Computational Biology and Bioinformatics
2023; 11(2): 19-32
http://www.sciencepublishinggroup.com/j/cbb
doi: 10.11648/j.cbb.20231102.11
ISSN: 2330-8265 (Print); ISSN: 2330-8281 (Online)

Visualizing Biclusters of Gene Expression Data and Their Overlaps Based on a Two-Dimensional Matrix Technique

Haithem Aouabed[1, *], Mourad Elloumi[2]

[1]Computer Science Department, Faculty of Economic Sciences and Management, University of Sfax, Sfax, Tunisia

[2]Computer Science Department, Faculty of Computing and Information Technology, University of Bisha, Bisha, Saudi Arabia

Emailaddress:

haithem.abdi@gmail.com (Haithem Aouabed), Mourad.Elloumi@gmail.com (Mourad Elloumi)

[*]Corresponding author

To cite this article:

Haithem Aouabed, Mourad Elloumi. Visualizing Biclusters of Gene Expression Data and Their Overlaps Based on a Two-Dimensional Matrix Technique. *Computational Biology and Bioinformatics*. Vol. 11, No. 2, 2023, pp. 19-32. doi: 10.11648/j.cbb.20231102.11

Received: September 24, 2023; **Accepted**: October 12, 2023; **Published**: October 30, 2023

Abstract: Biclustering is a data mining technique used to analyze gene expression data. It consists of classifying subgroups of genes that behave similarly under subgroups of conditions and can behave independently under other conditions. These discovered co-expressed genes (called biclusters) can help to find specific biological aims like finding characteristics of a specific disease. A large number of biclustering algorithms have been developed. Generally, these algorithms give as output a large number of overlapped biclusters. The visualization of these biclusters is still a non-trivial task. In this paper, we present a new approach to display biclustering results from gene expression data on the same screen. It is based on a two-dimensional matrix where each bicluster is represented as a column and each overlap between a set of biclusters is represented as a row. We illustrated the usefulness of our method with biclustering results from real and synthetic datasets and we compared it to other techniques that concentrate on biclustering overlaps issue. The method is implemented in a web-based interactive visualization tool called *VisBicluster* available at http://vis.usal.es/~visusal/visbicluster.

Keywords: Biclustering Visualization, Two-Dimensional Matrix, Filtering, Overlaps, InfoVis

1. Introduction

Genomic data analysis often involves identifying groups of biological entities (e.g., genes) that exhibit similar behavior under certain conditions. Traditional clustering algorithms group genes with similar expression patterns across all conditions [1-3]. However, biclustering algorithms can identify groups of genes that coexpress only under a subset of conditions, which can be more informative for certain biological questions [4]. Biclustering has two main theoretical advantages over traditional clustering: Bidimensionality; Biclustering groups both genes and conditions together, while traditional clustering only groups genes or conditions. This allows biclustering to identify more complex patterns in the data. Overlap; Biclustering allows genes to belong to multiple biclusters, while traditional clustering only allows genes to belong to a single cluster. This is more realistic, as genes can be involved in multiple biological processes.

Biclustering has been successfully applied to a wide range of genomic data analysis tasks, including identifying differentially expressed genes or discovering co-regulated genes. Biclustering is a powerful tool for genomic data analysis, and new biclustering algorithms are being developed all the time [5]. As genomic data becomes increasingly complex, biclustering is likely to play an even more important role in genomic data analysis in the future.

Visualization techniques are needed to facilitate the extraction of knowledge from the analyzed data since they provide abstract and mental models of information [6]. In fact, visualization exploits visual intelligence to ameliorate our abstract intelligence. By creating interactive visual representations, visualization can exploit human perceptual and cognitive capabilities for solving many kinds of problems [7]. However, resolving all visualization issues needs to consider many research fields at once among them we can cite Human-Computer Interaction (HCI), data mining, and others. Recently, Information Visualization (InfoVis) and Visual Analytics are well-considered in many application areas such as bioinformatics which is not beyond the scope of visualization, to fully achieve the visual representation based on the declared objectives.

Visualizing biclustering results is an interesting process to infer patterns from the expression data [8]. However, given the special characteristics of biclustering, its application to gene expression data often generates a large number of overlapping groups of biclusters, which are very hard to present in an informative way in a single view [9, 10]. In fact, mapping the biclustering results in one visual form is a non-trivial task. The most popular techniques to visualize a single bicluster are heatmaps and parallel coordinates [11-13]. The difficulty arises when a bioinformatician or an analyst wants to visualize a set of biclusters on the same screen [9, 10]. Also, sets and sets-relations visualization techniques have emerged as a possible solution to better visualize bioinformatics data such as biclustering results of gene expression data [14-16].

In this paper, we introduce a novel approach to visualize biclusters and their respective overlaps. Our method is based on a *two-dimensional matrix*, where each bicluster is represented by a column and each overlap between a group of biclusters is represented by a row.

