Coordinated Parallel Views for the Exploratory Analysis of Microarray Time-course Data

Paul Craig[1], Jessie Kennedy[1], Andrew Cumming[1]
[1] *School of Computing, Napier University, Edinburgh, Scotland, EH10 5DT*
{p.craig, j.kennedy, a.cumming}@napier.ac.uk

Abstract

Microarray time-course data relate to the recorded activity of thousands of genes, in parallel, over multiple discrete points in time during a biological process. Existing techniques that attempt to support the exploratory analysis of this data rely on static clustering views, interactive clustering views or coordinated clustering and graph views and are limited in that they fail to account for less dominant patterns in the data such as those that involve a subset of genes or a limited interval of the time-course. In this paper, we describe an alternative approach which avoids this limitation by using combined parallel views which present different complementary aspects of the data (i.e. timing, activity and change-in-activity). An example of how the views are combined to reveal significant patterns in the data (including those which cannot be found using clustering based techniques) is described and used to illustrate the benefits of combined parallel views to support exploratory analysis of this type of data.

Keywords--- **Information visualization, Coordinated views, Microarrays, Bioinformatics, time-series.**

1. Introduction

Microarray technologies [1, 2] are a recent development in the field of functional genomics. Allowing biologists to monitor the activity of thousands of genes (typically around 30,000) in parallel at different stages of a biological process [3-5], they produce large scale time-course data (Figure 1) which is, subject to proper analysis, valuable to a number of significant applications in various areas of biological and biomedical research.

The analysis of microarray time-course normally involves relating the activity of genes, and the timing of this activity, to biological processes. As there is relatively little information related to the function of genes, the timing of biological processes or even, in some cases, the existence of processes, this analysis will often require the biologist to reveal previously unknown or even unsuspected phenomena from the data [6]. To fully exploit the potential of microarray time-course to support this, biologists require analysis techniques that enable them to make unexpected discoveries and gain insights. These must support exploratory analysis of the data [6] while overcoming problems associated with its massive scale and complexity [7].

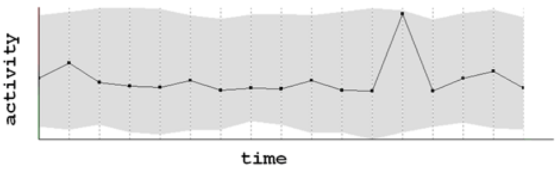


Figure 1. Microarray time-course data. A value against time graph view with the range of values at each time point defining a grey area and the recorded activity of a single gene highlighted.

The current tack of established techniques for the analysis of microarray time-course is one of ‘knowledge discovery’ where a complete visual overview of the data attempts to communicate unknown and unsuspected patterns [6, 8]. This is done by clustering the data so that genes with similar activity are grouped together and have their recorded activity displayed together. Figure 2 shows the most common of these representations – the combined heat-map/dendrogram display [9]. This uses the output of Hierarchical Clustering which is an algorithmic method that groups genes to produce nested clusters adhering to a hierarchical tree structure. This structure is represented using a dendrogram (top of Figure 2) with activity levels colour-coded (green for high activity and red for low) and displayed in a heat-map

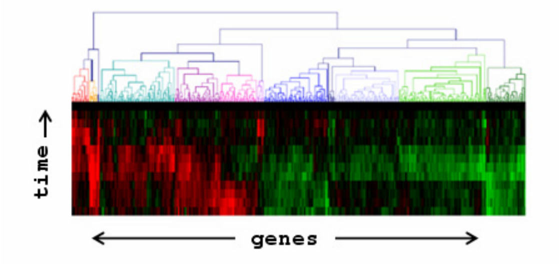


Figure 2. Heat-map/ dendrogram Hierarchical Clustering display.

grid where columns correspond to genes and rows correspond to time-points. The matching of heat-map columns to dendrogram branches causes them to be ordered so that genes with similar activity, and groups of genes with similar activity, are closer together.

A constriction of this type of clustering is that in combining the entire data into a single visual entity, gene representations must be compact and placed according to the relationship between their own recorded activity across the entire time-course and that of all other genes. As a direct consequence of this constriction, certain significant qualities are not well represented and significant patterns in the data are lost. Specifically, these are patterns of activity that are less dominant in the data and, in particular, patterns that are characterised over a limited interval of the experiments time frame [10] or, to a lesser degree, those that are contributed to by small proportion of the monitored genes (which, given the scale of the data, may still be a large number of genes).

An example of a significant pattern that would not be revealed by Hierarchical Clustering is illustrated in Figure 3. Here a rise then fall in activity found over a particular interval (P) could suggest that a group of genes are related to a particular biological process and that that process is associated with the experimental conditions. In this case, when the data is clustered, different patterns of activity before or after the interval cause the related genes to be assigned to different groupings with the significance of their common activity over the relevant interval lost. This is illustrated in Figure 4 using a heat-map clustering display. The arrows at the bottom of the figure show the columns which relate to genes with the characteristic rising and falling activity from P1 to P3. It can be seen that it would be impossible to detect this pattern without pre-knowledge of its existence using this particular type of clustering display.

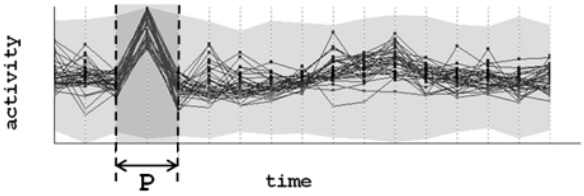


Figure 3. A significant pattern occurring exclusively over an interval (P).

2. Related work

In general, Hierarchical Clustering can be considered as an overview of microarray time-course directed at revealing dominant trends. There are, however, a number of other clustering techniques, and

