A Framework to Analyze Biclustering Results on Microarray Experiments

Rodrigo Santamaría, Roberto Therón and Luis Quintales

University of Salamanca

Abstract. Microarray technology produces large amounts of information to be manipulated by analysis methods, such as biclustering algorithms, to extract new knowledge. All-purpose multivariate data visualization tools are usually not enough for studying microarray experiments. Additionally, clustering tools do not provide means of simultaneous visualization of all the biclusters obtained.
We present an interactive tool that integrates traditional visualization techniques with others related to bioinformatics, such as transcription regulatory networks and microarray heatmaps, to provide enhanced understanding of the biclustering results. Our aim is to gain insight about the structure of biological data and the behavior of different biclustering algorithms.

1 Introduction

Biclustering methods are techniques that discover internal structure of data in a non-supervised way. In the last few years they have been extensively applied to bioinformatics, specially to extract knowledge from microarray experiments. The first effort was done by Cheng and Church [5]; many others are surveyed in [13, 23]. Nowadays, still new biclustering methods are developed [15, 4].

On the other hand, there are tools covering different aspects of biological and statistical analysis. BicAT [2] is a great tool focused in biclustering algorithms, implementing some of the most important ones along with traditional k-means and hierarchical clustering. BicAT presents the results as ordered lists of biclusters, that can be examined individually through heatmaps and parallel coordinates.

Expander [20, 18] is also a tool that implements clustering and biclustering methods. Although Expander implements less biclustering algorithms than BicAT, it has a great number of visualizations: heatmaps and boxplots to study microarray data matrices, dendrogram+heatmap visualization of hierarchical clustering results [7], clustering PCA displays and bicluster heatmaps. The PCA display may be the most interesting view because it allows a quick understanding of gene structure (coloring points depending on the cluster which groups them).

gCluto [16] makes use of more advanced information visualization techniques. The microarray data matrix is again represented by a heatmap but now the interaction with the representation is allowed, so rows and columns can be expanded, combined or grouped by hierarchical clustering. gCluto also uses 2D projections of clusters but in a 3D space called mountain maps, where perimeter, height, slope and color identify different properties of each cluster.

The Rank-by-feature framework [17] is another powerful tool for hierarchical and k-means clustering. In this case a great level of interaction is allowed, under a high number of views: heatmaps, dendrograms, histograms, scatter plots and parallel coordinates. Finally, Cytoscape [19] is a very different tool, focused in analyzing biomolecular interaction networks with an optimal degree of interaction (zooming, searching, changes of layout, coloring, database querying and lots more).

Although the aforementioned tools deal with clustering and/or biclustering results, they do not focus on the simultaneous visualization of them. BicAT visualizes biclustering results individually, and comparison must be done through navigation of lists, which makes difficult the discovery of relationships among biclusters. Expander and gCluto present different solutions to this but for clustering results. The representation of multiple biclustering results of one or more biclustering methods has not been treated.

To overcome these limitations, we have developed a visual analysis tool that allows the simultaneous display of all the biclustering results of different methods along with linked views of related information, such as microarray expression levels and transcription regulatory networks (TRNs). That way, a full framework to help in decision making has been implemented and tested.

The following sections are organized as follows. Section 2 exposes the visualization techniques implemented in the tool: definition of the structure, data, displays, user interactions implemented and linkages between views. Section 3 presents a full example of the use of the framework with a synthetic microarray data experiment. Finally, Section 4 draws the conclusions achieved and establishes future lines for expanding the tool.

2 Bicluster Visualization

The framework manages different data sources and display them by using a number of visualizations techniques. All the visualizations are interconnected by means of a session manager to allow flow of data and interactions among views (see fig. 1). Three data sources are distinguished. The most important is the Microarray Data Matrix, that contains information about gene names, condition details and gene expression levels. Following, TRN network, represented as an XML standard graph, provides information about genes and relationships between them (up or down-regulation). Finally, bicluster results are presented as an structured file with information about the type of biclustering algorithm, the dimension of the biclusters and the genes and conditions grouped by them.