2. Two-Dimensional Matrix Visualization Technique

We detail the main characteristics of our visualization method. We first introduce how biclusters and their corresponding overlaps are depicted. Then, we focus on the detailed view where elements (genes and conditions) of single biclusters or overlaps are represented as heatmaps [17]. Our technique is invented based on a sophisticated combination of a modified set visualization technique used to layout the generated biclusters in a two-dimensional matrix where each bicluster is represented as a column and each overlap between a set of biclusters is represented as a row and a traditional visualization technique which is heatmaps used to visualize single biclusters and overlaps between them as gene expression matrices [18, 14].

2.1. Matrix of Overlaps

In order to convey biclustering overlap in a scalable way, we focused on such overlaps as the main entity in our visualization (see Figure 1).

Three biclusters and their overlaps are represented based on the proposed technique. The three biclusters are represented as a Venn diagram [19] (Figure 1a). Zones 1, 2, and 3 depict exclusive elements (genes and/or conditions) of B1, B2, and B3, respectively. Zone 4 encodes the overlap (shared elements) between B1 and B2. Zone 5 illustrates the overlap between B2 and B3, while Zone 6 depicts the overlap between B1, B2, and B3. Gray circles depict elements of each zone. In addition, our described visualization technique is based on a two-dimensional matrix where individual biclusters are the columns and overlaps between biclusters are the rows (Figure 1b). Each bicluster that participates in any overlap is represented by a cell. The matrix is laid out based on all of the occurring overlaps between biclusters. Only biclusters that are present in an overlap will be represented with colored cells, which are encoded based on a white-to-black color scale. The more genes and conditions two or more biclusters share, the darker the cell color (Figure 1b). We chose the white-to-black color scale because it is the easiest scale for the human eye to perceive hue changes [7]. Rows with only one cell contain genes or conditions that are unique to a particular bicluster, while rows with two or more cells contain genes or conditions that are shared by multiple biclusters. For example, the cell of the intersection between row 1 and B1 depicts genes and conditions of B1 not shared with any other biclusters. It's the same case for rows 2 and 3 which encode exclusive genes and conditions for biclusters 2 and 3. Row 4 depicts the overlap between B1 and B3. Row 5 depicts the overlap between B2 and B3, while row 6 depicts the overlap between all three biclusters B1, B2, and B3. Based on the used color scale, we can note that B1 has the largest number of exclusive elements (genes and conditions), and the overlap between B1 and B2 is also the largest one. With this method of visualization, it is easier to know which biclusters are not overlapped. In our case, there is no exclusive overlap between B1 and B3.

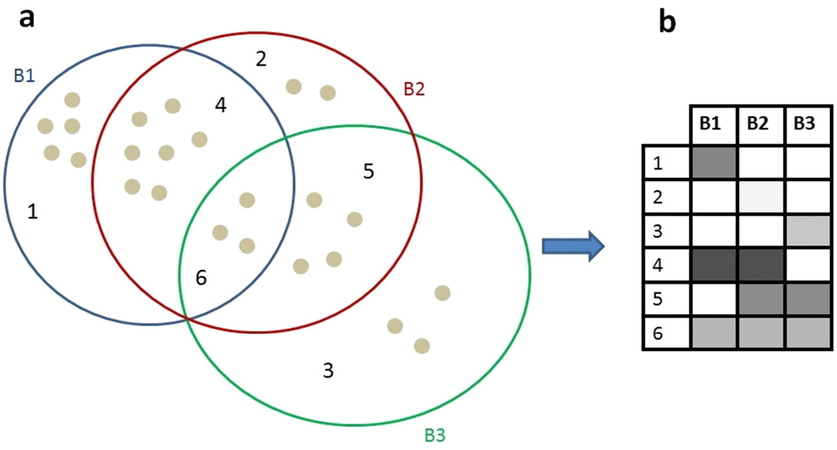


Figure 1. Bicluster visualization concept. (a) Venn diagram representation. (b) Two-dimensional matrix representation. Each column corresponds to a bicluster while rows of the matrix depict possible overlaps. Each intersection, if exists, between a column and a row is shown as a cell.

