Viewing the Larger Context of Genomic Data through Horizontal Integration

Matthew Hibbs[1,2], Grant Wallace[1], Maitreya Dunham[2], Kai Li[1], and Olga Troyanskaya[1,2]
[1]Department of Computer Science
[2]Lewis-Sigler Institute for Integrative Genomics
Princeton University, Princeton, New Jersey 08544
{mhibbs, gwallace, maitreya, li, ogt}@princeton.edu

Abstract

Genomics is an important emerging scientific field that relies on meaningful data visualization as a key step in analysis. Specifically, most investigation of gene expression microarray data is performed using visualization techniques. However, as microarrays become more ubiquitous, researchers must analyze their own data within the context of previously published work in order to gain a more complete understanding. No current method for microarray visualization and analysis enables biology researchers to observe the greater context of data that surrounds their own results, which severely limits the ability of researchers draw novel conclusions. Here we present a system, called HIDRA, that visually integrates the simultaneous display of multiple microarray datasets to identify important parallels and dissimilarities. We demonstrate the power of our approach through examples of real-world biological insights that can be observed using HIDRA that are not apparent using other techniques.

1 Introduction

Scientifically meaningful data visualization is vital for the advancement of knowledge in many fields, particularly molecular biology. Genomics is one of the fastest growing modern scientific disciplines, as it promises a better understanding of the inner workings of cells, is vital to understand diseases, elaborates our understanding of evolution, moves towards the era of personalized medicine, and reveals the root causes of cancer. One of the most powerful new tools molecular biologists wield to solve these problems are gene expression microarrays, and the majority of microarray analysis is done through visualization techniques[1, 2].

Gene expression microarrays simultaneously measure the activation or suppression of every gene in a genome at a particular point in time. These studies result in data matrices containing hundreds of thousands to millions of observations, and the majority of researchers rely on visualization tools to mine these data to discover new biological information. Biologists face the challenges of understanding not only the data that they generate, but also of comprehending their results in the broader context of previous studies. As microarray technology matures, decreases in cost, and becomes more accessible, the number of microarray studies produced is growing exponentially, which further complicates thorough analysis.

No existing method for microarray visualization enables researchers to directly understand and analyze their data within the greater context of previously published findings. This severely limits research capabilities by forcing users to focus on their own data during the initial analysis phase and to compare with other studies only at later stages to confirm or contradict their conclusions. Integrating the vast amount of available data into the analysis phase as early and seamlessly as possible will allow researchers to build upon previous results, observe inconsistencies, and form more powerful conclusions.

We propose a novel methodology for the analysis and exploration of multiple microarray datasets simultaneously. By leveraging visual paradigms that are commonly used for small-scale microarray analysis, our approach remains easily interpretable by researchers. Due to the sheer size of these datasets, we employ an "overview + detail" approach on a per-dataset basis to allow users to view specific genes as well as their context within the whole genome. However, we extend this paradigm to include the larger context of additional available datasets as well, which we call an "overview + detail + setting" paradigm.

We have implemented our approach into a system called HIDRA (Horizontally Integrated Dataset Relationship Analysis), and we have deployed this system to experimental genomics researchers both for individual use and for collaborative use on a large-format display device. In the next section we discuss existing microarray visualization approaches in more detail. We then outline our specific visualization goals and what techniques we use to achieve those goals. And finally we show two case studies where meaningful biological observations have been made by researchers using our system.

2 Related Work

Existing microarray visualization tools focus on the analysis of single datasets, and many of these tools are used on a daily basis by the research community[3]. The majority of visual displays of microarray data fall into two major categories: heat maps[4-6] and parallel coordinates[7]. Other approaches are also used, such as scatterplots, histograms, and spreadsheets, but these are generally complementary techniques used in conjunction with a heat map and/or parallel coordinate display[4, 8-10].

Heat map displays traditionally show a clustered data matrix of values represented as colors interpolated from red to green. This type of display allows a user to quickly identify prevalent patterns among genes in a dataset by looking for bands of data with similar profiles. These displays are often accompanied by a dendrogram created from hierarchical clustering, which dictates the order in which genes are displayed and visually encodes a distance metric relationship between genes.

