

UCLA IPAM RIPS 2004

Project Description

UCLA Faculty Advisor: Shawn Cokus

Industrial Partner: BioDiscovery, Inc.

Background

Gene microarrays are powerful tools for assessing cellular activity, simultaneously measuring the transcription levels of thousands of genes simultaneously. In brain cancer studies microarrays have the potential for improving prognostic accuracy and for finding targets for therapeutics. A common aim in microarray research has been to find a subset of the genome which is involved in the phenotype of interest. A basic approach is simply to select those genes whose expression levels are most greatly changed in the phenotype of interest. Several more novel approaches are explored in Mischel et al. A network of genes is created by building a graph structure whose nodes are genes and whose edges indicate high correlation between changes in transcription levels across tumor samples. In this paper, significant sub-networks (called “modules”) are derived using a topological overlap matrix (TOM) approach (Ravasz et al., 2002). Five modules are identified. The significance of these modules are verified by looking at the properties of known genes included in the modules, and noting that the properties are related to cancer development.

Research Tasks

The goal of this project is to explore certain refinements and alternatives that build on top of the network analysis described above. In particular:

1) The established approach depended on human inspection of the TOM cluster results to identify five modules. Develop an algorithm to identify modules without human intervention. If the algorithm requires any parameters they must be reasonable for a “user” to set, i.e. they should be meaningful to a non-graph theorist and their settings should be applicable to multiple problem instances. (E.g. it isn't OK to ask the user how many nodes a module should have.)

2a) The established approach creates an edge between two nodes when the inter-gene correlation is greater than or equal to 0.8. This begs the question of whether the result (the number and composition of modules) is sensitive to the choice of this threshold. The task is to recreate the results of the original analysis, then explore the sensitivity of the results to this and other algorithm parameters, as well as to any parameters in the algorithm you introduce in (1) above.

2b) Develop an algorithm for optimally choosing the threshold. This task includes working with the brain cancer researchers at UCLA to establish a measure of optimality. Analyzing module makeup/content may involve employing the statistical analysis of the Expression Analysis Systematic Explorer (EASE, <http://david.niaid.nih.gov/david/ease.htm>) tool discussed by Mischel et al.

2c) Instead of using “unlabeled” edges, include the correlation coefficients as edge weights to improve the clustering. Can improvement be obtained by removing the step of thresholding and including *all* edges (along with their weights)? How can intractability due to the completely connected network be avoided?

3) Gene networks can be created from a number of data sources. The cited work is based on direct measurement of gene expression. There is interest in building and analyzing gene networks based on known gene interactions, e.g. as described in published literature. Several efforts involve *text mining* the published literature to search for this information. A short cut is to use manually curated databases of gene-publication links maintained by the National Center for Biotechnology Information (NCBI). This task involves building a network in which edges indicate common publication references for a pair of genes (i.e. an *annotation-based network*), applying the module-discovery algorithms developed above, and comparing the resultant modules to those derived for the gene expression-based networks. A suggested validation step for these derived modules is to see to what extent genes which are differentially expressed in one or more tumor types cluster into the modules. Also, does the network have the same clustering properties found by Mischel et al. for the correlation-based network?

Data sources

- Expression data comes from the UCLA brain cancer research project, Prof. Stan Nelson, PI. They have tentatively agreed to allow access to these data for this project, and we acknowledge that they reserve the right to control the dissemination and use of their data.
- Annotation-based network information: This information is accessible by combining the microarray probe annotations available from the chip manufacturer (Affymetrix, www.affymetrix.com/) with gene annotation information maintained by the NCBI (www.ncbi.nlm.nih.gov/LocusLink/). For expediency, BioDiscovery has combined information from the two into a single annotation text file available for this project. Project members are encouraged to explore other sources of genomic information for the genes used in the study.

Output

The result of this work should be:

(1) a software application which takes as input a table whose rows are genes whose columns microarrays and whose entries are gene expression levels, which computes the inter-gene correlation coefficients (cc's), then determines the modules within the network implied by the cc's.

(2) a software application which takes as input a graph (e.g. derived from any source or mechanism) and determines the modules it contains.

Along the way, the project should provide explicit answers to the questions posed above, i.e.:

Is the result sensitive to the choice of cc threshold? If so, how shall the threshold be chosen?

Is it computationally tractable to analyze a network which hasn't been reduced by thresholding? If so, how can this be done? How are the results improved?

How do the results of a correlation-based network compare to those of an annotation-based network?

References

Mischel et al. "Network Properties of Gene Expression in Cancer: Do They Predict Survival?"

Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabasi AL (2002), Hierarchical Organization of Modularity in Metabolic Networks. Science 297, 1551-1555

Contacts

UCLA

Shawn Cokus, Cokus@math.ucla.edu

Marc Carlson, MRJCarlson@mednet.ucla.edu

Prof. Stan Nelson, snelson@ucla.edu

BioDiscovery

Bruce Hoff, hoff@biodiscovery.com

Anton Petrov, antonp@biodiscovery.com