Towards Visualising Temporal Features in Large Scale Microarray Time-series Data

Paul Craig, Jessie Kennedy and Andrew Cumming

School of Computing. Napier University, 10 Colinton road, Edinburgh, EH14 1DJ, UK
e-mail: {p.craig, j.kennedy, a.cumming}@napier.ac.uk

Abstract

Current techniques for visualising large-scale microarray data are unable to present temporal features without reducing the number of elements being displayed. This paper introduces a technique that overcomes this problem by combining a novel display technique, which operates over a continuous temporal subset of the time series, with direct manipulation of the parameters defining the subset.

1. Introduction

The genome is the complete set of instructions for making an organism, containing the master blueprint for all cellular structures and activities for the lifetime of the organism. The current initiative of microbiology is focused on advancing understanding of the organism by investigating the chemical structure and functioning of the genome.

A genome consists of several chromosomes, each of which is essentially a package for one long continuous strand of deoxyribonucleic acid (DNA). DNA is composed of building blocks called nucleotides consisting of a deoxyribose sugar, a phosphate group and one of four nitrogen bases – adenine (A), thymine (T), guanine (G) or cytosine (C). There have been several initiatives to map the precise chemical structure (the sequence of nitrogen bases) of the human genome and that of several model organisms. Sequence information is essentially a static view of the genome, telling us a lot about structure but relatively little about functioning. A better understanding of genome functioning can be reached by using microarray technology, which monitors the initial output of the genome by recording levels of messenger RNA (mRNA).

mRNA is the molecule that carries the code of a section of DNA into the cytoplasm surrounding the cell nucleus. Once in the cytoplasm, the mRNA encodes a protein or polypeptide specific to the section of DNA from which it was produced. This process is known as transcription, or expression. The sections of DNA, which are capable of transcription, are defined as genes. Following expression, the gene product interacts with a variety of other biomolecules, all primary or secondary gene products which in turn either directly or indirectly regulate the expression of genes through complex signalling cascades [1]. In effect we have a complex network of inter-gene reactions.

Microarrays facilitate the monitoring of gene expression for tens of thousands of genes in parallel [2], allowing a view of expression levels over a range of samples or over a period of time [3]. When working toward a better understanding of the functioning of the genome some of the questions typically asked of the data are:

- What genes, from the entire genome, are differentially expressed in a particular sample or cell state?
- What are the functional roles of genes and in which cellular processes do they participate?
- What are the mechanisms involved in these processes?

This paper gives an overview of microarray data and discusses some of the issues associated with its effective visualisation. We evaluate existing visualisation techniques and highlight their limitations with regard to uncovering certain aspects of the data. We conclude with our proposal for a new approach to representing and interacting with the data, which uncovers some aspects that are not revealed by existing applications.

2. Microarray Data

The output of any microarray experiment is in the form of a series of images where each gene is represented by a coloured dot. The colour of each dot depends on the level of mRNA in each sample or, in the case of temporal experiments, the control and the sample. Image processing software is used to translate these images into an expression matrix where columns relate to samples or time-points, rows relate to genes, and cells relate to relative mRNA abundances.

Before any analysis can proceed, a statistical procedure, known as normalization, is applied to the data. Normalization seeks to account for and remove sources of variation obscuring the underlying variation of interest, the level of gene expression [4]. Normalization adjusts for differences in labelling, detection efficiencies for florescent labels, and differences in the quality of RNA from the two samples examined in the assay. While expression values cannot be quantified due to the nature of the experimentation, which deals with gross cell populations, normalization makes values relative across genes and samples/times.

In the case of time-series experiments, the product of normalization can be considered as large-scale time-series data. An important aspect of the data, with regard to its analysis, is that it constitutes the output of a complex network and any investigation of the data will have the main objective of uncovering aspects of the underlying network's functioning.