Heat maps have seen near universal adoption amongst biologists, and their results are the canonical representation of gene expression used in the majority of microarray publications. While these displays allow the full matrix of data values to be viewed, the patterns and labels of individual genes are only visible by zooming into more detailed portions of the map. Many tools support this type of exploration by using an "overview + detail" paradigm[11], where users see the entire dataset, but can then select a smaller region to see in greater detail.

Parallel coordinate systems display genes as a collection of segmented lines overlaid on a measurement grid. These displays have the ability to show all of the available data in a relatively small area. This approach is also well suited for the identification of desired patterns, as users are able to select only those genes that pass through defined portions of the grid.

While parallel coordinate views show all of the available data, the results can be difficult to interpret. When viewing a large number of genes simultaneously, it is difficult to distinguish one expression profile from another. As with heat maps, this approach suffers from not being able to label individual genes within the total plot. The absence of a dendrogram created from hierarchical clustering presents both benefits and complications. The dendrogram visually indicates a quantitative distance metric between two genes in a dataset, but it also enforces an ordering and structure on the data that may be somewhat artificial. Parallel coordinate displays do not suffer from this imposition of ordering, but do not visually quantify arbitrary distances between profiles.

Many of the most successful microarray visualization approaches combine both heat map displays and parallel coordinates views, along with several other views of the same data[3]. We refer to these approaches as "vertically integrated" as they allow researchers to see the same data from many complementary angles. These methods have been very successful and have gained wide use among the microarray analysis community.

We propose extending the power of multiple simultaneous views in an orthogonal direction. Rather than displaying multiple viewpoints of the same data, our approach displays the same type of viewpoint on multiple datasets at the same time -- we refer to this paradigm as a "horizontally integrated" approach. This expansion of the amount of visualized data enables researchers to view a broader setting of known biology and place their own results within this larger context.

3 Design & Implementation

We established several goals for the design of our microarray visualization methodology that incorporates broader context. The following goals are a combination of our initial aspirations and the desires of our research collaborators that used our system:

- *Ease of use.* A successful system must be usable and intuitive for the target audience, in this case biology researchers.
- *Dynamic, consistent interaction.* The approach must be adaptive to user input as their desire to explore and observe information changes over time, but these adaptations must feel natural to the user.
- *Scalability.* Our approach must scale both with the amount of data visualized, and with available screen space.
- *Biologically meaningful.* Perhaps the most important criteria is that a microarray visualization system must enable researchers to explore their data in a way that facilitates biological observations and insights.

3.1 Single dataset visualization

In order to maintain a baseline of usability and comfort with the microarray analysis community, we have chosen to adopt the use of heat maps accompanied by dendrograms as the basis for our methodology. This approach is by far the most common presentation format for microarray data in biology literature. For individual dataset display we leveraged the codebase of the commonly used, open-source tool, JavaTreeView[5], which we then modified for our purposes. This provides the immediate advantage of utilizing pre-existing abilities and biases of the microarray research community. On the level of a single dataset we also utilize the "overview + detail" paradigm to allow users to view both the entire dataset, as well as a more detailed view of a subset of that data. An example of this visualization for a single dataset is shown in Figure 1.

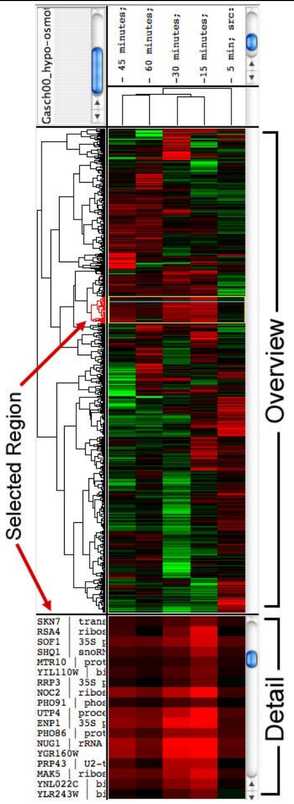

Figure 1 – A single dataset displayed using a heat map and dendrogram in an overview + detail format. Rows correspond to genes and columns to experimental conditions. Each intersection is colored on a continuous scale from green through black to red. The data was hierarchically clustered in both dimensions. A region of the dataset was selected in the overview and the corresponding section is shown in greater detail below.

