MaTSE: The Microarray Time-Series Explorer

Paul Craig[*]
Universidad Tecnológica de la Mixteca

Alan Cannon[**]
Edinburgh Napier University

Robert Kukla[†]
Edinburgh Napier University

Jessie Kennedy[‡]
Edinburgh Napier University

ABSTRACT

This paper describes the design, development and evaluation of the Microarray Time-Series Explorer (MaTSE), a novel information visualization application for the exploratory analysis of large scale microarray timeseries data. The software combines a variety of visualization and interaction techniques, which work together to allow biologists to explore their data and reveal patterns that would otherwise be impossible to find. These include a scatter-plot that can be animated to view different temporal intervals of the data, a multiple coordinated view framework to support the cross reference of multiple experimental conditions, a novel method for highlighting overlapping groups in the scatter-plot, and a pattern browser component that can be used with scatterplot box queries to support cooperative visualization.

Keywords: Information visualization, bioinformatics, timeseries, microarray, animation.

Index Terms: H.5.2 [Information Interfaces and Presentation]: User Interfaces – graphical user interfaces; J.3 [Computer Applications]: Life and Medical Sciences—Biology and Genetics.

1 INTRODUCTION

Microarray technologies generate massive amounts of experimental data, which provide a perspective on biological functioning that is recognized as having huge value for the diagnosis, treatment, and prevention of diseases [1, 2]. At the same time online data repositories are providing biologists with access to more potentially valuable data from others' experiments [3]. There are however significant challenges to properly exploiting these data resources. This is due to the scale and complexity of data with individual data sets typically including 30,000 to 100,000 genes with measurements for 1-4 conditions over 3-20 time-points.

Typical analysis of microarray data [4-6] involves clustering and filtering steps that dispose of large amounts of data in order to make the results 'manageable' before they can be visualized and explored. While these methods can be useful for finding general trends in the data they also make it difficult to find certain characteristic patterns. Specifically, patterns of changing activity over intervals of time that have the potential to lead to biological insights [7, 8]. An example of a significant pattern that would not be revealed by clustering is illustrated in Figure 1. 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, 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.

In order to allow biologists to discover these types of pattern we developed the Time-series Explorer technique [7, 9, 10] (see Figure 2). This uses two coordinated views of the data: a line-chart and a scatter-plot. The line-chart overlays value versus time representations of the recorded activity of all genes and allows the user to specify an interval. 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 positions of genes in the scatter-plot are recalculated to adjust for the change in temporal context. Repeated continuous 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.

This technique was evaluated with biologists in order to demonstrate its ability to reveal previously unsuspected patterns of temporal activity with potential biological significance [4]. There were, however, a number of limitations that would discourage biologists from using this technique on a more regular basis. These were investigated during a requirements analysis workshop undertaken with 8 expert biologists belonging to a number of distinct sub-disciplines. Problems included the fact that the patterns found were difficult to quantify. Here biologists wanted results to be easily quantifiable and to have some mechanism whereby they could store and share the patterns found during an exploration. The biologists also wanted some way to cross reference existing gene groupings with results. Likewise, they wanted to be able to cross reference with the results of some

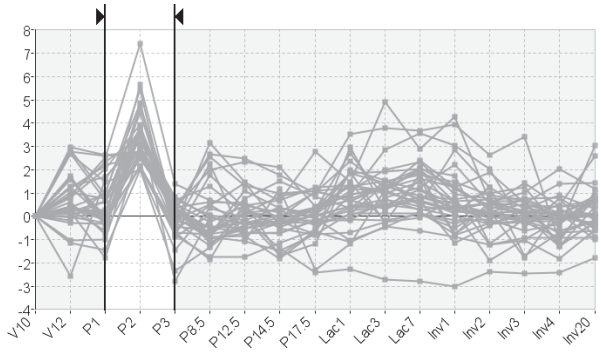

Figure 1: A significant pattern occurring exclusively over an interval. The highlighted rise then fall in expression could suggest that a group of genes are related to a particular biological process and that that process is associated with the experimental conditions.

[*] e-mail: p.craig@mixteco.utm.mx
[**] e-mail: a.cannon@napier.ac.uk
[†] e-mail: r.kukla@napier.ac.uk
[‡] e-mail: j.kennedy@napier.ac.uk

IEEE Symposium on Biological Data Visualization 2012
October 14 - 15, Seattle, Washington, USA
978-1-4673-4730-3/12/$31.00 ©2012 IEEE

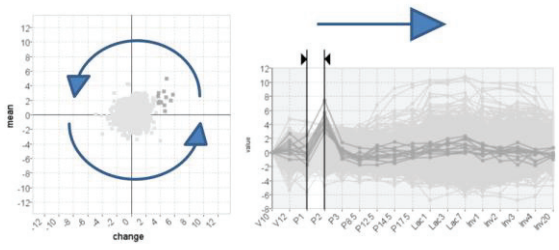

Figure 2: The basic Time-series explorer technique using two tightly coupled views of the data: a line-chart and a scatter-plot. The line-chart allows the user to select an interval. The scatter-plot summarizes the data within the selected interval by representing each gene as a single point with its y-position corresponding to mean activity and x position corresponding to change in activity. Adjusting the interval forward through time causes genes to move smoothly in a general anticlockwise direction as expression changes from low to rising to high falling back to low etc.

of the more established statistical methods such as clustering and t-tests variants. At the data level, they felt that the technique should be able to handle missing data, data with multiple conditions and different types of data rescaling.