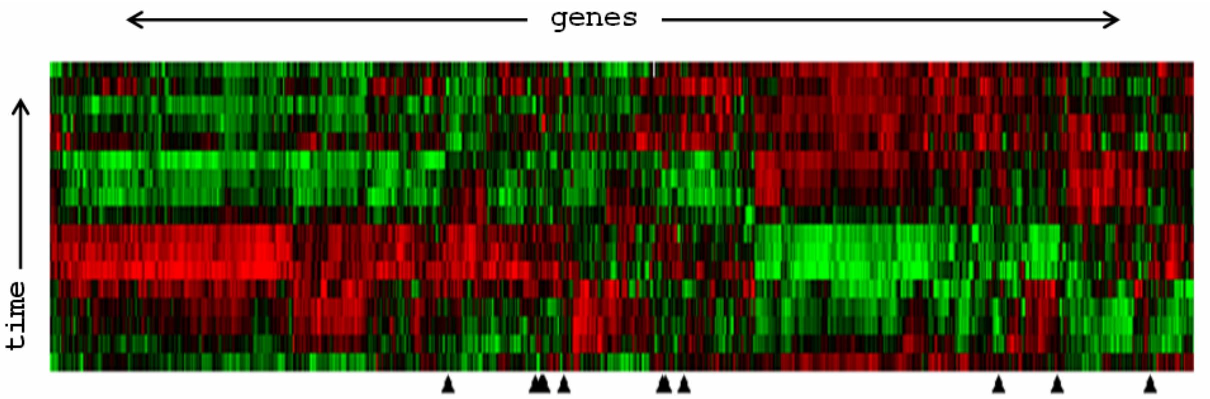


Figure 4. Heat-map clustering display of microarray time-course with arrows at the bottom of the figure highlighting genes which contribute to a pattern occurring exclusively over an interval of the data (see Figure 3).

associated techniques, which support the analysis of the data toward revealing certain less dominant patterns. These operate either by providing a more informative clustering view or allowing the results of multiple cluster views to be cross referenced and interactively explored. In the first case, the more informative clustering view is either one that allows genes to be assigned to multiple clusters or allows clusters to form across two or three dimensions (as opposed to the single dimension clustering provided by the ordering of gene representations in a hierarchical clustering heat-map).

The techniques that allow genes to be assigned to multiple clusters are Fuzzy k-means clustering [11, 12], Gene Shaving [13] and Plaid Modes [14]. Fuzzy k-means clustering and Gene Shaving produce groups of distinct clusters that can be represented as separate heat-maps (Figure 5). The disadvantage of this display is that, as genes can belong to multiple clusters, there will be a larger set of results and it will be even more difficult to perceive less dominant patterns unless they are specifically defined by the clusters. Given the scale of the data and subsequent high number of potentially ‘interesting’ patterns, for any given less-dominant pattern this will be extremely unlikely.

Unlike Fuzzy k-means clustering and Gene Shaving, the Plaid Models technique does not require for the activity of genes to be displayed more than once. Instead, the data is processed so that the bands of colour that link genes in a hierarchically clustered display are more predominant. This means that patterns of similar activity are more easily detected and the effort required to perceive associations between genes that are separated in the display is significantly reduced. It is, however, still the case that a pattern that is characterised over a limited interval of the experiments time frame will be dispersed along the gene-axis of the display and, if a pattern is contributed to by small proportion of the monitored genes (as is the biologically significant pattern in Figure 3) then it is unlikely that the pattern will detected.

Clustering across two or three dimensions can be performed using Principle Component Analysis (PCA) [15], Singular Value Decomposition (SVD) [16], Multidimensional Scaling (MDS) [17] or Self-Organizing Maps (SOM) [18-20]. The first three of these techniques are similar in that they communicate the activity of genes by presenting them as single points in a scatter-plot. From this type of display it is possible to perceive clusters of genes with similar activity from their proximity to areas where there is a higher density of gene representations (Figure 6). The two classes of clustering techniques which employ single point representations are PCA and SVD which use fixed axes to position gene representations and

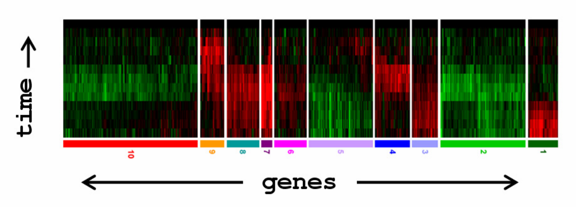


Figure 5. Separate heat-maps used to display the results of Fuzzy k-means clustering or Gene Shaving.

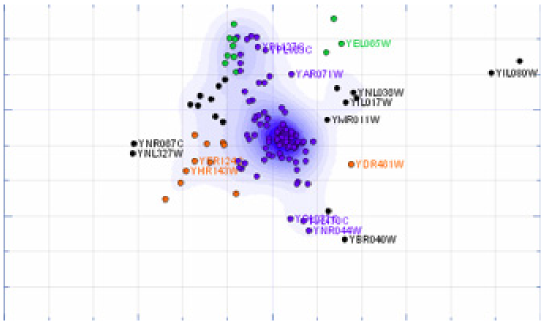


Figure 6. Scatter-plot clustering of microarray time-course data (areas with higher densities of gene representations are shaded to enhance clusters).

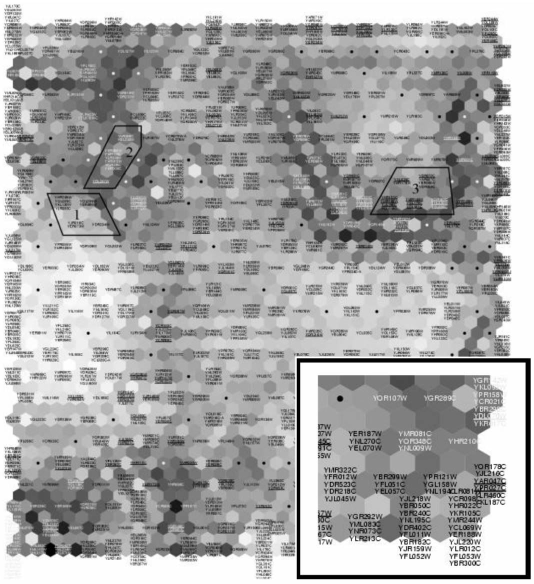


Figure 7. Self-Organizing Map clustering of microarray time-course data (cells with a higher number of genes assigned are lighter, inset - detail).

MDS which doesn't. In PCA and SVD these axes correspond to different measures of variation across the data so that the display characterizes the data along its lines of maximum variance. MDS forgoes axes employing an iterative algorithm that attempts to approximate inter-gene similarities in the display. Each of these approaches has its relative advantages and disadvantages. While a PCA or SVD display can be used to infer aspects of a gene's recorded activity by relating its position to the measure of variation communicated in the axes, it is also the case that variations in activity that are not represented in the axes will not be communicated in the display. Conversely, while MDS is more capable of communicating more subtle variations in the data, there are no axes from which to relate the position of a gene's representation to any aspect of its recorded activity.

SOM is similar to MDS in that it attempts to approximate inter-gene similarities in the display. However, rather than representing genes as single points in a scatter-plot display, genes are assigned to cells in a uniform lattice (Figure 7). When the cells of the lattice are shaded according to their relative gene populations, it is possible to perceive patches of dark and light which correspond to clusters. SOMs are more effective at grouping similar items while MDS is effective in preserving the structure of clusters [21]. This makes SOMs the preferred alternative in microarray data analysis where it is often the former objective that has higher priority.