Users have several options for interacting with this display of information. Subsets of genes to view in the detail portion can be selected by dragging a box on the heat map, or by choosing branches of the dendrogram. These selections can be refined by traversing up or down the dendrogram using the keyboard. This allows users to isolate particular desired areas of the larger dataset to view with greater scrutiny.

Due to differences in experimental technologies and personal preferences/abilities, it is also important for users to maintain control regarding the parameters of the heat map coloration. In general, microarray data lies in a broad, noisy range of values that depends on several laboratory factors. For this reason, values above/below a cutoff are saturated out to a maximum intensity, but this cutoff is not universal, and should thus default to a reasonable value, but be in the control of the user. Additionally, the color scheme used for display must be adjustable by the user. The red/green gradient is commonly employed because it has a direct link to the chemical dyes used in microarray experiments, however such a scheme is clearly unacceptable for color-blind researchers.

3.2 Multiple dataset visualization

Several factors are important to consider when incorporating additional datasets into microarray visualization. The common features of microarray datasets are genes, while the experimental conditions vary between datasets, which indicates that between dataset comparisons should be visible on a per-gene basis. However, microarray datasets are often created using disparate technologies or experimental practices, and individual datasets are generally targeted to investigate a specific area or process, which indicates that information such as clustering and normalization are appropriate only on a per-dataset basis.

In order to address these biological requirements, we have developed an approach we refer to as "overview + detail + setting". On the level of each dataset it is vital to observe both the entire dataset (overview) as well as more specific information (detail). However, for the larger goal of placing an individual researcher's data in the greater context of available data, datasets must be linked together (setting). In particular, we applied this approach to microarray data with the goal of making comparisons between datasets as intuitive as possible, while maintaining important per-dataset information.

The most common paradigm in microarray literature is for the expression of genes to correspond to rows of a visualized data matrix. As genes are the common element of interest between datasets, we place the datasets next to each other horizontally to preserve gene-row orientation across all data. However, the ordering of genes is determined by clustering, and the clustering process is biologically meaningful on the level of individual datasets. To address these issues, we have synchronized the detail views across all datasets to facilitate comparisons, while preserving the cluster order of individual datasets in the overviews.

By synchronizing the detail views, we preserve the expectation that gene measurements are aligned along rows, even across multiple datasets. The order of the genes shown in the detail views corresponds to the order of those genes in the dataset where the selection was made. To provide information about the per-dataset context of the selected genes, a thin line is displayed in