These data are visualized by means of five main visualization techniques: heatmaps, parallel coordinates, scatter plots, bubble maps and transcription graphs (Fig. 2). The first three visualizations represent microarray expression levels as multivariate data where each gene or sample is a variable and each condition or experiment is a dimension. The tool also allows the presentation of this data as a textual table. The bubble map represents biclustering results while the transcription graph represents a TRN of the organism studied in the microarray. For description purposes, we will use *gene* to address to a variable and *condition* for dimensions. We will have n genes $G = \{g_1, ..., g_n\}$ and m conditions $C = \{c_1, ..., cm\}$. A bicluster B is a subset of n_b genes ($G_b = \{g'_1, ..., g'_{nb}\}$) and m_b conditions ($C_b = \{c'_1, ..., c'_{mb}\}$).

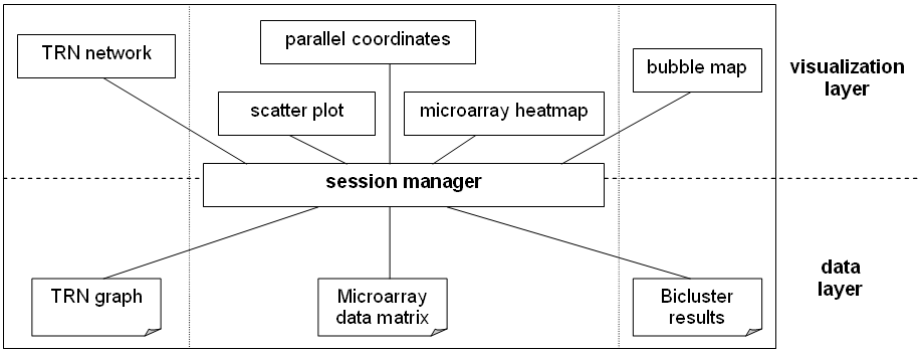

Fig. 1. Diagram of the structure of the framework. Three data sources can be used in the visualization of different displays by means of a session manager that interconnects them all.

2.1 Microarray data visualizations

Heatmaps (Fig. 2c) are the most usual representation of microarray data. In order to inspect genes or conditions individually, the heatmap implements bifocal distortion [12] by rows and/or columns, as well as zoom and navigation through expression levels. Selection of rows, columns or individual expression levels are linked to the other visualizations of the framework.

Parallel coordinates (fig. 2d) represent G as a set of lines of m-dimensional points. Selection of ranges of values on any condition can be done. Conditions also can be reordered as desired.

2.2 Bubble map

Bubble maps (fig. 2b) are related to gCluto mountain maps, but unlike gCluto maps, this visualization makes use of two dimensions to avoid 3D overlapping and improve time performance, allowing simultaneous comparison of a large number of biclustering results from different methods.

Each bicluster B is represented as a circle (bubble), where color identifies the biclustering method that computed it. The radius of the shape refers to the size of the bicluster, computed as $n_b m_b$. The transparency depends on bicluster homogeneity, defined as the inverse of the within variation described in eq. 1:

$$W_b = \frac{1}{n_b} \sum_{i=1}^{n_b} \sqrt{\sum_{j=1}^{m_b} (\bar{a}_j - a_{ij})^2} \tag{1}$$

where a_{ij} is the expression level of the gene g_i under the condition c_j and $\bar{a}_j$ is the mean of the expression levels of the genes grouped in B for condition j.

The position is determined by the genes and conditions grouped. The horizontal coordinate depends on conditions while the vertical coordinate depends on genes. To