The advantage of allowing clusters to form across two or three dimensions, as opposed to a single dimension, is that genes can be communicated as having less rigid associations with more clusters. For example, if a gene (x) is positioned on a one dimensional surface it can be placed between two clusters (A and B) to be associated with these two clusters. If a third cluster (C) is added then it becomes unclear as to degree by which the position of gene x is governed by its association between clusters A, B or C. If, however, the same gene is placed on a two dimensional surface we can determine, to a greater extent, its relative degree of association with three clusters A, B and C by observing its distance from each cluster. In general, however, an increased number of clusters tend to confuse the inference of which cluster which gene belongs to. This is particularly problematic when attempting to find patterns that are characterized over limited intervals of the time-course as the genes which contribute to these patterns are likely to be associated with other more dominant patterns and have their representations strongly associated with these patterns in the display.

In general the different types of cluster display are considered as being complementary. That is, there is a general acceptance that the relative advantages and limitations of the different techniques can be traded off against each other by combining them in multiple coordinated views of the same application interface [22-24]. For example, a cluster of associated genes perceived in a SOM display can be linked to a hierarchical clustering display so that their patterns of recorded activity can be identified, or Fuzzy k-means clustering could be used to find patterns not already found by, say, PCA.

To further compliment static clustering, information visualization techniques have been specifically developed to facilitate the interactive exploration of cluster results. These allow biologists to focus in on and labelling particular clusters [25] or query time-course to find patterns that may be less dominant in a clustering display. This second group of techniques (described here as visual queries) allow the user to specify a required pattern of activity over a limited interval of the time-course This can be an acceptable range of values over a given interval [26, 27], a change in values between time points [25, 27, 28] or a profile that the activity of genes must adhere to [25] (Figure 8). As this type of querying involves the specification of a limited time-interval, it can be seen as particularly appropriate for analysis which might involve the detection of less dominant patterns characterized by

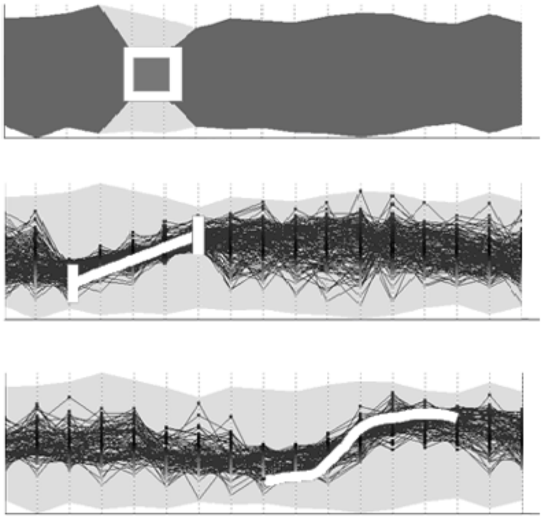


Figure 8. Visual queries: an acceptable range of values over a given interval (top), an acceptable change in values between time points (middle) and a profile that the activity of genes must adhere to (bottom).

trends in activity over such intervals. These techniques do not, however, provide an overview from which unknown or unsuspected patterns can be revealed and knowledge of a patterns existence is required before it can be formally identified.

Despite the range of clustering techniques, techniques that support clustering and the opportunity to combine and manipulate these different representations of the data, biologists are still unable to reveal a significant proportion of potentially relevant less dominant patterns in their data. Specifically, no technique or combination of these techniques are able to reveal previously unsuspected patterns that are characterized over limited intervals of an experiments time frame and are contributed to by a smaller proportion of the genes monitored (Figure 3). Given that to fully exploit the potential of the data biologists require analysis techniques that allow them to make unexpected discoveries and that a significant proportion of biological phenomenon will be related to less dominant patterns, new techniques are required that allow biologists to make unexpected discoveries of less dominant patterns.

3. Time-series Explorer

Our research to date has primarily focused on supporting the discovery of temporal patterns in microarray time-course and, specifically, the less dominant patterns that cannot be revealed using existing established techniques. This has included the development of two significant prototypes. The first of these allows biologists to relate scatter-plot representations of time-course intervals to a traditional graph view in order to distinguish the time-series of individual genes and groupings of genes from the background [29]. The second, allows biologists to query the activity of genes over time intervals by selecting gene representations in an interval scatter-plot view [30].

The Time-series Explorer [31] builds on our previous prototypes to facilitate the discovery of unsuspected patterns of temporal activity. This is accomplished by allowing biologists to explore their data using three coordinated parallel views of their data. These are an activity graph view, a change-in-activity graph view and an activity against change-in-activity interval scatter-plot (Figure 9). The layout of data in these views is as follows:

a) The activity graph view (top left of Figure 9) overlays activity versus time graphs for all genes, or a selected subset of the genes, to indicate the range of high or low activity at each time point (time-point labels are specified in the original data-file).

b) The change-in-activity graph view (bottom left of Figure 9) overlays change-in-activity against time graphs to indicate the range of rising or falling activity of selected genes between each pair of adjacent time points. Here change-in-activity defined as the relative change in recorded activity between time points for a gene.

c) A vertical bar overlaid onto each graph view is a visual representation of the current selected interval.

d) The activity against change-in-activity interval scatter-plot view (right-hand-side of Figure 9) summarizes the data within the specified interval by representing each gene as a single point. The translation of a gene representation along the Y-axis corresponds to its mean activity over the interval and the translation of a gene along the x-axis corresponds to its change-in- activity from the start of the interval to its end.

e) Each view is colour coded to indicate the density of overlaid gene representations (the coding used is adapted from a standard transparency composite and has the added benefit that it allows outliers to be distinguished from the background [31]). In certain selection modes genes are highlighted as red, in others un-selected genes are greyed out and do not contribute to the colour coding.

As the combined views of the Time-series Explorer use corresponding axes and can be directly related to each other they can be described as being parallel [32]. The x-axes of each graph view map to the same dimension (time) and are geometrically parallel in the interface. The y-axes of each graph view relate directly to each axes of the scatter-plot with the same rescaling and distortions applied [31]. As the rising/falling axes of the change-in-activity graph view and the scatter-plot are directly related but not geometrically parallel, clear labelling is used to enforce the association between these axes. For the sake of consistency, and to enforce the other relevant associations between axes, the same type of labelling is applied to all other related axes pairs.

