Animated Interval Scatter-plot Views for the Exploratory Analysis of Large Scale Microarray Time-course Data

Authors:
Paul Craig, Jessie Kennedy, Andrew Cumming

Correspondence:
Paul Craig
School of Computing
Napier University
Merchiston Campus
10 Colinton Road
Edinburgh
EH10 5DT
Scotland
Tel: +44 (0)131-455-2786
Fax: +44 (0)131-455-2727
E-mail: p.craig@napier.ac.uk

Running title:
Animated Views for Microarray Time-course

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

Supplementary material:
http://www.kcchosting.co.uk/~pc/?page=research

Abstract:
Microarrays technologies are a relatively new development that allow biologists to monitor the activity of thousands of genes (normally around 8,000) in parallel across multiple stages of a biological process. While this new perspective on biological functioning is recognised as having the potential to have a significant impact on the diagnosis, treatment, and prevention of diseases, it is only through effective analysis of the data produced that biologists can begin to unlock this potential. A significant obstacle to achieving effective analysis of microarray time-course is the combined scale and complexity of the data. This inevitably makes it difficult to reveal certain significant patterns in the data. In particular it is less dominant patterns and, specifically, patterns that occur over smaller intervals of an experiment's overall time-frame that are more difficult to find. While existing techniques are capable of finding either unexpected patterns of activity over the majority of an experiment's time frame or expected patterns of activity over smaller intervals of the time frame, there are no techniques, or combination of techniques, that are suitable for finding unsuspected patterns of activity over smaller intervals. h order to overcome this limitation we have developed the Time-series Explorer, which specifically supports biologists in their attempts to reveal these types of pattern by allowing them to control an animated interval scatter-plot view of their data. This paper discusses aspects of the technique that make such an animated overview viable and describes the results of a user evaluation assessing the practical utility of the technique within the wider context of microarray time-series analysis as a whole.

Keywords:
Information visualization, Bioinformatics, Microarray, Time-series, Animation, Evaluation

Introduction

The development of microarray technologies [1; 2] has revolutionized biological and biomedical research, specifically in the area of gene expression analysis. Gene expression is a key indicator of a gene's activity and where previous technologies only allowed biologists to monitor the activity of a few genes at a time, microarray experiments allow them to monitor the activity of thousands of genes in parallel across multiple conditions or [3], more commonly, across multiple stages of a biological process [4; 5; 6] to generate microarray time-course data. Figure 1 shows an expression against time plot of microarray time-course with the range of values at each time point defining a grey area and the time-series of a single gene highlighted.

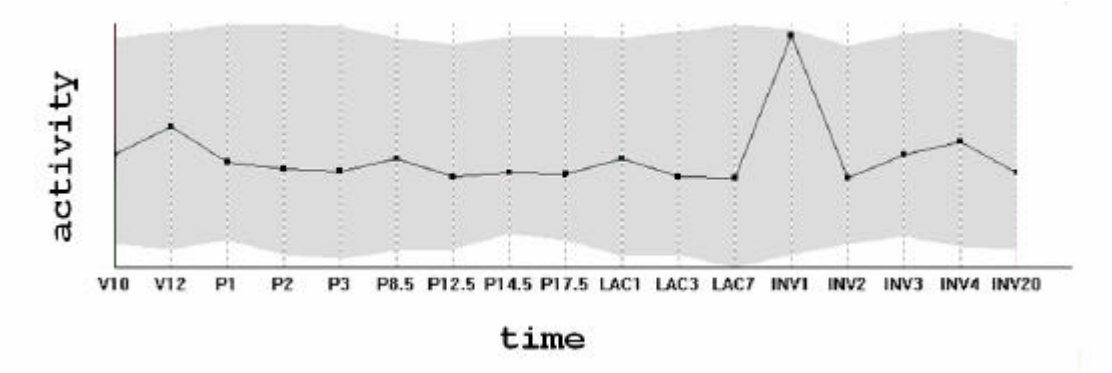

Figure 1 Microarray time-course data.

This new perspective on biological functioning has a number of advantages. Most significantly, the ability to monitor large number of genes in parallel allows biologists to investigate biological processes without the necessity of prior information indicating that any particular gene or group of genes are involved. Moreover, subject to proper analysis, the data produced by microarray experiments has the potential to reveal the existence and relationships between biological phenomena that combine to realise biological processes regardless of whether the existence of such phenomena are suspected or not. In effect, where previous technologies allowed biologists to test limited hypotheses involving a few genes at a time, microarrays provide biologists with data from which they can not only test hypotheses but also form new hypotheses as patterns in the data reveal previously unknown or unsuspected phenomena [7]. To fully exploit this potential biologists require analysis techniques that allow them to make unexpected discoveries and gain insights from their data. In order to meet these requirements, analysis techniques must support exploratory analysis of the data and overcome problems associated with its massive scale and complexity.

At present there are a limited range of techniques that are capable of revealing previously unsuspected patterns from microarray data. These techniques largely rely on procedures developed for the analysis of

multidimensional data and process the data to form clusters (groups) of genes based on the relative similarity of recorded expression (characteristic examples are [8, 9, 10]). Time-series data, of the type that is produced by microarray time-course experiments, can be conceptualised as a specialized subset of multidimensional data [11] with the distinguishing characteristic that dimensions (time-points) are ordered. Clustering techniques do not account for this aspect of the data and, as a consequence, are ill-suited to revealing certain significant patterns in the data [12]. Specifically, clustering tends to miss out patterns that occur exclusively over smaller intervals of an experiment's time frame.

An example of a significant pattern that would not be revealed by clustering is illustrated in Figure 2. Here a rise then a fall in expression found over a particular interval 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, if the data were clustered, any different patterns of expression before or after the interval would cause the related genes to be assigned to different groupings with the significance of their common activity over the relevant time period lost.

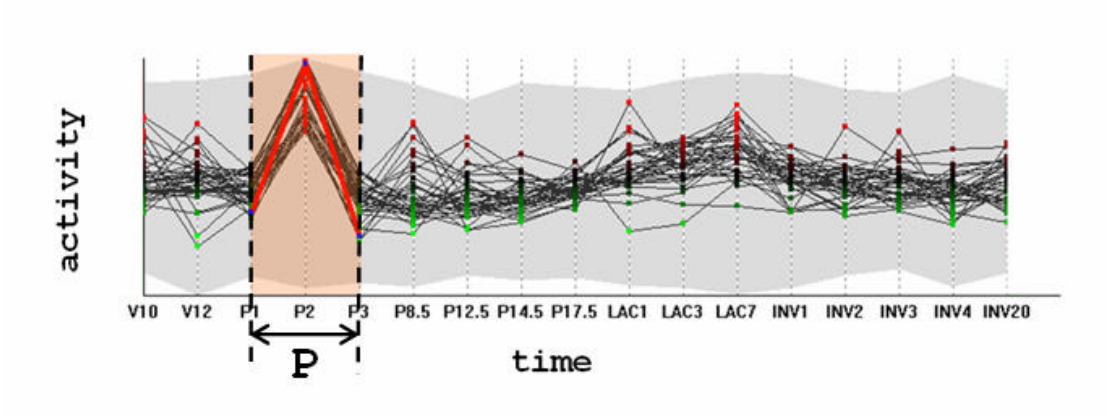

Figure 2 A significant pattern occurring exclusively over an interval (P).