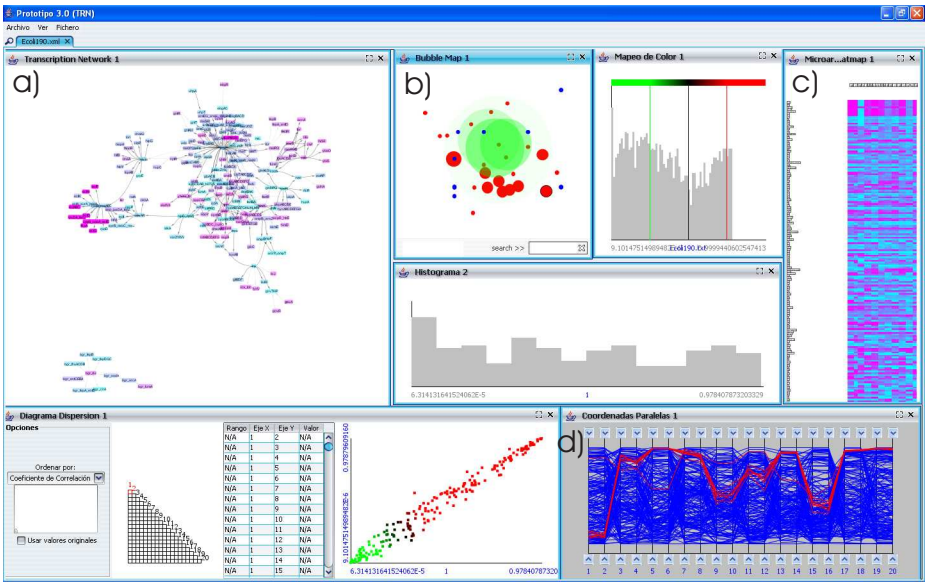

Fig. 2. Overview of the framework. Data belongs to the example discussed in Section 3. The most relevant visualizations are (a) TRN network, (b) bubble map, (c) microarray heatmap and (d) parallel coordinates.

compute the positions, bicluster B, grouping gene subset G_b and condition subset C_b, is mapped to the multidimensional points x_b and y_b as in eqs. 2 and 3.

$$x_b = (p_1, p_2, ..., p_n) \mid p_i = 1 \Leftrightarrow g_i \subset G_k,\ p_i = 0\ otherwise \tag{2}$$

$$y_b = (p_1, p_2, ..., p_m) \mid p_j = 1 \Leftrightarrow c_j \subset G_k,\ p_i = 0\ otherwise \tag{3}$$

These two points of n and m coordinates are projected to one dimension with either a classical metric [8] or non-metric [10] multidimensional scaling. This way both y-axis and x-axis components of the representation for each bicluster are obtained. Therefore, biclusters at the same horizontal/vertical line are expected to share genes/conditions, although this is not always precise due to the reduction of dimensionality, that obviously loses information.

The result is a set of distributed, colored, sometimes superposed circular shapes, where an analyst can easily identify biclusters distant from the trend, differences between biclustering methods or other relevant knowledge (Figs. 2b, 3a). The user can select any number of biclusters, a change that is transferred to other views to highlight the corresponding genes and/or conditions. Bubbles can be dragged to change their positions in case the user wants to reorder them using any other criterium.

2.3 TRN visualization

In a TRN, nodes represent the set of all genes G, while a directed edge from g_i to g_j means that g_i encodes for a transcription factor protein that transcriptionally regulates g_j [14]. It is important to distinguish at least two types of edges: activation and repression edges. When a gene up-regulated connects with an activation edge to another gene, this one is favored to up-regulation. If it connects with a repression edge, will be favored to down-regulation.

In our framework, TRNs have been represented as directed acyclic graphs led by forces (Fig. 2a). Nodes are labeled with gene names and edges are colored in dark or light grey depending if the interaction is activation or inhibition, respectively. To avoid edge cluttering, they are displayed with splines instead of straight lines. We also implement a gene search by name. The interacting forces display the nodes so the overlapping of nodes and edges is minimized.

2.4 Linked Visualizations

All the visualizations are linked so changes in a view are propagated to the rest of views (Fig. 3). The ability of visualizing changes in a representation because of interaction with another representation helps to reveal patterns. On the other hand, linkage limits the screen area because it has to be divided by different visualizations. All linkages implemented are bidirectional, so flow between visualizations can be followed at user's demand.

In our case, the usual flow of information that communicates views are subsets of genes and/or conditions. Thus, a selection of a node in the TRN will imply the flow of the gene represented by that node to other views, highlighting biclusters that contain this gene or focusing on the gene in the microarray heatmap, for example. The user can configure which visualizations to monitor simultaneously and if they are linked or not, thus adapting screen areas to her necessity.

3 Case Study

3.1 Example Dataset

In order to make the discussion simpler, we have chosen a reduced synthetic example obtained by SynTReN [6] from Shen-Orr's E. coli TRN [21]. From this network, with 424 nodes, SynTReN builds a synthetic TRN with 200 nodes, 190 nodes based in Shen-Orr's definition and 10 random nodes, without biological basis. SynTReN will also generate a microarray data matrix simulating 10 experiments, each one repeated two times.

We apply three different biclustering algorithms to the microarray data matrix: Bimax [15], Plaid models [11] and Spectral biclustering [9]. We have chosen methods that differ in its interpretation of biclusters, so it is expected that their results will be quite distinct.

