information filtering with minimal instruction).

References:

[1] M. B. Eisen, P. T. Spellman, P. O. Brown and D. Boststein. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95:14863-14868, Dec 1998.

[2] P. Tamayo et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *PNAS*, 96:2907-2912, Mar 1999.

[3] T. R. Golub et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286:531-537, Oct 1999.

[4] N. Friedman et al. Using Bayesian networks to analyze expression data. *Journal of computational biology*, 2000.

[5] M. P. S. Brown et al. Support vector machine classification of microarray gene expression data. *UCSC-CRL-99-09*, 1999.