Visual queries are commonly used to supplement existing clustering techniques in order to find certain patterns that exist over intervals. These allow the user to specify a required pattern of expression over a limited interval of the time-course This can be an acceptable range of values over a given interval [13] (top of Figure 3), a change in values between time points [13, 14] (middle of Figure 3) or a profile that the expression of genes must adhere to [14] (bottom of Figure 3). As type of querying involves the specification of a limited time-interval and it is particularly appropriate for analysis which might involve the detection of less dominant patterns characterized by trends in activity over such intervals. These techniques do not, however, allow biologists to reveal these type patterns if they are not already suspected. This is due to limitations in the overview provided (which is unable to reveal anything other than the range of values at individual time points) and means that if a biologist has no knowledge of a process's timing or the genes that participate in the process then they are required to execute

multiple speculative queries before patterns that relate that process to the experimental conditions can be revealed.

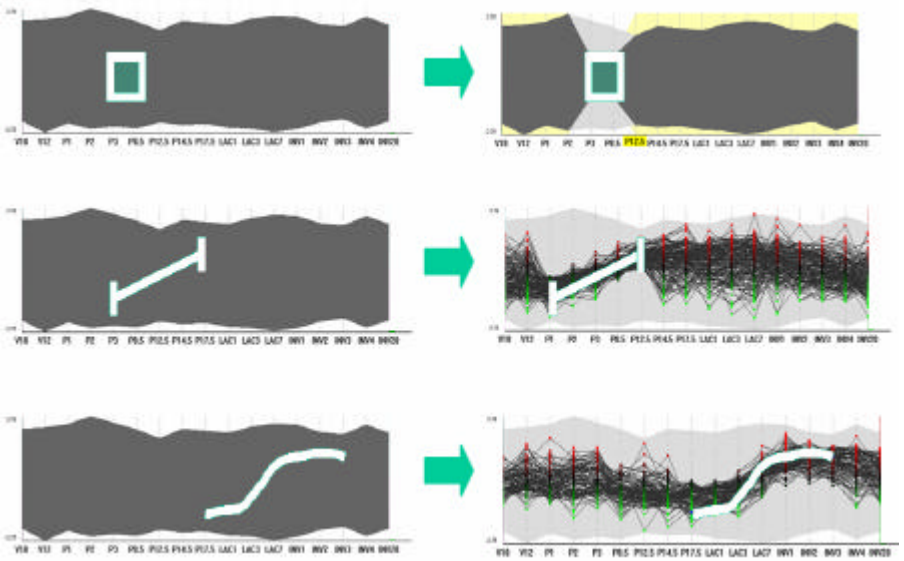

Figure 3 Visual queries (LHS query and RHS query results): an acceptable range of values over a given interval (top), an acceptable change in values between time points (middle) and a profile that the expression of genes must adhere to (bottom).

Our research to date has primarily focused on supporting the discovery of temporal patterns in microarray time-course and, specifically, the type of patterns that cannot be revealed using existing techniques. This has included the development of a technique that 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 [15]. This was followed by the development of a technique that allows biologists to query the activity of genes over intervals by selecting gene representations in an interval scatter-plot view [16]. In this paper we describe the Time-series Explorer, which builds on our previous work to facilitate the discovery of unsuspected patterns of temporal activity by allowing users to animate through scatter-plot representations of successive time-course intervals. This includes a discussion of the various display techniques used and describes the user interaction processes required to find the patterns of interest. The results of a user evaluation, which aimed at determining how the technique would be used in the exploration and analysis of real experimental data sets is discussed and we assess the Time-series Explorer's advantages over other existing techniques.

Time-series Explorer

A summary of the Time-series Explorer is presented in Figure 4. The technique uses two coordinated views of the data: a graph and a scatter-plot. The graph view overlays value versus time representations of the recorded activity of all genes and allows the user to specify an interval (p in Figure 4). The scatter-plot summarizes the data within the selected interval by representing each gene as a single point with its translation along the Y-axis

corresponding to its activity over the selected interval and its translation along the X-axis corresponding to its change-in-activity from the start to the end of the interval. As the graph view controls are manipulated and the selected interval is adjusted, the position of genes in the scatter-plot are recalculated to adjust for the change in temporal context. Repeated adjustments of the selected interval (where the start and end times of the selected interval are incremented independently or in parallel) cause the position of genes in the scatter-plot to be shifted with the resulting animation allowing the user to perceive patterns of gene activity over time.

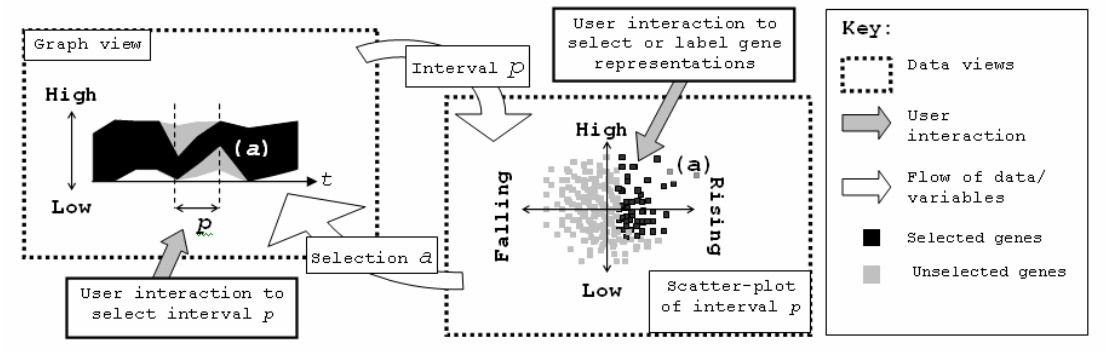

Figure 4 Summary of the Time-Series Explorer.

Figure 5 illustrates specifically how the positions of gene representations in the scatter-plot are calculated by describing how the activity of a single gene over an arbitrary interval (P) is used to generate its axes coordinates. Here, the Y-axis translation of a gene summarizes its average activity over all time points of the selected interval. This average is calculated as the area under the gene's rescaled time-series enclosed by the bounds of the selected interval divided by the number of time points enclosed. This ensures that as the interval selection is shifted by increments less that the space between two adjacent time-points, the Y-axes translation of a gene can be recalculated so that that gene's representation in the scatter-plot moves gradually along the Y-axis. The translation of a gene representation along the X-axis is calculated as its activity at the end of the selected interval divided by its activity at the start to give a measure of the relative change in activity. In this case when the interval is shifted by small increments, values are recalculated using linear interpolation so that gene representations shift gradually along the X-axis as well as the Y-axis.

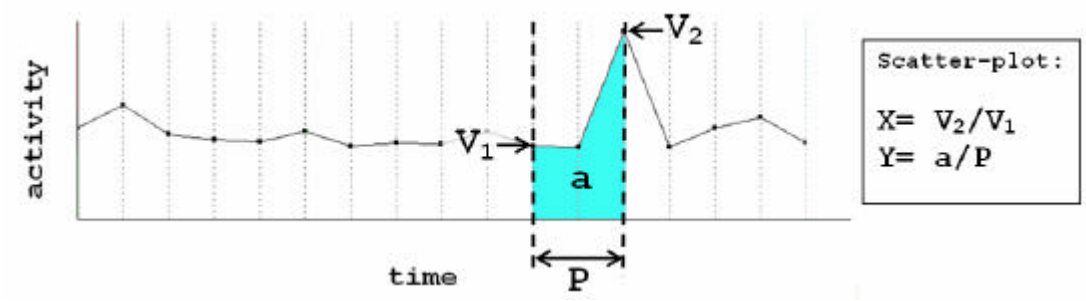

Figure 5 Attributes of a gene's time-series over an interval (P) used to determine the gene's scatter-plot coordinates.

