BiCluster Viewer: A Visualization Tool for Analyzing Gene Expression Data

Julian Heinrich, Robert Seifert, Michael Burch, Daniel Weiskopf

VISUS, University of Stuttgart

Abstract. Exploring data sets by applying biclustering algorithms was first introduced in gene expression analysis. While the generated biclustered data grows with increasing rates due to the technological progress in measuring gene expression data, the visualization of the computed biclusters still remains an open issue. For efficiently analyzing the vast amount of gene expression data, we propose an algorithm to generate and layout biclusters with a minimal number of row and column duplications on the one hand and a visualization tool for interactively exploring the uncovered biclusters on the other hand. In this paper, we illustrate how the BiCluster Viewer may be applied to highlight detected biclusters generated from the original data set by using heatmaps and parallel coordinate plots. Many interactive features are provided such as ordering functions, color codings, zooming, details-on-demand, and the like. We illustrate the usefulness of our tool in a case study where yeast data is analyzed. Furthermore, we conducted a small user study with 4 participants to demonstrate that researchers are able to learn und use our tool to find insights in gene expression data very rapidly.

1 Introduction

Modern research has developed promising approaches for analyzing gene expression data. The major bottleneck of such analyses is the vast amount of data. The behavior of single genes under different conditions such as different time points or tissue types are contained within these massive data sets. Such data poses a difficult task for analysis, as it might be incomplete and often has a low signal-to-noise ratio.

Biclustering is frequently used to analyze such data sets and to discover dependencies among genes and conditions. A single bicluster represents part of a table in which the corresponding genes and the conditions involved behave similarly in a certain way.

The immense and growing number of data sets as well as biclusters make it very difficult or even impossible to uncover all dependencies in a single static view and hence, exploit human perceptual abilities for a fast exploration of the data mapped to a visual form. Another problem is the fact that it is not clear a priori what an analyst hopes to detect in the data.

In this paper, we present an interactive visualization tool that can be used to visualize a large number of biclusters. In general, the tool follows the Information

Visualization Seeking Mantra: Overview first, zoom and filter, then details-on-demand proposed by Shneiderman [1]. The visualization techniques used in the tool are based on heatmap representations as used by Eisen et al. [2] and parallel-coordinate plots first introduced by Inselberg and Dimsdale [3].

In a case study we illustrate how a yeast data set can be explored for biclusters. Furthermore, we demonstrate the usefulness of our BiCluster Viewer by conducting a small user study with 4 participants. The visualized biclustering data is based on both an artificial data set and a real-world data set from gene expression analysis.

Our visualization tool is not restricted to bicluster representation in gene expression data but can easily be applied to any kind of matrix-like data containing real-valued numbers. Heatmap representations are ideal to visualize large tabular data sets such as gene expression data and to get an overview representation by mapping data values to color values. Parallel coordinate techniques are used to map the rows and columns of biclusters to vertical axes and connect them by direct lines. Visual clutter caused by line crossings in dense data sets can be reduced using interactive features such as filtering, transparency, and color coding. Linking and brushing techniques are used to allow different views on selected subsets of the data simultaneously.

2 Related Work

Most related applications use heatmaps [2], parallel coordinates [3] or node-link diagrams for the visualization of biclusters. While heatmaps use a matrix layout and color to indicate expression levels, parallel coordinates use polylines across many axes (representing columns of the matrix) to represent multivariate data. Both visualizations suffer from the ordering problem such that is generally not possible to represent more than two biclusters contiguously without row (line) or column (axis) duplication.

Grothaus et al. [4] represent overlapping biclusters in a single heatmap and allow row and column duplications if biclusters cannot be represented contiguously. While being optimal with respect to the number of duplications, such an automatic layout algorithm does not allow for interactivity. In our work, we allow overlapping biclusters and the user may decide which biclusters to show contiguously in order to minimize row and column duplications.

BicOverlapper by Santamaria et al. [5] is able to represent several overlapping biclusters simultaneously. But the techniques used there are based on graphs represented in a node-link visual metaphor which is different to our work where heatmaps and parallel coordinates are used.

Cheng et al. [6–8] use parallel coordinates to visualize additive and multiplicative biclusters. Visual clutter [9] caused by many intersecting lines is the main problem when drawing parallel-coordinate plots for dense data sets. We circumvent this drawback by using transparency and color coding of single lines and line groups. Furthermore, lines that are currently not in focus may be suppressed to further reduce the overlap and visual clutter. These clutter reduction

principles are also used in [10]. We further extend parallel-coordinate plots using bundling to show biclusters.

