GPX: Interactive Mining of Gene Expression Data [*]

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

Jian Pei
State University of New York
at Buffalo, USA
Simon Fraser University, Canada
jianpei@cse.buffalo.edu

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

Abstract

Discovering co-expressed genes and coherent expression patterns in gene expression data is an important data analysis task in bioinformatics research and biomedical applications. Although various clustering methods have been proposed, two tough challenges still remain on how to integrate the users' domain knowledge and how to handle the high connectivity in the data. Recently, we have systematically studied the problem and proposed an effective approach [3]. In this paper, we describe a demonstration of *GPX* (for Gene Pattern eXplorer), an integrated environment for interactive exploration of coherent expression patterns and co-expressed genes in gene expression data. *GPX* integrates several novel techniques, including the *coherent pattern index graph*, a *gene annotation panel*, and a *graphical interface*, to adopt users' domain knowledge and support explorative operations in the clustering procedure. The *GPX* system as well as its techniques will be showcased, and the progress of *GPX* will be exemplified using several real-world gene expression data sets.

1 Motivation

The DNA microarray technology has enabled measuring the expression levels of thousands of genes during important biological processes and across collections of related samples. It is often an important task to identify the co-expressed genes and the coherent expression patterns from the gene expression data. A group of *co-expressed genes* are the ones with similar expression profiles, while a *coherent expression pattern* characterizes the common trend of expression levels for a group of co-expressed genes. In practice, co-expressed genes may belong to the same or similar functional categories and indicate co-regulated families [9]. Coherent expression patterns may characterize important cellular processes and suggest the regulating mechanism in the cells [7].

To find co-expressed genes and discover coherent expression patterns, many gene clustering methods have been proposed, e.g., [9, 8, 1, 4, 5]. Each cluster is considered as a group of co-expressed genes and the coherent expression pattern can be simply the mean (the centroid) of the expression profiles of the genes in that cluster. While the clustering algorithms have been shown useful to identify co-expressed gene groups and discover coherent expression patterns, due to the specific characteristics of gene expression data and the special requirements from the biology domain, several great challenges for clustering gene expression data remained. Two of them are as follow.

Challenge 1: It is hard to integrate the domain- and user-knowledge to properly unfold the hierarchies of co-expressed genes and coherent patterns.

In a gene expression data set, there are usually multiple groups of co-expressed genes as well as the corresponding coherent patterns. Moreover, there is typically a hierarchy of co-expressed genes and coherent patterns in a gene expression data set. At the high levels of the hierarchy, large groups of genes approximately follow some "rough" coherent expression patterns. At the low levels of the hierarchy, the large groups of genes break into smaller subgroups. Those smaller groups of co-expressed genes follow some "fine" coherent expression patterns, which inherit some characteristics from the "rough" patterns, and add some distinct characteristics.

One subtle point here is that *there is no precise definition or objective standard to identify co-expressed gene groups*. The interpretation of co-expression depends on the knowledge from domain experts. For example, a microarray ex-

[*] This research is supported in part by NSF grants DBI-0234895, IIS-0308001 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.

Proceedings of the 30th VLDB Conference,
Toronto, Canada, 2004

periment typically involves thousands of genes. However, only a small subset (e.g. several hundred) of those genes may play important roles in the biological process under investigation. As an initial examination, biologists may prefer browsing the "rough" patterns in the data set. Then, they may choose the patterns of particular interest and decompose them into "finer" patterns in further analysis. In other words, biologists may have different requirements of "coherence" for different parts of the data set and at different stages of the analysis.

However, many clustering algorithms generate clusters at a single level. It is hard to see the inherent hierarchical relationship among the groups of co-expressed genes as well as the coherent patterns. Although some hierarchical approaches exist, it is usually hard to determine where to cut the resulted hierarchical dendrogram to meet the various clustering requirements for different subsets of genes.