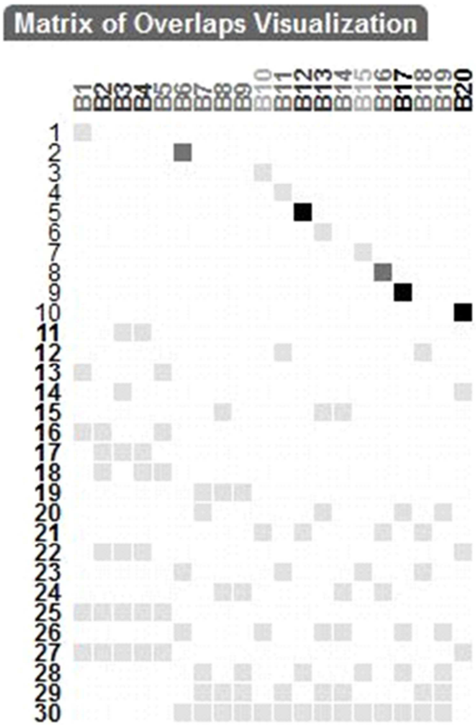

Matrix of Overlaps Visualization
B1 B2 B3 B4 B5 B6 B7 B8 B9 B10 B11 B12 B13 B14 B15 B16 B17 B18 B19 B20
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30

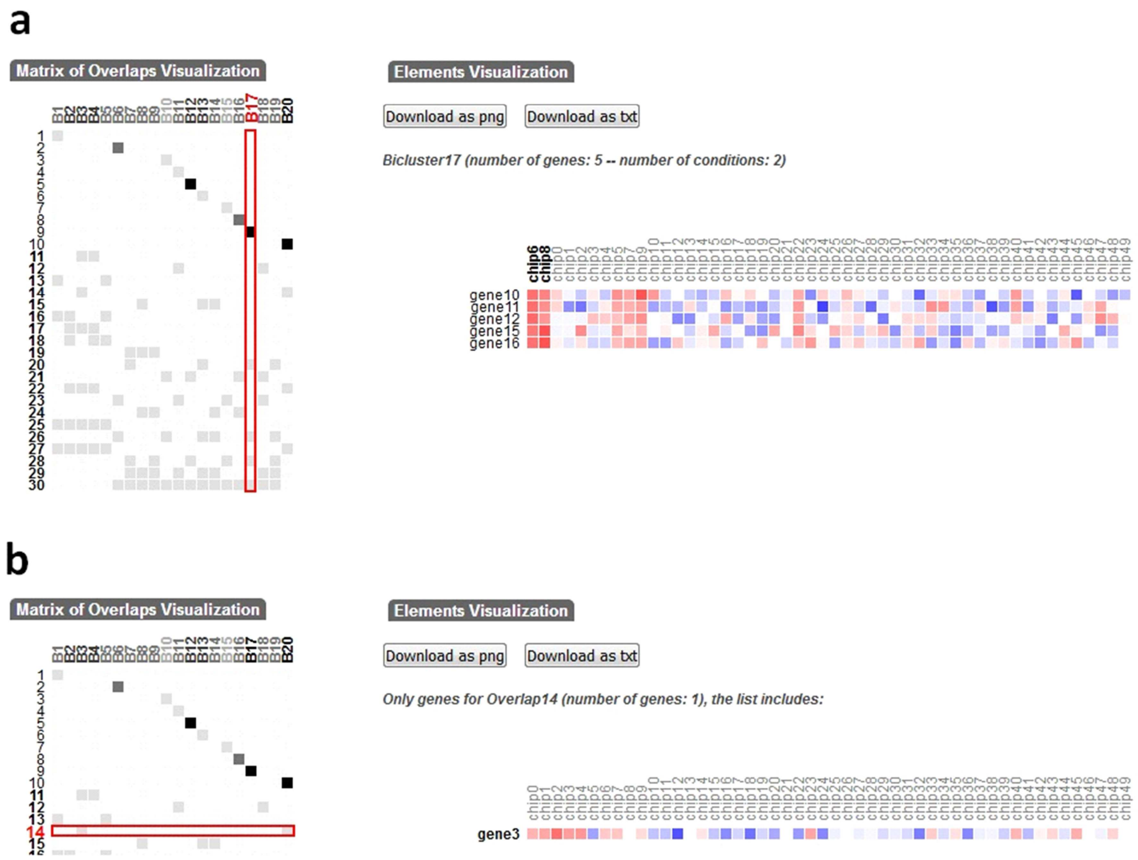
Figure 2. *Overview of the matrix of overlaps visualization showing biclustering results with 20 biclusters and a total number of overlaps, equal to 30. The Figure shows data from a synthetic microarray example [21] and the biclusters generated by Bimax biclustering algorithm [22].*

Two-dimensional matrix visualization of biclustering results is a clear and scalable approach, especially for large numbers of biclusters. It avoids the clutter of graph representations and depicts biclusters as simple columns with minimal space requirements [20, 16].

Figure 2 shows an example of a biclustering result with 20 generated biclusters. Cells are encoded based on a white-to-black color scale, as well as the names of columns (i.e., names of biclusters). So, the more genes and conditions a bicluster contains, the darker is its name.

