Pattern Browsing and Query Adjustment for the Exploratory Analysis and Cooperative Visualisation of Microarray Time-course Data

Paul Craig[1], Alan Cannon[1], Jessie Kennedy[1] and Robert Kukla[1],

[1] Center for Information and Software Systems, Edinburgh Napier University, 10 Colinton Road, Edinburgh, United Kingdom, EH10 5DT
{p.craig, a.cannon, j.kennedy, r.kukla}@napier.ac.uk

Abstract. This paper presents work to support collaborative visualisation and data analysis in the microarray time-series explorer (MaTSE) software. We introduce a novel visualisation component called the ‘pattern browser’ which is used to support the annotation and adjustment of user queries. This includes an explanation of why this component is required and how it can be used with our online pattern repository by biologists collaborating in the analysis of a microarray time-course data set. To conclude we suggest which other types of collaborative visualisation would benefit from the introduction of a component with comparable functionality.

Keywords: Cooperative Visualisation, Combined Multiple Views, Bioinformatics, Microarray Data Analysis

1 Introduction

In the past decade there has been a rapid increase in the amount of data generated by high-throughput genomic technologies [1], [2]. While this data shows great potential for allowing biologists to increase their knowledge of biological systems and processes, there remain significant challenges concerning effective exploitation. These relate to the scale and complexity of the data. In the first instance biologists need to be able to analyze data from their own laboratory and must contend with the large amount of data generated by individual experiments. There is also an increased need for biologists to be able to share data and analysis results in order to benefit from experiments and research undertaken outwith their own laboratory.

While published results of microarray experiments tend to focus on a small group of findings, the data on which any publication is based has the potential to reveal a range of findings related to a variety of biological processes. In order for the scientific community to better exploit this potential, biologists are encouraged to use online data repositories [2]. If a biologist who downloads data from one of these repositories has similar objectives to those of the original authors it is also likely that they will want to explore the results of the original authors’ analysis together with details of how that analysis was undertaken with a view to performing a similar analysis themselves. This process is not well supported in current analysis tools. While a number of tools allow users to share findings by saving and restoring application states (for example Spotfire DesisionSite and Agilent Genespring), analysis steps cannot be adjusted in a predictable way or with adequate feedback of results. The work we have undertaken with the MaTSE application and its pattern-browser component addresses this problem by providing users with a platform that allows them to discover and define patterns in their data using queries that can be shared, explored and adjusted in a manner that is both meaningful and informative.

2 Related work

The majority of software applications used for the analysis of microarray data rely on clustering algorithms which prescribe a fixed set of gene clusters based on gene-gene activity similarity scores. Information visualisation is used to explore the results of these algorithms. A finding from this type of interface tends to be an individual cluster exhibiting a pattern of gene activity with a gene population that is correlated with a predefined gene grouping or biological pathway sourced from the literature. For a user to alter a query on which a clustering finding is based they would need to select an alternative grouping, specify alternative clustering parameters (such as the distance metric) or use an alternative clustering algorithm. In the former case the results would be unrelated to the original result and in the latter two cases the outcome is highly unpredictable as an entirely new set of gene clusters will be generated with no relation to the original set [3], [4].