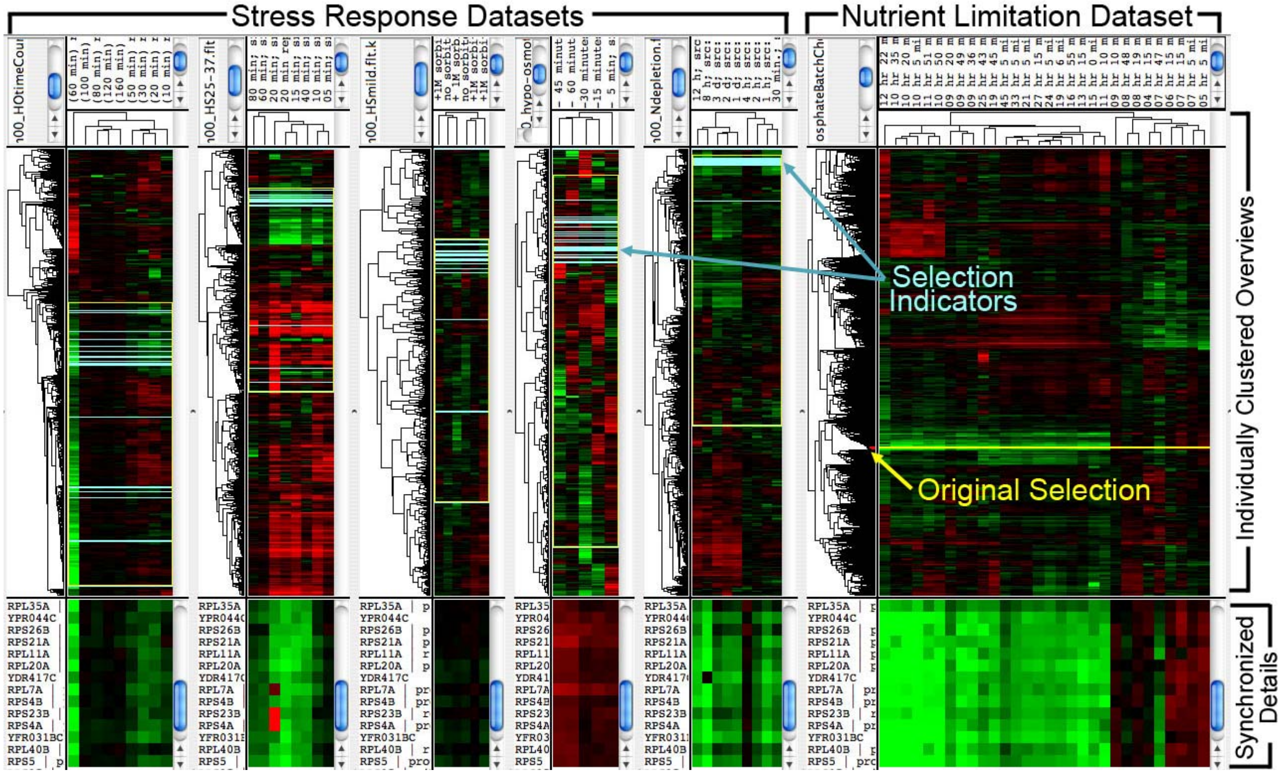

Figure 2 – A selection of disparate datasets viewed in HIDRA. Six different datasets are shown here tiled horizontally. Each dataset was individually hierarchically clustered in both dimensions. A selection has been made in the rightmost dataset (from a nutrient limitation study[12]) and the thin light blue lines in the left five datasets (from a stress response study[13]) indicate where these genes are located in their overviews. A user can quickly observe that the selected genes are non-randomly grouped in the clustering of the other datasets. Further inspection of the aligned genes in the detail views shows cases where these genes are behaving similarly/differently.

each dataset's overview to indicate where each selected gene falls within that dataset. An example of this multiple dataset visualization is shown in Figure 2. The gene-level synchronization of the detail views enables low-level comparisons of a gene's specific behavior in different datasets; while in the overviews, the selection highlights indicate a higher-level comparison of gene group relationships between datasets.

For example, a user can select a tight group of genes in one dataset, and immediately observe how those genes cluster together in every other dataset at a general level. A researcher can then examine the detail views to investigate the specific expression levels that led to the observed global patterns. This type of exploratory analysis across a large amount of diverse datasets is impossible with existing tools, but is vital for experimental microarray analysis, as we demonstrate in our validation.

3.3 Scalability, interactions, and interfaces

The inclusion of multiple datasets also requires addressing scalability, interaction and user interface concerns[14]. First, as more data is viewed simultaneously, screen space quickly becomes an issue. While several datasets can be viewed at once on even the smallest desktop/laptop displays, users may be in situations where they still feel limited. By default, when enough datasets are loaded to overflow the available display space, a scrollbar becomes active to pan between datasets. We also provide the ability to dynamically re-order, remove, and/or add new datasets as the researcher explores their data. In this manner users can choose the most relevant datasets to occupy the visible area as their needs change over time.

Another option to see more data is to move to large-scale display devices if they are available to the user[15-17]. Our approach scales very well to large-format devices by providing control over text size, column widths, row heights, etc (Figure 3). Using displays of this magnitude allows users to see as much as an order of magnitude more data at once. These very high-resolution displays are also helpful for collaboration, which is very common among microarray analysts.

Regarding the user interface, several visualization choices must be made on a per-dataset basis. In particular, the desired color scale, saturation cutoffs, dendrogram widths, etc. often vary greatly from one dataset to another, due to technological and experimental differences. We provide controls to alter all of these parameters for any selected subset of datasets, including the individual level. Further, we store these choices on a per-dataset basis, so that as users re-order, remove, and/or re-load data these per-dataset choices remain intact.

