REFINING THE PROTEIN-PROTEIN INTERACTOME USING GENE EXPRESSION DATA

SHANE T. JENSEN$^1$ and SIRA SRISWASDI$^2$

Abstract

Proteins interact with other proteins within biological pathways, forming connected subgraphs in the protein-protein interactome (PPI). However, proteins are often involved in multiple biological pathways which complicates inference about interactions between proteins. Gene expression data is informative here since genes within a particular pathway tend to have more correlated expression patterns than genes from distinct pathways. We provide an algorithm that uses gene expression information to remove inter-pathway protein-protein interactions, thereby simplifying the structure of the protein-protein interactome. This refined topology permits easier interpretation of multiple biological pathways simultaneously.

1 Introduction

The protein-protein interactome (PPI) is a large graph where proteins are nodes and edges between these nodes represent all known interactions between proteins. In cases where proteins interact in order to drive a particular biological process, the connected nodes of a PPI can represent an entire biological pathway. However, inferring a biological pathway from the PPI is complicated by the fact that many proteins are involved in multiple biological functions. Thus, a connected subgraph of the PPI must be viewed as a mixture of smaller graphs that each represent a particular pathway. It is the goal of this paper to refine the PPI by isolating these smaller graph components which are more likely to contain just a single pathway.

Our primary tool for this endeavor is gene expression data, which allows us to identify pairs of genes with highly correlated expression patterns. In general, gene pairs are more likely to have correlated expression if they belong to the same biological pathway, which gives us a mechanism for refining the PPI to isolate individual pathways. We introduce a procedure for reducing large connected com-

---

$^1$Department of Statistics, The Wharton School, University of Pennsylvania  
stjensen@wharton.upenn.edu  

$^2$Genomics and Computational Biology Graduate Group, University of Pennsylvania  
sirasris@mail.med.upenn.edu
ponents of the PPI into smaller groups with higher connectivity that represent a single pathway.

2 Methods

The input for our procedure is a protein-protein interactome and a set of gene expression profiles. Our data sources are given in the supplementary materials, which can be downloaded at:

http://stat.wharton.upenn.edu/~stjensen/research/ppi.html

Our algorithm focusses initially on proteins with the highest degree in the PPI as potential multi-pathway proteins. Expression profiles are used to infer interactions of this protein that span multiple pathways, and then the local topology of the PPI is edited to emphasize connectivity within single pathways. The overall framework of our algorithm is shown in Figure 1.

![Figure 1](http://stat.wharton.upenn.edu/~stjensen/research/ppi.html)

**Figure 1: Overall framework of our PPI refining procedure.**
Let $A$ be the protein that currently has the largest number of connections in the PPI. We denote $\mathcal{N}(A)$ as a set containing all proteins connected to $A$ via protein-protein interactions. For each pair of proteins $i$ and $j$ in $\mathcal{N}(A)$, we calculate the correlation $\rho_{ij}$ of their gene expression patterns. An agglomerative hierarchical clustering of the proteins in $\mathcal{N}(A)$ is performed using $d_{ij} = 1 - |\rho_{ij}|$ as the distance metric. A pre-specified threshold $\theta_{\text{cor}}$ is used to convert this hierarchical clustering into a partition of disjoint subsets $N_1, N_2, \ldots, N_m$ of highly correlated proteins, as well as an extra subset $N_{m+1}$ containing unclustered proteins.

We proceed under the assumption that each highly correlated subset $N_1, \ldots, N_m$ of $\mathcal{N}(A)$ is a group of proteins belonging to the same pathway. We remove inter-pathway connections within $\mathcal{N}(A)$ by replacing protein $A$ in the PPI with duplicates $A_1, A_2, \ldots, A_m$, where $A_i$ retains only the connections between $A$ and proteins in $N_i$. We also discard all connections between protein $A$ and proteins contained in the unclustered set $N_{m+1}$.

The expression clustering and network reduction steps are repeated for all highly-connected proteins in the protein-protein interactome. We terminate the algorithm when no protein in the refined PPI contains more connections than a pre-specified degree cutoff $\theta_{\text{deg}}$.

3 Evaluation using gene ontology

We examine the effects of our algorithm using the gene ontology (GO) database\footnote{The Gene Ontology Consortium, 2000} (The Gene Ontology Consortium, 2000), which is a multi-level collection of biological terms that are assigned to specific genes. The GO database contains three types of biological terms: cellular component, molecular function, and biological process. Molecular function is the most specific type, but many proteins either lack molecular function annotations or do not share a common annotation with other proteins. In contrast, most proteins are annotated with a cellular component GO term, but this feature is too broad to be particularly informative. We focus our analysis on biological process GO terms as the GO type most closely related to our goal of isolating biological pathways.

For a group of connected proteins in the protein-protein interactome, we define an evaluation metric called the GO distance. The GO distance is the depth in the GO hierarchy of the deepest GO term that is common to all proteins in a connected group. The GO hierarchy becomes more specific as the depth increases, so large GO distances are indicative of a group of proteins that have high coherence in

\footnote{Downloaded: April, 2009}
We quantify the improvement in coherence of the PPI by comparing the distributions of GO distances over all genes in the original and the refined versions of the PPI. Specifically, we focus on the mean GO distances over all genes in the original PPI compared to our refinement of the PPI. In Figure 2, we examine the refined PPI for multiple versions of our procedure corresponding to different choices of the absolute co-expression cutoff $\theta_{cor}$ parameter and the degree cutoff $\theta_{deg}$ parameter.

![Figure 2: Mean GO Distance for different parameter cutoffs](image)

We see that for all choices of input parameters $\theta_{cor}$ and $\theta_{deg}$, the refined PPI from our procedure shows a dramatically larger mean GO distance than the original PPI. This result demonstrates that the refined protein connections from our procedure have a much greater coherence in their biological processes compared to the original PPI. Although our procedure results in a PPI with greater biological
coherence, there is one sacrifice: some proteins are removed completely from the refined PPI due to all of their connections to other proteins being removed. We must balance our increase in biological coherence with the reduction in the number of proteins contained in the PPI. In Figure 3 we give the number of proteins in the original PPI as well as the number of proteins remaining in the PPI for each version of our procedure given in Figure 2.

Not surprisingly, stricter choices of the threshold parameters $\theta_{\text{cor}}$ and $\theta_{\text{deg}}$ lead to a PPI that has many proteins removed. Based on these results, we suggest parameter values of $\theta_{\text{cor}} = 0.8$ and $\theta_{\text{deg}} = 4$ as a good compromise that gives increased biological coherence without the removal of too many proteins from the PPI. We provide our refined PPI under these parameter settings along with code for producing refined PPIs under other parameter settings at:
Our procedure depends not only on input parameters $\theta_{\text{cor}}$ and $\theta_{\text{deg}}$ but also on the ordering of proteins in the iterative portion of the algorithm. We investigated the results of using a random ordering to our iterative algorithm compared to the default setting of process proteins in order of highest degree first. Details are given on our supplementary materials, but we found that our default choice of proteins ordered by highest degree was superior to random orderings.

In summary, we have provided an iterative algorithm that uses gene expression data to refine the protein-protein interactome. Our evaluation suggests that our refined PPI has greater biological coherence than the original PPI, at least in terms of the GO biological processes category.

**References**

The Gene Ontology Consortium (2000). Gene ontology: tool for the unification of biology. *Nature Genetics* **25**, 25–29.