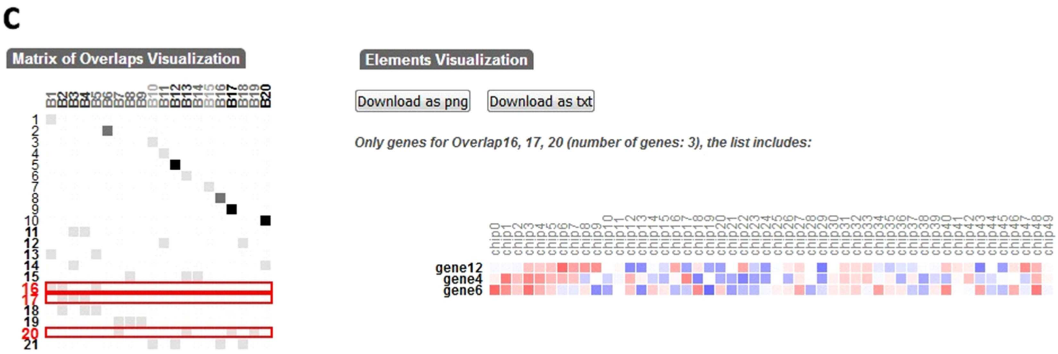
2.2. Selection and Highlighting

VisBicluster supports linking and brushing. Hence, selected columns or rows of the overlap matrix are shown on the right side of the interface as heatmaps. In fact, a click on column names shows the corresponding gene expression profile as a heatmap. Analysts can visualize the heatmap of an overlap or set of overlaps. To select rows, the analyst needs to click on each number of chosen rows. Then, double-clicking on one of the selected rows shows the corresponding heatmap (see Figure 3). To show the heatmap of one row, the analyst can either double-click on its number or click in any colored cell of the selected row. With such a technique, analysts can identify easily which biclusters or overlaps a single gene or condition is associated with.


a
Matrix of Overlaps Visualization
B1 B2 B3 B4 B5 B6 B7 B8 B9 B10 B11 B12 B13 B14 B15 B16 B17 B18 B19 B20
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Elements Visualization
Download as png
Download as txt
chip6 chip8 chip0 chip1 chip2 chip3 chip4 chip5 chip7 chip9 chip10 chip11 chip12 chip13 chip14 chip15 chip16 chip17 chip18 chip19 chip20 chip21 chip22 chip23 chip24 chip25 chip26 chip27 chip28 chip29 chip30 chip31 chip32 chip33 chip34 chip35 chip36 chip37 chip38 chip39 chip40 chip41 chip42 chip43 chip44 chip45 chip46 chip47 chip48 chip49
gene10
gene11
gene12
gene15
gene16
b
Matrix of Overlaps Visualization
B1 B2 B3 B4 B5 B6 B7 B8 B9 B10 B11 B12 B13 B14 B15 B16 B17 B18 B19 B20
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16
Elements Visualization
Download as png
Download as txt
chip0 chip1 chip2 chip3 chip4 chip5 chip6 chip7 chip8 chip9 chip10 chip11 chip12 chip13 chip14 chip15 chip16 chip17 chip18 chip19 chip20 chip21 chip22 chip23 chip24 chip25 chip26 chip27 chip28 chip29 chip30 chip31 chip32 chip33 chip34 chip35 chip36 chip37 chip38 chip39 chip40 chip41 chip42 chip43 chip44 chip45 chip46 chip47 chip48 chip49
gene3

Bicluster17 (number of genes: 5 -- number of conditions: 2)