Further, some interactions should be consistent from one dataset to the next. In order to preserve the gene-row alignment of the detail views, the heights of each panel are slaved to any panel being resized, such that all detail views maintain the same height. Additionally, scrolling in any detail view causes synchronous scrolls in all detail views to maintain consistency.

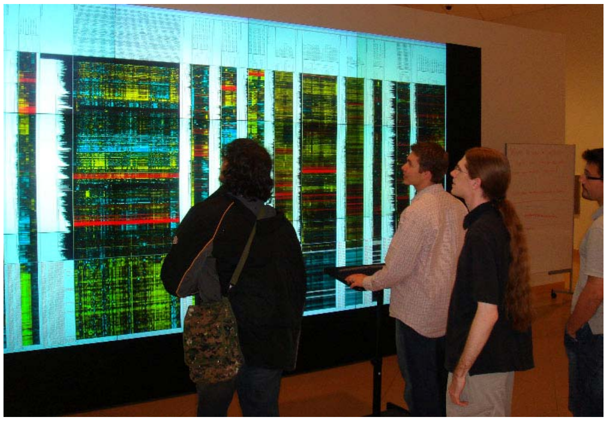

Figure 3 – A group of collaborators using HIDRA on the large-scale display wall at the Lewis-Sigler Institute for Integrative Genomics. This display is capable of simultaneously showing an order of magnitude more data than traditional desktop/laptop displays, which is helpful when dealing with very large data repositories.

3.4 Implementation

We have implemented these methods into a Java-based system called HIDRA. The use of Java as a development language allows us to more easily produce a cross-platform result, which is of particular importance to the biology community, who use a variety of operating system platforms. Among our immediate collaborators, individuals use Windows, Macintosh, and Linux operating systems to perform their analysis. The Java language also easily permits future expansion of our approach to include additional features, which is vital as genomic research is rapidly evolving.

4 Validation

The ultimate validation of a scientific data visualization approach is its usefulness and adoption within the research community. In particular, a successful approach should aid in the discovery of novel biology. We are working with many collaborators spread across five laboratories to assess how our multiple dataset visualization approach aids their research as well as how to improve HIDRA. We have deployed our system for these users both on their own desktop/laptop machines and on the large-scale shared display wall at the Lewis-Sigler Institute for Integrative Genomics at Princeton. While we are still receiving feedback from these users, here we discuss two of the user experiences that led to biological insights made using our approach. These examples demonstrate the power of our technique as these observations could not be easily made using any previously existing methodology.

4.1 User experience #1 – Stress response effects in yeast

One scientist using our system is interested in studying stress response and growth rate effects in yeast. By utilizing our multi-dataset visualization capabilities applied to several existing datasets, she was able to draw several novel, biologically meaningful conclusions. She was able to simultaneously examine the expression levels of genes in a set of standard stress response datasets[13] as well as results from a nutrient limitation study[12] and a collection of gene knockout experiments[18]. The biological question this user wished to examine is whether or not the traditional global stress response signal is present in other types of data.

Using our approach, she was able to easily find and select clusters of genes in the nutrient limitation and knockout studies that she suspected may be the result of a stress response effect, and then examine how those genes related to each other within the standard collection of stress datasets (see Figure 2). Performing this type of analysis is simple in our multiple dataset approach; however, using previously existing techniques we would need to launch over a dozen independent instances of a program and continually cut and paste selections between instances, rendering such analysis practically impossible.

Our collaborator identified several groups of genes in these datasets that exhibited a strong pattern of correlation within the stress response datasets as well. This suggests that the effect on gene expression of various nutrient limitations and gene knockouts may be superceded by the more general stress response effect. Our collaborators are currently performing further analysis, both in the lab and with our visualization system to better characterize this phenomenon. Thus, by observing the relationships between these very different datasets in HIDRA, this scientist quickly identified unexpected commonalities that may prove biologically interesting.