2 Related Work

At present there are a limited range of techniques that are capable of revealing previously unsuspected patterns from microarray data. The majority of these techniques 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 [5-7]). Time-series data, of the type that is produced by microarray time-course experiments, can be conceptualised as a specialized subset of multidimensional data [8] 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 [9]. 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 1. 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.

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 [10] (top of Figure 3), a change in values between time points [10, 11] (middle of Figure 3) or a profile that the expression of genes must adhere to [11] (bottom of Figure 3). As this type of querying involves the specification of a limited time-interval 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, they are required to execute multiple speculative queries before patterns that relate that process to the experimental conditions can be revealed.

Another group of techniques overlay temporal expression data onto graph views of a gene network [12-16]. These techniques are capable of allowing the user to find groups of co-expressed genes such as those shown in figure 1 only if those genes are already clustered together in the gene network diagram. This will only happen if genes are already known to have some type of association, in which case the pattern found will be suspected to some degree. Another limitation of these techniques is that the space required to properly visualise a gene network makes it impossible to have an effective overview of gene activity. These techniques are useful for investigating gene activity in known gene groupings but are not suitable for finding unexpected patterns of activity within the mass of data generated by a microarray experiment.

3 MaTSE

The first parts of the Microarray Time-series Explorer (MaTSE) application to be developed were the data-model and animated scatter-plot. Based on the evaluation of the Time-series Explorer application we knew that performance during scatter-plot animation would be a critical factor for usability. We also had an idea that the axis of the scatter-plot may need to be changed in order for results to be easily quantifiable. It was also important to implement the scatter-plot early-on since it was a prerequisite for much of the other planned MaTSE functionality, such as the visualisation of groups, missing data etc.

The next stage was to redesign the data model to support a wider variety of data (data with replicates, missing values, multiple conditions etc.). This would enable us to evaluate the technique with more biologists using their own data. From previous experience, evaluating Time-series Explorer prototypes, we found that biologists evaluating a technique working with their own data, or at least data directly related to their own work, tended to generate better results. There are obvious reasons for this. Biologists working with their own data find it easier to generate scenarios for evaluation since these can be based on their own objectives rather than simply imagined. They are also less

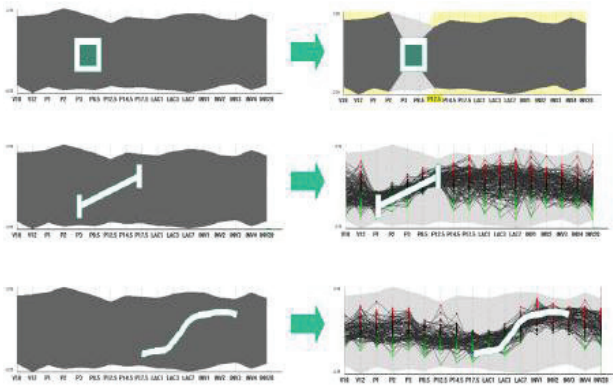

Figure 3: Visual queries (LHS query and RHS query results): an acceptable range of values over a given interval (top), an acceptable change in values between time points (middle) and a profile that the expression of genes must adhere to (bottom).

Table 1. Comparison of the scatterplot layouts in the Time-series Explorer and MaTSE (see figure 5 for attribute definitions)

Technique	Pre-processing	x-Axis represents	p>0	p=0	y-Axis represents	p>0	p=0	Limited to
Time-series Explorer	None	Ratio	$\frac{v_n}{v_0}$	n/a	n/a	$\frac{a}{p}$	n/a	Positive real numbers
MaTSE	None, log-rescaling and/or normalization	Difference or fold-change	$v_n - v_0$	Interpolated value	Mean value	$\frac{v}{v_{0\to n}}$	v	Real numbers

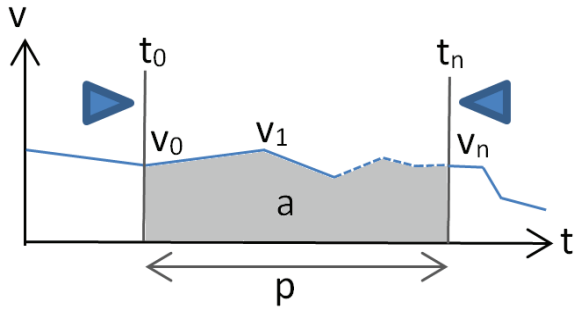

Figure 4: Attribute definitions for the Time-series Explorer and MaTSE scatterplot layouts.

likely to misunderstand the data and falsely interpret faults with the application. In general the process of evaluation is more natural when working with familiar data.

Our second MaTSE prototype dealt with the problems of visualising multiple experimental conditions and missing values. As with the improvements to the data-model, these features increased variety of data we could work with and therefore the number of biologists we could have evaluating the tool with their own data. Significant obstacles to be overcome for this prototype were to do with representing larger amounts of data on-screen and representing missing data. The third prototype dealt with the integration of statistics. These either constituted an alternative presentation of results (i.e. clustering) or lent a measure of confidence to groups of values (t-test, ANOVA). The forth prototype allowed biologists to view gene groupings with their data and the fifth allowed them to store and share results. The following sub-sections describe how different aspects of the new MaTSE interface where designed and implemented.