The different views of the Time-series Explorer are coordinated in two ways. Firstly, selecting genes in the scatter-plot or either graph view causes them to be highlighted in all views with immediate and continuous display of results (i.e. tight coupling [33] of the coordinated views). If a labelling tool is activated from the interface toolbar, moving the mouse over (brushing) gene representations in the scatter-plot view causes them to be labelled and have their expression patterns over the entire time-course coloured red the graph view. The labelling options are standard labelling, where only the scatter-plot gene representation directly under the mouse pointer is selected, and excentric labelling [34], where all gene

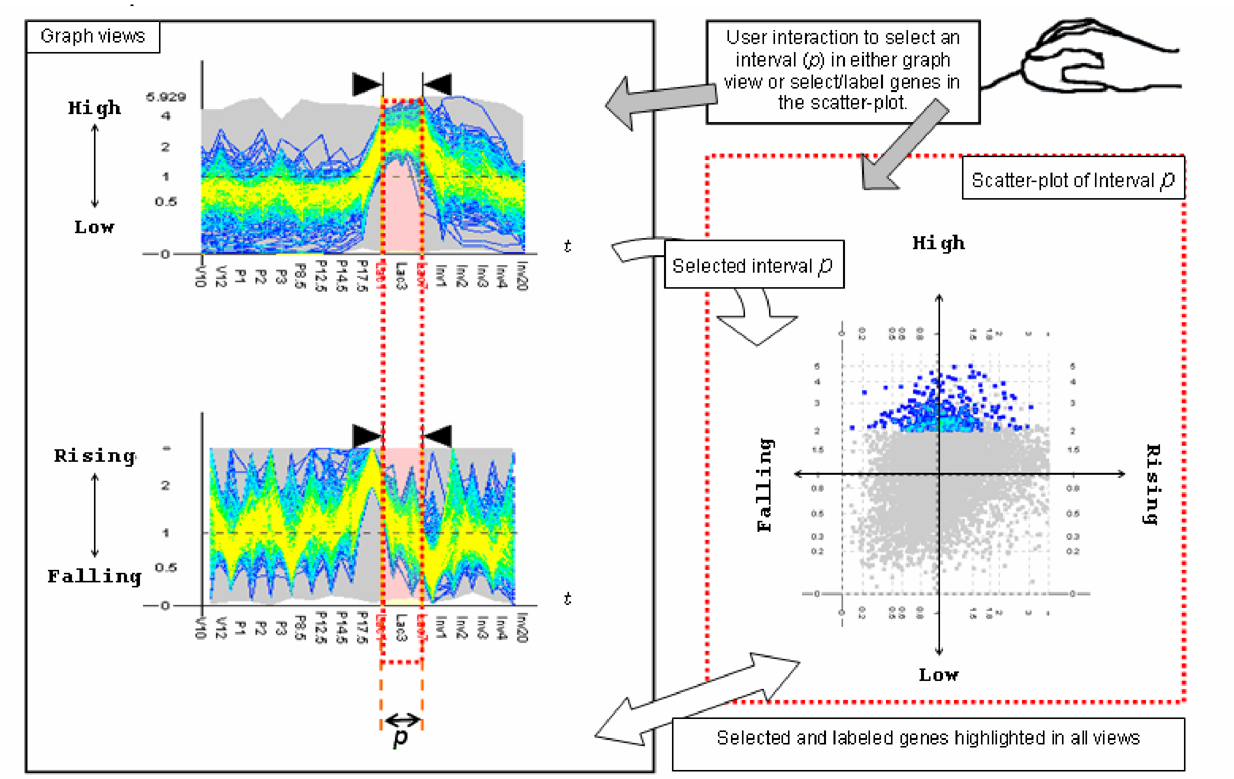

Graph views
High
Low
Rising
Falling
5.929
4
2
1
0.5
0
t
V10
V12
P1
P2
P3
P8.5
P12.5
P14.5
P17.5
Lac1
Lac3
Lac7
Inv1
Inv2
Inv3
Inv4
Inv20
p
User interaction to select an interval (p) in either graph view or select/label genes in the scatter-plot.
Scatter-plot of Interval p
Selected interval p
High
Low
Falling
Rising
Selected and labeled genes highlighted in all views

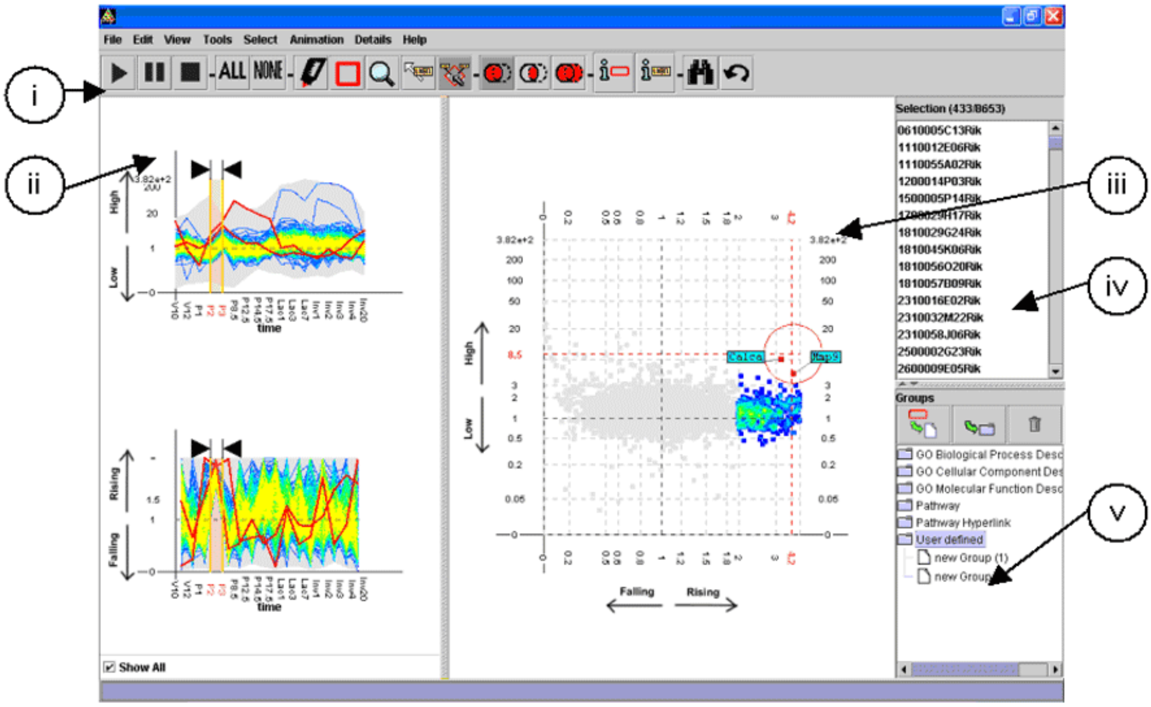
Figure 9. **Coordinated views of the Time-series Explorer (light colours represent a high density of overlaid elements, dark colours represent a low density of overlaid elements)**