4.2 User experience #2 – Cell cycle synchronization effects

A second example of an important observation was made by another biologist using HIDRA to investigate disparities among related datasets. In this case the scientist was examining several datasets all purportedly studying the same phenomenon, the yeast cell cycle. In particular, two studies used a variety of means to synchronize cell populations to create time courses of gene expression throughout the phases of the cell cycle[19, 20].

A group of genes in one of these time courses were tightly clustered with very high over-expression at early points of the time course. However, using HIDRA we could quickly see that these genes were largely unrelated in the other time courses, and during the early time points they were not over-expressed in the datasets produced from other means of synchronization (see Figure 4).

Upon further inspection, a significant number of these genes are known to be involved in cell conjugation and mating. The time course where these genes are tightly clustered was synchronized by exposing the cell population to a pheromone that induces a mating response, which halts cell cycle progression. Our collaborator quickly realized that the expression response seen in these early time points was an artifact of the synchronization method, rather than a change caused by the cell cycle. In this case, observing differences between datasets studying the same phenomenon helped focus efforts on important portions of the datasets.

4.3 Discussion

The two examples described above are representative of the types of interactions users have had with HIDRA. By quickly observing commonalities among disparate datasets collaborators have been able to identify common trends that could indicate meaningful relationships between experimental conditions. Conversely, by finding key differences between related datasets users can explore phenomena unique to particular assays. This type of exploration allows microarray researchers to quickly make key insights and form hypotheses that would be difficult to make viewing the data independently.

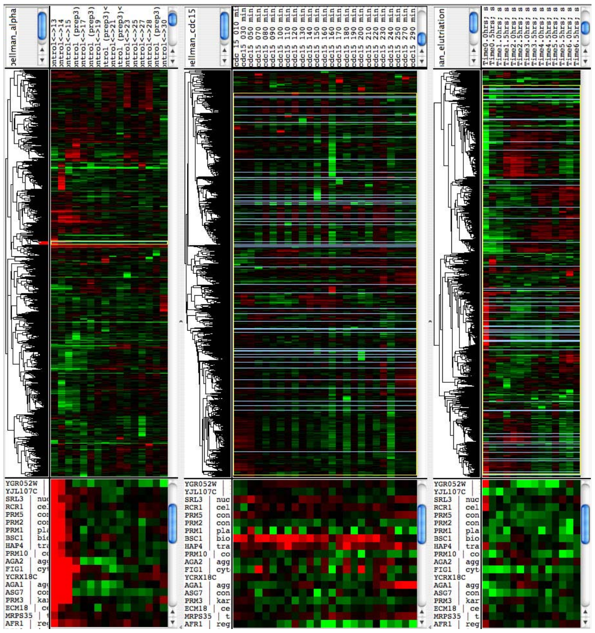

Figure 4 – Biological exploration of differences among three cell cycle datasets displayed in HIDRA. In this case, three time courses studying the same phenomenon from two studies [19, 20] are shown. A group of genes with very high over-expression at early time points is selected in the leftmost dataset, but these genes show little relationship to one another in the other two time courses. Further study revealed that the over-expression of these genes in the left dataset was an experimental artifact.

5 Conclusions

We have presented a novel methodology for the concurrent analysis of multiple gene expression microarray datasets. Our approach allows researchers to understand how their own data relates to data previously published in the literature, which is vital for continued analysis. By exploring the larger context of available data, users can overcome the limitations of existing approaches for higher-level analysis of their own data. Observing a more global view of expression data allows biologists to make more insights and formulate novel hypotheses.

Our approach to the inclusion of greater data context is based on expanding common visualization practices to create an "overview + detail + setting" system. We include the concept of the greater information setting by horizontally integrating and linking separate overview and detail views for individual datasets. This type of data integration – inclusion of multiple parallel views – is in contrast to the integration of a variety of viewpoints based on the same underlying data, which we call a vertically integrated approach.

Although we apply the concepts of including a broader setting of information through horizontal integration to a specific solution for microarray visualization, these principles are much more general. For example, a system similar to HIDRA for microarray analysis could be created based on parallel coordinates, rather than heat maps. Horizontally incorporating additional datasets into a system based on vertically integrated multiple views could potentially provide both the benefits of more complete understanding of single datasets and the benefits of understanding the greater information context.

