BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 11 2004, pages 1804–1805
doi:10.1093/bioinformatics/bth154

Genomic data visualization on the Web

Wei Wu* and William S. Noble

Department of Genome Sciences, University of Washington, 1705 NE Pacific Street, Seattle, WA 98195, USA

Received on September 23, 2003; revised on December 19, 2003; accepted on January 6, 2004
Advance Access publication February 26, 2004

ABSTRACT
Many types of genomic data can be represented in matrix format, with rows corresponding to genes and columns corresponding to gene features. The heat map is a popular technique for visualizing such data, plotting the data on a two-dimensional grid and using a color scale to represent the magnitude of each matrix entry. Prism is a Web-based software tool for generating annotated heat map visualizations of genome-wide data quickly. The tool provides a selection of genome-specific annotation catalogs as well as a catalog upload capability. The heat maps generated are clickable, allowing the user to drill down to examine specific matrix entries, and gene annotations are linked to relevant genomic databases.

Availability: http://noble.gs.washington.edu/prism

Contact: weiw@u.washington.edu

Humans are surprisingly good at visual pattern recognition. Consequently, converting complex data sets into a visual format often leads to useful insights. We describe a tool for visualizing genome-wide data sets of the kind generated by microarray expression experiments. The tool, called Prism, is available for interactive use via a Web interface.

Prism represents data via a heat map (Fig. 1), which was popularized in the context of gene expression analysis by the TreeView program (rana.lbl.gov/EisenSoftware.htm). As such, Prism takes as input any genomic data set that can be represented as a collection of fixed-length vectors of real numbers. Microarray expression data are the most common such data, but many other types of data can be cast in the same format: occurrences of domains in protein sequences, occurrences of protein orthologs across multiple genomes (phylogenetic profiles), protein–protein interactions, occurrences of binding site motifs in promoter regions, etc.

The input to Prism is a tab-delimited text file, in which the first row contains the headers for each column and each subsequent row corresponds to a single gene or protein. By default, Prism assumes that the first column contains the gene identifier and that subsequent columns contain data. However, this format is flexible, and the user can specify that the gene ID occurs in a different column, that some columns contain text to be printed next to the heat map and that some columns should be ignored entirely. In addition, the input file may contain 'flag' columns. A flag value in the n-th flag column indicates that the value in the n-th data column is missing or suspect.

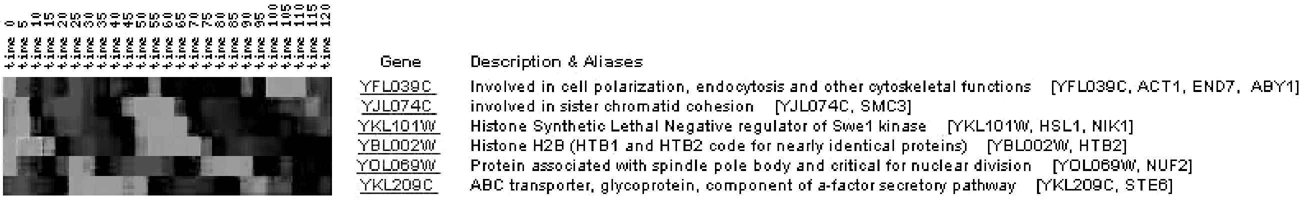
In addition to providing the data file, the user must select a catalog and a Web database for use in annotating the heat map representation. Two sets of catalogs are currently available. One set is based on gene ontology annotations (Gene Ontology Consortium, 2000), including catalogs for human, mouse, rat, zebrafish, fruitfly, *Caenorhabditis elegans*, *Arabidopsis* and budding and fission yeasts. The other set includes 34 catalogs for Affymetrix chips derived from NetAffx annotations (Liu *et al*., 2003). Each catalog contains multiple identifiers (delimited by commas) and a brief description for each gene, thereby allowing users to employ any standard naming scheme in their data files. Furthermore, users have the option of uploading their own data-specific catalog (in a simple, tab-delimited format). In addition to the catalog, the user-specified Web database is used to create hot links on each gene ID. When the user clicks the ID, the database is queried with that ID and the resulting page is displayed.

The primary Prism output is a single heat map representation of the entire data set (Fig. 1). Rows and columns of the matrix are labeled by gene IDs and column titles, respectively. In addition, each row may be labeled with additional, user-specified columns and with annotations from the selected catalog. The order of rows in the matrix is the same as in the input file unless the user requests sorting according to a given input column. The data rows can also be filtered by a threshold imposed on a given input column. The color scheme for the heat map can be selected by the user. Flagged data values are indicated by gray boxes in the heat map.

Each row in the heat map representation is clickable, allowing the user to drill down to examine the behavior of a specific gene (Fig. 2). These gene-specific pages are generated on the fly, and each one contains a line or histogram representation of the values associated with that gene. In these plots, flagged values are denoted by '+'.

Prism assigns an experiment number to every data file when it is uploaded for the first time. Using this number, the user can

*To whom correspondence should be addressed.



