GeneXPress: A Visualization and Statistical Analysis Tool for Gene Expression and Sequence Data

E. Segal[1], A. Kaushal[1], R. Yelensky[1], T. Pham[1], A. Regev[2], D. Koller[1] and N. Friedman[3]

[1]Computer Science Department, Stanford University, [2]Bauer Center for Genomics Research, Harvard University and [3]School of Computer Science and Eng., Hebrew University

ABSTRACT

Many algorithms have been developed for analyzing gene expression and sequence data. However, to extract biological understanding, scientists often have to perform further time consuming post-processing on the output of these algorithms. In this paper, we present *GeneXPress*, a tool designed to facilitate the assginment of biological meaning to gene expression patterns by automating this post processing stage. Within a few simple steps that take at most several minutes, a user of GeneXPress can: identify the biological processes represented by each cluster; identify the DNA binding sites that are unique to the genes in each cluster; and examine multiple visualizations of the expression and sequence data. GeneXPress thus allows the researcher to quickly identify potentially new biological discoveries. GeneXPress is available for download at http://GeneXPress.stanford.edu.

Contact: *E-mail*: eran@cs.stanford.edu

MOTIVATION

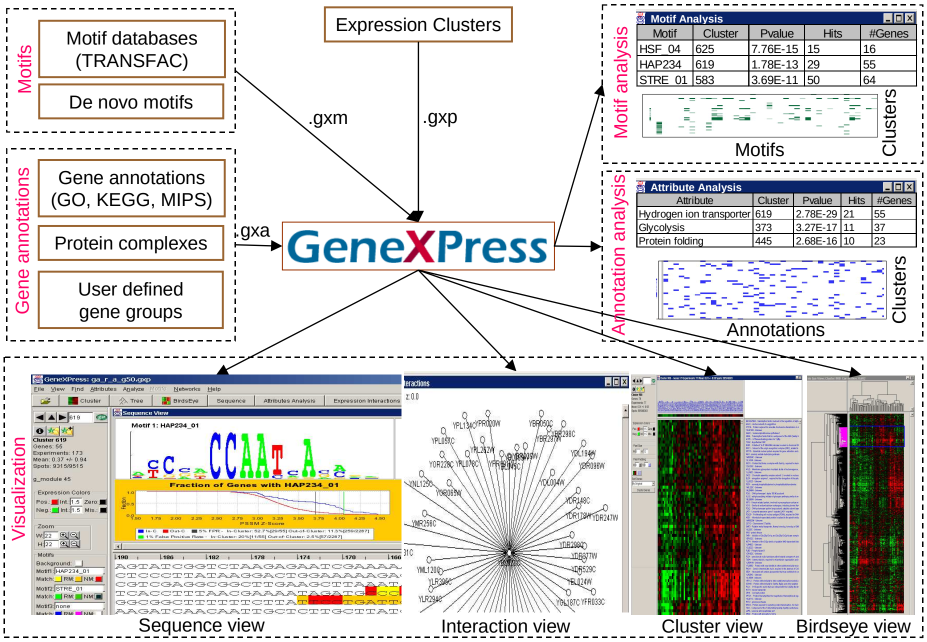
The availability of complete genomic sequences and genome wide measurements of gene expression provide us with the means to understand cellular processes and their regulation on a genome wide scale. Indeed, much recent work has been devoted to analysis of these data for this purpose. The most common method for analyzing gene expression data is clustering (e.g., (1)), that groups together genes with similar expression profiles. Genes that are similarly expressed often participate in the same cellular processes, so clustering suggests functional relationships between the clustered genes. Similarly, we expect co-clustered genes to be co-regulated by the same *cis*-regulatory mechanism, which can be revealed by searching for commonly occurring motifs in the promoter regions of the genes in a cluster (e.g., (5)).

The outputs of clustering and motif finding algorithms provide the basis for understanding the biological story underlying the data. However, to extract concrete biological understanding , further time-consuming post-analysis of such outputs is required. It is not rare for the analysis and post-analysis stages of a gene expression experiment to take several months of intensive manual work. This post-analysis usually focuses on relating gene expression patterns with other form of biological knowledge. During the analysis there is a need to answer questions such as: what biological processes are represented by each cluster; what *cis*-regulatory motifs are shared by genes within a cluster; how significant these associations are; and more. This type of analysis requires a multitude of scripts, visualizations, comparisons to multiple biological databases, and more. Currently, this work is duplicated many times, both within and between labs.