3.1 Scatterplot View

The scatter-plot view of the Time-series Explorer application [4] 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. Here activity and change in activity are measured in a particular way (see table 1 and figure 4). Activity over the interval is measured as the area under the graph bound by the interval divided by the length of the interval. 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. If we simply took the average of all values within the interval then genes would only move when the interval passed over a new time-point. Change in activity is measured as the value at the end of the interval divided by the value at the start. In this case when the interval is shifted by small increments, values are recalculated using linear interpolation so that gene representations can shift gradually along the X-axis.

There are a number of problems with the use of these axes that caused us to rethink our design for the first MaTSE prototype. Firstly, they rely on all values being positive. If the data were to include negatives values, a fall in activity between two positive values would be equivalent to a rise between negative values on the X axis. Similarly, negative or positive displacement would be equivalent when calculating a value for the Y axis. Critically, we would also have a situation whereby a second value (V_n) tending to zero would cause X to tend to infinity. Since values for V_0 and V_n are interpolated during animation, this would happen often (i.e. every time we interpolated a value between negative and positive values). Not being able to handle negatives was not such a problem for the Time-series Explorer since this application was limited to using un-normalised data recording all-positive mRNA levels. As, however, we wanted to develop MaTSE to work with a better variety of data, including data normalised to a control value or a value at a time-point, we could not apply the same restriction and this limitation would be problematic.

Another limitation of the Time-series Explorer scatter-plot is that it can't present gene activity at individual time points. If the user were to set the start and end of the time interval to the same point, V_0 would be same as V_n for every gene and displacement along the X axis would be 1 for every gene making it impossible to differentiate the points that represent genes. To avoid this happening, the duration of the selected interval has a lower limit of one. In the requirements for MaTSE however, biologists asserted that they wanted to store the results of queries based on measureable quantities. These types of queries would need to be based on activity and mean activity over an interval as well as change in activity.

Just like the Time-series explorer, MaTSE 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

$$f(v_0) = \log_2(v_0/c)$$

……f(v_0) expresses v_0 using it's fold change from the control value c.

$$\Rightarrow f(v_n) - f(v_0) = \log_2(v_n/c) - \log_2(v_n/c)$$

$$\Rightarrow f(v_n) - f(v_0) = \log_2(v_n c/v_0 c)$$

……using the logarithmic quotient identity $\log_b(x/y) = \log_b(x) - \log_2(y)$[22].

$$\Rightarrow f(v_n) - f(v_0) = \log_2(v_n/v_0)$$

……. v_n is now expressed using its fold change from v_o.

Box. 1. Mathematical proof (using the logarithmic quotient identity [22]) showing that the difference between two fold changes from the same control is equivalent to the fold change between the original values.

X-axis corresponding to its change-in-activity. However, activity and change in activity are measured using different attributes of the time-series (see table 1 and figure 4). This allows the technique to work with negative values and normalised data. In fact, the technique is tuned to work with normalised rescaled values and if the data is not already normalised and/or rescaled prior to being loaded into the tool, it can be normalised and rescaled within the tool before being visualised.

Normalisation of microarray time-series data is done on a per-gene basis where every value is divided by a control value. This could be a separate control value, an average of recorded values or the value at the first time point. The result of normalisation is that every value is expressed as a fraction of its control value. Rescaling of microarray time-course is a logarithmic transform used to make positive and negative fold changes equivalent. The resulting data has a normal distribution around zero where values are expressed using their fold change from the control value. A fold change of n indicates a multiplication by 2^n.

Since values are now expressed as fold changes from a control, it is more appropriate to use the difference between values instead of the ratio to express the change between two time points (i.e. $X=V_n-V_0$). This means that displacement on the X axis now corresponds to the fold change between the value at the start of the interval and the value at the end (see box 1 using identity [17]). In accordance with the users requirements to be able to form queries based on activity and mean activity, the Y axis now corresponds to the mean value of recorded expression for a gene over the time-points enclosed by the selected interval. When the interval covers a single time-point, displacement along the Y axis is equivalent to the value for that time-point and displacement along the X axis is equivalent to the average fold change in activity from the previous time-point to the next one. This avoids the limitation of the Time-series Explorer where genes would not be displaced about the X axis if the duration of the selected interval where to be set to zero. Interpolation is also handled differently in MaTSE. Here values for both X and Y axis displacements are interpolated using the values for the intervals directly before and after the selected interval. This means, for example, if the interval is from times 2.3 to 5.3, the X and Y values are interpolated using values for 2 to 5 and 3 to 6. If the interval is from times 3 to 5.3, values are interpolated using values for 3 to 5 and 3 to 6. This insures that genes always move smoothly in the scatterplot during animation when the selected interval is adjusted. To avoid interpolated values being used to form queries (and results being based on artefacts of the display rather than the data) the start and end of the interval selection automatically move to the nearest proper time-points (i.e. time-points for which expression is recorded) when the user is not adjusting the interval selection.

3.2 Visualisaing Missing Data

Figure. 5: Encoding for missing values in the scatter-plot: a) displacement along both axes is based on interpolated values b) y-axis displacement is based on interpolated values.

There are two types of scenario that require us to give a representation of missing data in MaTSE. These are values missing from the original data, and scatter-plot coordinates being based on interpolated values. In the first case the data is normally pre-processed and missing values can be attributed to variable replicate results (i.e. results with a low level of statistical confidence). In the latter case, scatter plot values are based on interpolated values when the duration of the selected interval is zero (see section 3.1) or values are missing in the data.