The BiClust package developed by Kaiser et al. [11] is an extension for the R environment [12]. The package proposes a variety of computation algorithms and also many visualization techniques to represent the biclustered data. In addition to the traditional heatmap representation and parallel-coordinate plots also bubble charts are provided. Our tool alllows to use BiClust or any other algorithm provided in R for the computation of biclusters.

BiCat [13] is another application that can be used to analyze biological data such as gene expression data. In contrast to R it is not based on command line arguments but provides a graphical user interface to manipulate and navigate in the data. All computed biclusters are shown in a list where they can be selected. A selection in the heatmap or parallel-coordinate plots is not supported by this tool. Also, a comparison of certain biclusters is not possible in this tool because only one bicluster at a time is represented.

ExpressionView [14] is another R package that allows heatmap-based browsing of biclusters obtained from gene expression experiments. The tool uses an ordering that maximizes the areas of the largest contiguous parts of biclusters. Again, the ordering is fix and the user is not allowed to change that.

3 Bicluster Viewer

In this paper we present the Bicluster Viewer for visualizing biclustering results based on gene expression data. We use and extend traditional heatmap representations and parallel-coordinate plots such that more than one bicluster can be visualized simultaneously. In addition, we support contiguous representation of selected biclusters by allowing row and column duplications.

3.1 Heatmap Visualization

Heatmaps are a good choice to represent a large portion of the data as an overview in a single static view. Many implementations of the heatmap use multi-hue colormaps to indicate up- and downregulated genes separately. In this work, we map data values to grayscale values using linear interpolation between the smallest and largest value of the data matrix, as changes in single-hue colormaps are perceived more accurately than red to green color scales for continuous data values. In general, it is not possible to represent more than two biclusters in a way that all of them are located in contiguous regions in the matrix using row and column permutations only (see Figures 1 (a) and (b)).

Bicluster Representation In order to achieve a contiguous representation of more than one (possibly overlapping) bicluster, we allow row and column reordering as well as duplication. Biclusters are highlighted in the heatmap by surrounding rectangles. If biclusters cannot be represented contiguously, they are represented by several rectangles. To distinguish different biclusters perceptually,

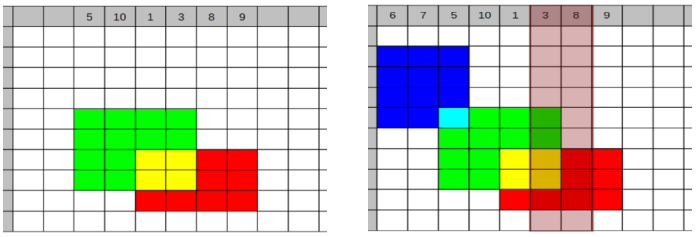


Fig. 1: Schematic example for the insertion of a bicluster without duplication: (a) The green and red colored biclusters are inserted in the matrix and their overlap is represented by a yellow color. (b) The dark blue colored bicluster is inserted in columns $\{3, 5, 6, 7, 8\}$. Without duplication, it is not possible to represent all biclusters in a connected way. The duplicated columns $\{3, 8\}$ are displayed in a transparent red color.

we use a nominal color scale that maps a unique color to each bicluster. For non-contiguous biclusters, only the largest area is displayed by default. The user can decide interactively if biclusters should be represented only by its major rectangle or if all rectangles of a bicluster should be displayed. For the latter, the user gets a representation where the overlap of disconnected biclusters is also visualized. In this mode, only the major rectangle of a bicluster is represented by continuous lines, all other rectangles are displayed with dashed lines.

Selected biclusters are color-coded with a transparent yellow color which is blended additively in overlapping regions. However, the user has the opportunity to choose colors for every bicluster individually. Figure 2 shows the different representations of the heatmap for an example dataset containing six biclusters.

The heatmap in the left part of Figure 3 shows that the light blue colored bicluster is displayed as disconnected rectangular regions. In the right part of the figure, two columns are duplicated to achieve connectivity. These are displayed in a transparent red color. To better distinguish duplicated columns from original columns we insert direct links to the header as shown at the bottom of Figure 3. These can be displayed on user demand. By clicking on the lines the user can interactively highlight all corresponding rows and columns.