Moreover, most clustering algorithms are "purely" unsupervised approaches. A user often cannot be involved in the clustering procedure. To derive a satisfactory result, a user may have to try different algorithms and/or different parameter values. Apparently, such a make-do-and-mend approach is deficient. Instead, if users can apply their domain knowledge at some critical points during the clustering process, it may become more effective and efficient to get meaningful results. However, how to involve users in the clustering process and integrate the domain knowledge is still an open problem.

Challenge 2: It is hard to handle the high connectivity in the gene expression data sets.

An interesting phenomenon in gene expression data sets is that *groups of co-expressed genes may be highly connected by a large amount of "intermediate" genes*. Technically, two genes g_i and g_j that have very different expression profiles in a data set may be bridged by a series of intermediate genes such that each two consecutive genes on the bridge have similar profiles. Our empirical study has shown that such "bridges" are common in gene expression data sets.

The high connectivity in the gene expression data raises a challenge: *It is often hard to find the (clear) borders among the clusters*. Many existing clustering methods use one of the following two strategies.

On the one hand, the data set is decomposed into numerous small clusters. While some clusters consist of groups of biologically meaningful co-expressed genes, many clusters may consist of only intermediate genes. Since there is no biologically meaningful criteria (e.g., size, compactness) to rank the resulted clusters, it may take a lot of effort to examine which clusters are meaningful groups of co-expressed genes. On the other hand, an algorithm may form several large clusters. Each cluster contains both the co-expressed genes and a large amount of intermediate genes. However, those intermediate genes may mislead the centroids of the clusters into going astray. The centroids then no longer represent the true coherent patterns in the groups of co-expressed genes.

To address the above challenges, we developed *GPX* (for Gene Pattern eXplorer), a research prototype system that supports interactive exploration of co-expressed genes and coherent expression patterns in gene expression data sets.

2 Features of GPX

2.1 Interactive Exploration Operations

A user can explore the co-expressed genes and their coherent patterns by unfolding a hierarchy of genes and patterns. The exploration starts from the root. To help a user to make decision to split the genes and detect the patterns, a *coherent pattern index graph* [3] is used at each node of the tree to illustrate the cluster structure in the corresponding subset of data at the node. Each pulse in the coherent pattern index graph indicates the potential existence of a coherent pattern, and a higher pulse represents a stronger indication. Based on the index graph, GPX supports several exploration operations [6]. Two essential ones are *drill-down* and *roll-up*.

Drill-down. A user can select the pulse(s) in the index graph, and the system will split the genes accordingly. Each split subset of genes becomes a child node of the current node. If the user does not specify any pulse, the system will choose the highest pulse by default and split the genes accordingly.

Roll-up. A user can revoke any drill-down operation. The user can select a node and undo the drill-down operation from this node. All descendants of this node will be deleted. The user can also roll up one node A to its parent P. This equals to skipping the selection of the pulse in P's index graph that corresponds to A, and undoing the drill down operation for P.

2.2 A Robust Model for Clusters and Patterns

Most existing clustering methods try to find the clusters (i.e., groups of co-expressed genes) based on some global criteria, and then derive the coherent patterns as the centroids of the clusters. Such strategies may be sensitive to a large amount of intermediate genes in data sets. Contrast to those methods, *GPX* adopts a novel strategy: it first explores the hierarchy of coherent patterns in the data set and then finds the groups of co-expressed genes according to the coherent expression patterns.

In *GPX*, a cluster of co-expressed genes is modeled as a dense area in the multidimensional gene space. Genes at the "center" of the dense area have relatively high density and present the coherent pattern of the whole cluster. Genes at the periphery of the dense area have relatively low density and will be "attracted" toward the center area level by level.

Through this density-based model, *GPX* can distinguish co-expressed genes from intermediate genes by their relative density. The coherent expression pattern in a dense area is represented by the expression profile of the gene that has the highest local density in the dense area. Other genes in the same dense area can be sorted in a list according to the similarity (from high to low) between their expression profiles and the coherent expression pattern. Since the intermediate genes have low similarity to the coherent pattern, they are at the rear part of the sorted list. Users can set up a similarity threshold and thus cut the intermediate genes from the cluster.

