Intensive course on modeling biological networks
Summer 2007
Computer exercise: Co-expression and clustering

1 First part

Co-expression, similarity of expression profiles of two genes, has been used as a computationally simple sign of co-regulation or relatedness of gene function. Correlation of expression over a set of conditions (treatments, tissues, even organisms) is an obvious criterion of relatedness.

One obvious weakness is that for N genes there are N^2 pairwise interactions. This is a large number and, more importantly, considering the pairs separately will lose similarities among several genes. It would be nice to be able to summarize the interactions somehow, and clustering is an obvious choice.

Try to find co-regulated clusters of yeast genes with the simple K-means clustering (function `kmeans` of Matlab). File `NODE_DATA.dat` contains yeast gene expression profiles, one per row, collected by [1]. Crucial steps in clustering are the choice of distance measure and number of clusters, and validation of the results.

Try at least the commonly used Euclidean distance and correlation, and several cluster numbers. How do the results differ and which are better? How do you know?

Some suggestions include robustness (compute several clusterings with resampled data), conformity with prior biological knowledge, and various goodness indices. Try out some and discuss your choice.

Are the results informative? Why/why not? The file `GOclasses.txt` contains (some of the) functional classes of the genes from the Gene Ontology (GO, ontology "biological process"), collected in a binary matrix between genes and the functional classes.

2 Second part

Co-expression gives hypotheses for co-regulation. The hypotheses would be more convincing if there was even more in common with the co-expressed genes. One suggestion has been to use protein-protein interaction (PPI) data

(`LINKS.dat`) as further constraints: If two genes are co-expressed and furthermore the proteins they encode interact, it is more likely that they are functionally related.

Study this idea using clustering. An easily implementable way is to use the Matlab function `clusterdata` and integrate both expression data and PPI data to the pairwise distance matrix. Is the result better than with expression data alone?

3 Third part

What weaknesses does this whole approach have? How could they be solved? What is the tradeoff in each solution?

4 Suggestions for project works

1. Graph-based clusterings would be particularly justified here, and able to combine several distance measures.
2. K-means clustering is a simplified variant of mixture models, which could be applied here.
3. Bi-clustering has been suggested as a solution to the problem that regulation may be condition-specific.

References

[1] Igor Ulitsky and Ron Shamir. Identification of functional modules using network topology and high-throughput data. *BMC Systems Biology*, 1:8, 2007.