While other automatic layout algorithms [4] produce one optimized solution, we developed an algorithm that allows the user to choose one bicluster that should be represented contiguously. Subsequent biclusters are then inserted iteratively into the matrix based on the total size of row- and column-overlap. In every iteration, the bicluster with the largest overlap to the last inserted bicluster will be added to the current set of biclusters, and rows and columns are duplicated to ensure a contiguous representation. By default, the first bicluster that is chosen is the bicluster with the largest total overlap (i.e. the sum of overlaps with all other biclusters).

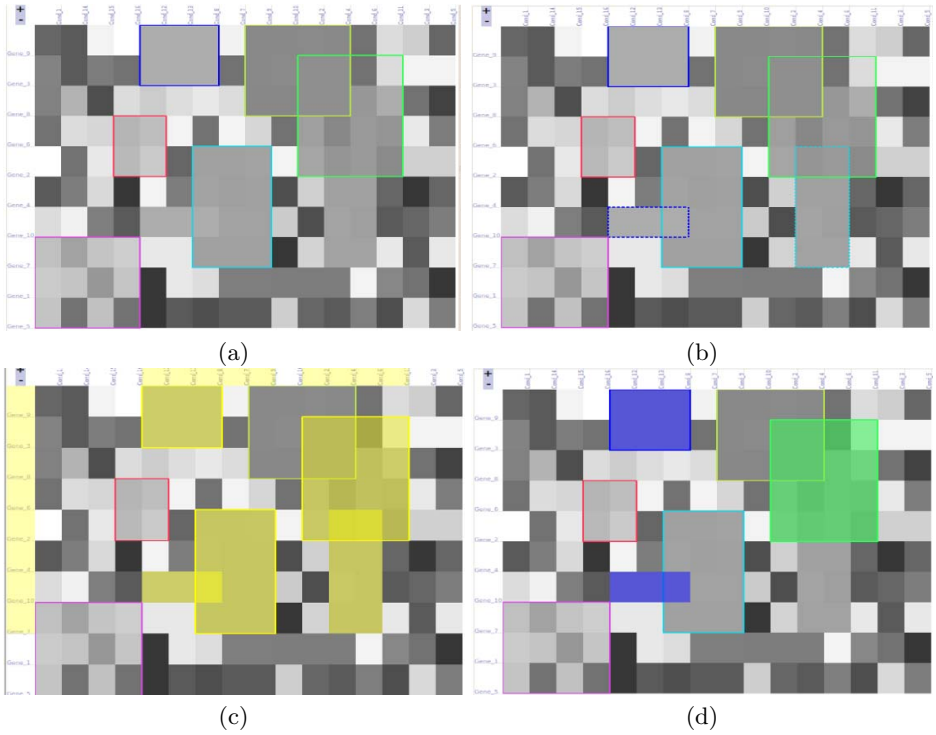


Fig. 2: Different representation modi for the heatmap: (a) Default view: each bicluster represented by its major rectangle only. (b) All biclusters are represented. (c) Representation with three highlighted biclusters. (d) Representation with permanently highlighted biclusters.

3.2 Parallel-Coordinate Plots

We further use parallel-coordinate plots to display biclusters. Our visualization tool exploits linking and brushing techniques to link heatmap and parallel-coordinate plots. This allows a user to explore the data from different points of view.

In the parallel-coordinates plot, each polyline represents the expression of a gene over all conditions (which are represented by vertical axes). Genes belonging to a bicluster are rendered using the same color as the corresponding bicluster in the heatmap. The axes of the parallel-coordinates plot are arranged in the same order as the columns in the heatmap representation. In order to visualize the conditions (axes) of genes belonging to a bicluster, we compute the average vertical position of all lines of a bicluster halfway between adjacent conditions if at least one of the conditions is part of the bicluster. Then, the corresponding lines are forced to cross this point, which we call the centroid. Figures 4 (a)-(c) illustrate both visual metaphors in a heatmap representation and a parallel-coordinates plot for two biclusters.