Other scatter-plot layouts considered for the Time-series Explorer were a multi-dimensional anchor layout [15] and a plot of the activity at the start of the selected interval against the activity at the end. While both of these layouts produced interesting results with smaller scale time-course data including the recorded expression of around 200 genes, neither could cope satisfactorily with larger scale data. The layout algorithm of the multi-dimensional approach was too complex to allow for a satisfactory animation frame rate and tended to overlay genes with diverse patterns of activity when larger intervals were selected. When the plot of the activity at the start time-point against the activity at the end time-point was animated with larger numbers of genes, more complex patterns of activity were difficult to perceive due to difficulties in associating the same genes with multiple trends in activity as the animated paths of genes with outlying activity tended to cross dense clusters of inactive genes.

Design Rationale

The design of the Time-series Explorer is based on two primary assumptions. The first of these is that an animated representation of the microarray time-course data will be able to reveal more of its detail and, therefore, more of the less dominant patterns of the type that are not already revealed by existing techniques. The second assumption is that, as the data is temporal, it is appropriate to present such an animation across time so that the user will be able to relate the visualisation to the data with changes over time in the data represented as changes over time in the visualisation. As a quantity of the patterns our users wishes to find are less dominant and not determined by any pre-knowledge of the data, it was undesirable to irrevocably filter any genes from the data-set to be visualised. This made it necessary to make the representation of genes in an animation of the data compact. As the most compact distinct visual entities are single-points and these combine to form a scatter-plot, a scatter-plot type display was chosen to represent interval gene activity in the Time-series Explorer visualisation.

The potential disadvantages in using such an animated scatter-plot to analyse this type of data are:

1) The increased delay in seeing the data as it is animated.
2) The transient nature of pattern perception.
3) The inability to compare data at multiple time points simultaneously.
4) The large degree of variation in recorded expression for genes between time points which could make it difficult to track genes between frames.
5) The time taken to become familiar with a new alternative view of the data.

These are accounted for in a number of different ways.

Firstly, to reduce the delay in seeing the data as it is animated, the user is given tight control over the direction and pace of the animation so that they can animate slowly over intervals where interesting patterns appear and

quickly over the remainder of the time course. In this regard the Time-series Explorer can be though of as a kind of genomic video cassette player where the user can play, fast forward, rewind, slow motion, pause and stop the animation of gene activity to examine, and re-examine, the more interesting intervals of the data.

Tight control over the direction and pace of the animation also makes the second listed potential disadvantage of our animated display, the transient nature of pattern perception, somewhat less of a problem. If the animation played at a fixed rate from start to finish then it would be possible for the user to first see a pattern then, waiting for the animation to finish and viewing other patterns in the data, forget what they saw. With a high level of control over the animation, as soon as a pattern is seen the animation can be stopped, rewound and the pattern can be viewed again or the user can select and store the relevant genes' names for further reference.

The third potential disadvantage of an animated time-course scatter-plot is the inability to compare data at multiple time points simultaneously. While a comparison of all time-points simultaneously (for the activity of all genes) would be impossible for the biologist to digest, there are certain situations where it is of particular value to compare a smaller number of time-points. Specifically, these are situations where recorded activity at a smaller number of adjacent time points define a pattern which occurs over a limited interval of the experiment's time frame. To better facilitate the finding of such patterns we have designed the Time-series Explorer so that each frame of the animation represents an interval of the time-course rather than an instantaneous time-point. While the interval selection cannot comprehensively summarise the information contained in an interval containing any more than two time-points it is assumed that comparisons between multiples of two time points when such a view is animated is enough for biologists to initially perceive the majority of interesting patterns. Once such an animation is stopped, the relevant genes can be selected based on their activity over the selected interval and comparisons with other time points can be made either by further animations or, if there is a sufficiently small number of selected genes, the linked expression versus time graph view.

Next, the course granularity of the data is potentially problematic in that the expression of genes can vary dramatically between time points. This means that if an animated scatter-plot of the data were only to include frames relating to intervals starting and ending at time-points at which expression is recorded, the single point representations of genes would shift dramatically between frames and it would be impossible to track genes between time points to identify patterns of activity over more than two time points. To account for this potential pitfall, the expression of genes is interpolated between the time points in the data and the interval selection is incremented by quantities independent of the space between time-points for which expression is recorded. This allows for an animation where the motion of gene representations is smooth and they can be tracked by the user to reveal more sophisticated patterns. To prevent the user inferring undue significance from interpolated values, the start and end of the interval selection automatically

move to the nearest time-points for which expression is recorded when the user is not adjusting the interval selection. In effect the interval selection scale is only continuous when the selected interval is in motion, so that genes can be tracked to discern more complex patterns of activity, and discrete when the selected interval is fixed for the user to interact with the static scatter-plot by making selections and forming queries.

The fifth and final potential weakness of an animated time-course scatter-plot is the time taken to familiarise with a new alternative view of the data. While biologists are generally familiar with scatter-plot representations of their data, two specific aspects of the Time-series Explorers design reduce the time they take to become familiar with its own particular scatter-plot layout. Firstly, the activity Y-axis of the scatter-plot is placed parallel to the expression axis of a coordinated graph view. These axes can be thought of representing complimentary notions of high or low activity. The graph representation is already familiar to the biologists and the parallel coordination of high/low axes encourages familiarity in the scatter-plot by association. As the points in the scatter-plot move at the same time as the time selected interval overlay on the graph view, the user will associate the scatter-plot with the selected interval and familiarise themselves with the Y-axis of the scatter-plot layout. The change-in-activity X-axis layout also becomes apparent when the time-interval is moved forward. Here genes with rising activity over time have their representations rise in the scatter-plot and are shifted to the right. Genes with falling activity over time have their representations fall in the scatter-plot and are shifted to the left. The association is simple – right rising, left falling.

Rescaling and Distortion

In order for the technique to provide an animated overview of the data from which previously unsuspected patterns of temporal activity can be revealed, it is necessary to ensure that a biologist can not only relate quickly between successive frames of an animated view but also interpret individual frames to quickly detect the most relevant information. This means that the most relevant aspects of the data must be predominant with the spread of data representations appropriate to communicate the extent of outliers and realistically portray general trends.

In order to effectively communicate biological activity from microarray data it is necessary to account for the fact that any changes in a gene's expression are proportional to its current level of expression and that the activity of a gene is indicated by relative changes in expression (either between time-points or from some base level) rather than absolute values. While relative changes can be derived from absolute values, the fact that changes in expression are proportional to expression levels causes the data to have log-normal skewed distribution with a large number of low values and a small number of high outliers. As low values with low changes may hold valuable biological information it is necessary to adjust the spread of the data in any visual representation so that the significance of these features is appropriately communicated.

The majority of existing techniques (for example [8; 9; 10; 13; 14; 17]) account for this characteristic by rescaling the data using a log-transform [18; 19]. This makes relative changes comparable across genes and the distribution of values close to normal distribution [18]. Eq. (1) describes the transform as it is applied to each gene's time-series where *V* is the original time-series of a gene, *LS* is the rescaled time-series and *base(V)* is a derived base value (for example, the mean value of expression for that gene or the expression of the gene in a control experiment).

$$LS(t) \leftarrow \log_2(V(t)/base(V)) \tag{1}$$

The disadvantage of this form of rescaling is that as *v(t)* tends to zero *LS(t)* tends to negative infinity. This means that the rescaled data cannot be completely displayed using a regular finite scale, such as the axes of a graph or scatter-plot as required by the Time-series explorer, without overemphasizing small changes in very small values.