2.3 Graphical Interface and Gene Annotation Panel

A graphical interface provides users a direct impression of the trends of gene expression levels. Users may interpret the underlying biological process and decide whether the co-expressed genes should be further split into finer subgroups. In *GPX*, we use the *parallel coordinate* to illustrate the expression profiles of genes. We also visualize the whole hierarchical structure of the co-expressed genes and coherent patterns. Users can browse the hierarchical tree, select a node and apply the exploration operations (Figure 1(b)).

There exist very rich literatures about the functions and regulation mechanisms of genes. It would be very helpful to integrate such domain knowledge into the system. For example, given a group of co-expressed genes, if the well studied genes in this group are significantly populated in a certain gene function category, biologists may postulate that the functions of the novel genes in the group fall in the same function category. On the other hand, if a group of co-expressed genes scatter into diverse functional categories, biologists may further split this group or roll back to the parent node and choose a different splitting path. To meet this need, we design a *gene annotation panel*. Given a specific node on the hierarchical tree, the panel sorts the genes belonging to the node as we described in Feature 2, and displays the name and the annotation (if any) for each gene (Figure 1(c)). The gene annotations are downloaded from some public databases, such as the Gene Ontology Consortium (http://www.geneontology.org).

3 Major Components of GPX

3.1 Data Preprocessor and Pattern Manager

Usually the original data obtained from microarray experiments contains missing values and variations arising from the experimental procedure. Data pre-processing is indispensable before any cluster analysis can be performed. In *GPX*, we apply the K-nearest neighbor approach described in [10] to estimate the missing values. Then the system performs a logarithmic transformation of each expression level and standardizes each gene expression profile with a mean of zero and a variance of one. Finally the system calculates the pairwise distance between gene expression profiles and stores the pre-processed data and the pairwise distances. Once a data set has been pre-processed, users can explore the coherent patterns in the data set and save/load the coherent patterns through the pattern manager (Figure 1(a)).

3.2 Interactive Exploration Environment

GPX has a *working zone* (Figure 1(b)), which integrates the parallel coordinates, the coherent pattern index graph, and a *tree view*. An example tree view is shown in Figure 2, which illustrates the hierarchical structure of co-expressed genes and coherent expression patterns in the well known Iyer's data set [2]. Users can select a node in the tree view, then the working zone will display the corresponding expression profiles and coherent pattern index graph. Users can click on the coherent pattern index graph to split the node or roll back previous split operations. The tree structure is adjusted dynamically according to the exploration operations.

3.3 Gene Annotation Panel

Once users select a node on the tree structure, the gene annotation panel will sort the genes and display their names and annotations accordingly (Figure 1(c)).

4 About the Demonstration

Our demo consists of three major parts.

First, we present the *techniques* to interactively clustering and analyzing gene expression data. We will analyze the rationale of our designs, as well as their advantages and disadvantages. We will also illustrate the effects of our method using real data sets.

Second, we present a set of *real case studies* on the proposed techniques. The experimental results on several real data sets will be exhibited.

Last, we showcase our *prototype system*, including a data analysis engine and an interactive user interface. The audience is encouraged to play with the prototype system and experience the exciting tour over real data sets.

After the demonstration, the integrated *GPX* system will be available on web for public access.