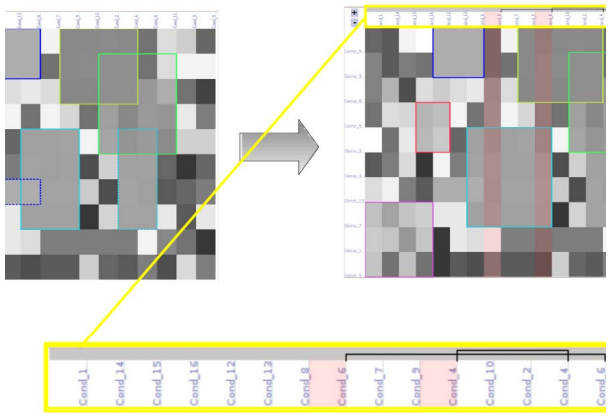
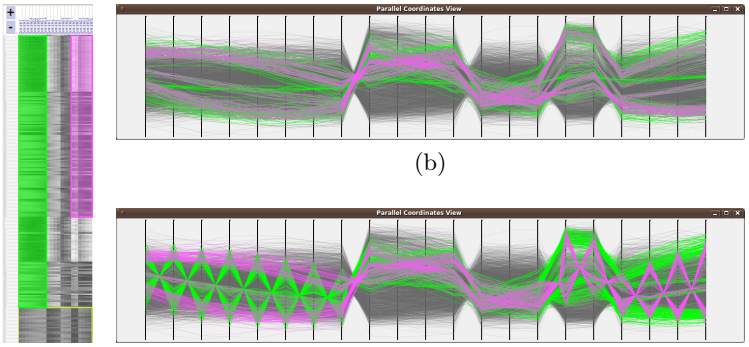


Fig. 3: Column duplication to achieve bicluster connectivity.

3.3 Interactive Features

The BiCluster viewer supports many interactive features. Figure 6 shows how the graphical user interface of the BiCluster viewer is structured and Figure 5 (a) shows a screenshot of the visualization tool in its biclustering mode.

- **Navigating the heatmap** In the upper part of the heatmap, column descriptions are displayed and row descriptions are represented in the first column. A zooming function can be applied by using the *Plus(+)* or the *Minus(-)* sign in the upper left part of the frame to select the zooming factor. The zooming factor can also be changed by pressing the *Ctrl* key on the keyboard and moving the mouse wheel. If the heatmap cannot be represented completely, scroll bars are displayed automatically.
- **Selection of biclusters** By clicking the left mouse button within a highlighted bicluster rectangle, the corresponding biclusters are selected and highlighted with a transparent yellow box. Overlapping biclusters are represented using additive blending. The corresponding rows and columns are also highlighted in yellow. Additionally, an information is shown about which biclusters are currently selected, see Figure 5 (b).
- **Bicluster navigation list** The navigation list in the right part of the application serves as an overview of all selected biclusters in the heatmap. Furthermore, it highlights selected biclusters in yellow. Additionally, an overview is shown with information about which bicluster is currently duplicated. This is colored in red. It is also possible to select a bicluster from the navigation list and selections are linked to the heatmap (see Figure 5 (c)).
- **Duplication of rows and columns** In some cases it is impossible to display all connected biclusters in a single view. For this reason we allow a duplication of involved rows and columns of single biclusters. To show a certain bicluster as a connected entity, a context menu has to be opened by



(a) (b) (c)

Fig. 4: Mapping biclusters from a heatmap representation to parallel-coordinate plots: (a) Two biclusters in a heatmap. (b) The same two biclusters in a parallel coordinates plot without centroids. (c) The same two biclusters in a parallel coordinates plot with centroids.

rightclicking in the corresponding rectangle. By selecting *expand duplicated* the required rows and columns are displayed as duplicated. These rows and columns are then highlighted in a transparent red color, see Figure 5 (d).

– **Menu entries** If not all biclusters can be represented in a connected way the menu entry *show all biclusters* can be used to display rectangles with dashed lines, see Figure 5 (e). All cells that belong to a bicluster are represented by a rectangle with dashed lines in the corresponding bicluster color. The layout of the heatmap is generated automatically by the tool. It may happen that certain biclusters cannot be displayed as connected entities without row or column duplication. By selecting the menu entry *setup fit policy* a dialog will be opened where the user has an impact on the layout, i.e. on the order of rows and columns. The order of the heatmap is built by successively adding all biclusters into the heatmap representation.