representations within the bounds of a visible circle are selected. When freehand or box selection tools are activated genes can be selected more permanently. With the box selection tool, genes are selected by clicking and dragging to draw a box round their representations in the scatter-plot. The freehand selection tool allows the user to select genes by clicking and dragging a freeform shape around their representations. In either case, the density colour coding is reapplied according to only to those genes which remain selected (the representations of un-selected genes are greyed out) and genes remain selected until another selection is made.

The second coordination between views involves the interval selection in the graph views. The interaction mechanism that allows the interval selection to be adjusted is essentially the same as that of a multi-range dynamic query slider [35] utilizing the internal slider space for a visual representations of data in a manner similar to that of data-visualization sliders [36]. Dragging the edges of the vertical bar overlaid onto the graph views to represent the selected interval allows the user to adjust its start and end times independently. Dragging the center of the bar changes the start and end times with the duration remaining

constant to shift the selected interval. During these interactions the interval adjusts in small steps with activity levels interpolated so that changes are gradual and the motion of gene representations in the scatter-plot, which are dependent on the range of the selected interval, is fluid. This allows the scatter-plot to be animated so that it is possible to track genes, and correlated groups of genes, according to changes in their recorded activity across time.

Manual adjustments of the selected interval, actioned by interacting with the graph view, give users tight control over the pace and direction of the animation so it can be slowed down as interesting features become apparent, reversed when they want to look at something again and stopped, when appropriate, to focus in on an interesting interval and investigate patterns occurring over that interval in more detail by interacting with the scatter-plot view. Alternatively, the animation can be progressed automatically in a regular fashion using a play button located on the Time-series Explorer toolbar.


i
ii
iii
iv
v
File Edit View Tools Select Animation Details Help
ALL
NONE
3.82e+2
200
20
1
0
High
Low
Rising
Falling
1.5
1
time
V10 V12 P1 P2 P3 P8.5 P12.5 P14.5 P17.5 Lac1 Lac3 Lac7 Inv1 Inv2 Inv3 Inv4 Inv20
Show All
Calca
Mmp9
8.5
Falling
Rising
Groups
GO Biological Process Desc
GO Cellular Component Des
GO Molecular Function Desc
Pathway
Pathway Hyperlink
User defined
new Group (1)
new Group

Selection (433/8653)

0610005C13Rik
1110012E06Rik
1110055A02Rik
1200014P03Rik
1500005P14Rik
1700029H17Rik
1810029G24Rik
1810045K06Rik
1810056O20Rik
1810057B09Rik
2310016E02Rik
2310032M22Rik
2310058J06Rik
2500002G23Rik
2600009E05Rik

Figure 10. A screen-shot of the Time-series Explorer interface (i. toolbar, ii. graph view, iii. scatter-plot, iv. selected gene list and v. grouping panel).

The interface of the Time-series Explorer is shown in Figure 10. The five panels combined in the interface are a toolbar, the graph views, the scatter-plot, a list of the selected genes and grouping panel which allows users to store selections and select genes that belong to predefined classifications (a more comprehensive description of the specific interaction mechanisms and the functionality of each panel is described in [31]).

3.1 Combining animated and static views to find patterns of temporal activity

In order to determine whether or not the Time-series Explorer technique was indeed capable of supporting the analysis of microarray time-course in a manner that allowed the biologists to reveal the specific type of pattern that could not be revealed using existing established techniques, we conducted an formal user evaluation of the tool. The latest phase of this evaluation involved an experienced biologist attempting to analyse a familiar set of microarray time-course data. This related to the recorded activity of around 8,500 genes over 17 time points belonging to 4 concurrent stages of mouse development: virgin (activity recorded at days 10 and 12 of this stage), pregnancy (activity recorded at days 1, 2, 3, 8.5, 12.5, 14.5 and 17.5), lactation (activity recorded at days 1, 3 and 7) and involution (activity recorded at days 1, 2, 3, 4 and 20) [4]. While the original data contained three replicates (separate runs of the experiment under identical conditions allowing results to be statistically verified), the data used in the evaluation combines these replicates using the average recorded value for each gene at each time-point to optimally exploit the existing functionality of the Time-series Explorer.

The evaluation allowed us not only to assess the relative benefits of using the Time-series Explorer to analyse microarray time-course but also to determine specifically how the different views of the data would be combined in order to specifically find an unsuspected pattern of temporal activity. In this section we describe three characteristic patterns found and the manner in which views were combined by the biologists in order to find them.

The first characteristic pattern involves a combination of two distinct trends in activity over different intervals of the time-course. Here the biologist wished to find which genes and groups of genes having rising activity at the start of lactation and falling activity at the end of lactation. The first stage in finding this pattern was to use the activity graph view to select the interval at the start of lactation. After this the scatter-plot was used to select all the genes with high rising activity over this interval. Next, the interval at the end of lactation was selected and the query was refined by selecting all the genes with falling activity over this interval. The results of the selections were then viewed over the entire time-course in the graph views. This revealed two groups of outliers that either had significant rising or falling activity over different periods of the pregnancy stage (see Figure 12) and prompted the biologist, who was concerned that the activity of these genes would not adhere to the required profile, to investigate further by adjusting the selected interval so that the representations of these genes could be labelled and their activity over the entire time-course highlighted. This revealed that a significant proportion of the genes failed to fit the original profile and the original query was adjusted (by making the required level of activity at the start of lactation higher) to exclude them from the results. Finally the resultant gene listing was cross-referenced with existing predefined gene classifications in a pop-up window so that biological significance of the pattern could be properly assessed. This indicated a number of interesting genes and gene groupings which the biologist exported from the Time-series Explorer to be cross referenced with the results of other related experiments at a later date.