As an alternative to logarithmic rescaling we use a more basic linear rescaling and distort the axes onto which the values are represented. While linear rescaling makes values comparable across time-series, distortion improves their distribution in the display and represses the dominance of large outlying values. Eq. (2) describes the linear rescaling as it is applied to each gene's time-series with *V* the original time-series of a gene, *median(V)* the median of all values for that gene and *MS* the rescaled time-series.

$$MS(t) \leftarrow V(t)/median(V) \tag{2}$$

The median [20] of a time-series can be considered as a statistical measure of what can be considered as a *normal* value accounting for a skewed distribution of values (this being the primary advantage of using median rescaling over other popular transformations such as z-scores and mean rescaling that would not account for this aspect of the data). Each value in the rescaled time-series of a gene is equal to the proportion of its corresponding value in the original time-series to the median of all its values. This means that, in the rescaled time-series, anything below 1 is below normal activity and anything above 1 is above normal activity.

Eq. (3) describes the distortion applied to the graph view Y-axis for rescaled values. Here *V* is a value, *Ydisp* the position for plotting the value on the axis and C_1, C_2 and C_3 are derived so that the maximum and minimum values are at the top and bottom of the allowed display space with *V*=1 (the *normal* value) at its mid-point.

$$Ydisp \leftarrow \log_2(V \times C_1 + 1) \times C_2 + C_3 \tag{3}$$

An advantage of combining linear rescaling and logarithmic distortion in this manner is that when two values for a gene are divided to calculate the relative change in value between time-points the factors used to rescale the data (both being the gene's median value) cancel out. This means that the rescaled data can be used to calculate values for the X-axes of the scatter-plot which is used to display relative change. The log transform distortion simply improves the distribution of the values in the display and does not impact on any further calculations. This means that it can be adjusted (specifically by adding one to the argument of the logarithm) to prevent small values dominating the display.

In the Time-series explorer scatter-plot each gene is represented as a single point with its translation along the Y-axis corresponding to its activity over the selected interval and its translation along the X-axis corresponding to its change in activity from the first time point to the second. As the distribution and range of values along the Y-axis of the scatter-plot is roughly equivalent to the distribution of values in the rescaled data, in order to make the distribution normal it is appropriate to use the same transform as that used for the Y-axis of the graph view (3). While the distribution of values along the X-axis is also similar to the distribution of values in the rescaled data, the same transform cannot be used. This is because when V_1 tends to zero *Xdisp* will tend to infinity and logarithmic transforms cannot translate values tending to infinity onto a finite range. Instead, an alternative distortion transform was constructed using the hyperbolic tangent function. This function is similar to the logarithmic function with the notable exception that as a number tends to infinity its hyperbolic tangent tends to one. The transform as applied is described in Eq. (4) where X is the derived value of change in expression over the selected interval, *Xdisp* is the position for plotting on the axis and C_1, C_2 and C_3 are derived so that the value for no change (X=1) is in the centre of the display space and the values for biologically significant halving or doubling of expression (X=0.5 and X=2) are one and three quarters along the display space.

$$Xdisp \leftarrow TanH(X \times C_1 + 1) \times C_2 + C_3 \tag{4}$$

The resultant spread of data in the scatter-plot view is illustrated in figure 6. It can be seen that from this representation that it is possible to perceive the activity of genes with outlying high, low, falling and rising activity over the selected interval. If the data were not rescaled or distorted, the majority of gene representations would cluster around the bottom left hand side of the plot and it would be impossible to interpret anything other than the activity of a few genes with high outlying activity over the selected interval.

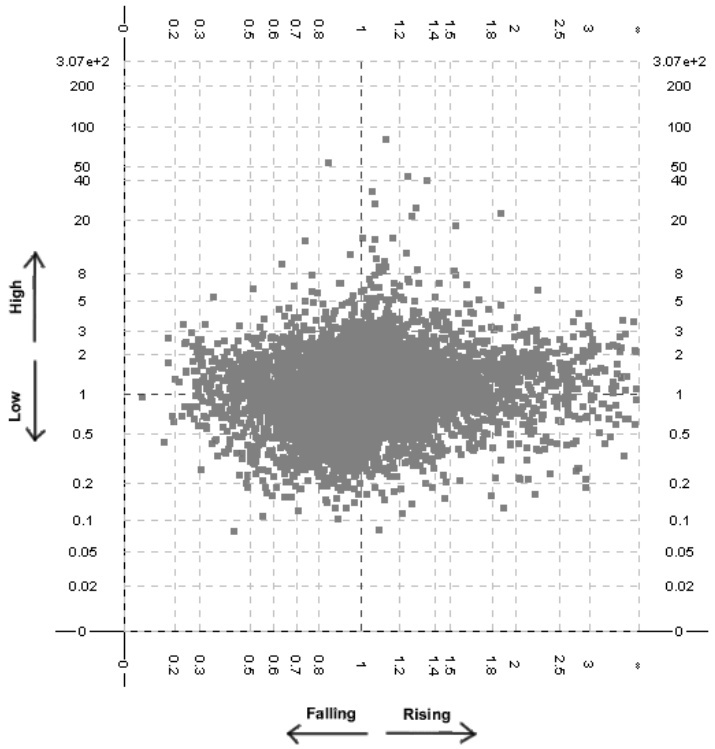

Figure 6 The spread of data in the distorted scatter-plot view.

Colour mapping

While rescaling and distortion improve the spread of the data in an interval scatter-plot view allowing for better detection of outlying patterns of temporal activity, the sheer volume of gene representations creates a general grey mass of gene representations in the centre of the scatter-plot that makes it impossible to perceive the majority of more general trends (Figure 6). While a transparency composite [21; 22] would allow a user to more accurately perceive general trends by indicating the relative density of overlaid gene representations, it would also make it harder to distinguish outliers from the background as smaller numbers of overlaid elements are represented with their colour closer to that of the background. As an alternative, we have developed a colour mapping composite that communicates the density of genes through a colour-scale where outliers are significantly different in colour from the background and, therefore, easy to distinguish from the background.

Our colour composite is similar to a standard transparency composite in that each pixel of the display has an alpha value that is increased by the alpha value of overlaid elements. The essential difference in our approach is that alpha values are translated into a colour scale rather than used to combine the colours of overlaid elements. The scale used (described in Figure 7a) ranges from dark-blue, for small numbers of overlaid genes, through blue, cyan and green to yellow for larger numbers of overlaid gene representations. This attempts to utilize as much of the visible spectrum as possible without using reds or greys, which are more appropriately used to represent

highlighted and deselected genes. The ordering of colours is such that light colours represent a high density of genes and dark colours represent a low density of genes. This ensures that dark colours surround light colours, which would be otherwise hard to distinguish from the background. Colour-coded graph and scatter-plot views are shown in Figures 7b and 7c.

Interaction

The interaction mechanisms of the Time-series Explorer are best described with reference to a screen-shot of its interface (Figure 7d). This contains five main panels with which the user can interact in order to manipulate the representations of their data. These are the toolbar, graph view, scatter-plot, gene list and grouping panel.

The toolbar (i. in Figure 7d) contains 17 buttons in five groups with various different functions such as animating the scatter-plot view, changing the selection mode on the scatter-plot and viewing details for a selection (see Table 1).

Table 1 Functionality of the Time-series Explorer toolbar.

Group	Name	Action
Animation	Play	Animates the scatter-plot by increasing the start and end times of the selected interval at regular intervals of time.
	Pause	Pauses the animation.
	Stop	Stops the animation.
Selection	Select all	Selects all genes.
	Select none	Deselects all genes.
Scatter-plot tools	Freehand selection	When selected allows the user to select genes by dragging a freehand shape around their representations in the scatter-plot.
	Box selection	When selected allows the user to select genes by dragging a box around their representations in the scatter-plot.
	Zoom tool	When selected left clicking on the scatter-plot zooms in, right clicking zooms out and double right clicking zooms out fully.
	Labelling tool	When selected moving the mouse over gene representations in the scatter-plot causes them to be labelled and have their expression patterns highlighted (over the entire time-course) in the graph view.
	Excentric labelling tool	Similar function to that of the labelling tool (above) with the exception that all gene representations within the bounds of a circle are labelled. Right clicking increases the size of the circle and left-clicking decreases its size.
Selection mode	Replace selection mode	Successive selections replace the previous selection.
	Refine selection mode	Successive selections refine the previous selection (equivalent to combining the two results using a logical AND operation).