There are a number of methods for dealing with missing data in information visualisation applications [18]. These include the use of dedicated visual attributes, annotation and animation. Since animation is already used to communicate the change in expression for genes in our data and we have too much data to

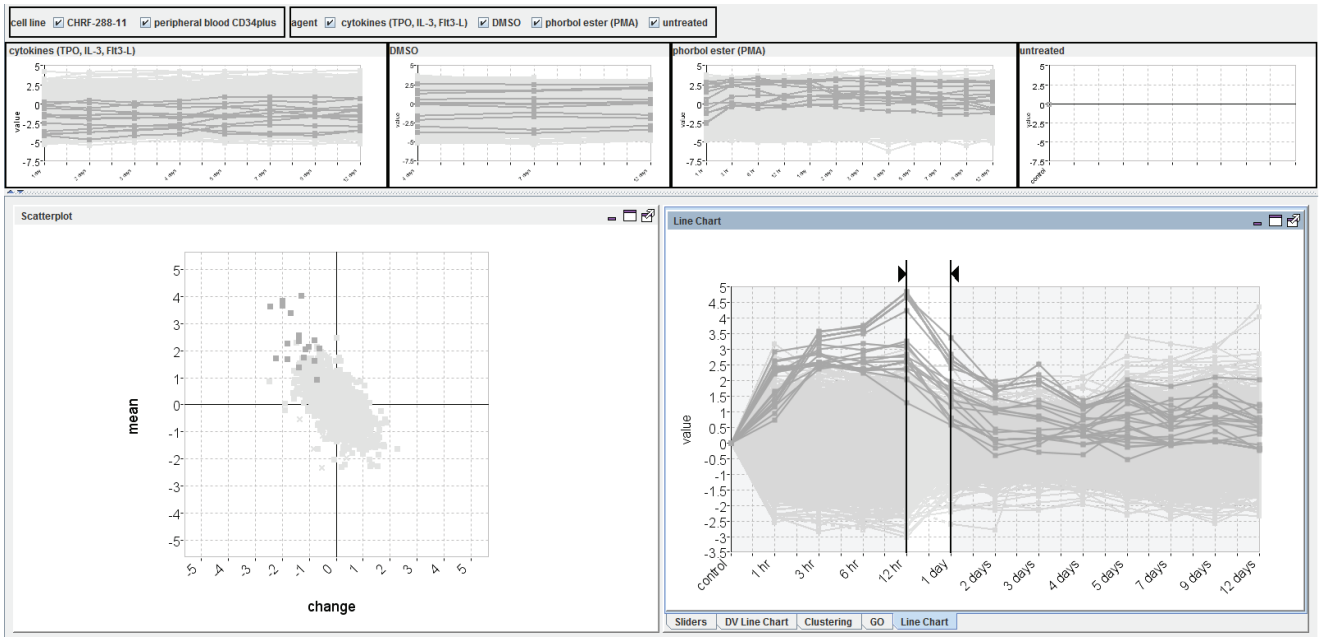

Figure.6: Representing multiple conditions in MaTSE. One principal scatterplot and line-graph showing data for a single condition and an overview line-chart for each other condition.

annotate genes individually, we decided to use dedicated visual attributes to deal with our missing values. Specifically, we use dotted lines in the line chart when interpolating over missing values and dots with missing segments to indicate interpolation in the scatter-plot (see figure 5). Here the visualisation uses less 'ink' to display scatter-plot positions or line-chart lines with less evidence.

3.3 Visualizing Multiple Conditions

Repeating an experiment two or more times with a single variable changed is a common practice in experimental biology. In the case of microarray experiment the variable that changes can be an environmental factor (e.g. temperature or humidity), exposure to different treatments or some form of genetic modification (e.g. gene knockout). Here biologists want to know about the effects of the new condition by comparing the data generated with that generated by other conditions or a control. The effect of multiple conditions on microarray data is to add another dimension and have the size of the data multiplied by the number of conditions.

In an earlier MaTSE prototype we allowed the user to view two or more conditions at the same time using a different set of linked scatterplot and line-chart views for each condition. Each of these separate condition views could have its selected interval moved independently or the selected intervals coordinated. While this gave the users a lot of freedom enabling then to compare the activity of genes with different conditions at different times, at also caused a significant degree of confusion. Specifically, users were confused as to how to control the multiple views, and had difficulty in comparing side-by-side multiple animated scatterplots in any meaningful manner. Furthermore, users found that with multiple sets of linked views sharing the limited screen space, it became difficult for them to make out enough detail in either scatterplot.

In later MaTSE prototypes we redesigned the interface to have one principal scatterplot and line-graph showing data for a single condition and an overview line-chart for each other condition (see figure 6). Clicking on any of the overview line charts causes that data to be shown in the principle scatterplot/line-chart. Views are coordinated so that selecting genes in one view causes them to be highlighted in all views. This new design is based on our finding that users generally preferred to compare patterns gene activity across conditions using line-chart views and explore patterns using both the line-chart and scatterplot. During evaluation we found that users did not miss the freedom to compare the activity of genes with different conditions at different times and found the new layout a lot easier to use.

3.4 Statistical Views of Data

Statistics are a method of summarising and analysing data for the purpose of drawing conclusions. Statistical methods can be categorised as either descriptive or inferential. Descriptive statistics are methods used to summarize or describe a collection of data while inferential statistics model patterns in the data in a way that accounts for randomness and uncertainty in the observations. Both types of method are used with microarray data and the biologists involved in MaTSE requirements analysis wanted to view both types of statistics in our prototype application.

