Biological Pathway Prediction from Multiple Data Sources Using Iterative Bayesian Updating

Corey Powell and Joshua Stuart
Department of Biomolecular Engineering
University of California, Santa Cruz
1156 High Street, Santa Cruz, CA 95064
{cpowell,jstuart}@soe.ucsc.edu

Abstract

There is a diversity of functional genomics data, such as gene expression data from microarray experiments, phenotypic data from gene deletion experiments, protein-protein interaction data, and data from manually curated databases of gene function. Each data source finds certain types of relationships between genes and misses other types of relationships. A method that can combine multiple data sources might then be able to uncover more relationships than a method that depends on a single data source. This paper presents a method that uses an iterative Bayesian updating technique to combine data from multiple sources, represented as undirected weighted graphs, in order to estimate the probability that a gene is part of a given biological pathway. This method improves performance over a simple neighbor based approach for several well characterized biological pathways.

1 Data

This study uses microarray based gene expression data and data from protein interaction experiments to create two undirected weighted graphs, denoted by the abbreviations **GEN** (Gene Expression Network) and **PIN** (Protein Interaction Network), respectively. The vertices of these graphs are **metagenes**, or collections of genes from multiple species that are inferred to have functional similarity based on sequence similarity as determined by BLAST [1]. The GEN connects a pair of metagenes with an edge if the correlation between the expression profiles of the corresponding genes over thousands of microarray experiments is higher than expected by chance. The weight placed on the edge is $-\log(p)$, where p is the $p-$value of chance correlation between the expression profiles. For more information about the GEN and its construction, see [5]. The PIN takes protein interactions from the GRID database for yeast (Saccharomyces cerevisiae), fly (Drosophila melanogaster), and worm (Caenorhabditis elegans) [2] and from the Vidal lab [4] and DIP [6] for additional interactions in worm. The weight placed on an edge in the PIN is the number of species in which the protein interaction has been detected in a high-throughput assay. The GEN and PIN consist, respectively, of $3,416$ and $3,920$ vertices and $22,163$ and $16,092$ edges, with $2,571$ vertices and 411 edges in common.

The biological pathways tested in this study are the Cell cycle, Oxidative phosphorylation, Proteasome, and Ribosome pathways from the KEGG database [3], consisting of $60, 51, 31$, and 82 metagenes, respectively, in the GEN. The study also uses "Random" pathways 25_Random, 50_Random, 75_Random, and 100_Random, created by randomly picking $25, 50, 75$, and 100 metagenes, respectively, from the GEN.

2 Methods

2.1 A Simple Neighbor Based Method

If the data sources above contain functional information, then it would be reasonable to expect that a metagene g is more likely to belong to a pathway P if a high percentage of its neighbors N_g are in the pathway. This leads to a straightforward criterion that classifies g as belonging to P if

$$S(g) = \frac{\sum_{h \in P \cap N_g} w_{gh}}{\sum_{h \in N_g} w_{gh}} > c,$$

where w_{gh} is the weight of the edge from g to h and c is a predefined cutoff value. This method generalizes in a natural way to more than one weighted graph by classifying g as belonging to P if and only if g satisfies the criterion above in every graph where g is a vertex. This algorithm is implemented on the GEN and on the GEN together with the

PIN. The Results and Conclusions section compares cross validation results from both methods.

2.2 An Iterative Bayesian Updating Method

Suppose that $G_1, \cdots, G_n$ are weighted graphs with vertices $V(G_i)$, and designate one graph G_R as the **r**eference graph. This graph should be the most complete of the graphs, and all or almost all of the vertices in the pathway should be in G_R. For the implementation for this paper, G_R is the GEN. Let P be the pathway and denote by $p(g)$, $N_g^{G_i}$, and $w_{gh}^{G_i}$ the probability that $g \in G_R$ is in P, the set of neighbors of g in G_i, and the weight of the edge from g to h in G_i, respectively. The symbol $\#$ refers to the number of elements of a set. The following is a high level description of the algorithm.