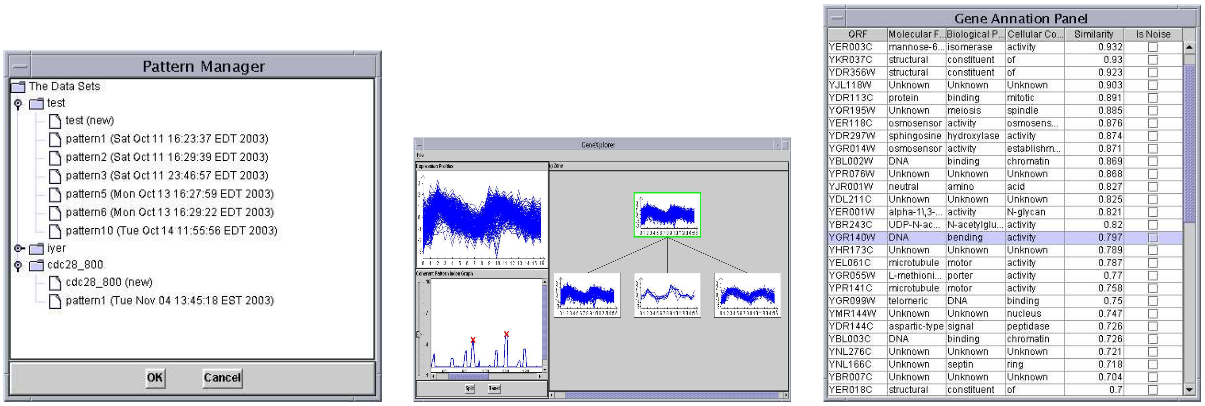
References

[1] Eisen M.B., Spellman P.T., Brown P.O. and Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95(25):14863–14868, December 1998.

[2] Iyer V.R., Eisen M.B., Ross D.T., Schuler G., Moore T., Lee J.C.F., Trent J.M., Staudt L.M., Hudson Jr. J., Boguski M.S., Lashkari D., Shalon D., Botstein D. and Brown P.O. The transcriptional program in the response of human fibroblasts to serum. *Science*, 283:83–87, 1999.

[3] Jiang D., Pei J. and Zhang A. Interactive exploration of coherent patterns in time-series gene expression data. In *Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'03)*, Washington, DC, USA, August 24-27 2003.

[4] Jiang D., Pei J. and Zhang A. Towards interactive exploration of gene expression patterns. *ACM SIGKDD Ex-*



Pattern Manager

- The Data Sets
  - test
    - test (new)
    - pattern1 (Sat Oct 11 16:23:37 EDT 2003)
    - pattern2 (Sat Oct 11 16:29:39 EDT 2003)
    - pattern3 (Sat Oct 11 23:46:57 EDT 2003)
    - pattern5 (Mon Oct 13 16:27:59 EDT 2003)
    - pattern6 (Mon Oct 13 16:29:22 EDT 2003)
    - pattern10 (Tue Oct 14 11:55:56 EDT 2003)
  - iyer
  - cdc28_800
    - cdc28_800 (new)
    - pattern1 (Tue Nov 04 13:45:18 EST 2003)

OK Cancel

Gene Annation Panel

ORF	Molecular F...	Biological P...	Cellular Co...	Similarity	Is Noise
YER003C	mannose-6...	isomerase	activity	0.932	□
YKR037C	structural	constituent	of	0.93	□
YDR356W	structural	constituent	of	0.923	□
YJL118W	Unknown	Unknown	Unknown	0.903	□
YDR113C	protein	binding	mitotic	0.891	□
YOR195W	Unknown	meiosis	spindle	0.885	□
YER118C	osmosensor	activity	osmosens...	0.876	□
YDR297W	sphingosine	hydroxylase	activity	0.874	□
YGR014W	osmosensor	activity	establishm...	0.871	□
YBL002W	DNA	binding	chromatin	0.869	□
YPR076W	Unknown	Unknown	Unknown	0.868	□
YJR001W	neutral	amino	acid	0.827	□
YDL211C	Unknown	Unknown	Unknown	0.825	□
YER001W	alpha-1\,3-...	activity	N-glycan	0.821	□
YBR243C	UDP-N-ac...	N-acetylglu...	activity	0.82	□
YGR140W	DNA	bending	activity	0.797	□
YHR173C	Unknown	Unknown	Unknown	0.789	□
YEL061C	microtubule	motor	activity	0.787	□
YGR055W	L-methioni...	porter	activity	0.77	□
YPR141C	microtubule	motor	activity	0.758	□
YGR099W	telomeric	DNA	binding	0.75	□
YMR144W	Unknown	Unknown	nucleus	0.747	□
YDR144C	aspartic-type	signal	peptidase	0.726	□
YBL003C	DNA	binding	chromatin	0.726	□
YNL276C	Unknown	Unknown	Unknown	0.721	□
YNL166C	Unknown	septin	ring	0.718	□
YBR007C	Unknown	Unknown	Unknown	0.704	□
YER018C	structural	constituent	of	0.7	□