3.4.1 Inferential statistics

The first statistical measurements incorporated into the application were p-values using a standard t-test [19] or ANOVA test [20]. These were shown on an interactive data visualisation slider [21] control (see figure 7), which displayed gene values as vertical lines and could be used to filter the coordinated view by moving the slider to set a threshold p-value. These values were

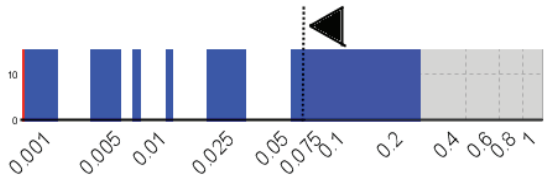

Figure. 7: The p-value data visualization slider allows users to filter.

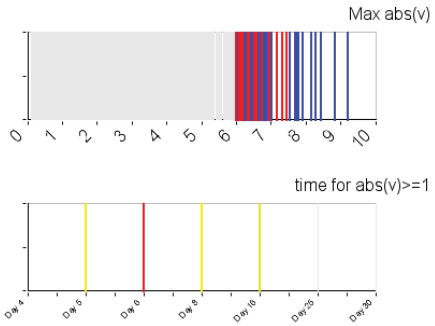

Figure. 8: *Maximum fold change* and *time for significant change* data visualization sliders.

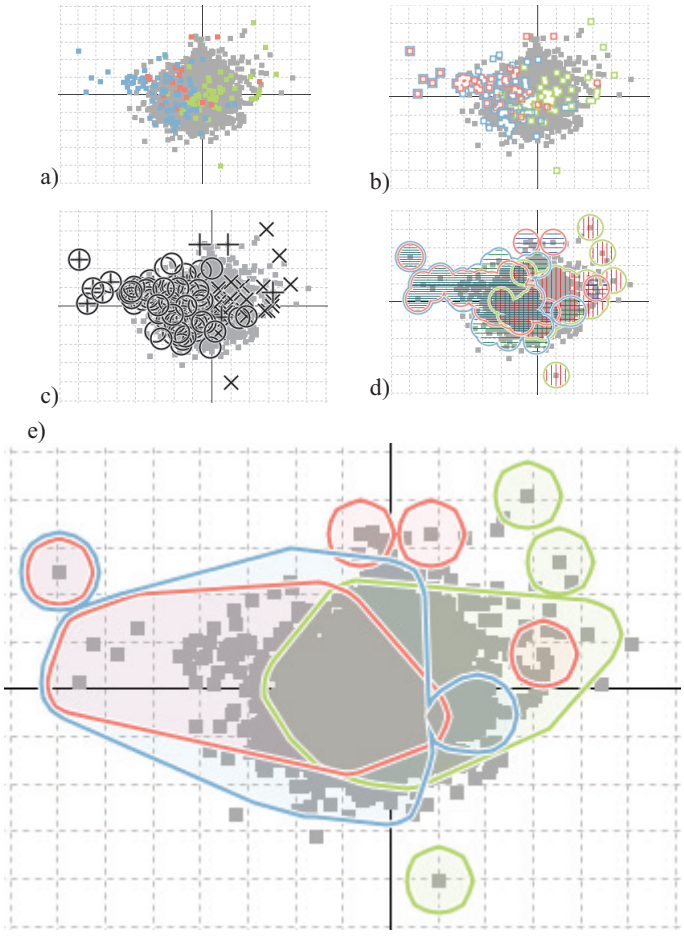

Figure. 9: Methods available for displaying gene groupings in the scatter-plot. a) colour coding, b) outlined colour coding, c) symbols, d) areas with texture and colour, and e) smoothed outline shapes with transparent shading.

also displayed alongside the gene id when data points were brushed on the scatterplot.

An evaluation of the t-test slider revealed that users who utilized statistics had many variations of p-value tests that they might wish to use in different circumstances and would prefer to run their own tests in a known package after using MaTSE to find patterns. They concluded that they would rather use a component such as the p-value slider for descriptive statistics such as a measure of maximum fold change for a gene over the time-course or the first time point where a genes activity has a fold change above 1. For these sliders we modified the design slightly by using a colour gradient to communicate the density of overlapping genes and allowed genes under the mouse cursor to be brushed (see figure 8).

3.4.2 Descriptive statistics

A dendrogram [5] showing the results of various hierarchical clustering methods is available in MaTSE in order to provide an alternative overview of the data. The methods for clustering include various variants of top-down and bottom up methods implemented using a link to the statistical package R [22]. While the Time-series Explorer and MaTSE techniques were initially developed to overcome the limitations of clustering methods for exploratory analysis (see section 1), the biologists felt that a view of clustering results could be used to complement the functionality of MaTSE. Here it was acknowledged that clustering would provide a more prescriptive type of analysis that was unbiased by the variable qualities of human perception. This would be used either confirm that patterns found by exploration are dominant in the data, or provide an initial set of gene groups that the user can explore using the more flexible MaTSE scatter-plot and line charts. In either case, all views are coordinated so that selections can be compared in different views. Genes are selected in the dendrogram view by clicking on or moving the cursor over branches.