3. Mircroarray Data Analysis

The primary objective of Microarray data analysis, a better understanding of the genome functioning, can be addressed by considering a number of lower level objectives relating to the data produced. These are:

- Representing the experimental results: A natural first step in extracting some of the biological information tied up in Microarray data is to examine the extremes by viewing the differential expression [5]. A representation for individual gene expression patterns is required which, when used to represent all of the gene expressions measured in the experiment can combine to provide a more complete view of the genome. A single model such as this facilitates an assessment of differential expression between samples or across time.
- Inferring associations: This allows us to group genes with regard to the particular sample or cellular process, which leads to information about each gene's functional role and cellular process participation. Grouping genes can also be thought of as the first stage in inferring interactions.
- Inferring interactions: As the genome mechanism consists of a network of gene interactions, the uncovering of such interactions is necessary to understand the genome. The timing of interactions is a crucial aspect of cellular functioning with regard to a number of significant biological processes, such as the switching between alternate process pathways. It is therefore also important to consider the temporal aspect of the data. Moreover, observing the timing of events is necessary to infer certain mechanisms, such as combinatorial regulation, where the expression of a single gene is affected by that of more than one other gene.

To address these objectives, a variety of statistical methods and visualisation techniques can be used. This paper is specifically concerned with the challenges of developing a visualisation technique.

4. Challenges of Microarray Data Visualisation

There are a number of significant challenges associated with the objectives of microarray data analysis, many of which apply specifically to information visualisation approaches.

When representing microarray data the number of individual data-elements (genes) being considered in any one experiment can be anything up to around ten thousand. For time series experiments the quantity of data can be further multiplied by the number of time points. With such large amounts of data, representative visualisation is a significant challenge.

When modelling an unknown process it is advisable to observe as many parameters of the system as possible. This is reflected in the current initiative to measure the expression of more and more genes [5]. In considering associations between expression patterns, the same logic leads us to consider all pairwise associations. The number of possible associations is equal to the number of genes raised to the power of two. Displaying such a large quantity of information, anything up to around 10^8 associations, is also problematic.

While the number of possible interactions is equivalent to the number of possible associations, there are also a number of other challenges pertaining more specifically to the complexity of the underlying network. Some features of this complexity that prove particularly problematic are the variety of gene-gene interaction types and the existence of combinatorial regulation.

A gene may act to inhibit or activate the expression of another gene. The time lag between event and reaction is variable, depending on the route of the signal, as are the relative concentrations of mRNA in each gene. As the number of inferred interactions will rise with the variety of possible interactions it follows that actual reactions will be harder to detect.

Combinatorial regulation occurs when the expression of one gene is controlled by the expression of more than one other gene. These types of interaction are crucial for many of the more subtle mechanisms within the cell, such as pathway switching, yet they are particularly hard to detect with existing visualisation methods.

5. Existing techniques

This section describes some established techniques employed in the visualisation of microarray data. The techniques are categorised by the primary objectives that they achieve i.e. representing the data, inferring associations, and inferring interactions.

5.1. Representing the data

Before the data can be presented in a single visualisation, a representation for each expression pattern is required.

When the data has been produced by a time-series experiment, visualising the expression pattern of an individual gene is fairly intuitive. As both expression and time are ordered quantities they can be represented in a simple graph like the example shown in Figure 1.

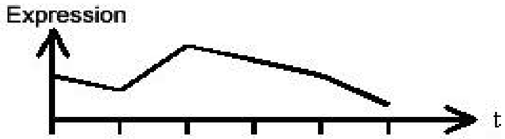

Figure 1. Graph representation of expression versus time.

With multi-sample experimental data there is no intrinsic ordering of samples making it inappropriate to use a graph for display. Instead, expression levels are usually represented by adjacent colour-coded squares. Negative values are green and positive values are red, with the colour intensity linearly proportional to the expression (or log ratio of the expression). This approach has the advantage of being more compact than mapping to a graph, and as such, is often also used to represent time-series data when screen space is at a premium. The problem with this approach is that a colour representation of expression has fewer distinguishable steps than a planar representation. This will make small differences in expression between cells harder to detect. This problem will be exacerbated for the sizeable minority of the population who are colour-blind or have difficulty distinguishing between green and red.