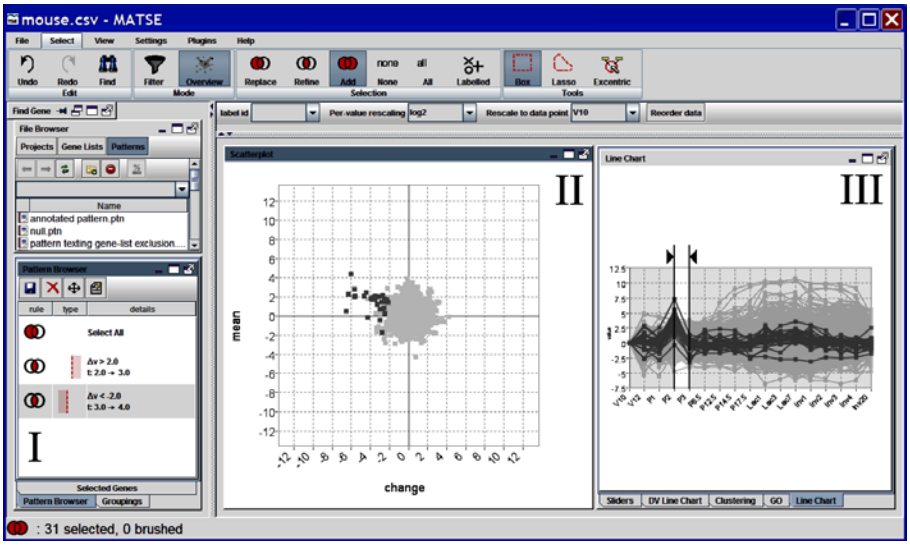
Other software applications for the analysis of microarray time-course data allow users to query their data by using a line-chart representation to specify a required pattern of expression, such as an acceptable range of values, over a given interval of the time-course [5], [6]. They allow queries to be adjusted but do not support data sharing since the attributes upon which queries are based are not particularly useful for most biologists who prefer to quantify patterns in their data using fold-changes in differential expression [1]. These techniques also fail to provide adequate feedback of results when queries are adjusted since the overview provided is unable to reveal anything other than the range of values at individual time points [7]. Changes in activity are represented by angled lines between time-points in a line-chart and the majority of lines are occluded.

3 MaTSE

MaTSE [7] is a different type of analysis tool which forgoes the clustering step prior to visualisation using a more direct representation of the data that allows selections to be explored and adjusted. Instead of prescribing a fixed set of gene clusters MaTSE allows the user to explore their data using a scatter-plot that can be animated across time-points and time-intervals to display measurable attributes of the data.

The MaTSE technique uses two coordinated views of the data: a line-chart and a scatter-plot (see figure 1). The line-chart view overlays value versus time representations of the recorded activity for all genes and allows the user to control the interval of the data for displayed in the scatter-plot. 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 fold change-in-activity from the start to the end of the interval. As the line-chart view controls are manipulated and the interval selection is adjusted, the positions of genes in the scatter-plot are recalculated to reflect the change in temporal context. Continuous adjustment of the selected interval (where the start and end times of the selected interval are moved 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 [7].

Evaluation of an early version of MaTSE demonstrated that it was capable of allowing users to find previously unexpected patterns of temporal activity [7]. However, further requirements analysis in preparation for the design and development of the current version of the software revealed that users wanted to be able to quantify the queries used to uncover and define those patterns so they could share their findings with other users. This led us to incorporate a new component, the pattern browser, into our latest version of MaTSE.



label id

Per-value rescaling log2

Rescale to data point V10

annotated pattern.ptn
null.ptn
pattern texting gene-list exclusion....

Select All

Δv > 2.0
t: 2.0 → 3.0

Δv < -2.0
t: 3.0 → 4.0

: 31 selected, 0 brushed

Fig. 1. A screenshot of the MaTSE interface. Labeled components are **I)** The pattern-browser, **II)** scatter-plot and **III)** line-chart. The current pattern is the result of two separate queries.

4 The Pattern Browser

The primary functions of the MaTSE pattern-browser component are to store query components, combine query components, allow queries to be explored and allow query components to be adjusted. The browser is allocated its own separate panel within the MaTSE interface and is coordinated with other MaTSE components. These components are adapted to allow query parameters to be adjusted and ensure that stored queries are concise without the inclusion of superfluous values or values a user might find difficult to interpret.

4.1 Composing box queries

MaTSE allows users to compose queries in the scatter-plot by clicking on a point and dragging a box around the genes they want to select. Since these queries will be revisited via the browser panel it is in the users' interest that they are as concise as possible. This makes it important to insure that stored queries omit superfluous parameters. 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 figure 2. Here it can be seen that the user uses one edge of the 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 query parameters which have no effect on the overall query result. These redundant parameters are removed before any query is stored.