1. For each $g \in G_R$, initialize

$$p(g) = 1 - \frac{\#V(G_R) - \#P}{\#V(G_R)\#P}$$

if g is in the initial list, and $p(g) = 1/\#V(G_R)$ otherwise. This prior reflects a high, but not absolute, level of confidence in the original pathway assignments, and is chosen so that the expected number of pathway elements before the first iteration is $\#P$.

2. Repeat j times:

   (a) For each graph G_i and each vertex $g \in G_i \cap G_R$, calculate the following score functions, which are essentially the weighted averages of the probabilities of the neighbors of g in G_i that are also in G_R.

$$S^{G_i}(g) = \frac{\sum_{h \in N_g^{G_i} \cap G_R} w_{gh}^{G_i} p(g)}{\sum_{h \in N_g^{G_i} \cap G_R} w_{gh}^{G_i}}.$$

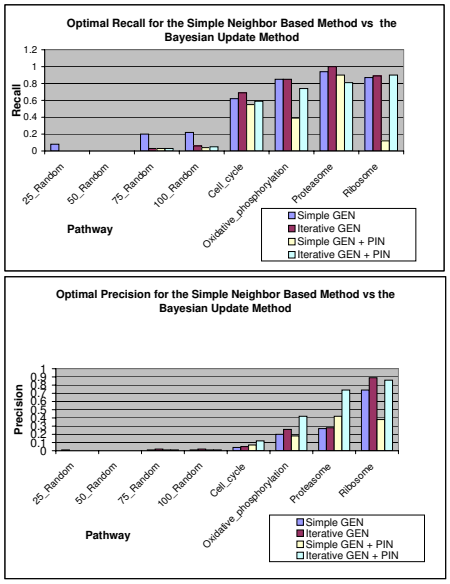
   (b) For each graph G_i, construct a positive distribution $p(S^{G_i}(g) \mid g \in P)$ and a negative distribution $p(S^{G_i}(g) \mid g \notin P)$ by considering g as $p(g)$ positive examples and $1-p(g)$ negative examples for each $g \in G_i \cap G_R$.

   (c) Use the Naïve Bayes assumption together with the positive and negative distributions to update $p(g)$ for each $g \in G_R$.

3 Results and Conclusions

The two major performance indicators for both methods are the **recall**, which is the percentage of pathway metagenes that are classified as being in the pathway, and the **precision**, which is the percentage of positively classified genes that are in the pathway. The following bar graph gives recall and precision results for both methods. The iterative Bayesian updating method using just the GEN displays modest improvement in both recall and precision over the simple neighbor based method using just the GEN. When the PIN is added, the precision of the iterative Bayesian updating improves dramatically with a small cost in recall, while the performance of the simple neighbor based method degrades. The probabilistic framework of the iterative Bayesian method give it the flexibility to integrate multiple data sources, while the rigid "and" logic of the simple neighbor based method makes it less adaptable.

Figure 1. Optimal recall and precision for the simple neighbor based method and the iterative Bayesian updating method.



References

[1] S. Altschul, W. Gish, et al. Basic local alignment search tool. *Journal of Molecular Biology*, 215:403–410, 1990.

[2] Available on web site `http://biodata.mshri.on.ca/grid/servlet/Index`.

[3] M. Kanehisa and S. Goto. KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28:27–30, 2000.

[4] S. Li et al. A map of the interactome network of the metazoan c. elegans. *Science*, 303(5657):540–543, January 2004.

[5] J. M. Stuart, E. Segal, D. Koller, and S. K. Kim. A gene-coexpression network for global discovery of conserved genetic modules. *Science*, 302:249–55, 2003.

[6] I. Xenarios, D. Rice, L. Salwinski, M. Baron, E. Marcotte, and D. Eisenberg. DIP: The database of interacting proteins. *Nucleic Acids Research*, 28:289–291, 2000.