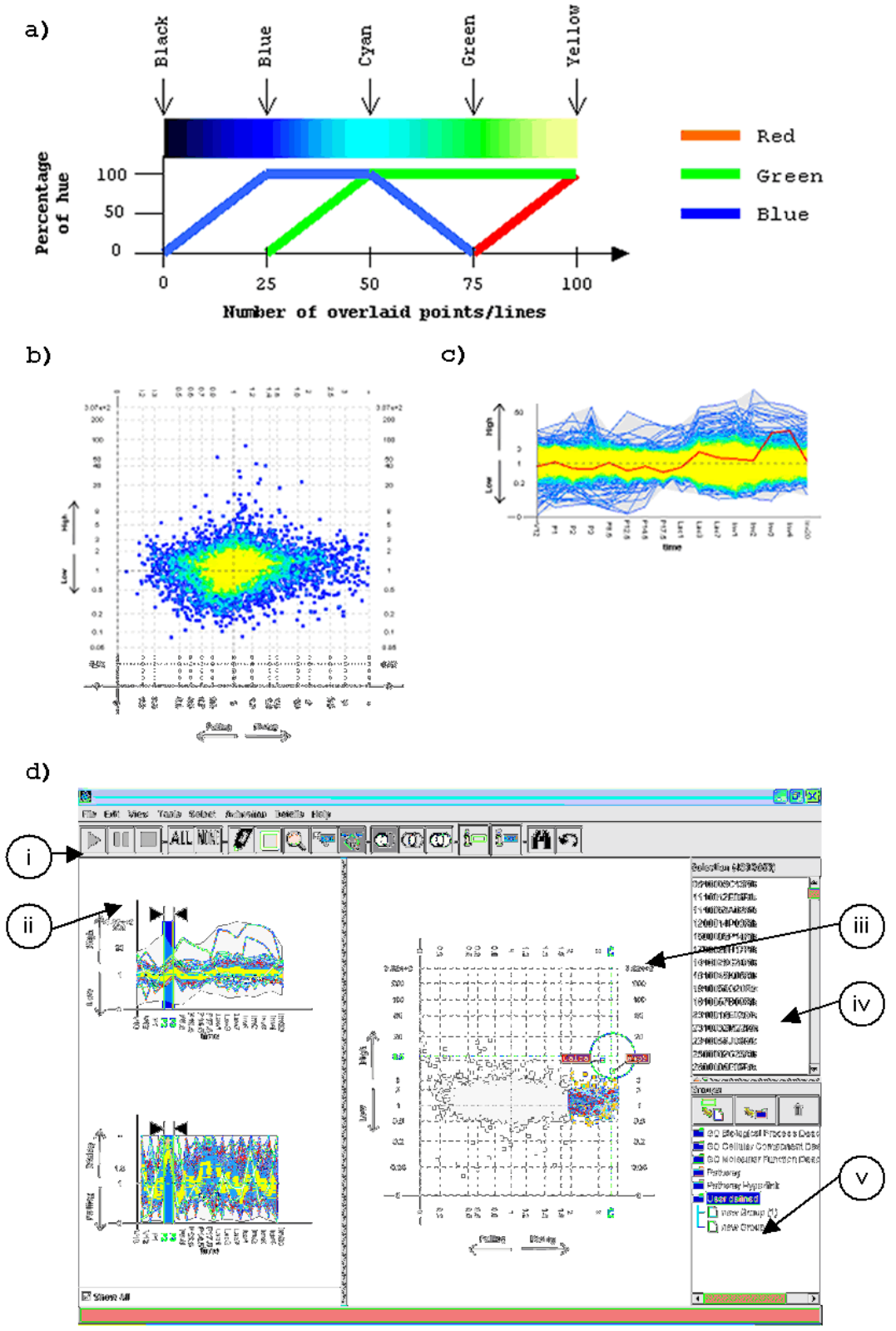

Figure 7 a) The colour mapping used for the Time-series Explorer with the percentage of each hue corresponding to the number of overlaid points or crossing lines, b) the scale applied to the scatter-plot view, c) the scale applied to the graph view and d) a screen-shot of the Time-series Explorer interface (i. toolbar, ii. graph view, iii. scatter-plot, iv. selected gene list and v. grouping panel).

	Add-to selection mode	Successive selections add to the previous selection (equivalent to combining the two results using a logical OR operation).
Details	Selected gene details	Activates a pop-up window with details-on-demand for the selected genes including a cross-reference with other groupings and a list of the genes selected with reference to groupings.
	Labelled gene details	Activates a pop-up window with details-on-demand for the labelled genes including a list of the groupings to which the gene belongs and it's original recorded expression values.
Miscellaneous	Find gene	Activates a pop-up window that allows the user to find a gene by typing its name.
	Undo	Undoes the previous selection.

The graph view (ii. in Figure 7d) allows users to adjust the selected interval to focus [23] in on a specific interval or animate the interval scatter-plot to reveal general trends and outliers across the time-course. The interaction mechanism of the graph view is essentially identical to that of a multi-range dynamic query slider [24] utilizing the internal slider space for a visual representation of data in a manner similar to that of data-visualization sliders [25]. Dragging the edges of a vertical bar overlaid onto the view to represent the selected interval allows the user to adjust its start and end times independently. Dragging the centre of the bar changes the start and end times with the duration remaining constant to shift the selected interval. During this interaction if the selected interval is shifted in the positive direction from earlier to later time-points changes across time in the animated scatter-plot convey changes across time in the data.

During an animation across time, which can be initiated either by interacting with the graph view or using the play button in the toolbar for an animation with a regular frame rate, genes with outlying high, low, rising or falling interval activity remain on the periphery of the scatter-plot and move smoothly in a predictable anticlockwise rotation around the axes origin. This effect is best explained by relating the axes of the scatter-plot to cardinal points of a compass (Figure 8). Genes with low expression have rising before high expression moving from south to east to north and genes with high expression have falling before low expression moving from north to west to south. The rescaling of the gene expression (described above) ensures that genes with outlying interval activity kept to the periphery of the scatter-plot and persistent outlying (high, low, rising or falling) interval activity will move gene representations around the axes origin crossing points of the compass somewhere in the sequential order of south, east, north, west, south…. (i.e. anticlockwise) for as long as their interval activity is distinct from that of the other genes. Conversely, if the animation of scatter-plot is reversed by shifting the selected interval backwards gene representations will move clockwise. Whichever the direction of animation, this uniformity of rotational direction will make it easier to interpret more complex patterns of activity. This is because with uniformity of rotational direction the representations of genes are less likely to have crossing paths and more likely to remain distinct during the animation which, in turn, reduces the ambiguity in relating the representation of genes from one selected interval to another and allows genes to be tracked

across time with multiple trends in activity related to the same genes or gene groupings.