3.5 Visualizing Gene Groupings

In bioinformatics genes can be grouped according to a number of factors such as functional similarity, the type of protein they encode or how they express in a particular type of experiment. These groupings can consist of small or large numbers of genes and often overlap. In MaTSE, gene groupings can be imported, visualised, exported, and created from a list of the currently selected genes. Figure 10 shows the five methods available for displaying gene groupings in the scatter-plot. These are colour coding, outlined colour coding, symbols, areas with texture and colour, and smoothed outline shapes with transparent shading. In each case the colours used for colour coding were drawn from the qualitative schemes suggested by Brewer [23].

Simple colour coding of genes in the scatterplot (figure 9a) was the easiest option to implement. There is, however, a problem when genes belonging to different groups overlapped each other in the scatterplot or genes belong to overlapping groups. In either case the colour for an overlapped grouping membership would need to be omited. This encoding can also be problematic due to the quality of human perception which tends to view small areas of colour differently depending on the colour of the background [24, 25]. The outlined colour coding method (figure 9b) avoids the problem of overlapping colours by displacing the different grouping outlines slightly so they don't overlap. The problem with this solution is that the area covered by colours becomes smaller and they become more difficult to properly identify against the background. This problem was slightly alleviated when the outlines where filled with a lighter colour but the overall effect was still far from satisfactory. Our third solution, encoding gene groupings using symbols (figure 9c), tended to work well with scatterplot outliers but not when genes were grouped together or belonged to multiple groupings since overlapping symbols were difficult or impossible to decipher. The encoding of groupings using texture and colour (see figure 9d) was part of a series of experiments designed to see if pre-alternative visual variables [26] other that shape or colour, such as texture and line orientation, might be used to identify groupings. These tended not to work well since overlapping textures or lines tended to look cluttered and were difficult to visually decode.

Our final encoding outlined gene groupings using smoothed outlines shapes with transparent shading (figure 9e). Here outlines are generated using a border at a fixed distance from the genes. This formed a shape that was smoothed to remove concave internal angles. The resultant simplified shapes were found to be easier to interpret and internal shading increased the area covered by colours making them easier to identify and differentiate. Grouping outlines are also slightly displaced from each other to avoid overlap. During evaluation this encoding was considered

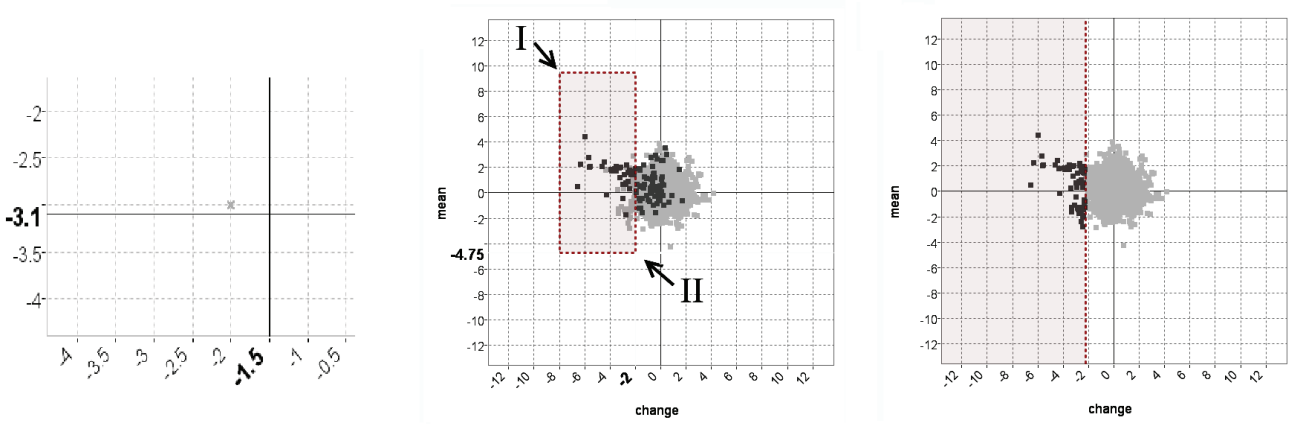

Fig. 10. Cooperative visualization in MaTSE: a) Cross-hair positioned at a rounded-value approximation of the mouse cursor position. The coordinates of the cursor are used to when forming queries. Bold font labels on the axes describe the cross-hair position to inform the user before and during query specification. b) A users attempt to specify a threshold on the value of a single axis by dragging a box query. The user clicks on point I and drags to point II to form the box-query illustrated with dotted lines. c) The dotted line indicates the threshold the user wants to set and the threshold sent to the MaTSE pattern browser as the recorded query. This is also what the user sees when they elect to refine the pattern.

the most effective and aesthetically pleasing. A potential disadvantage of roughly outlining a grouping of genes in this way, rather than highlighting individual genes (as in the methods shown in figs 12a to c), is the false positive effect where genes *not* in a grouping are inadvertently circled due to their position in the scatterplot. The biologists involved in our evaluation did not however consider this to be a significant disadvantage. This was because a) they expected this to happen, and b) they could easily select the gene grouping if they specifically wanted to see individual genes highlighted in the scatterplot. The outlined shape was interpreted as being similar to a line drawn around a group of points on printout of an academic paper or a projection over a whiteboard. These are classic examples of users drawing metaphors from their working environment in order to interpret abstract interface concepts. The technique worked well with up to 4 groups, and the biologists let us know would be sufficient for the vast majority of their requirements.