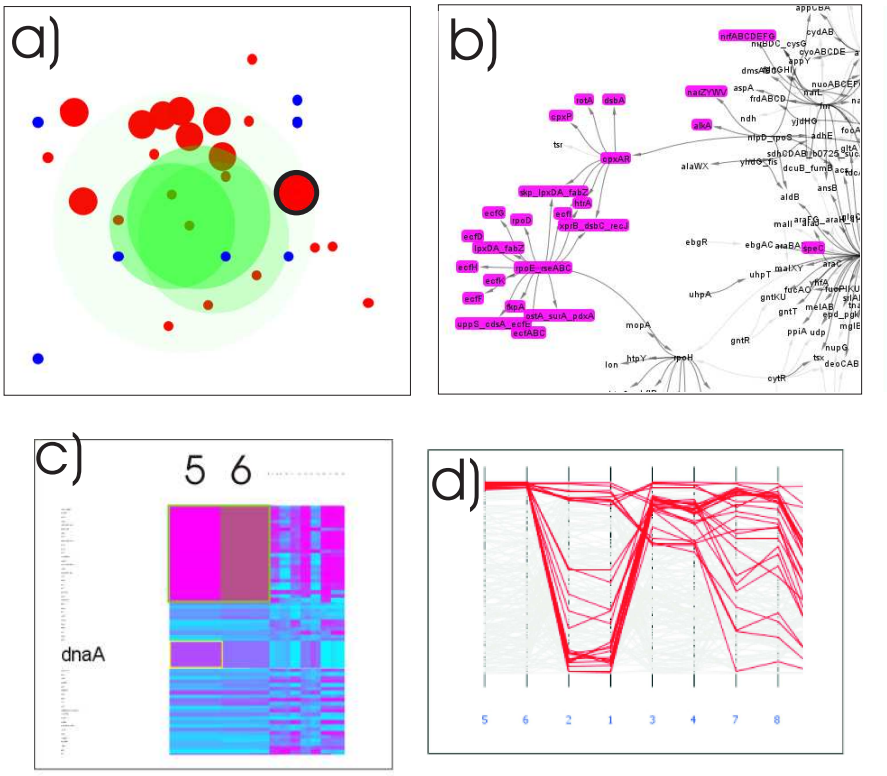

Fig. 3. Example of how linkage works. A Bimax (red) bicluster is selected in bubble map (a) and this provokes gene highlighting in TRN (b), reordering of rows and columns in heatmap (c) and highlighting of lines and reordering of axis in parallel coordinates (d). Similar flows can be followed by interaction with other visualizations.

3.2 Objectives

The framework has been designed in such a way that analysis will naturally follow the Information Visualization Mantra: ”Overview first, zoom and filter, details on demand” [22]. This way, it will start with a general overview of our problem, to continue with filterings by biclusters, genes and conditions. With these flows, supported by linked views, we will prove the potence of the framework to analyze the mentioned dataset regarding the following: 1) detecting relationships between the two replications of each experiment, 2) determining characteristics in the biclusters computed by different methods, 3) checking if related groups in TRN are grouped by biclusters and 4) detecting random genes, and determining if they appear in the biclusters computed.

Additionally, we want to discover: 1) new relationships between genes not related in the TRN, 2) biclusters deviated from the trend and 3) differences and similarities of the three biclustering methods and its performance for this example.

3.3 Overview

A simple overview using different visualizations gives interesting information. The TRN layout (fig. 2a) shows how genes are related according to existing biological knowledge. A group of random genes is easily detected as a separate graph at the bottom-left. The bubble map (fig. 2b, 3a) shows biclusters for Bimax (red), Plaid model (green) and Spectral (blue). With just a glance, we can tell than Plaid model gives bigger, heterogeneous (transparent) biclusters (due to some extent by a reported problem of this algorithm [24]), while Spectral biclustering gives very small ones and are displayed linearly, revealing the checkerboard structure of Spectral biclustering. Bimax returns middle-size, homogeneous (solid) biclusters. Also, biclusters deviated from the trends and groups of neighbor biclusters are easily detected, possibly worth a deeper study with the tool. The microarray heatmap and parallel coordinates are not very helpful on an overview, being the expression level information overwhelming without previous filtering. Finally, a scatter plot comparing expression levels of different replications of the same experiment (fig 2, bottom left) reveals its correlation.

3.4 Bicluster-oriented analysis

Once the overview has given us a context to draw preliminary analysis, deeper exploration is needed. This usually starts with biclusters, displayed with different colors depending on their method of biclustering. Interesting biclusters because of their homogeneity, size or position are salient in the bubble map visualization and can be selected, provoking changes in other visualizations that give us insight about what is grouped in the bicluster and why.