Our users also wanted to limit query parameters to rounded values with a smaller number of significant digits. To ensure that only these values are included in stored queries a cross-hair, positioned at a rounded approximation of the mouse position, is used for query formation rather than a direct mapping of the mouse position. Oversized labels detail the cross-hair position on each axis to inform the user of the current cross-hair position (see figure 3a). Once a query is executed a text and icon representation of query parameters is displayed in the pattern-browsing panel.

4.2 Combining queries

MaTSE allows users to combine queries to find and specify more complex patterns of activity in the data. An example of such a pattern is illustrated in figure 3b. Here a group of genes have activity rising over one interval and falling over another. Patterns can also include queries that select gene-groupings (imported from the Gene Ontology online database) or lasso queries formed by drawing a line-loop around genes in the scatter-plot. The user can specify how queries are combined using options in the 'select' tab of the main menu. The options are to replace, refine or add to the results of the previous queries, select all genes or select no genes.

The selected option for combining queries affects how the pattern browser stores query parameters. When the 'refine' or 'add' options are selected the result of any new query is combined with the previous result. This makes it necessary to keep a record of how the previous result was obtained so new queries are added bellow existing queries to form a list in the pattern-browser panel. Conversely, when the ‘replace’ option is selected and a new query is performed the results of previous queries are discarded so their entries are removed from the panel. The ‘select all’ and ‘select none’ options also remove the results of previous queries but do this as soon as either button is pressed.

4.6 Annotating, storing and restoring patterns

Adding and editing pattern annotation is a relatively straight-forward process. When the user wants to view the annotation they simply need to click on the note-pad icon in the pattern-browser menu. This causes a pattern-annotation dialogue to appear with text of the current annotation. The text in the dialogue is editable so it can also be used to add new annotation or amend existing annotation.

MaTSE patterns can be saved either offline to the user’s hard-drive or online to the MaTSE website. Patterns are stored offline by clicking on the save button in the

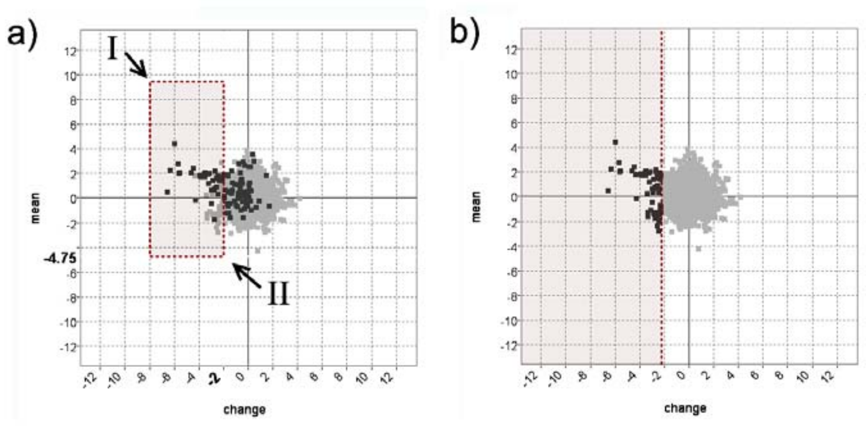


Fig. 2. a) A users attempt to specify a threshold on the value of a single axis by dragging a box query. The user clicks on point I and drags to point II to form the box-query illustrated with dotted lines. **b)** The dotted line indicates the threshold the user wants to set and the threshold sent to the MaTSE pattern browser as the recorded query.

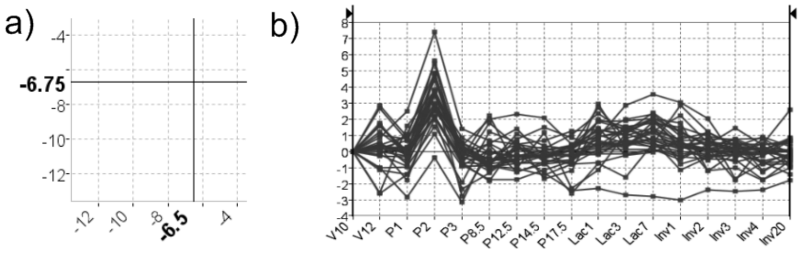


Fig. 3. a) Cross-hair positioned at a rounded-value approximation of the mouse cursor position. The coordinates of the cursor are used to when forming queries. Oversized labels on the axes describe the cross-hair position to inform the user before and during query specification. **b)** Line-chart representation of a compound query result.