In this paper, we present *GeneXPress*, a general-purpose visualization and analysis tool that is designed to support extensive post-analysis of gene expression experiments. GeneXPress contains a suite of tools to automatically answer questions such as the ones we described above, through visual and statistical analysis of the outputs of clustering and motif finding algorithms. GeneXPress has several different visualizations that allow both global and detailed views of expression profiles, promoter regions, and motifs. Through statistical analysis of the clusters relative to databases of gene annotations (e.g., GO — http://geneontology.org), GeneXPress can associate each cluster with one or more biological processes. Through similar analysis for motifs, GeneXPress can identify the motifs that are present in the promoter regions of the genes in each cluster. The discovered associations are statistically benchmarked by p-values that are automatically computed for each association.

GeneXPress uses simple and extensible XML-based file formats. It is easy to convert the output of clustering and motif finding algorithms to such format, and use them within GeneXPress. In addition, GeneXPress supports files generated for viewing with TreeView (http://rana.lbl.gov), the most commonly used software for visualizing expression data. GeneXPress implements all the views provided by TreeView, but enhances them to include additional convenient features.

GeneXPress is freely available at
http://GeneXPress.stanford.edu. The web site also provides sample files, detailed tutorials, and gene annotation and sequence motif files from existing databases that can be loaded to GeneXPress and used for analyzing



Motif Analysis

Motif	Cluster	Pvalue	Hits	#Genes
HSF_04	625	7.76E-15	15	16
HAP234	619	1.78E-13	29	55
STRE_01	583	3.69E-11	50	64

Attribute Analysis

Attribute	Cluster	Pvalue	Hits	#Genes
Hydrogen ion transporter	619	2.78E-29	21	55
Glycolysis	373	3.27E-17	11	37
Protein folding	445	2.68E-16	10	23

Fig. 1. The GeneXPress architecture: clustering results (top), sequence motifs (left) and gene annotations (left) can be loaded into GeneXPress. The expression and sequence data can then be viewed using several different views (bottom). Clusters can be analyzed for their enrichment for the loaded gene annotations and sequence motifs (right).

clustering outputs.

We have used GeneXPress extensively for analyzing the results of algorithms that we developed for detecting regulators of biological processes (3) and for identifying the sequence motifs that play a role in controlling the expression of genes (4).

MAIN PROGRAM FEATURES

Visualization of Gene Expression. GeneXPress supports three formats for loading expression data and corresponding clustering information: tab-delimited; TreeView; and its own .gxp (XML) format. The gxp format can represent gene hierarchies as well as mutually exclusive gene groupings, and can thus support the output of many types of clustering algorithms. GeneXPress has several customizable and interactive views for the loaded gene expression data: a *birdseye view* of the entire expression data and the hierarchy in which the genes and experiments are organized; a *cluster view* that shows in detail the expression of the genes in the currently selected cluster, along with their descriptions and the experiment descriptions; and an *interaction view* that places one gene from the currently selected cluster in the center and shows all the other genes from the cluster that have a Pearson correlation above a selected threshold with the center gene.

Statistical Analysis of Clusters. GeneXPress can associate gene annotations with clusters. Annotations are loaded either as tab-delimited files or as .gxa (XML) files, and can come from a variety of sources including gene functional annotations (e.g., GO), protein sequence motifs (e.g., INTERPRO), protein complex data, or any user defined file that specifies annotations for genes. GeneXPress can then identify all the cluster-annotation pairs in which the fraction of genes in the cluster having the annotation is higher then expected, and associate a hypergeometric p-value for this this association. For a cluster with n genes, of which k are annotated with a certain annotation that exists in K of the N genes in the database, the hypergeometric p-value is given by:

$$P(X \geq k) = \sum_{i=k}^{n} \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}}$$

where $P(X \geq k)$ represents the probability that the cluster has k or more genes with the annotation. The results can then be viewed in two different representations: an excel-like table sortable by annotations, clusters, or p-values; and a matrix of clusters versus annotations where entries for which the annotation is significantly enriched for the cluster (a p-value under some specified threshold) are colored and the color intensity is used to represent the fraction of genes in the cluster that have the annotation.

Visualization and Statistical Analysis of Motifs. Sequence motifs are loaded into GeneXPress from .gxm (XML) files, which can represent several motifs and also points to a sequence file in the *fasta* format containing the promoter regions of all genes. Each motif is encoded using the common *Position Specific Scoring Matrix* (PSSM) representation. GeneXPress uses the PSSM to score each putative k-mer binding site for its fit to the motif. These scores are standardized to z-scores relating to the mean and variance of the scores of random