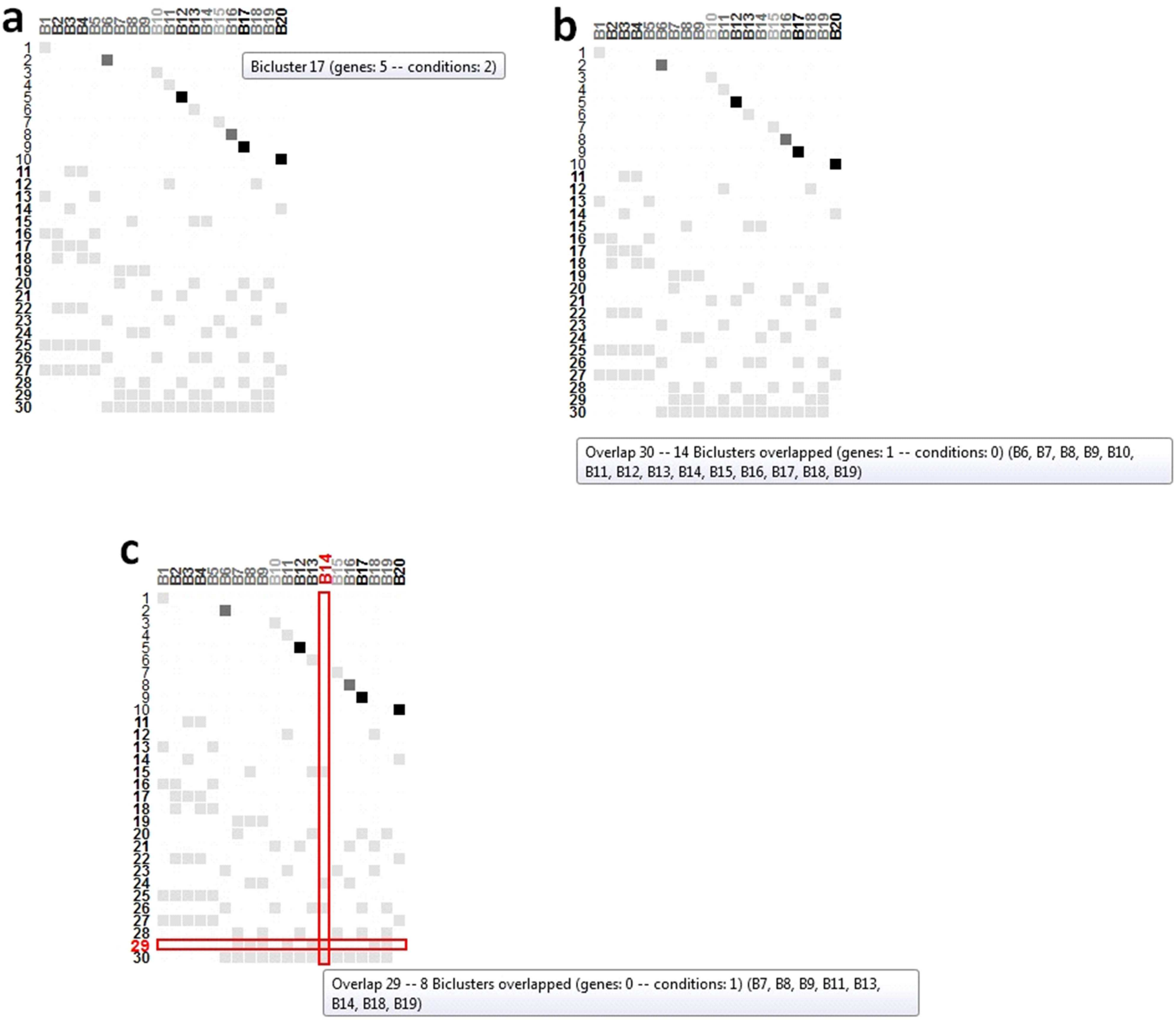
Only genes for Overlap14 (number of genes: 1), the list includes:



Only genes for Overlap16, 17, 20 (number of genes: 3), the list includes:

Figure 3. *Selection results. (a) Heatmap of bicluster17. (b) Heatmap of overlap14. (c) Heatmap of overlaps 16, 17, and 20.*

Highlighting biclusters and overlaps on hover provides analysts with valuable insights into the data like knowing the number of genes, the number of conditions, the participated biclusters in an overlap, the number of participated biclusters in an overlap, and the name of a bicluster. When hovering over a column name, the analyst can see as a tooltip the bicluster name as well as its size (number of genes and conditions). When hovering over a row number, the analyst can see the number of overlaps, the total number of overlapped biclusters, the overlap size, and a list of the participating biclusters. Similarly, hovering over a filled cell shows the overlap details and highlights the corresponding column and row of the matrix (see Figure 4).



Bicluster 17 (genes: 5 -- conditions: 2)

Overlap 30 -- 14 Biclusters overlapped (genes: 1 -- conditions: 0) (B6, B7, B8, B9, B10, B11, B12, B13, B14, B15, B16, B17, B18, B19)

Overlap 29 -- 8 Biclusters overlapped (genes: 0 -- conditions: 1) (B7, B8, B9, B11, B13, B14, B18, B19)