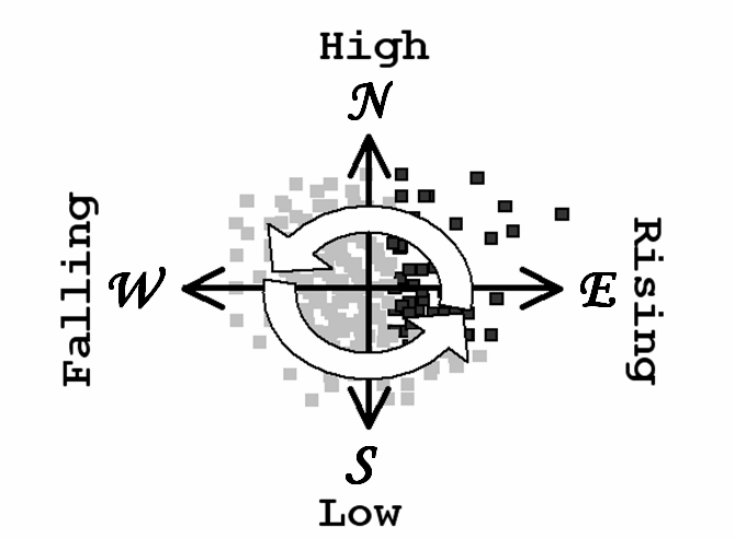

Figure 8 General movement of gene representations as the scatter-plot is animated forward through time: genes rotate anticlockwise through low ? rising ? high ? falling ? low

To complement the benefits of uniform rotational, direction tight control of selected interval manipulation using the graph view gives users control over the pace of the animation. This allows them to slow down as interesting features become apparent, reverse the animation when they want to look at something again and stop the animation, 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.

Once the user ceases interacting with the graph view there are a number of different options for interacting with the scatter-plot (iii. in Figure 7d). The majority of these interactions employ standard brushing and linking [23] information visualisation operations. If the labelling tool is activated from the interface toolbar, moving the mouse over gene representations in the scatter-plot view causes them to be labelled and have their activity over the entire time-course highlighted in the graph view. The functioning of the excentric labelling tool (adapted from [26]) is similar to that of the labelling tool with the exception that all gene representations within the bounds of a visible circle are labelled and highlighted. The additional information revealed by labelling and the subsequent coordination between scatter-plot and graph views allows the user to rapidly perform a more informed assessment of a pattern's significance. If the user is interested in a smaller number of genes or wishes to investigate a sample of selected genes in more detail, double-clicking on gene representations in the scatter-plot allows them to view a pop-up details-on-demand [11] window describing the un-scaled recorded intensities for the subject gene and a summary of the groupings to which the gene belongs. This, again, will lead to a more informed assessment of a pattern's significance.

As an alternative to labelling, when the freehand or box selection tools in the toolbar are activated genes can be selected. 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 representations of un-selected genes are greyed out in both the graph and scatter-plot views allowing users to focus in on selections which are colour-coded, labelled, animated and selected again (using logical AND or OR rules) independent of the un-selected data. This allows the user to find groupings within groupings and combine selections to uncover or investigate more complex patterns in the data.

The operation of the grouping panel (v. in Figure 7d) is similar to that of the Microsoft Windows file explorer tree-pane. Imported groupings are stored within folders that correspond to grouping categories. Clicking on folders causes their contents to be expanded or collapsed and clicking on a grouping name causes the genes which belong to that grouping to be selected. Buttons on the grouping panel mini-toolbar allow new grouping categories to be added, grouping categories to be deleted, stored groupings to be deleted or new groupings to be generated from the genes that are currently selected.

Evaluation

The aim of our user evaluation was to assess the practical utility of the Time-series Explorer within the wider context of microarray time-series analysis as a whole. In particular we wanted to assess the extent to which Time-series Explorer could be considered as a specialised microarray time-course exploration tool which complements and adds to the functionality of existing clustering techniques. To do this we needed to discover if the technique was capable of overcoming the limitations of existing techniques to reveal previously unsuspected patterns of temporal activity and, to a lesser degree, assess its ability to uncover certain suspected patterns of temporal activity in order to highlight the areas where it may be advantageous for a biologist to use the Time-series Explorer in preference to other more established techniques.

To achieve the objectives of our analysis we did *not* require to evaluate the technique exhaustively with regard to patterns that are found with other tools (for a more comprehensive evaluation of microarray data analysis tools see [27]) but rather assess the potential advantages of this technique over the other techniques considered and find the areas where the functionality of the techniques overlapped. This reflects the fact that the technique was specifically developed to support the analysis of time-course and an assessment of its ability to analyse other types of microarray data would be irrelevant. Given the current preference of biologists' to switch between different techniques for the analysis of their data and the limitations of existing techniques, we believe this to be an appropriate tack for our evaluation especially since any replication of the core functionality of existing techniques would not necessarily promote the use of a new technique as users of

microarray analysis software will inevitably have an ingrained preference for more familiar techniques and applications.

With reference to the objectives stated, the questions that we posed in our evaluation were:

1) Is the Time-series Explorer capable of allowing the biologist to find previously unsuspected patterns of temporal activity?
2) Which of the patterns found can be revealed using other existing techniques, or combinations thereof, and what are the patterns that can only be found using the Time-series Explorer?
3) Of the patterns that can be found using other techniques what are the advantages, if any, of using the Time-series Explorer?
4) Are the patterns that can be found by the Time-series Explorer of sufficient significance to justify its use?

In order to answer some of these questions it was necessary for participants in our evaluation to make discoveries and find unexpected patterns. As these patterns are, by their very nature, more difficult to find, it was necessary to minimise any factors that could detract from the natural processes of exploration that would lead to their discovery. It was, therefore, inappropriate to restrict the participants by instructing them to follow pre-defined tasks or operate in an alien environment while trying to find these patterns. Instead, the main active session of our evaluation procedure was to be relatively informal. Operating in their normal workspace the users were encouraged to operate the tool in a manner that was appropriate to their own working objectives with minimal disruption.

The evaluation proceeded in three stages. The first of these was a training session involving a short tutorial guiding the users through the basic functionality of the technique and allowing them to become familiar with its interaction mechanisms and data representations. This was followed by a session where the users were asked to explore their data in order to find new patterns as per their normal working procedures. Lasting approximately one hour, interactions were recorded and users were encouraged to think-aloud so that patterns revealed could be identified for further analysis in the third and final stage of the evaluation where they where asked to compare the results obtained using the Time-series Explorer with those obtained using other techniques.

In order to present a coherent case study, the results presented in this paper describe the final run of our evaluation which involved an experienced biologist (who was independent from the development of the technique) analyzing familiar data from experiments that he himself had designed. While the need for experienced biologists analysing their own data during this run of the evaluation severely limited our pool of potential participants (and this final run of our evaluation procedure was to involve only one biologist), it has been shown that domain experts are significantly more motivated to find patterns in data of this type [27]. Moreover, it is only specific domain experts that are capable of assessing the relevant biological significance of the patterns found in microarray data or indeed finding any number of patterns with any

substantial biological significance. This became clear after a few preliminary evaluation sessions with biologists working on data from unfamiliar experiments and domain novices. While these sessions were able to provide feedback on the Time-series Explorer tool's usability, they told us relatively little about its core functionality. The biologist participating in the final stage of our evaluation also had extensive experience of analyzing their data using a range of established software tools such as Time-Searcher [13], Hierarchical Clustering Explorer [14] and GeneSpring [17]. As these tools implement a range of existing clustering and visual query techniques and have been evaluated as being most effective at revealing patterns from microarray time-course data [27], this qualified the biologist to properly assess the relative advantages, and disadvantages, of using the Time-series Explorer as an alternative.

This data under analysis in the evaluation recorded the expression of around 8,500 genes over 17 time points belonging to 4 successive stages of mouse development: virgin (days 10 and 12), pregnancy (days 1, 2, 3, 8.5, 12.5, 14.5 and 17.5), lactation (days 1, 3 and 7) and involution (days 1, 2, 3, 4 and 20) [4].