While graph and colour coded representations have the advantage of revealing the timing of events they are inadequate for presenting the large number of expression patterns available from microarray experiments. While selection and filtering techniques may reduce the number of patterns that require to be displayed at any one time, a global view of gene expression is often necessary. To facilitate this global view, the expression pattern of a gene is often encoded into a single pixel or a small square. The position of the representation on the screen, with regard to that of other representations, corresponds to aspects of the gene's expression pattern. These techniques often apply some measure of association or interaction between gene pairings and will be discussed in the following sections.

5.2. Inferring Associations

The inference of associations between genes is normally preceded by the creation of a similarity matrix. The similarity matrix compares all possible gene pairings using some predefined distance measure. There are a number of different distance measures that account for the different associations that may exist between genes. Some popular distance measures are Euclidean distance, Pearson's linear dissimilarity, Mutual information [6], Correlation metric [7] and Edge detection [8].

'Euclidean distance measure' is used to measure direct correlation between expression patterns. 'Pearson's linear dissimilarity' is similar to Euclidean distance, with the addition that it accounts for variable expression amplitude between the genes it associates. 'Mutual information' groups genes according to shared information content, picking up negative and positive correlation. 'Correlation metric' groups genes according to their maximum phase-shifted correlation. 'Edge detection' scores pairs of genes with regard to slopes between significant maximum and minimum expression levels that have a time lag below a set threshold. The resulting measure has amplitude of 1, with the sign indicating positive or negative correlation.

The similarity matrix visualisation is a direct visualisation of the similarity matrix with similarity values colour-coded. If genes are ordered according to functional groupings, then the vertical and horizontal bands that define the groupings can be analysed, with outlying genes easily identifiable.

The most common display of microarray data is based on the results of hierarchical agglomerative clustering [5]. The output of this clustering is a type of binary tree known as a dendrogram. For display, gene expression patterns are colour-coded and stacked. This part of the display is known as an expression mosaic. A tree type graphic at the top and/or sides is a direct visual representation of the dendrogram. This shows the groupings, which have been imposed by the clustering algorithm. An example of this visualisation method is shown in Figure 2.

Parallel Plots can be used to combine the results of different clustering algorithms and scientific information such as the functional grouping of genes [9].

Principal component analysis is a linear mapping of data points in n-dimensional space to d-dimensional space, where usually d<<n. When dimensionality is reduced, the intersection with maximum variation is used so as to preserve that aspect of the data. This is principal

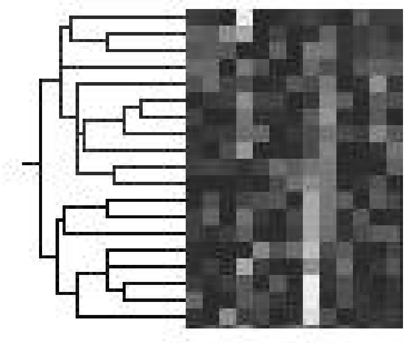

Figure 2. Combination colour mosaic and dendrogram visualisation of expression data.

component 1 (PC1), which can be thought of as describing most of the data. Principal component 2 (PC2) lies perpendicular to PC 1 and can be thought of as describing most of the rest of the data. Normally PC1 is plotted against PC2 in any visualisation.

Multidimensional scaling (MDS) is another dimensionality reduction technique. Individual elements are laid out so that the distance between any two elements is approximate to their dissimilarity. As the dissimilarity matrix exists in a higher dimensional space than the display space, which is normally two or three dimensional, it is inevitable that there will be some stress in the resultant display. Stress is measured as the total of all differences between scaled inter-point distances and dissimilarity measures between elements. The major advantage of multidimensional scaling over principal component analysis is that it can work with a number of different (dis) similarity measures in order to reveal more subtle inter-gene associations, such as the inverse correlation of expression patterns.

Self organizing maps [10] are similar to MDS but have the advantage that clusters may be seeded to incorporate biological knowledge. While evidence shows that self-organizing maps are more effective at grouping similar items [11] it is also evident that SOMs are less effective than MDS in preserving the structure of clusters [12].