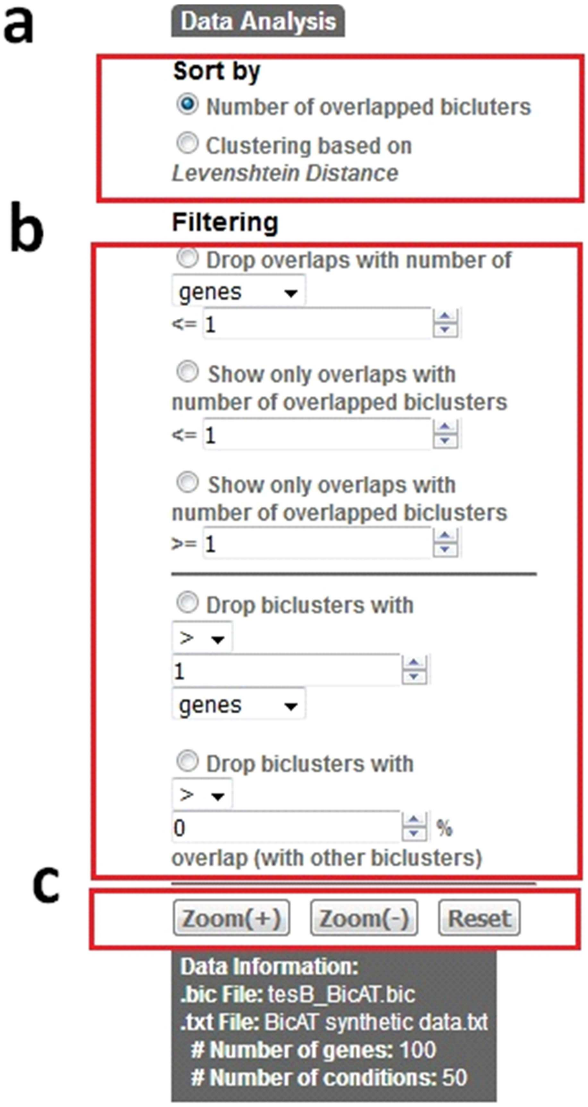
Figure 4. *Details after hovering over. (a) A column name. (b) A row number. (c) A cell. Hovering over a filled cell highlights the column and row that belong to them.*

2.3. Filtering, Sorting, and Zooming

Interactive visualization tools like VisBicluster allow analysts to explore and analyze data in new and innovative ways. By applying filters to the visualized data, analysts can simplify the layout, focus on subsets of results, and gain insights that would be difficult to see otherwise (see Figure 5). VisBicluster offers five kinds of filters that can be used to modify the layout of biclustering results:

1. Minimum genes/conditions number of overlaps: If this filter is chosen with parameter *n*, only overlaps with more than n genes or conditions are drawn.
2. Minimum number of overlapped biclusters: If this filter is chosen with parameter *n*, only overlaps with more or equal to *n* overlapped biclusters are drawn.
3. Maximum number of overlapped biclusters: If this filter is chosen with parameter *n*, only overlaps with less or equal to *n* overlapped biclusters are drawn.
4. Size of biclusters: If this filter is chosen with parameter n, biclusters more/less/equal to *n* genes or conditions are dropped and the matrix is built again with the rest of biclusters.
5. Rate of overlaps between biclusters: If this filter is chosen with parameter *n*%, biclusters with overlap rates with other biclusters above/below/equal to *n*% are dropped and the matrix is built again with the rest of biclusters.

For example, when the analyst fixes the minimum number of overlapped biclusters to 2 and the maximum one to 5 from the 20 biclusters overlapped matrix, the result is a matrix with overlaps between 2 and 5 biclusters (see Figure 6).



Sort by

● Number of overlapped bicluters

○ Clustering based on *Levenshtein Distance*

Filtering

○ Drop overlaps with number of
genes
<= 1

○ Show only overlaps with number of overlapped biclusters
<= 1

○ Show only overlaps with number of overlapped biclusters
>= 1

○ Drop biclusters with
>
1
genes

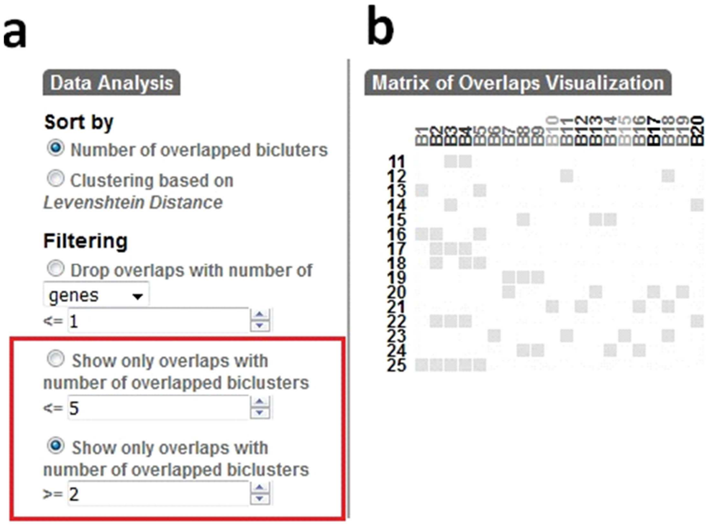
○ Drop biclusters with
>
0 %
overlap (with other biclusters)

Data Information:
.bic File: tesB_BicAT.bic
.txt File: BicAT synthetic data.txt
# Number of genes: 100
# Number of conditions: 50

Figure 5. *Data analysis part. (a) Two sorting criteria. (b) Different possible filters. (c) Zoom controls.*

Our approach also supports sorting and zooming. By default, the data is sorted based on the number of overlapped biclusters. The analyst can define a secondary sorting criterion using a clustering algorithm based on the *Levenshtein distance* [23], which represents a string metric for measuring the difference between two sequences. This allows the analyst to explore the data from different perspectives and identify patterns and trends that may not be visible otherwise (see Figure 7). The analyst can control the zoom level of the biclustering results matrix using the zoom controls in the data analysis section of the global visualization interface. The analyst can zoom in, zoom out, or reset the visualization, which will remove all zoom and filter interactions.