The second pattern relates to general trends across the entire time-course. To investigate these general trends the biologist selected an interval fixed at its minimum value (an interval constrained by two time-points for which activity is recorded) and shifted it across the entire time frame of the experiment to animate the scatter-plot view. At various stages during this animation the spread of gene representations in the scatter-plot became horizontally elongated. This occurred primarily during transitions between stages of development (i.e. virgin to pregnancy, pregnancy to lactation and lactation to involution) and indicated large numbers of genes with significant changes in their level of activity.

The majority of these trends were unsurprising to the biologist as they reflected changes in the essential functioning of cells within the sample that would largely be detected by the observing the general activity patterns of groupings formed by clustering. Somewhat more interesting were the more subtle trends, such as the increased number of genes with changes in activity during pregnancy in relation to lactation. It was later verified that these particular

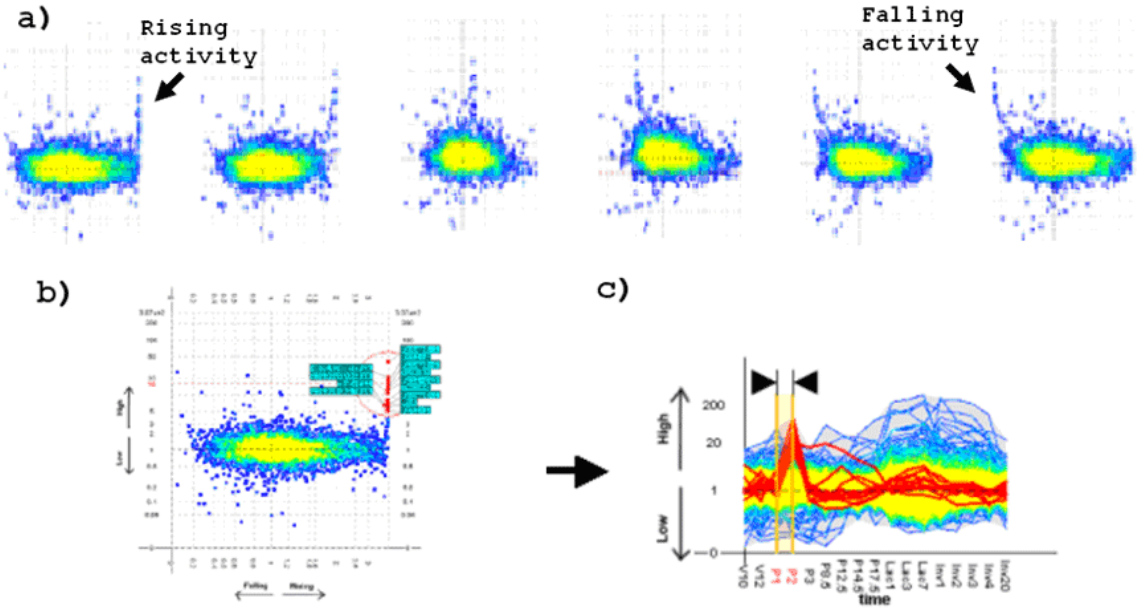


Figure 11. An unexpected pattern of temporal activity found using the Time-series Explorer: a) Animating the scatter-plot reveals a group of outlying genes with rising then falling expression over a small interval of the time-course, b) moving the mouse over the gene representations in the scatter-plot view allows them to be labelled and c) have their expression patterns over the entire time-frame highlighted in the standard graph view.

patterns would not be revealed by clustering.

The third, and most significant, pattern was found while animating the scatter-plot to investigate general trends. As the scatter-plot animated through days 1 to 3 of the pregnancy stage (an interval for which there are three time-points for which activity is recorded) an outlying group of gene representations showed significant rising then falling activity. To investigate this further the relevant interval was animated again, and then stopped so that the outlying genes could be labelled by moving the mouse over their representations in the scatter-plot. As the labelled genes were highlighted in the graph view, this revealed that the majority of the genes also shared low activity over the remainder of the time-course. Next the genes were selected and cross-referenced with pre-defined gene classifications. Significantly the selection was found to contain a high proportion of Keratin associated genes. Figure 11 illustrates this pattern showing selected frames of the original animation from interval P1 to P2 through to interval P2 to P3, the labelled scatter-plot at P1 to P2 and the effect of labelling in the coordinated graph view where the genes are highlighted.

The main outcome of our user evaluation was to verify that the Time-series Explorer is uniquely capable of revealing previously unsuspected patterns of temporal activity and that the patterns found were of sufficient relevant biological significance to encourage a biologist to use the technique in the analysis of data from other experiments. Moreover, the biologist felt the tool to be more flexible than the other techniques used. This stemmed from the fact a large number of the valuable patterns in microarray data are combinations of patterns of temporal activity and that a biologist can use the technique to investigate different types of pattern without having to readjust between different tools and multiple unrelated representations of the data.

4 Discussion

Analysis of the biologist's interaction with the Time-series Explorer interface during our user evaluation showed that each of its combined views has its own specific utility with regard to the finding of unsuspected patterns of temporal activity and that the direct relationships between these different views allowed them to be used together in order to realise a number of valuable information seeking tasks. In general, the Time-series Explorer was found to be used in three complementary modes of operation which utilized the functionality, and combined functionalities, of the views in different ways.

The first mode of operation is a type of general overview mode. This occurs when a large group of genes with dissimilar patterns of activity (or the entire data set) are selected and the graph views cannot communicate any trends or outliers due to the density of crossing lines. Here the biologists will interact with the graph views to animate across the entire time-course in order to uncover general trends across time or outliers at smaller intervals by focusing their attention on the animated scatter-plot.

The second mode of operation involves queries that return a large set of results. Here the biologist will roughly select a grouping of genes in a static scatter-plot view and repeatedly refine their selection by cross referencing it in either graph view until their selection appears to match a certain profile. The biologist's idea of the profile they wish to query is generally prompted by either pre-knowledge of the data (e.g. the notion that a significant group of genes will rise, fall, be high or be low over a significant interval) or a general pattern perceived by animating the scatter-plot over the entire time-course. Most notably, it is during this mode of analysis that the change-in-activity graph view seems to be of most value. This is because there are still a large number of selected genes and crossing lines can mask significant trends in change-in-activity in the activity graph view without necessarily masking the same trends in the change-in-activity graph view (Figure 12).

The third mode of operation is where the biologist attempts to identify genes contributing to a less dominant pattern occurring over a limited interval of the time-course. These types of pattern are normally revealed unexpectedly during either of the other analysis modes as a smaller group of genes appear as coherent outliers when the scatter-plot is animated. Once found, the genes contributing to such a pattern are normally labelled in the scatter-plot and highlighted in the standard graph view. This allows the biologist to asses any significance that may be derived from the relevant gene symbols and identify whether or not activity is suitably coherent over the remainder of the time-course. If this information reveals that the genes are of sufficient interest to the biologists, they are selected (normally using the freehand selection tool on the scatter-plot) and cross referenced with pre-defined groupings.