An extension of this standard display of hierarchical clusters is the cluster tree produced by the FITCH software [13]. The length of the branches joining endpoints (genes) is approximate to their dissimilarity. The extra freedom provided by using connected-line length rather than direct inter-point distance should allow a significant reduction in any measure of stress as compared with MDS. However, tracking lines to assess dissimilarity can be a complicated visual operation [14].

Mutual information relevance networks [15] use a mutual information distance measure to display genes in a graph where joins represent associations above a given mutual information threshold. Relationships with higher mutual information are drawn with a thicker line. Unlike MDS and SOM the positioning of genes in the diagram has no significance as to their expression pattern, the technique uses a standard graph layout algorithm to position gene representations.

5.3. Inferring interactions

Many of the methods that infer associations can be thought of as working toward the inference of interactions e.g. if two genes have closely correlated expression it may be inferred that they are functionally related. These associations do not, however, assign any form of causality. In order to assign causality, the timing of events must be accounted for. In order to do this, the time structure of the data must be preserved.

Visualisation techniques that compress the expression pattern into a single point do so by destroying the time structure of the data and therefore they cannot be used to assign causality. Another disadvantage of these techniques is that the distance measures employed can only detect singular associations over the entire time period being considered. This is inconsistent with the reality of the biological system, where an individual gene's expression may be regulated by that of a number of different genes at different times [16].

Visualisation techniques that fully accommodate the inference of interactions and combinatorial relationships are those that preserve the time structure of the data. These include the time versus expression graph display and the colour coding of the expression pattern.

5.4. Summary

At present there is no clear consensus as to the best method for revealing patterns of gene expression. Indeed it is becoming increasingly clear that there might never be a 'best' approach and that the application of various techniques will allow different aspects of the data to be explored [17]. Important aspects of the data which are not effectively revealed by any current microarray data visualisation technique, at least not when a large number of patterns are combined in a single image, are temporal features such as the timing of events and the effects of combinatorial relationships.

Graph and colour coded representations of expression patterns, which preserve the time structure of the data, are too bulky to be included in an entire genome display. Genome wide visualisations destroy the time structure of the data and have little scope to reveal temporal features.

6. Visualising Temporal Features in Large Scale Microarray Time-series Data

We propose a technique for visualising microarray data that facilitates mining for temporal features. Our strategy is to visualise a continuous temporal subset of the data allowing the user direct manipulation of the subset parameters. The subset parameters are; t_0, the start-time of the subset, and Dt, the subset duration. This allows the user to view events only when they occur within the subset, giving some indication of the timing of events. Altering the subset parameters will reveal multiple events within a single expression pattern as distinct entities. This will allow the user to mine for more complex mechanisms within the network, such as combinatorial regulation.

We have devised a novel display technique, from which the expression pattern subset of each gene can be inferred. This display includes all genes from the experiment and is combined with a freehand selection technique that facilitates the association of gene subset representations across time as the subset parameters are adjusted.

6.1. Displaying the subset

The requirements for displaying each subset are as follows:

1) When t_0 is incremented, it would be beneficial for the representation of individual genes to have minimal displacement in order to reduce the perceptual complexity of relating the representation of a single gene through successive iterations of t_0.
2) The visualisation should consider all genes that are examined in the experiment within any single image. Exclusive filtering of the data may hinder innovative hypotheses and should be avoided.

In order to fulfil the first requirement it would be beneficial if t_0 could be incremented with granularity smaller than that of the experiment. Although the time series is discrete, it can be presumed that expression changes smoothly over time [18] so, t_0 can be adjusted in small steps and expression levels at time points between actual measurements derived using linear interpolation.

