Mining Cross-graph Quasi-cliques in Gene Expression and Protein Interaction Data *

Jian Pei
Simon Fraser University
jpei@cs.sfu.ca

Daxin Jiang
State University of New York at Buffalo
djiang3@cse.buffalo.edu

Aidong Zhang
State University of New York at Buffalo
azhang@cse.buffalo.edu

1 Problem Description and Model

A protein is the product of a gene. From the gene expression data, we can find co-expressed genes, which are groups of genes that demonstrate coherent patterns on samples. On the other hand, from the protein interaction data, we can find groups of proteins that frequently interact with each other. If we can conduct a joint mining of both gene expression data and protein interaction data, then we may find the clusters of genes that are co-expressed and also their proteins interact.

Such clusters found from the joint mining are interesting and meaningful for at least two reasons. First, both the gene expression data and the protein data are very noisy. The clusters confirmed by both data sets will strongly indicate the correlation/connection among the genes in a cluster. In other words, the clusters found from the joint mining are more reliable. We may thus have the high confidence that the genes in a cluster found as such are regulated by the same mechanism or belong to the same biological process.

Second, although highly related, gene expression data and protein interaction data still carry different biological meaning. The coincidence of co-expressed genes and interacting proteins is biologically significant. As indicated in [5], many pathways exhibit two properties: their genes exhibit a similar gene expression profile, and the protein products of the genes often interact.

1.1 Model

Technically, a gene expression data set is a matrix $W = \{w_{ij}\}$ for a set G of n genes and a set S of m samples, where $w_{i,j}$ $(1 \leq i \leq n, 1 \leq j \leq m)$ is the expression level of gene g_i on sample s_j.

Two genes g_1 and g_2 are called *coherent* if they show similar expression patterns on the set of samples. There are different methods to measure the similarity (or distance) between gene expression patterns as required by the application domain, such as Euclidean distance, Pearson's correlation coefficient, KL-distance [2], and pattern-based similarity measures [1, 6]. Without loss of generality, in this paper, we simply assume that a similarity measure $sim(\cdot)$ is specified, and the higher the similarity value, the more similar the genes.

We can define a binary relation $\sim$ on the set of genes. For genes g_1 and g_2, $g_1 \sim g_2$ if $sim(g_1, g_2) \geq \delta$, where δ is a user-specified *minimum similarity threshold*.

Naturally, the relation $\sim$ can be represented as *gene expression graph* $geneG = (G, E)$: the genes are the vertices, and $(g_1, g_2) \in E$ if $g_1 \sim g_2$.

Similarly, for a set of proteins P, if we have the data about the interactions between proteins, we can define a *protein interaction graph* $proteinG = (P, I)$: the proteins are treated as vertices, and $(p, p') \in I$ if proteins p and p' interact with each other.

For gene expression data, a subset of genes forms a perfect cluster if each gene in the subset is similar to all the others in the same subset. For protein interaction data, a subset of proteins forms a perfect cluster if each protein in the subset interacts with all the others in the same subset. To generalize, in gene expression/protein interaction graphs, a *perfect cluster* is a clique[1].

However, due to the noise in the data sets, we may not be able to expect perfect clusters. Instead, a user may be interested in a subset of genes/proteins as a cluster such that each gene in the subset is similar to most of the other genes in the cluster, and each protein in the subset interacts with most of the other proteins in the cluster.

To quantify, for a user-specified threshold γ $(0 < \gamma \leq 1)$, a subset C of k genes forms a γ-quasi-cluster if each gene $g \in C$ is similar to at least $\gamma \cdot (k-1)$ other genes in C. Similarly, we can define γ-quasi-cluster for protein interaction data. Clearly, a maximal γ-quasi-cluster is a γ-quasi-clique in the corresponding gene expression/protein interaction graph.

*This research is partly supported by the Endowed Research Fellowship and the President Research Grant from Simon Fraser University, NSF grants IIS-0308001, DBI-0234895, and NIH grant 1 P20 GM067650-01A1. All opinions, findings, conclusions and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

[1] In this paper, we follow the terminology usage that a clique is a *maximal* subset of mutually adjacent vertices in a graph.

Since a protein is a product of a gene, there is a mapping f from the set of proteins P to the set of genes G: $f(p) = g$ if protein p is the product of gene g.