The microarray heatmap will reorder and highlight genes and conditions on the bicluster, giving a quick way to identify what is in the bicluster. Also heatmaps, along with parallel coordinates, help to understand why these genes and conditions are grouped together by the algorithm in terms of their expression levels. For example, when selecting a Bimax bicluster as in fig. 3, genes highlighted in heatmap and parallel coordinates present high and constant expression levels through the corresponding conditions. These are two of the features of Bimax algorithm, and therefore the information helps us to confirm that the results are correct or (if the biclustering method is not well known) to learn about the biclustering behavior. On the other hand, when a bicluster is selected, the corresponding genes highlight in the TRN network. Usually, as in fig. 3b, groupings are reflected in previously biological relationships (left bunch of genes) but in some cases previously unrelated genes are grouped, as it is the case of the gene at the right of the figure. Thanks to the force layout of the TRN graph, genes unrelated (very separated) can be easily detected.

Various biclusters can be selected simultaneously, thus highlighting in other visualizations the intersecting genes and conditions. This is interesting when clouds of biclusters are detected in the bubble map.

3.5 Gene and condition-oriented analysis

Studying the biclusters, some genes appear grouped without direct (or obvious indirect) relation in the TRN. These genes could be actually related or be misgrouped by biclus-

tering algorithms. If that kind of genes are grouped by a large number of biclusters, the probability of them being really related increases, justifying further analysis. The same is valid with conditions.

To analyze these interesting objects, we can change the scope and flow of the navigation through the tool and start by selecting particular genes. Picking those genes in the TRN will highlight all the biclusters that groups them together. If a high number of biclusters is highlighted, it is possible that the genes are truly related and we have discovered relevant knowledge (Fig. 4).

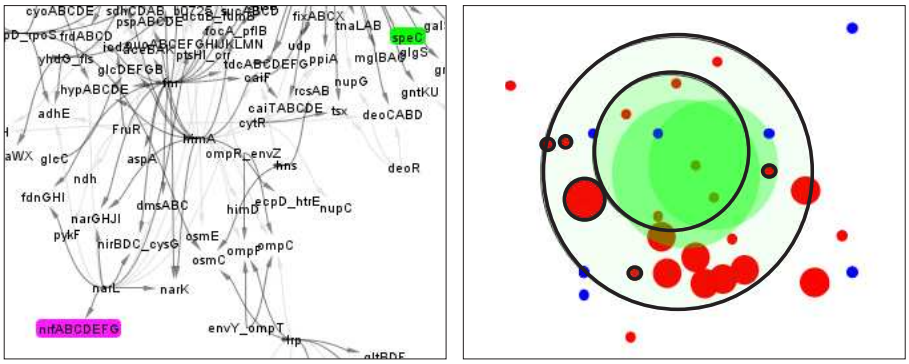

Fig. 4. Genes named *spec* and *nrfABCDEFG* are grouped together in seven biclusters from two different biclustering methods, without known biological evidence. The framework helps to discover it quickly.

4 Conclusion and future work

A framework to study biclustering methods in terms of its results by different visualizations, including biological knowledge with TRNs, is presented. The use of this framework, along with benchmark datasets and statistical and biological validation techniques can shed more light on performance of biclustering methods. It also will help analysts in the study of the usually large number of biclusters given by biclustering algorithms, decreasing analysis time and helping in the detection of relevant results. The tool discussed has relevant advantages over other current tools:

- Visualization of all biclusters *simultaneously* by means of the bubble map. This visualization also allows the representation of biclusters from different biclustering algorithms simultaneously. Only gCluto and Expander implements simultaneous visualization of simple clusters form a single method, without interaction.
- Incorporation of biological information from transcription regulatory networks to the visualization of microarray data and biclusters, allowing their communication.

This is an unusual feature, only implemented by Expander (by means of visualization of transcription binding sites in gene sequences) and Cytoscape (coloring of TRNs by expression levels).
- Simultaneous visualization and linking between different views. This is a key concept to increase the user's insight on the problem, witnessing the changes that interaction with a visualization causes in other views.
- Use of statistical measures such as coherence and variance by means of bubble map, thus including another relevant aspect of biclustering analysis: validation metrics.