In order to fulfil the second requirement, including all genes in each static image of the visualisation, it was decided to represent each expression pattern subset as a pixel or small square. Given the more compact nature of this strategy and the fact that the visual representation of each gene is based only on a limited temporal subset, there is a greater capacity to place individual representations with regard to their absolute expression pattern rather than inter-pattern relationships, while maintaining the distinction between unrelated patterns. This has the advantage of allowing us to design a layout where there is minimum displacement of gene representations when t_0 is incremented. Moreover, the position of a representation within the visualisation will yield information relating to its expression pattern over the specified time period (i.e. whether it is high or low, rising or falling) regardless of the state of other genes.

As a precursor to placing the gene representations on a 2-dimensional plane, we took the approach of conceptualising the range of possible expression levels at each time-point as axes defining an n-dimensional space.

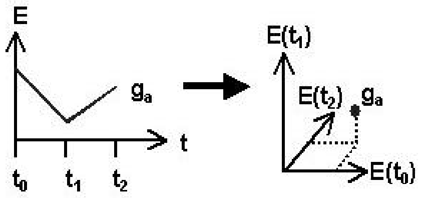

Figure 3. Translating a three time-point expression pattern subset to a point in n-dimensional space.

Figure 3 demonstrates this approach with the translation of an expression pattern subset, for which there are three time-points, into a point in three-dimensional space.

As the range of possible expression levels at all time points are equivalent, we can bound the n-dimensional space in an n-dimensional cube that contains all the expression patterns.

In order to cluster genes according to absolute shape, the gene representations are positioned according to the distance of their n-dimensional mapping to the corners of the n-dimensional cube which bounds all the expression patterns. This involves the creation of a dissimilarity matrix consisting of all gene-corner dissimilarities rather than all inter-gene dissimilarities.

For mapping the corners of the cube to a two-dimensional surface, the layout chosen was circular with the points representing corners distributed evenly around the rim. The following algorithm was employed to assign corners to points.

1) Opposing points around the circle should correspond to opposing points in the n-dimensional cube. This will partially preserve the symmetry of the representation, allowing for better detection of converse correlation.
2) The representation of a gene should have minimal displacement when t_0 is advanced, regardless of whether its expression has changed or not.

We employed this algorithm for three and four time-point subsets. The resulting layouts are illustrated in Figures 4 and 5. Each node corresponds to a corner of, the n-dimensional cube that bounds the n-dimensional mappings of all expression pattern subsets. The

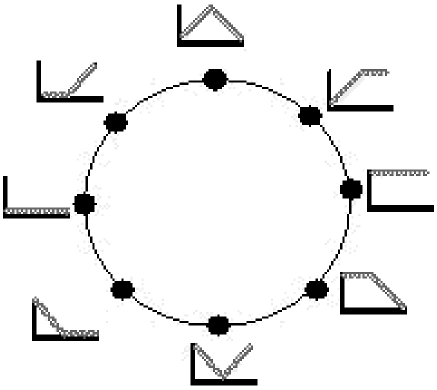

Figure 4. Corner mappings for 3 time points.

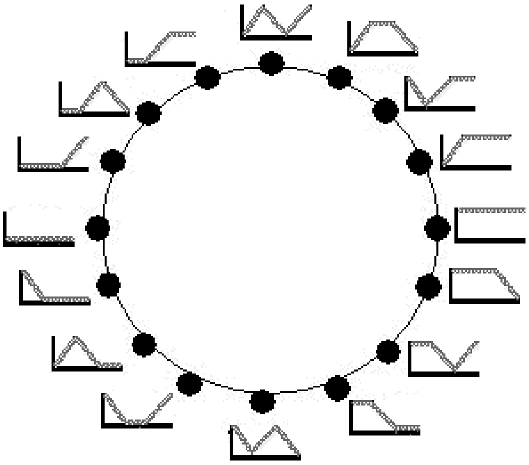

Figure 5. Corner mappings for 4 time points.