Gene	Description & Aliases
YFL039C	Involved in cell polarization, endocytosis and other cytoskeletal functions [YFL039C, ACT1, END7, ABY1]
YJL074C	involved in sister chromatid cohesion [YJL074C, SMC3]
YKL101W	Histone Synthetic Lethal Negative regulator of Swe1 kinase [YKL101W, HSL1, NIK1]
YBL002W	Histone H2B (HTB1 and HTB2 code for nearly identical proteins) [YBL002W, HTB2]
YOL069W	Protein associated with spindle pole body and critical for nuclear division [YOL069W, NUF2]
YKL209C	ABC transporter, glycoprotein, component of a-factor secretory pathway [YKL209C, STE6]

Fig. 1. A small heat map produced by Prism, showing yeast time series data. Each gene identifier is represented as a hyperlink pointing to the gene-specific page in the yeast genome database. Gray boxes for gene YBL002W at times 5 and 10 indicate missing data.

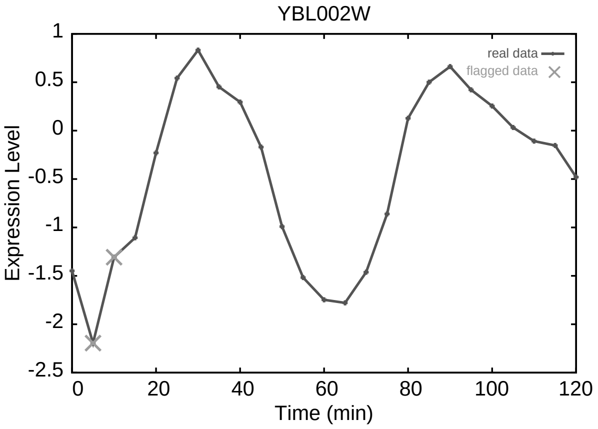


Fig. 2. The gene-specific plot that is produced by clicking on the fourth row of the heat map in Figure 1. This plot shows time series data; for non-time series data, a histogram would be produced.

return later to the same data set or share experimental results with a colleague. Prism stores each data file, along with the user-specified display options, for seven days.

Prism is written in Perl using the CGI module. Heat maps are generated using matrix2png (Pavlidis and Noble, 2003), and gene-specific plots are produced by gnuplot. Prism is freely available under a GNU public license. Two Perl programs for making the catalogs are also available: Go2Prism for making gene ontology-based catalogs and NetAffixer for making catalogs for Affymetrix chips.

Many software tools exist for visualizing and analyzing microarray expression data. These range from high-end commercially supported software costing thousands of dollars to freely available toolkits. Among these, TreeView and GeneXPress (genexpress.stanford.edu) are popular, free programs for viewing the results of gene expression clustering performed by the associated Cluster program. GeneXPress also includes additional analytical functions such as analyzing gene clusters for enrichment of transcription factor binding sites (e.g. TRANSFAC motifs) or functional annotations (e.g. Gene Ontology). These programs must be installed locally and are targeted specifically toward analyzing the results of clustering algorithms. Several, more complex Web-based platforms are available. For example, the EPCLUST module of the Expression Profiler package (ep.ebi.ac.uk) is a generic Web-based data clustering and visualization tool. Users can upload data files along with gene annotations and perform various clustering procedures. The Web-based GEDA (Gene Expression Data Analysis Tool) package (bioinformatics.upmc.edu/GE2/GEDA.html) also provides data normalization and clustering functions, and tests for differentially expressed genes. However, neither EPCLUST nor GEDA provides heat map visualizations. The BioArray Software Environment (base.thep.lu.se) is a free, Web-based platform for comprehensive management, storage and analysis of microarray data. In general, these tools are much more complex than Prism and accomplish a much broader range of analytical tasks.

Prism is targeted toward novice, casual or computer-shy users who are unwilling or are unable to install software tools on their local computers. Prism is useful, e.g. for many biologists who perform much of their microarray analyses in Microsoft Excel, which cannot produce heat maps visualizations but which can export tab-delimited text files for use by Prism. Prism is by no means a sufficient tool for microarray analysis, but by exploiting the powerful human visual system, Prism's graphical output can provide a valuable complement to traditional statistical analyses.

ACKNOWLEDGEMENTS

The authors thank Paul Pavlidis for advice and help on this project. This work is supported by a Sloan Research Fellowship to W.S.N. and by National Science Foundation grants DBI-0078523 and ISI-0093302.

REFERENCES

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

Liu,G., Loraine,A., Shigeta,R., Cline,M., Cheng,J., Valmeekam,V., Sun,S., Kulp,D. and Siani-Rose,M. (2003) Netaffx: Affymetrix probesets and annotations. *Nucleic Acids Res.*, **31**, 82–86.

Pavlidis,P. and Noble,W.S. (2003) Matrix2png: a utility for visualizing matrix data. *Bioinformatics*, **19**, 295–296.