4 Case Study

We conducted a small case study to show how the BiCluster Viewer can be used to visualize biclusters. We loaded the yeast data and biclusters from [15] into our tool and used the bicluster navigation list to sort all biclusters by score. After resorting the heatmap according to this ordering, we selected the highest ranking bicluster B_{87}. While the heatmap nicely shows the dimensions of the bicluster, its pattern cannot be seen from the heatmap. Highlighting the bicluster and displaying it in parallel coordinates resolves this issue. Figure 6 shows a screenshot of the tool with the heatmap in the background and parallel coordinates on top. As can be seen, B_{87} exhibits a pattern only for the five rightmost conditions, which is nicely illustrated in parallel coordinates.

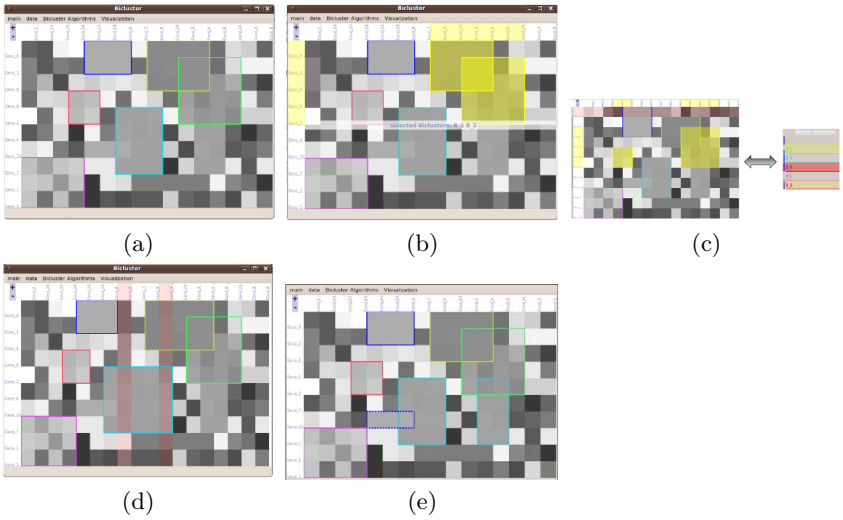


Fig. 5: The BiCluster Viewer in its: (a) Biclustering mode. (b) Biclustering selection mode. (c) Navigation list mode. (d) Duplication mode. (e) Dashed rectangles mode.

5 Pilot Study

To show the usefulness of our technique we conducted a small pilot study with 4 participants. All subjects were researchers from our institute. The pilot study is also used to demonstrate the interactive features that are supported by the visualization tool.

5.1 Study Design

The study was conducted with an Intel Core2 Duo notebook at a frequency of 2 GHz and a 2 Gigabytes RAM. The functionality of our visualization tool was explained to the participants and they got a short introduction about biclusters. They were shown printed tutorials to understand the tool and the single steps in the study. After reading the tutorial, participants were introduced by working with the interactive features of the visualization tool. In the training phase they could work with the tool as long as they wanted and they were allowed to ask questions to the experimenter. Finally, subjects were asked questions to uncover if they understood the visualization and if they could navigate in the visualization tool to answer the given tasks correctly.

The goal of the pilot study was to test the interactive visualization and analysis tool for the bicluster representation. For this reason, participants were presented different data sets. The overall task for the participants was to explore preliminary defined biclusters based on these data sets. We outline four major categories of phenomena to be tested:

– How accurately are biclusters distinguished from each other?

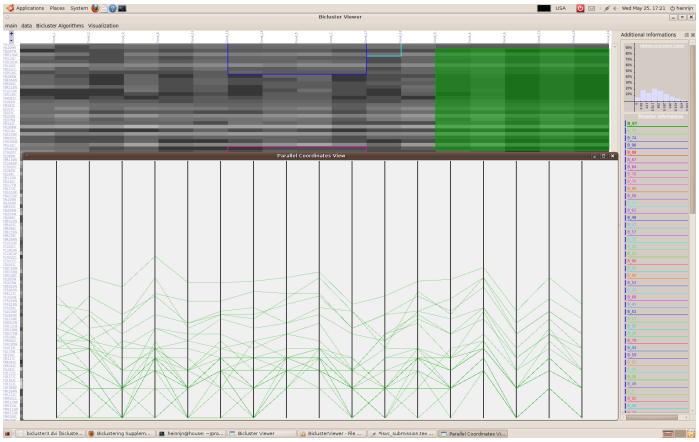


Fig. 6: The GUI of the Bicluster Viewer proposes many interactive features to explore gene expression data and to navigate in it.

- How accurately are overlaps between biclusters detected?
- How accurately are biclusters mapped to their rows and columns?
- How accurately are single biclusters analyzed?