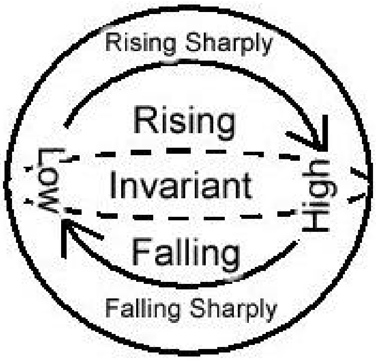

Figure 6. General layout of subset representations

The single point representations of genes were positioned according to a least stress direct mapping of their Euclidean distance from each corner. The general layout of subset representations is shown in Figure 6.

In addition to the patterns that can be revealed from static frames of the display, there are also some interesting features revealed by incrementing t_0. For example, if there is a pulse in expression then the representation will rise into the top hemisphere then fall sharply into the bottom hemisphere.

It was found that gene representations would be more evenly dispersed, with clusters more distinct, if their expression patterns were rescaled and adjusted so that maximum and minimum values equalled ±1. This is a standard pre-processing procedure that allows the analysis to focus on shape of the expression patterns rather than absolute expression. The disadvantage of this technique is that it may amplify noise for patterns where the expression amplitude is low and may require pre-filtering of the data to remove genes with invariant expression patterns.

6.2. Data Exploration

Figure 7 shows a screen shot of our initial prototype. The visualisation comprises three co-ordinated panels, one representing traditional gene expression graphs, another providing a list of gene names, and the main visualisation panel. The main panel employs the visualisation technique described above to reveal temporal features in the microarray data. Genes are selected by either clicking on their names in the list panel, or enclosing an area on the main visualisation panel. Once selected, the gene representations are highlighted in the main panel; their names are highlighted in the list panel, and their expression patterns are displayed as traditional expression graphs in the graph display panel. A control bar, at the bottom of the screen, allows the user to adjust the parameters of the continuous temporal subset considered in the main visualisation panel. Using a slider or a complementary set of play, pause, stop, rewind and fast-forward buttons to control t_0, the user is able view how the expression of any selected gene grouping evolves through time. An additional slider can be used to adjust Dt, changing the period of time that is considered in each static frame of the display.

7. Conclusions and Further Work

Given the intrinsic complexity of the system and subsequent variation in expression patterns, it is evident that the mapping to a lower dimensional space of any subset can never be truly representative. However, the technique employed was found to reveal some unique patterns in the data leading to information regarding the

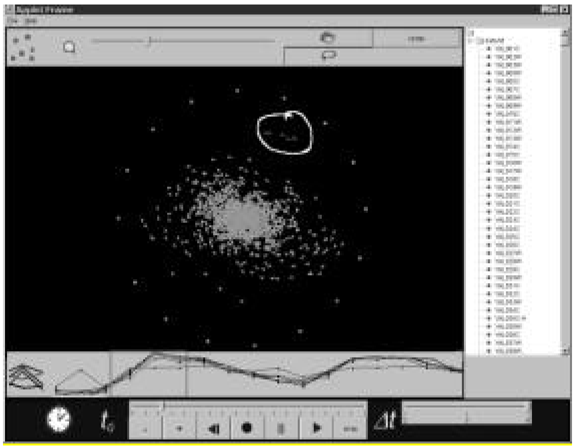

Figure 7. Screen shot of initial prototype.

timing of events and the evolution of gene clusters. Moreover, while it is accepted that in any individual static view of the data unrelated pattern subsets may be clustered together (e.g. a rising pattern and a pulse) it is extremely unlikely that this clustering will persist while the subset parameters are adjusted.

So far we have tested our technique with a small data set (~200 genes). The tool proved effective at displaying differential expression and gave a clear indication of how clusters of genes evolve throughout the period of the experiment. A more comprehensive evaluation, involving members of our target user group, will proceed once we have adapted our tool to accommodate larger data sets (~10^4 genes).

Possible extensions for the tool include:

- A Boolean selection mechanism, so that the user can more effectively mine for cluster events.
- A mechanism by which the user can save selections to file for cross experiment comparison.
- Integrated pre-processing of the data, with some measure of confidence integrated into the visualisation, to accommodate for the amplification of noise.

In summary, it was found that our technique has the potential to assist the mining of Microarray data for features that are not revealed by existing technologies. In this capacity, once fully developed, the tool should prove a valid addition to the existing arsenal of Microarray visualisation techniques already available.