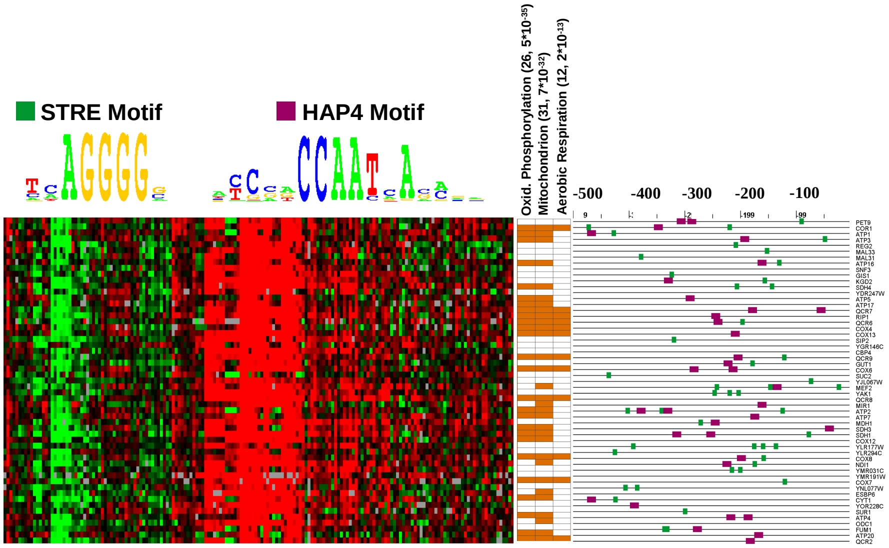


Fig. 2. Visualization of the example cluster mentioned in the text. From left to right, shown is the expression of the genes in the cluster, the significant annotations enriched in the genes in this cluster (e.g. oxidative phosphorylation), and the promoters of the genes in the cluster along with the significant motifs enriched in the genes in the cluster. The logos of the motifs (also generated by GeneXPress) are shown in the upper part of the figure.

k-mers sampled from the background nucleotide distribution in promoter sequences. Genes whose promoter sequence contains at least one k-mer with z-score above a user-specified threshold, are annotated as bound by the motif. These annotations are then analyzed as standard gene-annotations by computing the p-value of the motif's uniqueness for the cluster. As for gene annotations, the results of a global analysis of motif-cluster pairs can be viewed in either a sortable table or a matrix. When zooming in to a specific cluster, we can view how specific motifs relate to the selected cluster. GeneXPress helps illustrate the uniqueness of the motif to the cluster with a chart that shows, for each possible PSSM z-score s, the fraction of genes in and out of the cluster whose promoter region contains at least one k-mer with score above s. The promoter regions of all genes in the cluster are shown and all occurrences of the selected motifs above the chosen z-score threshold are indicated on the promoter.

EXAMPLE CASE STUDY

We now demonstrate the ease with which GeneXPress can be used to give biological interpretation to clustering results. We applied the method of (3), which identifies modules of co-regulated genes and their regulators, to the yeast expression dataset of (2) and loaded the results into GeneXPress. We also loaded a gxm file of known sequence motifs from the TRANSFAC database (6) and a gxa gene annotation file of known gene annotations from GO. We performed a global statistical analysis of both motifs and annotations as described above. All of the above steps took approximately 3 minutes. We note that the output of any standard clustering program can be similarly loaded into GeneXPress.

At this point, the biological story of the results started to emerge. For example, in one of the clusters, 26 of the 55 genes were annotated in GO as oxidative phosphorylation from a total of 31 genes annotated as such in GO (p-value $5 \cdot 10^{-34}$). Moreover, 39 of the 55 genes had the Hap4 motif in their promoter region (p-value $2 \cdot 10^{-13}$); as Hap4 is a known activator of the oxidative phosphorylation pathway, its presence confirms our analysis of the cluster. We also visualized the cluster and saw exactly which genes were part of the pathway according to GO, which genes had the motif, and under what conditions these genes were activated (see Fig. 2).

Using a similar process, we found other clusters that spanned a wide variety of annotations, including cell-cycle, RNA processing, stress responses, amino acid metabolism, cell wall organization, protein degradation, chromatin remodeling, and glycolysis. For many clusters, known motifs associated with the cluster confirmed the annotation analysis.

REFERENCES

[1] M. Eisen et al. Cluster analysis and display of genome-wide expression patterns. *PNAS*, 95:14863–68, 1998.

[2] A.P. Gasch et al. Genomic expression program in the response of yeast cells to environmental changes. *Mol. Bio. Cell*, 11:4241–4257, 2000.

[3] E. Segal, M. Shapira, A. Regev, D. Pe'er, D. Botstein, D. Koller, and N. Friedman. Module networks: Discovering regulatory modules and their condition specific regulators from gene expression data. *Nature Genetics*, 34(2): 166–76, 2003.

[4] E. Segal, R. Yelensky, and D. Koller. Genome-wide discovery of transcriptional modules from DNA sequence and gene expression. *Bioinformatics*, 19(Suppl 1): 1273–82, 2003.

[5] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church. Systematic determination of genetic network architecture. *Nature Genetics*, 22(3):281–5, 1999.

[6] E. Wingender et al. The TRANSFAC system on gene expression regulation. *NAR*, 29:281–283, 2001.