

Gene Expression Mural: Visualizing Gene Expression Databases

Mathew Clement, Margaret Ellis, Josh Steele, Yuying Tian, Chris North
Department of Computer Science
Virginia Tech, Blacksburg, VA 24061
north@cs.vt.edu

Abstract

The Gene Expression Mural is a tool designed for managing the vast amount of information produced by the rapidly growing field of Bioinformatics. The advantages of this tool are that it not only provides an overview of many experiments for an entire genome, but also promotes discovery with its zooming and navigation capabilities. A “tier array” data structure and an averaging algorithm were implemented to facilitate efficient interaction with enormous amounts of data from experiment databases.

1. Introduction

Bioinformatics is an emerging field due to the development of new technologies. The rapidly accelerating advances in high-throughput technologies, including screening, robotics, and combinatorial biology makes bioinformatics an extremely data-rich environment. Microarray experiments produce large amounts of numerical data quantifying the expression level of gene sequences in a number of conditions [1]. Software tools integrating analysis methods with interactive visualization are essential for scientists to recognize gene expression patterns and uncover new knowledge [2].

Roughly, a genome consists of a set of chromosomes. Each chromosome is a linear sequence of up to 10^8 units. One gene expression experiment consists of numeric values for $\sim 10^4$ sub-sequences, each of $\sim 10^2$ units in length, within the chromosome. The values represent activity level ratios of the chromosome subsequences (“gene expression”).

The Gene Expression Mural provides a uniform approach to visualize gene expression levels across entire genomes. The key advantages the Gene Expression Mural provides are:

- overview of entire genome for analysis of gene expression patterns across a chromosome.
- investigation and comparison of many experiments simultaneously.

- interaction capabilities for zooming on genome regions of interest, filtering experiments for display, and display of detail information.

Users can potentially view over 100 experiments simultaneously. Multiple experiment results are displayed for each chromosome on the screen for users to directly view and compare data across experiments. This visualization design alleviates users from navigating through cumbersome windows to find data, and allows users to quickly locate patterns and anomalies in the data based on color and location.

2. Display

The Gene Expression Mural displays each experiment as a horizontal row representing a chromosome and its measured expression level values. Many experiments can be listed vertically, and are horizontally aligned.

The Gene Expression Mural has two different formats for users to view the expression levels: histogram view (Figure 1) and mural view (Figure 2). The mural view maps the expression ratio to a color value and represents each region by a small rectangle. In addition to mapping the ratio to color, the histogram format maps ratios to vertical height of the rectangle. The legend for the color values is shown on the bottom left.

Since length is perceived more readily than color [3], the histogram view is more efficient. However, the histograms require more vertical space than the murals. Hence, the mural view is effective for displaying an overview of an entire database of experiments, whereas the histogram view is better when users focus in on a few experiments of interest.

3. Interaction

Initially, the Gene Expression Mural displays an overview of the data for an entire chromosome containing millions of expression values by aggregating values over large regions [4]. This enables the histogram or mural to fit in one screen width. Users can then zoom in on interesting regions of a

chromosome (horizontal zooming) to expose more detailed, less aggregated data [5]. Users can also vary the number of experiments shown on the screen (vertical zooming) to make more space for a histogram-style view. Furthermore, additional information for each experiment can be displayed by clicking the individual experiment buttons. Users can select among the chromosomes of a genome using the buttons on the bottom. By using these visualization tools, scientists can investigate the gene expression patterns at any level of scale, including: the entire genome, a particular chromosome, a specific area of a chromosome, or across potential multi-gene families.

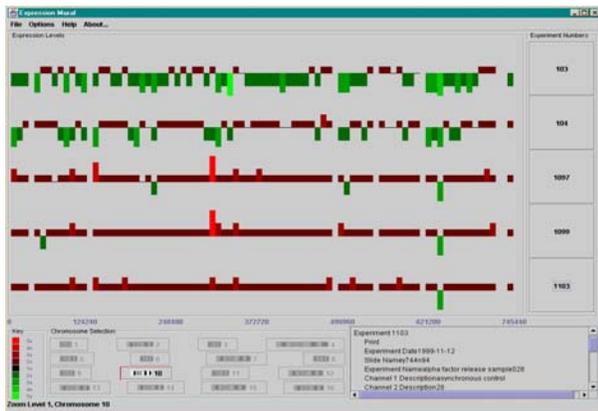


Figure 1: Histogram view of 5 experiments.

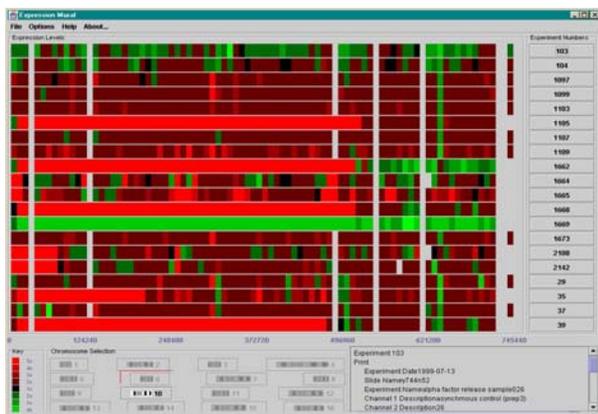


Figure 2: Mural view showing 22 experiments.

4. Algorithm and Data Structures

The interaction capabilities are implemented using a “tier array” data structure and an averaging algorithm, which are the key components of Gene Expression Mural’s advantageous design. The data is inserted into Java’s TreeMap data structure according to the gene sequence starting point. For each zoom level, a tier in

the tier array is constructed from the original TreeMap to preserve as much data integrity as possible and prevent averaging loss. As the user zooms in on an area of a chromosome, more precise expression values are displayed. As the user zooms out to analyze more of the chromosome, average weighted ratios are visualized. The average weighted ratios represent the expression level for a specific region of a chromosome. The averaging algorithm handles the complex nuances of Microarray data, such as:

- the expression values are ratios,
- the gene sequences can overlap, and
- some ranges have no data values.

6. Conclusions and Future Work

The Gene Expression Mural provides an integrated environment for visualizing data from gene expression databases. The methods and data structures implemented are successfully applied to the analysis of gene expression data. The tier array data structure approach to visualize hierarchical structures enables meaningful representations of huge amounts of data. This overview combined with navigation and zooming capabilities provides the foundation for a powerful scientific tool to discover unknown patterns in databases of experimental results.

Continued work focuses on improved methods for visualizing aggregated data, smooth zooming of the histogram and mural views, automated and user-controlled vertical sorting of experiments based on similarity, and integration with additional data sources.

7. References

- [1] Hamadeh, H., Afshari, C., "Gene Chips and Functional Genomics", *American Scientist*, vol. 88, pp. 508-515, (Nov. 2000).
- [2] Dysvik B., Jonassen I., “J-express: exploring gene expression data using Java”, *Bioinformatics Applications Note*, 17(4), pp. 369-370, (2001).
- [3] Cleveland W., McGill, R., “Graphical Perception and Graphical Methods for Analyzing Scientific Data”, *Science*, 229, pp. 828-833, (1985).
- [4] Jerding, D., Stasko, J., “The Information Mural: A Technique for Displaying and Navigating Large Information Spaces”, *Proc. IEEE Symposium on Information Visualization*, pp. 43-50, (1995).
- [5] Bederson, B., Hollan, J., “Pad++: A Zooming Graphical Interface for Exploring Alternate Interface Physics”, *Proc. ACM UIST’94 - User Interface Software and Technology*, pp. 17-26, (1994).