The Goal

Create a model of the transcriptome using factor analysis

Create model that simplifies and reveals structure in observed data by describing it as the result of hidden factors.

Simple Example of Factor Analysis

\

Observed Data – Grades for a class of students
4 problem set scores, 3 quiz scores, and 3 exam scores, 10 total variables

Factor Models - There are many possible factor models to describe this dataset

Factor Model A – 1 Factor, describing scholastic aptitude

Factor Model B – 2 Factors, one describes problem sets, one describes exams+quizzes

Factor Model C – 3 Factors, one each for problems sets, quizzes, and exams

Factor Model D – 2 Factors, one for first half of course one one topic, one for second half of course on another topic

Each model simplifies that data while maintaining much of the original information.

The same thing can be done for gene expression data.

The Gene Expression Data

Gene Expression Omnibus Database

Affymetrix U133A microarray platform

2,866 samples

22,215 transcripts per sample

This is the data that is used to generate a factor model

Data Preparation

Download Data:
2.0 GB – approximately 65 million data points

Parse data:
Not exciting, but non-trivial

Clean Data:
Remove samples and transcripts that are missing data anywhere in the dataset. This only removes a small fraction of the data

Normalization

Normalization is a big issue in microarray analysis

Affymetrix GeneChips use a single color hybridization which makes the data fundamentally different from two color data

Rank normalization works well and is simple

The expression level of a transcript is based on the rank of that transcripts signal compared to the signal of all of the other transcripts from that sample.

Transcripts were assigned to one of 16 expression levels

The Factor Model

Sparse Linkage:
In this factor model, there is a sparse association between the factors and the transcripts. Each of the transcripts is associated with a restricted subset of the factors. In this analysis the transcripts were associated with as few as zero and at most three factors. A factor may be associated with a large number of transcripts, or with as few as one. The transcript-factor associations are constant across all of the samples. This sparse association is biologically realistic and differs from most factor analytic techniques such as principal components analysis

Mutual information:
Mutual information is a good metric for quantifying the ability of a factor to predict the expression of an associated transcript because it can detect non-linear dependencies between two variables and it is generally insensitive to outlying data. The mutual information content between the expression profile of each transcript and the profile of that transcript's associated factors was used as a metric to score how well a candidate factor model described the expression dataset.

Optimizing the Model

No analytic method exists to determine the optimal factor model of the type described for a given expression dataset. Consequently, a maximization approach was used to search for the factor model that best described the expression data.

The model has two distinct sets of parameters that can be optimized.

1) The subset of factors associated with each transcript can be changed.

2) The factor profile of each sample can be changed.

These parameters were changed to maximize the mutual information scoring metric.

Computation Run-Time Problems

Calculating the mutual information between a factor model and the expression data can be performed in less than a minute, but optimizing the model adds orders of magnitude in complexity.

Consequently, the optimization cannot currently be completed on the full data set.

While the python code is relatively efficient, performance could be improved by recoding parts of the algorithm in C++.

The an optimal model can be generated for a subset of the data containing 250 transcripts using 50 factors

Results for 250 transcript dataset

Multiple probes for the same gene were usually associated with the same factors

Biologically coherent factors were found, for example a set of transcripts that were associated with the same set of factors included

tumor protein p53
immediate early response 3
Src homology 2 domain transforming protein 1
3-hydroxy-3-methylglutaryl-Coenzyme A reductase
cytochrome P450, family 1, subfamily B, polypeptide 1
Treacher Collins-Franceschetti syndrome 1
FAT tumor suppressor homolog 1 (Drosophila)

What can the model be used for once complete?

The factor states that come from the optimized model could be used as a substitute for the expression measurements themselves. Samples could be clustered or categorized based on factor profiles. Comparison of factor states before and after experimental manipulation could reveal the cellular systems that are affected by the experiment.

The association between transcripts and factors that comes from the optimized model will be a concise approximation of the connectivity of regulation in the transcriptome. Trancripts that are associated with the same factors can be infered to be co-regulated and involved in related biological processes.