With reference to these different modes of operation the benefits of each of the combined Time-series Explorer views can be described as follows:

- **Activity graph view:** This is the most common representation of microarray time-course and, in many cases, it is found to be the most intuitive. It communicates the range of values at different time points. This which makes it useful to indicate the activity of genes over intervals for which the activity patterns of the selected genes are similar. It is, however, unable to communicate the activity of genes over intervals of time where activity is dissimilar. In general this view becomes useful when an initial query causes the selected group of genes to share some common pattern of activity over one or more intervals of the time-course.

- **Change-in-activity graph view:** This view is less intuitive than the standard graph view but is useful when crossing lines mask trends in change-in-activity in the activity graph view (Figure 12). This occurs when there are a large number of genes with similar changes in activity but the activity levels are dissimilar.
- **Scatter-plot (static):** The chief benefit of the static scatter-plot view is that it provides an interface from which genes can be selected or labelled according to aspects of their activity over a limited interval. The activity of these genes can then be viewed over the entire time-course as it is highlighted in the more intuitive activity graph view (this linking also helped the biologists to familiarise themselves with the scatter-plot in static and animated form [30]).
- **Scatter-plot (animated):** When animated, by adjusting the interval selection using the graph views, the scatter-plot can reveal trends and outliers over the entire time-course. This animated overview is required when trends and outliers are occluded by crossing lines in the graph views.

On the whole the biologists seemed to shift focus between the different views of the data with relative ease. This was due to the fact that the views were parallel and easily related to each other. The coordination of views as the biologists interacted with the data enforced this understanding with a resultant effect that they were able to combine views to realise a number of information seeking objectives and explore their data in a manner that allowed them to uncover several significant patterns.

The Time-series Explorer can be considered as unique in its application of parallel coordinated views for the exploratory analysis of microarray time-course. While existing techniques often coordinate clustering and graph views of microarray time-course, the relationship between views cannot be considered as direct due to the extent of the alternate processes which are applied to the data before visualisation. These techniques are better described as multiform visualisations [32] as they represent alternative realisations of the same data.

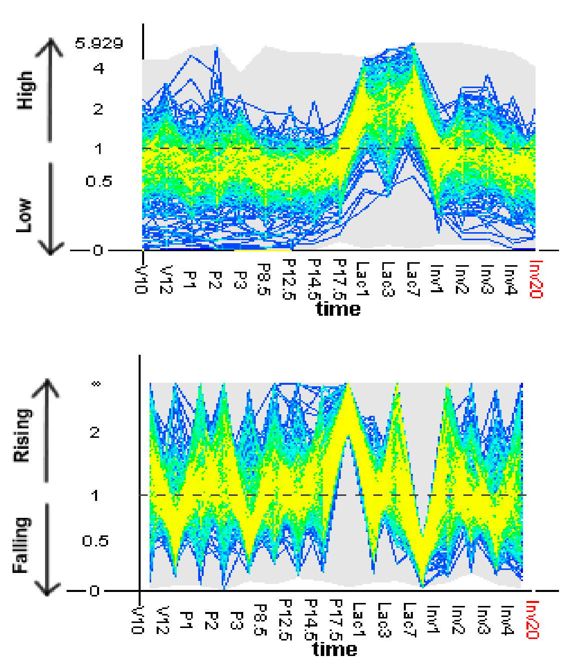


Figure 12. Change-in-activity graph view (bottom) can be used when significant trends in change-in-activity (from V12 to P1 and P3 to P8.5) are obscured in the activity graph view (top).

5 Conclusions and further work

In this paper we have described the Time-series Explorer's combination of parallel views for the exploratory analysis of microarray time-course data. The specific views combined in the Time-series Explorer are an activity graph view, a change-in-activity graph view and an activity against change-in-activity interval scatter-plot. These views are coordinated so that the scatter-plot summarises an interval which is selected using either graph view and selected genes are highlighted in all views. Having determined the types of pattern that cannot be revealed using existing established techniques as being less dominant patterns in the data, such as those that occur over limited intervals of the time-course, we evaluated the Time-series Explorer and verified that it is capable of effectively supporting biologists in their efforts to find these types of pattern. Analysis of the biologist's interaction with the Time-series Explorer interface during this evaluation showed that each of its different views has its own distinct benefits and the direct relationship between views allowed them to be combined in such a manner that they could be used together to explore the data toward revealing the less dominant patterns. This indicates that, while existing established techniques (which either use a single clustering view of the data or multiform visualisations that combine multiple clustering and/or graph views may be more appropriate for revealing general trends in microarray time-course), a parallel view approach is more appropriate for revealing detail in the data and supporting the discovery of less dominant patterns.

Due to the success of our user evaluation the tool has been adopted for the analysis of two ongoing experiments. As these experiments will involve multiple time-courses under different conditions we plan to adapt the technique so it is capable of comparing the results of multiple related experiments. In order to improve the quality of patterns found by such an extended technique and increase the biologists ability to make informed decisions as to the significance of their results we also plan to incorporate measures of statistical confidence, into both the visual representations and interaction mechanisms of the technique, and increase the possibilities for linking the data and results to external databases of pre-defined classifications, gene annotations and pathway information. The expected result is that we can adapt the existing Time-series Explorer technique exploiting the benefits of the animated interval scatter-plot and complimentary parallel views to develop a more complete microarray time-course analysis tool.

Acknowledgements

Edward K. Wagner
The Center for Virus Research (CVR)
The University of California at Irvine

Peter Ghazal, Thorsten Forster, Paul Dickinson
Scottish Centre for Genomic Technology and Informatics
The University of Edinburgh

Torsten Stein
Division of Cancer Sciences and Molecular Pathology
Western Infirmary
University of Glasgow

Donald Dunbar
Molecular Physiology Group
University of Edinburgh Medical School

References

[1] Schena, M., et al., *Parallel human genome analysis: microarray-based expression monitoring of 1000 genes.* Proc. Natl. Acad. Sci. U.S.A. 93,, 1996. **93**(20): p. 10614-10619.

[2] Schena, M., et al., *Quantitative monitoring of gene-expression patterns with a complementary-DNA microarray.* Science, 1995. **270**(5235): p. 467-470.