Our approach also supports sorting and zooming. By default, the data is sorted based on the number of overlapped biclusters. The analyst can define a secondary sorting criterion using a clustering algorithm based on the *Levenshtein distance* [23], which represents a string metric for measuring the difference between two sequences. This allows the analyst to explore the data from different perspectives and identify patterns and trends that may not be visible otherwise (see Figure 7). The analyst can control the zoom level of the biclustering results matrix using the zoom controls in the data analysis section of the global visualization interface. The analyst can zoom in, zoom out, or reset the visualization, which will remove all zoom and filter interactions.


a
b
Matrix of Overlaps Visualization
B1 B2 B3 B4 B5 B6 B7 B8 B9 B10 B11 B12 B13 B14 B15 B16 B17 B18 B19 B20
11 12 13 14 15 16 17 18 19 20 21 22 23 24 25

Data Analysis

Sort by

● Number of overlapped bicluters

○ Clustering based on *Levenshtein Distance*

Filtering

○ Drop overlaps with number of

genes ▾

<= 1

○ Show only overlaps with number of overlapped biclusters

<= 5

● Show only overlaps with number of overlapped biclusters

>= 2

Figure 6. *Result from two kinds of filters. (a) The minimum number of overlapped biclusters is fixed to 2 and the maximum is fixed to 5 (red rectangle). (b) The resulting matrix of overlaps.*

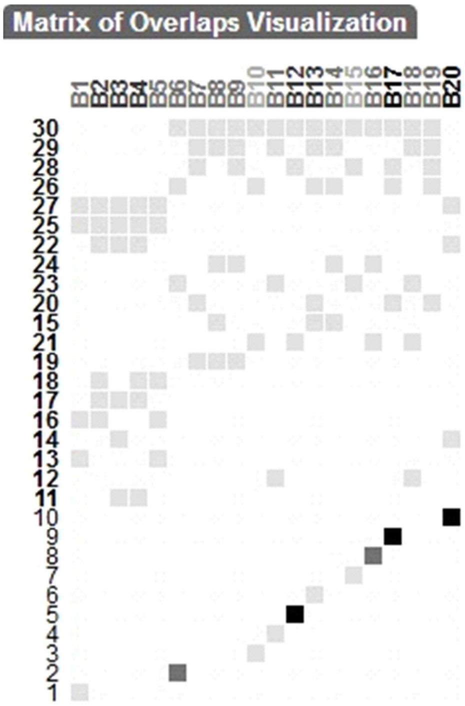
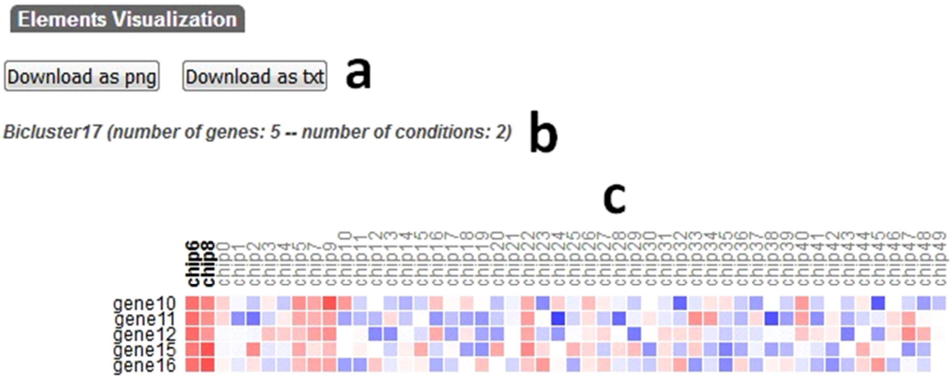

Matrix of Overlaps Visualization
B1 B2 B3 B4 B5 B6 B7 B8 B9 B10 B11 B12 B13 B14 B15 B16 B17 B18 B19 B20
30 29 28 26 27 25 22 24 23 20 15 21 19 18 17 16 14 13 12 11 10 9 8 7 6 5 4 3 2 1

Figure 7. *Matrix overlaps sorted by Levenshtein distance [23]. Three groups of similar rows (overlaps) can be released from visualization. Group 1 contains rows 29 and 30, group 2 contains rows 22, 25, and 27, and group 3 contains rows 16, 17 and 18. This means that these three groups are very similar on the side of overlapped biclusters.*