We are particularly interested in subsets of proteins C such that C is a γ_1-quasi-cluster in the protein interaction data and $\{f(c)|c \in C\}$ is a γ_2-quasi-cluster in the gene expression data, where γ_1 and γ_2 are user-specified parameters. We call C a *cross-data set cluster*. Moreover, C is particularly interesting if it is maximal.

1.2 Why Is the Problem Challenging?

One may ask, "*Can we solve the joint mining problem by a simple extension of the existing techniques?*" Unfortunately, the answer is no.

A natural thinking may be as follows. We can integrate the multiple graphs into one based on a similarity function between data objects. The integrated similar function combines the similarity between data objects in different data sets in some weighted manner. Then, we can find quasi-cliques in the integrated graph.

However, the above naïve method does not work at all. The key is that vertices of cross-graph quasi-cliques can be connected in different ways in individual graphs. Therefore, the integrated graph cannot capture the cross-graph quasi-cliques. It is easy to come up with a counter example to show that a cross-graph quasi-clique is not a quasi-clique in the integrated graph.

2 Experimental Results

We use the cell-cycle gene expression data CDC28 and the corresponding protein-protein interaction data from DIP as the data set. We found $4,668$ matched gene-protein pairs between CDC28 and DIP. For CDC28 data set, we set the coherence threshold $\rho = 0.5$ (using Pearson's correlation coefficient as measure). As a result, the gene graph G_E contains $865,080$ edges whose both endpoints (genes) appear in the matched gene-protein pairs. After removing the self-interacting protein pairs, the protein graph G_P contains $15,115$ edges whose both endpoints (proteins) appear in the matched gene-protein pairs.

In our experiments, we find the complete set of quasi-cliques across the gene graph G_E and the protein graph G_P. We set $\gamma_E = 1$ for G_E, $\gamma_P = 0.5$ for G_P, and $min_s = 5$. That is, we are interested in a subset of at least 5 genes whose expression patterns are coherent with each other, and the corresponding proteins frequently interact with each other.

Figure 1 shows an example pattern Q ($\gamma_E = 1$ and $\gamma_P = 0.4$). The induced graph of G_E (the gene expression graph) on Q is a perfect clique, so we only show the induced graph of G_P (the protein interaction graph) on Q here. The pattern contains 11 vertices. We use the ORF (Open Reading Frame) names to identify the corresponding genes and proteins.

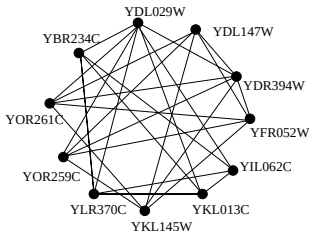

Figure 1. A cluster of 11 proteins.

Although the exact biological meaning of this pattern is still under investigation, it is very interesting in biology since these 11 genes are highly coherent and the corresponding 11 proteins are intensively interacting.

3 Related Work

For more examples on joint mining of multiple sources, Page and Craven [3] surveyed the biological applications of mining multiple tables, such as pharmacophore discovery, gene regulation, information extraction from text and sequence analysis.

Recently, joint mining of multiple biological data sets has received intense interest. As a pioneer work, Segal et al. [5] proposed a unified probabilistic model to learn the pathways from gene expression data and protein interaction data. However, their method requires the users to input the number of pathways that is usually unknown in advance.

4 About the Full Version of This Paper

In the full version of this paper [4], we built a general model, investigated the properties of the problem and the computational complexity, and developed an effective and efficient algorithm to tackle the problem. A systematic performance study was also reported.

References

[1] Cheng, Y. and Church, G.M. Biclustering of expression data. *Proceedings of ISMB'00*, 8:93–103, 2000.

[2] Kasturi, J., Ramanathan, M. and Acharya, R. An information theoretic approach for analyzing temporal patterns of gene expression. *Bioinformatics*.

[3] Page, D. and Craven, M. Biological Applications of Multi-relational Data Mining. *SIGKDD Explorations*, 5(1):69–79, July 2003.

[4] Pei, J., Jiang, D. and Zhang, A. On Mining Cross-graph Quasi-cliques. *Technical Report TR 2004-15, School of Computing Science, Simon Fraser University.*

[5] Segal, E., Wang, H. and Koller, D. Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, 19:i264–i272, 2003.

[6] Wang, H., Wang, W., Yang, J. et al. Clustering by Pattern Similarity in Large Data Sets. In *Proceedings of SIGMOD'02*, pages 394–405, 2002.