[3] Wen, X., et al., *Large-Scale Temporal Gene Expression Mapping of CNS Development.* Proc Natl Acad Sci USA, 1998. **95**(1): p. 334-339.

[4] Stein, T., et al., *Involution of the mouse mammary gland is associated with an immune cascade and an acute-phase response, involving LBP, CD14 and STAT3.* Breast Cancer Research, 2004. **6**(2): p. R75 - R91.

[5] Hughes, T.R., et al., *Functional discovery via a compendium of expression profiles.* Cell, 2002. **102**(1): p. 109- 126.

[6] Brown, P.O. and D. Botstein, *Exploring the new world of the genome with DNA Microarrays.* Nature Genetics, 1999. **21**(1 Suppl): p. 33-37.

[7] Brazam, A. and J. Vilo, *Gene expression data analysis.* FEBS letters, 2000. **480**(1): p. 17-24.

[8] Bassett, D.E., M.B. Eisen, and M.S. Boguski, *Gene expression informations- it's all in your mine.* Nature Genetics, 1999. **21**(1 Suppl): p. 51- 55.

[9] Eisen, M.B., et al., *Cluster analysis and display of genome-wide expression patterns.* Proc. Natl. Acad. Sci. USA, 1998. **95**(25): p. 14863-14868.

[10] Segal, E., et al., *Rich probabilistic models for gene expression.* Bioinformatics, 2001. **17**(Suppl 1): p. 243-52.

[11] Gasch, A.P. and M.B. Eisen, *Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering.* Genome Biology, 2002. **3**(11): p. research0059.1–research0059.22.

[12] Gath, I. and A.B. Gev, *Unsupervised Optimal Fuzzy Clustering.* IEEE Transactions on Pattern Analysis and Machine Intelligence, 1989. **11**(7): p. 773 - 780.

[13] Hastie, T., et al., *Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns.* Genome Biology, 2000. **1**(2): p. research0003.1-0003.21.

[14] Lazzeroni, L. and A. Owen, *Plaid Models for Gene Expression Data*. 2000, Stanford University.

[15] Raychaudhuri, S., J.M. Stuart, and R.B. Altman. *Principal Components Analysis to Summarize Microarray Experiments: Application to Sporulation Time Series*. in *Pacific Symposium on Biocomputing*. 2000.

[16] Alter, O., P.O. Brown, and D. Botstein, *Singular value decomposition for genome-wide expression data processing and modeling.* PNAS, 2000. **97**(18): p. 10101–10106.

[17] D'haeseleer, P., et al. *Mining the Gene Expression Matrix: Inferring Gene Relationships from Large Scale Gene Expression Data*. in *Information Processing in Cells and Tissues*. 1998: Plenum Publishing.

[18] Kaski, S., et al., *Analysis and visualization of gene expression data using self-organizing maps.* Proceedings of NSIP-01, IEEE-EURASIP Workshop on Nonlinear Signal and Image Processing, 2001. **15**(8-9): p. 953- 966.

[19] Kohonen, T., *The Self-Organizing Map.* Proceedings of the IEEE, 1990. **78**(9): p. 1464-1480.

[20] Tamayo, P., et al., *Interpreting patterns of gene expression with self-organizing maps.* Proceedings of the National Academy of Sciences of the United States of America, 1999. **96**(6): p. 2907-2912.

[21] Flexer, A., *Limitations of Self-Organizing Maps for Vector Quantization and Multidimensional Scaling*, in *Advances in Neural Information Processing Systems 9*, M.M.C.e. al., Editor. 1997, MIT Press/Bradford Books. p. 445-451.

[22] Saraiya, P., C. North, and K. Duca. *An Evaluation of Microarray Visualization Tools for Biological Insight*. in *IEEE Symposium on Information Visualization (INFOVIS'04)*. 2004. Austin Texas.

[23] *GeneSpring*. 2004, Silicon Genetics www.silicongenetics.com.

[24] *Spotfire Decisionsite for Functional Genomics*. 2002, Spotfire.

[25] Seo, J. and B. Shneiderman, *Interactively Exploring Hierarchical Clustering Results.* IEEE Computer, 2002. **35**(7): p. 80-86.

[26] Hochheiser, H., Shneiderman, B., *Visual Specification of Queries for Finding Patterns in Time-Series Data*, in *Proceedings of Discovery Science 2001*. 2001, University of Maryland, Computer Science Dept.

[27] Hochheiser, H. and B. Shneiderman, *Dynamic query tools for time series data sets: Timebox widgets for interactive exploration.* Information Visualisation, 2004. **3**(1): p. 1-18.

[28] Hauser, H., F. Ledermann, and H. Doleisch. *Angular Brushing of Extended Parallel Coordinates*. in *IEEE Symposium on Information Visualization (InfoVis'02)*. 2002. Boston, Massachusetts, USA.

[29] Craig, P., J.B. Kennedy, and A. Cumming. *Towards Visualising Temporal Features in Large Scale Microarray Time-series Data*. in *6th International Conference on Information Visualisation - IV2002*. 2002. University of London, London, GB: IEEE Press.

[30] Craig, P. and J.B. Kennedy. *Coordinated Graph and Scatter-Plot Views for the Visual Exploration of Microarray Time-Series Data*. in *I2003 IEEE Symposium on Information Visualization*. 2003. Seattle WA: IEEE Computer Society.

[31] Craig, P., J.B. Kennedy, and A. Cumming, *Animated Interval Scatter-plot Views for the Exploratory Analysis of Large Scale Microarray Time-course Data.* Information Visualisation, 2005. **4**(3): p. (accepted for publication, to appear Autumn 2005).

[32] Roberts, J.C. *Multiple-View and Multiform Visualization*. in *Visual Data Exploration and Analysis VII, Proceedings of SPIE*. 2000.

[33] Ahlberg, C. and B. Shneiderman. *Visual Information Seeking: Tight Coupling of Dynamic Query Filters with Starfield Displays*. in *ACM Conference on Human Factors in Software, CHI '94*. 1994. Boston, MA, USA: ACM Press.

[34] Fekete, J. and C. Plaisant. *Excentric Labeling: Dynamic Neighborhood Labeling for Data Visualization*. in *Conference on Human Factors in Computing Systems*. 1999. Pittsburgh, Pennsylvania, United States.

[35] Shneiderman, B., *Dynamic Queries for Visual Information Seeking.* IEEE Software, 1994. **11**(6): p. 70 -77.

[36] Eick, S.G. *Data Visualization Sliders*. in *UIST '94*. 1994. Marina del Ray, California, USA: ACM Press.