Aside for the aforementioned advantages, new paths to improve the tool are opened:

- The bubble map, although useful, is based in projections that reduce dimensionality at the cost of discarding details. The result is that the overlapping of bubbles does not exactly convey the real overlapping of biclusters. Another technique is being currently studied to solve this.
- More biological knowledge will be, specially network motifs [14] identified in TRNs and GO [1] and MIAME annotations [3], increasing the details-on-demand.
- Gene and bicluster-oriented analysis discussed here are just two ways of revealing new knowledge. Testing of the tool by analysts will reveal new requirements in both visualization and genomic/transcriptomic areas.

5 Acknowledgements

The authors wish to thank Javier Molpeceres its contributions to programming this tool. This work was supported by the Education and Science Ministry of Spain under project TIN2006-06313 and by a grant from the Junta of Castilla y León.

References

1. M. Ashburner, C. A. Ball, J. A. Blake, D. Bolsteing, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. *Nature Genetics*, 25:25–29, 2000.
2. S. Barkow, S. Bleuer, A. Prelic, P. Zimmermann, and E. Zitzler. Bicat: a biclustering analysis toolbox. *Bioinformatics*, 22(10):1282–1283, 2006.
3. A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, C. Stoeckert, J. Aach, W. Ansorge, C. Ball, H. Causton, T. Gaasterland, P. Glenisson, F. Holstege, I. Kim, V. Markowitz, J. Matese, H. Parkinson, A. Robinson, U. Sarkans, S. Schulze-Kremer, J. Stewart, R. Taylor, J. Vilo, and M. Vingron. Minimum information about a microarray experiment (miame)-toward standards for microarray data. *Nat Genet*, 29(4):365–371, 2001.
4. P. Carmona-Saez, R. D. Pascual-Marqui, F. Tirado, J. M. Carazo, and A. Pascual-Montano. Biclustering of gene expression data by non-smooth non-negative matrix factorization. *BMC Bioinformatics*, 7(78), 2006.
5. Y. Cheng and G. M. Church. Biclustering of expression data. *Proc. Int'l Conf Intell Syst Mol Biol.*, 8:93–103, 2000.
6. T. V. den Bulcke, K. V. Leemput, B. Naudts, P. van Remortel, H. Ma, A. Verschoren, B. D. Moor, and K. Marchal. Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics*, 7(43), 2006.

7. M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, 1998.
8. J. C. Gower. Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53(3-4):325–338, 1966.
9. Y. Kluger, R. Basri, J. T. Chang, and M. Gerstein. Spectral biclustering of microarray data: Coclustering genes and conditions. *Genome Research*, 13:703–716, 2003.
10. J. B. Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, March 1964.
11. L. Lazzeroni and A. Owen. Plaid models for gene expression data. Technical report, Standford University, 2002.
12. Y. K. Leung and M. D. Apperley. A review and taxonomy of distortion-oriented presentation techniques. *Transactions of Computer-Human Interaction*, 1(2):126–160, 1994.
13. S. Madeira and A. Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions of Computational Biology and Bioinformatics*, 1(1):24–45, 2004.
14. R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs: simple building blocks of complex networks. *Science*, 298:824–827, 2002.
15. A. Prelic, S. Bleuer, P. Zimmermann, A. Wille, P. Bhlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129, 2006.
16. M. Rasmussen and G. Karypis. gcluto: An interactive clustering, visualization and analysis system. Technical Report 04-021, University of Minnesota, 2004.
17. J. Seo and B. Shneiderman. A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections. *IEEE Symposium on Information Visualization*, pages 65–72, 2004.
18. R. Shamir, A. Maron-Katz, A. Tanay, C. Linhart, I. Steinfeld, R. Sharan, Y. Shiloh, and R. Elkon. Expander - an integrative program suite for microarray data analysis. *BMC Bioinformatics*, 6(232):1471–2105, 2005.
19. P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13:24958–2504, 2003.
20. R. Sharan, A. Maron-Katz, and R. Shamir. Click and expander: a system for clustering and visualizing gene expression data. *Bioinformatics*, 19(14):1787–1799, 2003.
21. S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nature Genetics*, 31:64–68, 2002.
22. B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *IEEE Visual Languages*, number UMCP-CSD CS-TR-3665, pages 336–343, College Park, Maryland 20742, U.S.A., 1996.
23. A. Tanay, R. Sharan, and R. Shamir. Biclustering algorithms: A survey. *Handbook of Computational Molecular Biology*, 2004.
24. H. Turner, T. Bailey, and W. Krzanowski. Improved biclustering of microarray data demonstrated through systematic performance tests. *Computational Statistics and Data Analysis*, 48:235–254, 2003.