3.6 Cooperative Visualization

An advantage of having scatterplot coordinates based on measurable parameters in MaTSE (see section 3.1) is that it allowed us to adapt the interface to support some of the processes of collaborative visualization [27]. These processes are based on what we call 'patterns'. Each time a user selects genes by clicking and dragging a box around their representations in the scatterplot, the query parameters are stored as a pattern. These patterns can be combined, stored, refined, restored, annotated or passed to other users as a file or through a prototype online pattern repository.

In order to better facilitate the understanding of recalled or shared patterns, a number of mechanisms are in place to insure that stored patterns are concise and do not include superfluous parameters. The first of these is to ensure that only rounded values (i.e. values with smaller numbers of decimal places) are included in stored queries. Here a cross-hair, positioned at a rounded approximation of the mouse position, is used to generate the values for query formation rather than a direct mapping of the mouse position. Bold-font labels detail the cross-hair position on each axis to inform the user of the current cross-hair position during query composition (see figure 10a) and once a query is executed a text and icon representation of query parameters is displayed in a specialised pattern-browsing panel (see [27]). Superfluous parameters are included in a box query when a user attempts to specify a threshold on the value of a single axis. This situation is illustrated in figures 10 b and c. Here it can be seen that the user uses one edge of the dragged box to separate the genes to be included in their query from the rest of the data. The other edges of the box specify additional parameters which have no effect on the overall query result. These redundant parameters are removed before any query is stored. When a user returns to view part of a pattern, edges of the box that define the query can be dragged to adjust parameter values.

4 Evaluation

Usability and user satisfaction are crucial factors for the effectiveness of information visualization applications [28]. For the development of MaTSE we employed a user-centred cyclical process from requirements gathering through iterative development phases to final evaluation. This involved continuous evaluation and testing with industrial and academic partners.

This evaluation was designed to address working practice aimed at long-term adoption and support an iterative user-centred design process were biologists would evaluate successive prototypes with evaluation results fed-back into progressive stages of design. The focus of evaluation was on in-depth expert opinion to ensure reliable feedback and elicit domain expertise to inform development. Since our target users included academic users, industrial users, bioinformaticians, experimental biologists, and users with a general biological knowledge, it was evaluated accordingly with all these different types of users.

In actuality a mixture of techniques [29] were used to elicit qualitative feedback. These included interviews, demonstration, hands-on observation of real world exploration using the speak-aloud protocol and heuristic evaluations. A final evaluation involved interviews and observation of real-world exploration. To remove bias, this was confined to users not involved in evaluation sessions during the earlier stages of development cycle. In general, users expressed satisfaction and were positive about the power of the application for exploration. Of the three main case studies undertaken, two sets of academic users expressed the belief that they would like to use MaTSE in the exploratory analysis phase of future experiments. The other study involved biologists working in a commercial environment. These users commented that while the technique had potential to be "very desirable" for those researchers able to focus on an experiment to generate biological knowledge, it was less useful for work which involved investigating a limited set of possibilities relating to a subset of the data over a short period of time. These findings were in accordance with our original presumption that the approach would be of best use with exploratory analysis for hypothesis generation rather than for investigating known phenomena. All users saw MaTSE as one of a suite of applications and services that they would use to analyse microarray data as part of their iterative analysis process. As such it was seen as a complimentary tool for time-series work rather than a general alternative to their current analysis software. It was suggested that future versions of the software could benefit from harnessing the growing and highly valuable knowledge bases and annotation services such as the Gene Ontology [30], as well as supporting the use of the large sets of gene group lists built up by experienced biologists, which would make analysis more convenient for the users.

Conclusion

In this paper we describe MaTSE, a new information visualization application for the exploratory analysis of microarray time-series data. This is based on the earlier Time-series Explorer technique which uses two tightly coupled views of the data: a line-chart and a scatter-plot. The line-chart allows the user to select an interval. The scatter-plot summarizes the data within the selected interval by representing each gene as a single point with its y-position corresponding to mean activity and x position corresponding to change in activity. Adjusting the interval forward through time causes genes to move in a general anticlockwise direction. This allows the user to find previously unsuspected patterns of temporal activity that could not be found using other techniques.

MaTSE also overcomes a number of Time-series Explorer limitations with an improved scatterplot that displays more useful quantities (fold change in activity, activity and mean activity) and supports negative values, missing values and selected intervals with a duration of zero to show instantaneous activity. The new application can also display multiple experimental conditions using coordinated views and uses data-visualisation sliders to give a view of various statistical measures. It also allows up to four different gene groupings can be viewed at a time in the scatter-plot using a novel method using smoothed outline shapes with transparent shading, and allows queries to be stored as patterns to be combined, refined, restored, annotated or passed to other users as a file or through a prototype online pattern repository.

Functionality was evaluated throughout the development process allowing us to refine requirements and tailor development according to the needs of potential users. A final evaluation provided us with evidence of the tools utility for the processes of exploratory data analysis. Likely opportunities for future work

based on the results of this paper might include; improving the visualisation technique for microarray data (e.g. supporting a view of gene network data or linking to gene ontology data), adapting the technique for other types of time-series data (e.g. economic data or social statistics) or adapting parts of the technique for other types of data (e.g. using the method used to outline gene groupings to outline, for example, different types of event labelled on a map).

5 Acknowledgments

The authors would like to thank CXR Biosciences; Institute for Stem Cell Research, Edinburgh; Scottish Crop Research Institute; and Queen's Medical Research Institute, Edinburgh for participating in the requirements analysis and evaluation of MaTSE. This project was funded by the Scottish government funding body Scottish Enterprise.