8. References

[1] B. Bryant, A. Milosavljevic, and R. Somogyi, "Gene Expression and Genetic Networks (Session Introduction)," *Pacific Symposium on Biocomputing*, vol. 3, pp. 3-5, 1998.

[2] M. Schena, D. Shalon, R. Heller, A. Chai, P. O. Brown, and R. W. Davis, "Parallel human genome analysis: microarray-based expression monitoring of 1000 genes," *Proc. Natl. Acad. Sci. U.S.A. 93*,, pp. 10614-10619, 1996.

[3] D. J. Duggan, M. Chen, P. Meltzer, and J. Trent, "Expression profiling using cDNA microarrays," *Nature Genetics*, vol. 21, pp. 10-14, 1999.

[4] A. Hartemink, D. Gifford, T. Jaakkola, and R. Young, "Maximum likelihood estimation of optimal scaling factors for expression array normalization," *Microarrays: Optical Technologies and Informatics, Proceedings of SPIE*, vol. 4266, 2001.

[5] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Bostein, "Cluster analysis and display of genome-wide expression patterns.," *Proc. Natl. Acad. Sci. USA*, vol. 95, pp. 14863-14868, 1998.

[5] P. D'Haeseleer, R. Somogyi, and S. Liang, "Gene expression data analysis and modeling," *Pacific Symposium on Biocomputing*, 1999.

[6] S. Fuhrman, P. D'Haeseleer, and R. Somogyi, "Tracing genetic information flow from gene expression to pathways and molecular networks," 2000.

[7] A. Arkin, P. Shen, and J. Ross, "A test case of correlation metric construction of a reaction pathway from measurements," *Science*, vol. 277, pp. 1275- 1279, 1997.

[8] V. Filkov, S. Skiena, and J. Zhi, "Analysis Techniques for Microarray Time-Series Data," *RECOMB 2001: Proceedings of the Fifth Annual International Conference on Computational Biology*, 2001.

[9] X. Wen, S. Fuhrman, G. S. Michaels, D. B. Carr, S. Smith, J. L. Barker, and R. Somogyi, "Large-Scale Temporal Gene Expression Mapping of CNS Development," *Proc Natl Acad Sci USA*, vol. 95, pp. 334-339, 1998.

[10] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, E. Lander, and T. Golub, "Interpreting patterns of gene expression with self-organizing maps.," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, pp. 2907-2912, 1999.

[11] J. Venna and S. Kaski, "Neighborhood preservation in nonlinear projection methods: An experimental study," presented at International Conference on Articial Neural Networks, 2001.

[12] A. Flexer, "Limitations of Self-Organizing Maps for Vector Quantization and Multidimensional Scaling," *Neural Information Processing Systems*, 1997.

[13] J. Felsenstein, "FITCH," 3.5c ed. Washington: University of Washington, 1993.

[14] D. B. Carr, G. S. Michaels, and R. Somogyi, "Templates for Looking at Gene Expression Clustering," *Statistical Computing & Graphics Newsletter*, vol. 9, pp. 20-29, 1997.

[15]A. J. Butte and I. S. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," *Proceedings of the Pacific Symposium on Biocomputing*, 2000.

[16] F. C. P. Holstege, E. G. Jennings, J. J. Wyrick, T. I. Lee, C. J. Hengartner, M. R. Green, T. R. Golub, E. S. Lander, and R. A. Young, "Dissecting the regulatory circuitry of a eukaryotic genome," *Cell*, vol. 95, pp. 717-728, 1998.

[17] J. Quackenbush, "Computational Analysis of Microarray Data," in *Nature Reviews*, vol. 2, 2001, pp. 418- 427.

[18] P. D'Haeseleer, X. Wen, S. Fuhrman, and R. Somogyi, "Linear Modeling Of mRNA Expression Levels During CNS Development And Injury," *Pacific Symposium on Biocomputing*, pp. 41-52, 1999.