(a) Pattern manager (b) Working zone (c) Gene annotation panel

Figure 1: Screen snapshots of *GPX*

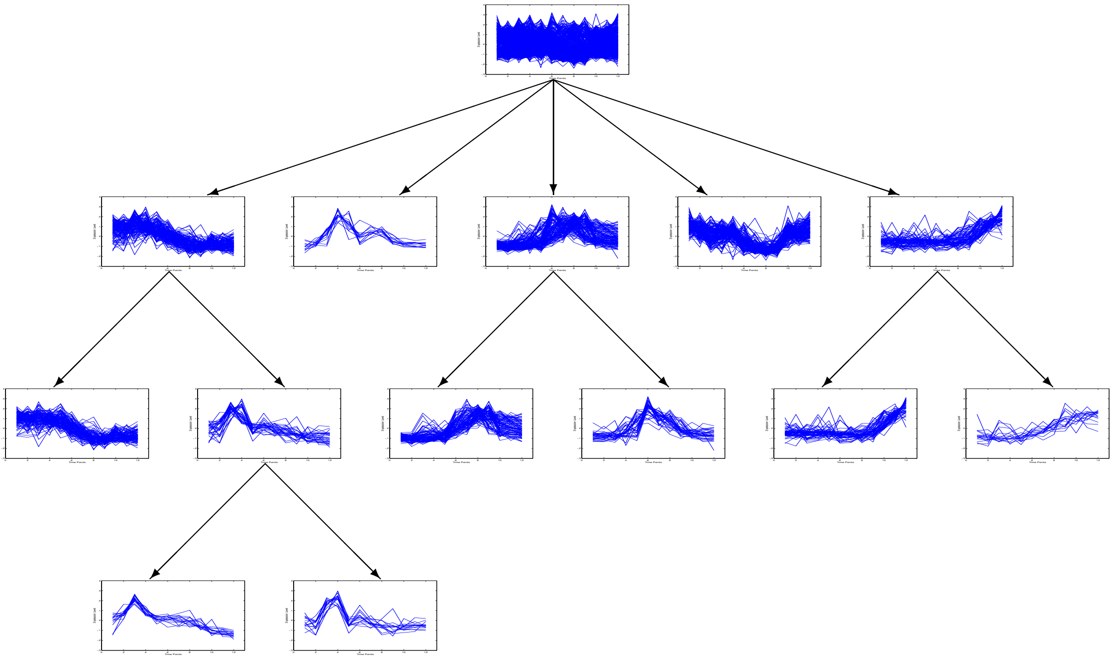


Figure 2: The tree view of co-expressed gene groups on Iyer's data set [2]

plorations (Special Issue on Microarray Data Analysis), 5(2):79–90, 2004.

[5] Jiang D., Pei J., Ramanathan M., Tang C. and Zhang A. Mining coherent gene clusters from three-dimensional microarray data. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04)*, Seattle, Washington, USA, August 22-25 2004.

[6] Pei, J. A general model for online analytical processing of complex data. In *Proceedings of the 22nd International Conference on Conceptual Modeling (ER'03)*, Chicago, IL, October 13-26 2003.

[7] Spellman P.T., Sherlock G., Zhang M.Q., Iyer V.R., Anders K., Eisen M.B., Brown P.O., Botstein D. and Futcher B. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell*, 9(12):3273–3297, 1998.

[8] Tamayo P., Solni D., Mesirov J., Zhu Q., Kitareewan S., Dmitrovsky E., Lander E.S. and Golub T.R. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA*, Vol. 96(6):2907–2912, March 1999.

[9] Tavazoie S., Hughes D., Campbell M.J., Cho R.J. and Church G.M. Systematic determination of genetic network architecture. *Nature Genet*, pages 281–285, 1999.

[10] Troyanskaya O., Cantor M., Sherlock G., Brown P., Hastie T., Tibshirani R., Botstein D. and Altman R. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, Jun 2001.