pattern-browser panel menu. Once the pattern is named it is saved in XML format within the current project's directory. This file includes the pattern's queries, annotation and any transforms applied to the data. Patterns can be renamed and organized into sub-directories using MaTSE's file browser panel. Double-clicking on a pattern file in the file browser restores the pattern to the pattern browser and sets the gene selection to match the result of the final query combination. Patterns are uploaded to and downloaded from the MaTSE website in a similar manner. The advantages of uploading patterns to the MaTSE website are that they are made public and can be downloaded, further explored and discussed by other users. The ability to share the result of an exploration while also giving access to the intermediate steps leading up to it is a unique feature of MaTSE that can further stimulate cooperation between biologists. The quick turn-around time that is achieved by direct access to the online repository from the application guarantees a smooth interchange between users in different locations.

4.3 Browsing patterns and adjusting queries

MaTSE patterns can be explored and adjusted at any point during their formation or after they are restored to the application. Patterns are explored by clicking on the visual representations of individual queries in the patter browser. If the query has an associated time-interval context the scatter-plot animates to that particular interval before the original query parameters are highlighted on the scatter-plot. The highlighting remains for as long as the mouse-pointer hovers over the pattern-browser panel. The highlighting uses the same visual encoding of the original query formation tool (see figure 2b). Use of automated animation between query intervals allows users to rapidly alternate between queries to gain a better idea how individual query parameters relate to the underlying data.

Patterns can also be adjusted by clicking on the visual representations of individual box queries in the patter browser. Clicking on a query then the 'delete' button in the pattern-browser menu deletes a query from the pattern and removes its influence on the set of selected genes. Clicking on a query followed by the 'adjust' button allows queries to be adjusted incrementally. After this button is pressed the selected query remains highlighted on the scatter-plot. Clicking and dragging on the edges of this representation resizes it and updates the associated query parameters with the set of selected genes updated continuously to provide visual feedback of the adjustment's effect on the data-set. This type of dynamic feedback is useful if, for example, a user sees that a threshold is set high and they want to see the result if it is set lower. On setting the threshold lower and revealing that this has little or no effect on the overall result the user can immediately return the parameter to its original value by continuing their dragging action in the opposite direction.

5 Evaluation

User evaluation of the MaTSE pattern-browser functionality was undertaken with five different biologists working in the areas of toxicology, stem cell research, plant biology and bioinformatics. The evaluation included sessions with and without the pattern-browser functionality. The previous method for storing patterns was for users to save selections as gene-lists which included annotation but did not include query parameters. Evaluation sessions included a tutorial where the biologists where taught how to use new functionality and free exploration sessions where users were asked to use the tool to analyze their own data using the functionalities of the tool they felt most appropriate, in the way in which they felt most comfortable. We applied a think-aloud protocol during the free sessions and conducted follow up interviews to record results.

Overall, we found that reaction to the pattern-browser functionality was positive from all users and all users expressed a strong preference for using the pattern-browser rather than the previous method for storing patterns. The pattern-browser was found to have a number of clear benefits and tended to be used in a distinctive way across the different user groups. As users formed compound queries they tended to monitor the pattern-browser panel to confirm their query was committed. If a user moved away from the MaTSE interface during their MaTSE session (to, for example, answer the phone or look up the function of a particular gene) their first task on returning would be would be to quickly review the list of query parameters in the pattern-browser. Occasionally this involved clicking on queries in the browser to view them in the scatter-plot. The users expressed a degree of satisfaction with the pattern browser and its ability to help them recall their previous selections. The converse situation was apparent in our evaluation of previous prototypes where, on returning to MaTSE from another task, users often became frustrated when they forgot what their previous selections had been.