References

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

[2] P. O. Brown and D. Botstein, "Exploring the new world of the genome with DNA Microarrays," *Nature Genetics,* vol. 21, pp. 33-37, 1999.

[3] Barrett*, et al.*, *NCBI GEO: archive for functional genomics data sets—10 years on*. Oxford, ETATS-UNIS: Oxford University Press, 2011.

[4] P. Craig*, et al.*, "Animated Interval Scatter-plot Views for the Exploratory Analysis of Large Scale Microarray Time-course Data," *Information Visualization,* vol. 4, pp. 149-163, Sept. 2005.

[5] M. B. Eisen*, et al.*, "Cluster analysis and display of genome-wide expression patterns.," *Proc. Natl. Acad. Sci. USA,* vol. 95, pp. 14863-14868, December 1998 1998.

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

[7] S. Raychaudhuri*, et al.*, "Principal Components Analysis to Summarize Microarray Experiments: Application to Sporulation Time Series," in *Pacific Symposium on Biocomputing*, 2000, pp. 452-463.

[8] B. Shneiderman, "The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations," in *IEEE Visual Languages '96*, Boulder, Colorado, USA, 1996, pp. 336-343.

[9] E. Segal*, et al.*, "Rich probabilistic models for gene expression," *Bioinformatics,* vol. 17, pp. 243-52, 2001.

[10] H. Hochheiser and B. Shneiderman, "Dynamic query tools for time series data sets: Timebox widgets for interactive exploration," *Information Visualisation,* vol. 3, pp. 1-18, 2004.

[11] B. Shneiderman and J. Seo, "Interactively Exploring Hierarchical Clustering Results," *IEEE Computer 35,* vol. 7, pp. 80-86, 2002.

[12] M. A. Westenberg*, et al.*, "Interactive Visualization of Gene Regulatory Networks with Associated Gene Expression Time Series Data," in *Visualization in Medicine and Life Sciences, Visualization and Mathematics*, H. H. L. Linsen, and B. Hamann, Ed., ed Berlin, Germany: Springer Verlag, 2007, pp. 293-312.

[13] B. Kim*, et al.*, "GeneShelf: A Web-based Visual Interface for Large Gene Expression Time-Series Data Repositories," *IEEE Transactions on Visualization and Computer Graphics,* vol. 15, pp. 905-912, 2009.

[14] D. H. Jeong*, et al.*, "Interactive visual analysis of time-series microarray data," *Vis. Comput.,* vol. 24, pp. 1053-1066, 2008.

[15] R. Bourqui and M. A. Westenberg, "Visualizing Temporal Dynamics at the Genomic and Metabolic Level," presented at the Proceedings of the 2009 13th International Conference Information Visualisation, 2009.

[16] M. Westenberg*, et al.*, "Visualizing Genome Expression and Regulatory Network Dynamics in Genomic and Metabolic Context," *Computer Graphics Forum,* vol. 27, pp. 887-894, 2008.

[17] "Logarithmic Function," in *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables, 9th printing*, M. Abramowitz and I. A. Stegun, Eds., ed New York: Dover, 1972, pp. 67-69.

[18] C. Eaton*, et al.*, "Visualizing Missing Data: Classification and Empirical Study," in *INTERACT*, Rome, Italy, 2005, pp. 861-872.

[19] X. Cui and G. Churchill, "Statistical tests for differential expression in cDNA microarray experiments," *Genome Biology,* vol. 4, 2003.

[20] G. A. Churchill, "Using ANOVA to analyze microarray data.," *Biotechniques.,* vol. 2, pp. 173-5, 177, Aug 2004.

[21] S. G. Eick, "Data Visualization Sliders," in *ACM UIST*, Marina del Ray, California, USA, 1994, pp. 119-120.

[22] R. F. f. S. Computing, *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2011.

[23] C. Brewer, "Guidelines for use of the perceptual dimensions of color for mapping and visualization," in *Color hard copy and graphic arts III, Proceedings of the international society for optical engineering (SPIE), San Jose, February 2004.*, 1994, pp. 54-63.

[24] A. Kirschmann, "Ueber die quantitativen Verhältnisse des stimultanen Helligkeits- und Farben-Contrastes.," *Philosophische Studien,* pp. 417-491, 1891.

[25] C. H. Graham and J. L. Brown, "Color contrast and color appearance: Brightness constancy and color constancy," *Vision and visual perception,* pp. 452-478, 1965.

[26] A. Treisman, "Preattentive processing in vision," *Comput. Vision Graph. Image Process.,* vol. 31, pp. 156-177, 1985.

[27] P. Craig*, et al.*, "Pattern browsing and query adjustment for the exploratory analysis and cooperative visualisation of microarray time-course data," presented at the Proceedings of the 7th international conference on Cooperative design, visualization, and engineering, Calvia, Mallorca, Spain, 2010.

[28] M. Graham*, et al.*, "Towards a methodology for developing visualisations," *International Journal of Human-Computer Studies,* vol. 53, pp. 789-807, 2000.

[29] J. Nielsen, *Usability Engineering*. Boston: Academic Press Professional, 1993.

[30] M. Ashburner*, et al.*, "Gene Ontology: tool for the unification of biology," *Nat Genet,* vol. 25, pp. 25-29, 2000.