5.2 Tasks

Participants had to perform six tasks where each of the tasks consisted of a number of subtasks. Tasks 1 to 3 have to be solved without using the BiCluster Viewer and its interactive features whereas tasks 4 to 6 had to be answered using the tool. The following tasks and subtasks were performed in the study:

- **Task 1:** Biclustering results are shown for three different data sets, see Figure 7 (a)-(c).
  1. How many biclusters can be found in each of the three representations?
  2. In which representation is the largest overlap, i.e., the most cells belonging to more than one bicluster?
- **Task 2:** Biclustering results are shown for three different data sets. Now, the option *Show All Biclusters* is set, see Figure 7 (d)-(f).
  1. How many biclusters can be found in each of the three representations?
  2. In which representation is the largest overlap, i.e., the most cells belonging to more than one bicluster?
- **Task 3:** Biclustering results are shown for three different data sets. Now, the option *Show All Biclusters* is not set but all biclusters were selected beforehand, see Figure 7 (g)-(i).
  1. How many biclusters can be found in each of the three representations?
  2. In which representation is the largest overlap, i.e., the most cells belonging to more than one bicluster?

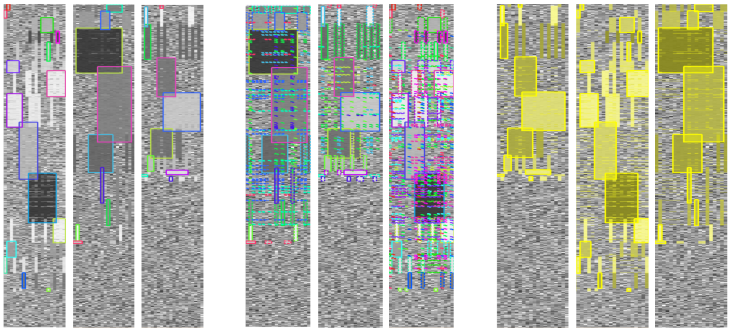


Fig. 7: Biclustering results for three different data sets: (a) 15 biclusters. (b) 9 biclusters. (c) 10 biclusters. (d) 9 biclusters with option *Show All Biclusters*. (e) 10 biclusters with option *Show All Biclusters*. (f) 15 biclusters with option *Show All Biclusters*. (g) 10 biclusters with option *Show All Biclusters* not set but all clusters selected beforehand. (h) 15 biclusters with option *Show All Biclusters* not set but all clusters selected beforehand. (i) 9 biclusters with option *Show All Biclusters* not set but all clusters selected beforehand.

- **Task 4:** A listing of all shown biclusters with the corresponding color codings is given. Write down the bicluster ID and the row and column number.
  1. Which is the largest bicluster?
  2. Are there overlapping biclusters? If so, which?
  3. Which column contains the most biclusters?
  4. Are there columns that do not contain any biclusters? If so, which?
- **Task 5:** A listing of all shown biclusters with the corresponding color codings is given. Write down the bicluster ID and the row and column number.
  1. Which bicluster has the most overlappings with bicluster B_8?
  2. Find a cell (give row and column) that is at least inside three biclusters. Write down the IDs of the biclusters in which the cell is located in.
  3. A certain bicluster has to be analyzed more accurately. Determine all biclusters with which B_9 has rows **AND** columns in common. Write down the corresponding row and column descriptions.
- **Task 6:** A listing of all shown biclusters with the corresponding color codings is given. Write down the bicluster ID and the row and column number.
  1. Find the bicluster that has an overlap with most other biclusters. Write down the IDs of the biclusters.

5.3 Study Results

The results of our pilot study are represented in Figure 8. The blue colored bars show the sum of the correct answers of all participants for each task. The maximal number of correct answers is limited by 4 (because of the 4 participants). Partially correct answers have not been taken as correctly answered. The red

line shows the average completion time for each task. Because of the limited number of participants we cannot deduce any statistical results from the study but we can see that there are already some trends in the different tasks. The additionally recorded comments of the participants are a good basis for a first analysis of the applicability of our interactive visualization tool.

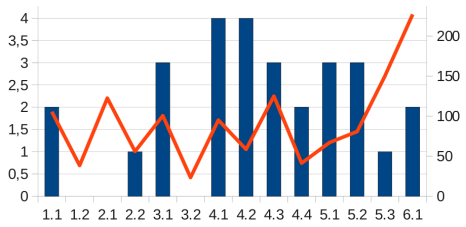


Fig. 8: Correct answers and average completion times of the pilot study.