The first two patterns found by the biologist in our case study involved genes with a high level of activity at the early stages of lactation and genes with activity rising at the start and falling at the end of lactation. While these two patterns where found using queries formulated using pre-knowledge of the data and could to a large extent already be found using visual query type techniques, the biologist expressed a preference for the Time-series Explorer method to find these patterns due to the fact that they where able to distinguish the activity patterns of individual genes from the background according to aspects of their activity over the selected interval. This additional functionality prompted them to adjust their original query to select what they felt to be natural groupings of genes rather than groupings defined by some arbitrary cut off. While visual queries can also be incrementally adjusted and a user could find natural groupings by observing the number of genes that fall in or out of the results after each increment, the user indicated that the Time-series Explorer method required less interaction and the visual indicators where more natural.

The next pattern found was a combination of general trends in activity for all genes over the entire time-course. Here the biologist selected an interval fixed at its minimum value (an interval constrained by two time-points for which expression is recorded) and shifted it across the entire time frame of the experiment to animate the scatter-plot view. Selected frames of this animation are illustrated in Figure 9. At various stages of the 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 expression (top of Figure 9). 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 expression 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 (bottom of Figure 9). It was later determined, by re-examining the results of clustering already applied to the data (SOM and hierarchical clustering) and predicting the likely results of other forms of clustering based on the biologist's knowledge of clustering algorithms in general, that it would be unlikely for these particular trends to be revealed by established clustering techniques.

The final, and most significant, pattern found in the evaluation was discovered, in part, when investigating general trends across the entire time-period. 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 expression is recorded) an outlying group of gene representations showed significant rising then falling expression. To investigate this further the relevant interval was animated again, then stopped so that the outlying genes could be labelled by moving the mouse over their representations in the scatter-plot. This revealed the majority of the genes also shared low expression 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 10 illustrates this pattern showing selected frames of the initial 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.

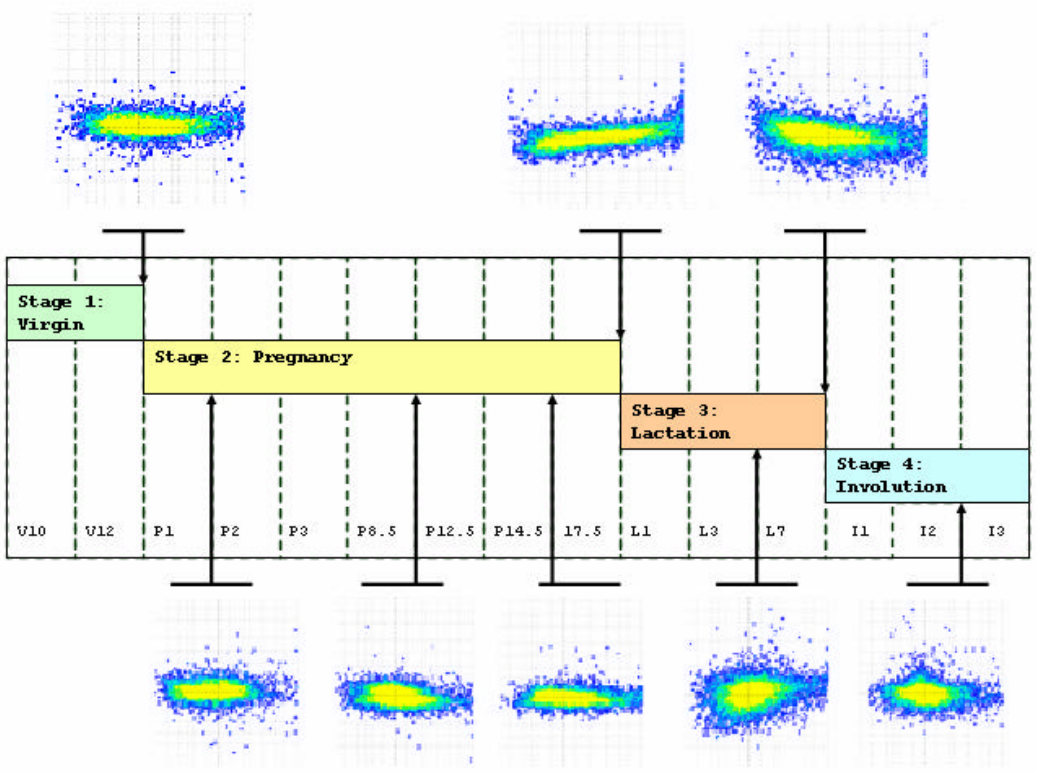

Figure 9 Selected frames of an animation across the entire time-course (time proceeding left to right with stages of development indicated using a Gantt chart): Horizontally elongated scatter-plots indicate large numbers of genes with significant changes in their level of expression. This occurs primarily during transitions between stages of development (top) although more subtle trends, such as the increased number of genes with changes in activity during pregnancy in relation to lactation (bottom), can also be interpreted from the animation.

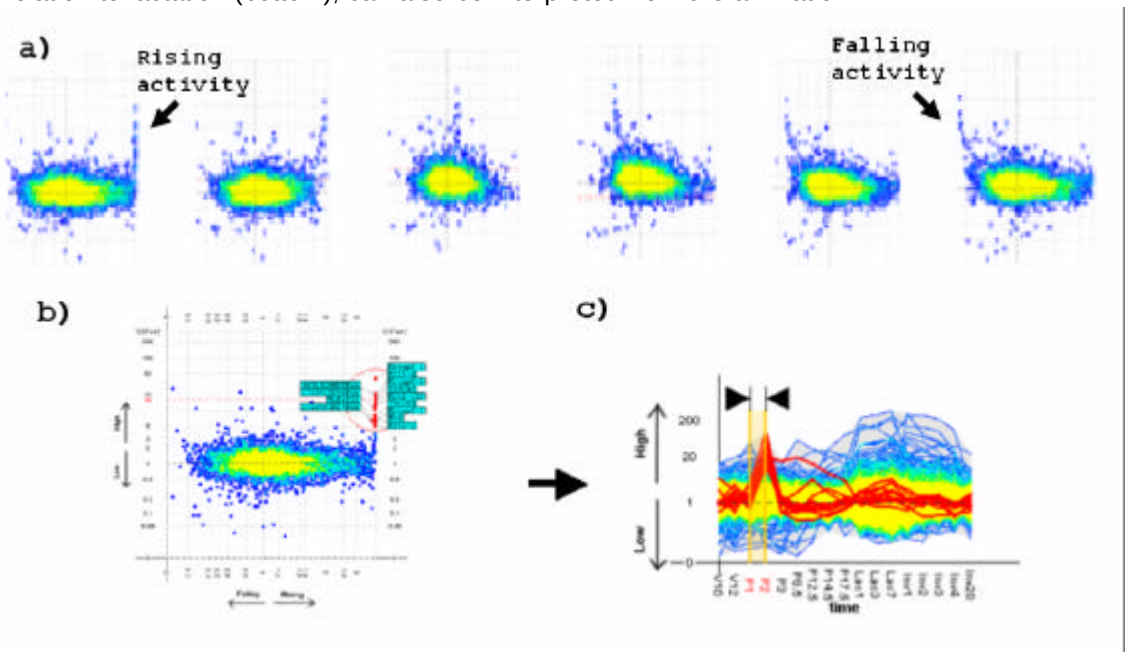

Figure 10 An unexpected pattern of temporal activity: a) Animating the scatter-plot reveals a group of outlying genes with rising then falling activity over a small interval of the time-course, b) moving the mouse over the gene representations in the scatter-plot view allows them to be labelled and c) have their expression patterns over the entire time-course highlighted in the graph view.