While all users stored their patterns and restored them to the tool from time to time, it was only the biologists working in a research orientated environment who tended to adjust query parameters. The toxicologists who operated in a commercial environment worked in a more rigid fashion tending to set query thresholds to a fixed level with little motivation to adjust query parameters during their exploration of the data. The other biologists would often adjust parameters as part of their exploration of the data. Here, the continuous feedback of results in all views was found to be useful with biologists adjusting parameters to see how the overall selection changed or if a particular gene became selected before deciding on a final query parameter value.

The online storage and annotation of MaTSE patterns has not as yet been formally evaluated with biologists. The biologists in our other evaluations have however expressed enthusiasm for the idea of being able to share their results with colleagues using the pattern-browser functionality. They also believed that they would be able to understand and make good use of stored patterns if they were submitted by experts in their own particular field.

7 Conclusion and further work

We have adapted the MaTSE software to support cooperative visualisation of microarray time-course data with the addition of a new component called the ‘pattern-browser’. The adapted interface stores meaningful query parameters allowing the user to define patterns which can be recalled, explored and adjusted. Query parameters are restricted to rounded values and superfluous parameters are removed as queries are formed. Animation is used when patterns are explored and there is continuous feedback of results as query parameters are adjusted. In our evaluation of the new interface, users have been able to identify a number of benefits. These related to the ease in which query parameters could be recalled and the additional information that could be gleaned from the data when adjusting parameters. Our users also believe that the new functionality will serve them well to share findings during the process of cooperative visualisation.

In MaTSE the use of a pattern browser panel for exploration of query parameters is necessitated by the fact that the scatter-plot is animated and the relative position of data point's changes according to the selected time-frame. This makes it impossible for all selections to be presented in the scatter-plot at the same time and the pattern browser is needed to allow the user to access selections individually. Including a separate scatter-plot for each selection would be another solution if not for the fact that screen space is at a premium and a suitable resolution is necessary for the scatter-plot to be functional. The results of this paper should be of use to designers of other visualisations using multiple projections of the data where all user selections cannot be summarized at the same time by overlaying them on top of the visualisation. Examples of such visualisations include visualisations for data-cubes [8] and similar data with a large numbers of dimensions.

For future-work we aim to demonstrate the MaTSE technique in a published case study with data from an original experiment so we can promote the general technique and encourage use of the online pattern-repository. We also plan to adapt the pattern browser technique for use with other visualisations developed in our research group.

References

1. Allison, D.B., Cui, X., P. Page, G., Sabripour, M.: Microarray data analysis: from disarray to consolidation and consensus. Nature Reviews Genetics. 7 (2006) 55-65
2. Barrett, T., Troup, D., Wilhite, S., Ledoux, P., Rudnev, D., Evangelista, C., Kim, I., Soboleva, A., Tomashevsky, M., Edgar, R.: NCBI GEO: mining tens of millions of expression profiles--database and tools update. Nucleic Acids Res. (2007) 760-765
3. Thalamuthu, A., Mukhopadhyay, I., Zheng, X., Tseng, G.C.: Evaluation and comparison of gene clustering methods in microarray analysis. Bioinformatics. 22 (2006) 2405-2412
4. Kerr, G., Ruskin, H., Crane, M., Doolan, P.: Techniques for clustering gene expression data. Computers in Biology and Medicine. 38 (2008) 283-293
5. Seo, J., Shneiderman, B.: Interactively Exploring Hierarchical Clustering Results. IEEE Computer. 35 (2002) 80-86
6. Hochheiser, H., Shneiderman, B.: Dynamic query tools for time series data sets: Timebox widgets for interactive exploration. Information Visualisation. 3 (2004) 1-18
7. Craig, P., Kennedy, J.B., Cumming, A.: Animated Interval Scatter-plot Views for the Exploratory Analysis of Large Scale Microarray Time-course Data. Information Visualisation. 4 (2005) 149-163
8. Stolte, C., Tang, D., Hanrahan, P.: Multiscale Visualization Using Data Cubes. IEEE Transactions on Visualization and Computer Graphics. 9 (2003) 176-187