The concept of visualizing the broader setting of available data is vital for future analysis and comparisons within the biology community. We have shown real-world examples of insights that can be made using our approach for microarray visualization that are difficult or impossible to discover using existing techniques. We believe integrating additional datasets into visualization systems is a powerful paradigm not only for genomics data, but potentially for many other disciplines as well.

6 Acknowledgements

The authors would like to thank the members of the Troyanskaya, Botstein, Coller, Kruglyak, and Dunham laboratories for using the system and providing feedback and validations. Matthew Hibbs is supported by NSF grant DGE-9972930. This work was supported in part by NSF grants CNS-0406415, EIA-0101247, and IIS-0513552, by NIH grant R01 GM071966, and by the NIGMS Center of Excellence grant P50 GM071508.

References

[1] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A,* 95(25):14863-8, 1998.

[2] Quackenbush J. Computational analysis of microarray data. *Nat Rev Genet,* 2(6):418-27, 2001.

[3] Saraiya P, North C, Duca K. An Evaluation of Microarray Visualization Tools for Biological Insight. *The IEEE Symposium on Information Visualization,* 1-8, 2004.

[4] Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, Sturn A, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovich E, Borisovsky I, Liu Z, Vinsavich A, Trush V, Quackenbush J. TM4: a free, open-source system for microarray data management and analysis. *Biotechniques,* 34(2):374-8, 2003.

[5] Saldanha AJ. Java Treeview--extensible visualization of microarray data. *Bioinformatics,* 20(17):3246-8, 2004.

[6] Seo J, Shneiderman B. Interactively exploring hierarchical clustering results [geneidentification]. *Computer,* 35(7):80-86, 2002.

[7] Hochheiser H, Baehrecke EH, Mount SM, Shneiderman B. Dynamic querying for pattern identification in microarray and genomic data. *Proceedings of the International Conference on Multimedia and Expo,* 3, 2003.

[8] Spotfire Decisionsite for functional Genomics. *http://www.spotfire.com*.

[9] GeneSpring. *http://www.silicongenetics.com.*

[10] Hibbs MA, Dirksen NC, Li K, Troyanskaya OG. Visualization methods for statistical analysis of microarray clusters. *BMC Bioinformatics,* 6115, 2005.

[11] Baldonado MQW, Woodruff A, Kuchinsky A. Guidelines for using multiple views in information visualization. *Proceedings of the working conference on Advanced visual interfaces,* 110-119, 2000.

[12] Saldanha AJ, Brauer MJ, Botstein D. Nutritional homeostasis in batch and steady-state culture of yeast. *Mol Biol Cell,* 15(9):4089-104, 2004.

[13] Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO. Genomic expression programs in the response of yeast cells to environmental changes. *Mol Biol Cell,* 11(12):4241-57, 2000.

[14] North C, Schneiderman B. A Taxonomy of Multiple Window Coordinations. *Technical Report CS-TR-3854,* , 1998.

[15] Li K, Chen H, Clark D, Cook P, Damianakis S, Essl G, Finkelstein A, Funkhouser T, Klein A, Liu Z, Praun E, Samanta R, Shedd B, Singh JP, Tzanetakis G, Zheng J. Building and Using a Scalable Display Wall System. *IEEE Comput Graph Appl,* 20(4):29-37, 2000.

[16] Wallace G, Anshus OJ, Bi P, Chen H, Chen Y, Clark D, Cook P, Finkelstein A, Funkhouser T, Gupta A, Hibbs M, Li K, Liu Z, Samanta R, Sukthankar R, Troyanskaya O. Tools and applications for large-scale display walls. *IEEE Comput Graph Appl,* 25(4):24-33, 2005.

[17] Wei B, Silva C, Koutsofios E, Krishnan S, North S. Visualization Research with Large Displays. *IEEE Comput Graph Appl,* 20(4):50-54, 2000.

[18] Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, King AM, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburtty K, Simon J, Bard M, Friend SH. Functional discovery via a compendium of expression profiles. *Cell,* 102(1):109-26, 2000.

[19] Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW. A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol Cell,* 2(1):65-73, 1998.

[20] Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol Biol Cell,* 9(12):3273-97, 1998.