We found out that the representation with the dashed lines is not suited to distinguish or count single biclusters (no correct answers). Surprisingly, the degree of overlap has been interpreted incorrectly in most cases. The reason for that is the fact that participants weighted the connected overlaps more than those that are highlighted by the dashed rectangles. The representations from Tasks 1 and 3 (standard and select all) are suited well to distinguish biclusters. The wrong answers are based on the fact that the representations are interpreted incorrectly.

6 Conclusion and Future Work

In this paper we introduced the BiCluster Viewer, an interactive visualization tool for analyzing gene expression data. Biclusters are extracted from tabular data by applying a biclustering algorithm and a layout is computed by allowing a minimal number of row and column duplications. We use a heatmap representation in which all generated biclusters can be displayed simultaneously in a connected way by allowing these kinds of duplications. parallel-coordinate plots are used to have another point of view to the same data set and linking and brushing techniques additionally support a viewer to link both views together with the goal to get even more insights in the data than by using a single visual metaphor alone. The BiCluster Viewer contains many interactive features such as ordering functions, color codings, zooming, or details-on-demand to explore the data. We demonstrated the usefulness the tool in a case study by showing insights from a yeast data set. A small user study with 4 participants illustrates which features of the tool are easily understood and used accurately and efficiently.In future we plan a more sophisticated user study with a larger number of participants.

Acknowledgment

The authors would like to thank the German Research Foundation (DFG) for financial support of the project within the Cluster of Excellence in Simulation Technology (EXC 310/1) at the University of Stuttgart.

References

1. Shneiderman, B.: The eyes have it: A task by data type taxonomy for information visualizations. In: Proceedings of the IEEE Symposium on Visual Languages. (1996) 336–343
2. Eisen, M., Spellman, P., Brown, P., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences **95** (1998) 14863–14868
3. Inselberg, A., Dimsdale, B.: Parallel coordinates: A tool for visualizing multi-dimensional geometry. In: Proceedings of IEEE Visualization. (1990) 361–378
4. Grothaus, G., Mufti, A., Murali, T.: Automatic layout and visualization of biclusters. Algorithms for Molecular Biology **1** (2006)
5. Santamaria, R., Theron, R., Quintales, L.: A visual analytics approach for understanding biclustering results from microarray data. Bioinformatics **9** (2008)
6. Cheng, K., Law, N., Siu, W., Liew, A.C.: Biclusters visualization and detection using parallel coordinates plots. In: Proceedings of the International Symposium on Computational Models for Life Sciences. (2007)
7. Cheng, K., Law, N., Siu, W., Lau, T.: BiVisu: Software tool for bicluster detection and visualization. BMC Bioinformatics **23** (2007) 2342–2344
8. Cheng, K.O., Law, N.F., Siu, W.C., Liew, A.: Identification of coherent patterns in gene expression data using an efficient biclustering algorithm and parallel coordinate visualization. BMC Bioinformatics **9** (2008) 210–238
9. Rosenholtz, R., Li, Y., Mansfield, J., Jin, Z.: Feature Congestion: A Measure of Display Clutter. In: Proceedings of SIGCHI Conference on Human Factors in Computing Systems, ACM Press (2005) 761–770
10. Dietzsch, J., Heinrich, J., Nieselt, K., Bartz, D.: SpRay: A visual analytics approach for gene expression data. In: IEEE Symposium on Visual Analytics Science and Technology. (2009) 179–186
11. Kaiser, S., Santamaria, R., Theorn, R., Quintales, L., Leisch, F.: Bicluster algorithms. http://cran.r-project.org/web/packages/biclust/biclust.pdf (2009)
12. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. (2011) ISBN 3-900051-07-0.
13. Barkow, S., Bleuler, S., Zitzler, E., Prelic, A., Frick, D.: BicAT: Biclustering analysis toolbox, ETH Zürich. http://www.tik.ethz.ch/sop/bicat/?page=bicat.php (2010)
14. Luscher, A.: ExpressionView. http://www2.unil.ch/cbg/index.php?title=ExpressionView (2010)
15. Cheng, Y., Church, G.: Biclustering of expression data. In: Proceedings of International Conference on Intellegent Systems for Molecular Biology. (2000) 93–103