The final stage of the evaluation was a follow up meeting where the biologist was asked to assess the degree of relevant biological significance to which each of the patterns revealed could be attributed and identify the extent to which the patterns could be uncovered using other techniques. These results were combined to produce a summary (Table 2) describing the specific areas and extent to which the utility of the technique contributes to the support of microarray time-course analysis. Here the patterns in the data were categorized, described as suspected or unsuspected, rated from one to five according to the extent to which existing techniques are already capable of uncovering them and assigned a measure of biological significance. Here a suspected pattern is defined as one for which the biologist has had some pre-knowledge of the genes which contribute to the pattern or the interval of time over which significant changes in activity relating to the pattern occur prior to its discovery. As the suspected/unsuspected nature of a pattern is in many cases integral to the investigation or discovery of that pattern, all other fields of the table relate to patterns in their listed suspected or unsuspected form. The ratings that describe the extent to which existing techniques are already capable of uncovering patterns are based on previous analysis of the data using the techniques indicated and, where appropriate, supplementary analysis involving the listed techniques. When the subject was unsure of whether or not a pattern could be found using another technique we re-applied that technique to try and assess whether or not the pattern could indeed be found using that technique. The ratings are; 0 (the pattern absolutely cannot be found using this technique), 1 (the pattern will be found but the results will be less satisfactory), 2 (the pattern can be found but with significant difficulty) and 3 (the pattern can be found without any problem). The measures of biological significance indicated are: high (biological significance relevant to the specific objectives of the experiment), medium (biologically significant but not relevant) or low (not significant).

Table 2 Results of the user evaluation

Patterns found using the Time-series Explorer	**Type**	**Suspected**	**Can be found using alternative technique (0 - 3)**		**Significance**
			Clustering	Visual queries	
Genes associated with Milk Proteins with very high expression from L1 to L3.	Outliers over an interval.	✔	0	2.5	Medium
309 genes belonging to various interesting groupings with expression rising at L1 and falling L7.	General trend over an interval.	✔	1	2	High
Large changes in gene activity during known transitional phases.	General trends over the entire time-course.	✔	3	0	Medium
Increased number of genes with changes in activity during pregnancy in relation to lactation.	General trends over the entire time-course.	✘	0	0	Medium
Keratin associated genes with expression rising sharply at P2 and falling sharply at P3.	Outliers over an interval.	✘	0	0	High

The main outcome of our user evaluation was to verify that the Time-series Explorer is uniquely capable of revealing certain 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 their data (positively answering questions 1, 2 and 4 posed by our evaluation). Moreover, the technique also proved capable of revealing suspected patterns of temporal activity and the evaluation uncovered significant advantages in using the Time-series Explorer over other more established techniques (positively answering questions 3 and 4). Specifically, when the technique was used to uncover general trends occurring over limited periods of the time-course the user had the advantage (over clustering techniques that would also allow biologists to find such patterns) of being able to quickly identify interesting sub-groupings, when identifying suspected outliers over smaller intervals the technique offered the biologists the ability to perceive distinct groupings of outliers and when looking for general trends across the entire time-course the biologists found it easier to assess more subtle patterns of general activity.

Conclusion

We have developed a novel technique for the analysis of microarray time-course data. This technique specifically focuses on allowing biologists to reveal previously unsuspected patterns of gene activity over smaller intervals of an experiments time frame by allowing them to control an animated interval scatter-plot view of their data. This alternative representation of the data is supported by the combination of multiple new, and existing, information visualisation techniques. Most notably we have introduced a unique combination of linear rescaling, distortions and colour coding to improve the display of data in our linked graph and scatter-plot views. This ensures that the animation is smooth and that biologists can perceive outlying patterns and general trends of temporal activity. An evaluation, involving biologists working with real data, tested the extent of the tools desired functionality and assessed the technique's practical utility within the wider context of microarray time-series analysis as a whole. This proved the technique not only capable of revealing previously unsuspected temporal patterns but also, in certain cases, more appropriate for finding previously suspected patterns and patterns that occurred over the majority of the time-frame.

References

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

2. Schena M, et al. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc. Natl. Acad. Sci. U.S.A. 93,* 1996: 10614-10619.

3. Duggan DJ, et al. Expression profiling using cDNA microarrays. *Nature Genetics* 1999; **21**: 10-14.

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): R75 - R91.

5. Hughes TR, et al. Functional discovery via a compendium of expression profiles. *Cell* 2002(102): 109- 126.

6. Wen X, et al. Large-Scale Temporal Gene Expression Mapping of CNS Development. *Proc Natl Acad Sci USA* 1998; **95**: 334-339.

7. Brown PO and Botstein D. Exploring the new world of the genome with DNA Microarrays. *Nature Genetics* 1999; **21**: 33-37.

8. Eisen MB, et al. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* 1998; **95**: 14863-14868.

9. 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**: 2907-2912.

10. Raychaudhuri S, Stuart JM, and Altman RB. Principal Components Analysis to Summarize Microarray Experiments: Application to Sporulation Time Series. *Pacific Symposium on Biocomputing* 2000; 452-463.

11. Shneiderman B. The eyes have it: A task by data type taxonomy for information visualizations. *Proceedings IEEE Visual Languages* 1996: 336-343.

12. Segal E, et al. Rich probabilistic models for gene expression. *Bioinformatics* 2001; **17**: 243-52.

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

14. Seo J and Shneiderman B. Interactively Exploring Hierarchical Clustering Results. *IEEE Computer* 2002; **35**: 80-86.

15. Craig P, Kennedy JB, and Cumming A. Towards Visualising Temporal Features in Large Scale Microarray Time-series Data. *6th International*

Conference on Information Visualisation - IV2002 2002 (University of London, London, GB), IEEE Press.

16. Craig P and Kennedy JB. Coordinated Graph and Scatter-Plot Views for the Visual Exploration of Microarray Time-Series Data. *I2003 IEEE Symposium on Information Visualization* 2003 (Seattle WA), IEEE Computer Society.

17. *GeneSpring*. 2004, Silicon Genetics www.silicongenetics.com.

18. Nadon R and Shoemaker J. Statistical issues with microarrays: processing and analysis*. TRENDS in Genetics* 2002(18): 265-271.

19. Quackenbush J, *Computational Analysis of Microarray Data*, in *Nature Reviews* . 2001. p. 418- 427.

20. Weisstein EW, *"Statistical Median." From MathWorld--A Wolfram Web Resource. http://mathworld.wolfram.com/StatisticalMedian.html*. 2005.

21. Fekete J and Plaisant C. Interactive Information Visualization of a Million Items. *IEEE Symposium on Information Visualization* 2002 (Boston, Massachusetts, USA), IEEE Computer Society; 117-126.

22. Colby G and Scholl L. Transparency and blur as selective cues for complex visual information. *SPIE Image Handling and Reproduction Systems Integration* 1991 (San Jose, California, USA), SPIE; 114-125.

23. Buja A, et al. Interactive Data Visualization Using Focusing and Linking. *IEEE Visualization '91* 1991 (San Diego, California, USA), IEEE Computer Society Press; 156-163.

24. Shneiderman B. Dynamic Queries for Visual Information Seeking*. IEEE Software* 1994; **11**(6): 70 -77.

25. Eick SG. Data Visualization Sliders. *UIST '94* 1994 (Marina del Ray, California, USA), ACM Press; 119-120.

26. Fekete J and Plaisant C. Excentric Labeling: Dynamic Neighborhood Labeling for Data Visualization*.*

27. Saraiya P, North C, and Duca K. An Evaluation of Microarray Visualization Tools for Biological Insight. *IEEE Symposium on Information Visualization (INFOVIS'04)* 2004 (Austin Texas); 1-8.