2.4. Detail Visualization

In addition to the matrix of overlapped bicluster overview, the proposed technique allows the visualization of single biclusters or overlaps as a heatmap in the detail view part. We chose heatmap representation as it is the most intuitive and widely used technique to visualize this type of biological data. Heatmaps provide a dense and visually appealing way to represent the complex relationships between genes and samples in bicluster data. In fact, heatmap visualization is a must-be in any visual analytics approach to gene expression analysis. Analysts are used to it and it directly conveys the idea of microarray: color intensities on arrayed spots [24]. Cells of heatmaps, which depict transcription levels of genes under each condition, are drawn based on a blue-white-red color scale since the typical green-black-red scale is not suitable for perceiving changes in hue [7]. Also, to improve time performance and to save screen real estate, gene expression matrices are not fully displayed although we fix the maximum number of rows to 2000 in order to display large groups of genes, which can change with the selections performed along the analysis. The overall aspect of the data can be perceived by just a sample of the original matrix, and the real utility of a heatmap comes with the filtering and reordering of rows and columns based on analysis results.

By highlighting a cell in the heatmap, analysts can view the corresponding transcription level, gene name, and condition name. Additionally, analysts can bring a bicluster, a single overlap, or a set of overlaps into the focus of the heatmap visualization, which can help to identify patterns and relationships in the data. Heatmaps can be rearranged to better understand the context of gene profiles, and analysts can export the data visualized as heatmaps either as text or image files (see Figure 8).



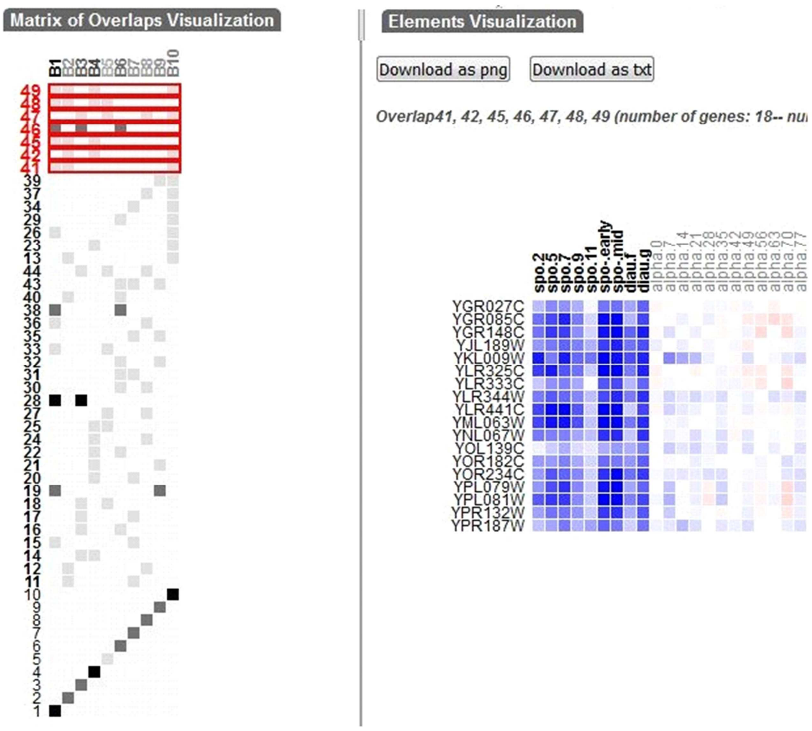
Bicluster17 (number of genes: 5 -- number of conditions: 2)

Figure 8. *Detail visualization of bicluster17. (a) Two buttons are used to download the heatmap data as text or image files. (b) Bicluster number and its size (number of genes and conditions). (c) Heatmap visualization. Only genes of the bicluster are represented. Conditions of the bicluster are rearranged on the left and are in bold.*

3. Results from a Real Dataset

We demonstrate the usefulness of our bicluster visualization method by a representative analysis of real dataset encoding results from yeast *Saccharomyces cerevisiae* microarray data [1]. We use this dataset because it has been broadly studied by biologists and images of heatmap clustering are available. Also, it is well cited in the literature and available from well-known web repositories such as ArrayExpress or Gene Expression Omnibus (GEO).

The dataset contains the gene expression values of 2467 transcripts for 79 samples that have been analyzed under Plaid model biclustering algorithm [25] which uses a statistical model to identify the distribution parameters and generate the data by minimizing a certain criterion iteratively to find the best biclusters, if they exist. It is executed under its default parameters, only the verbose parameter was fixed to false. Ten biclusters are yielded. Figure 9 shows the results as two-dimensional matrices.



Overlap41, 42, 45, 46, 47, 48, 49 (number of genes: 18-- nu

Figure 9. *Plaid model result visualization for yeast Saccharomyces cerevisiae expression data. Overlaps are depicted as a two-dimensional matrix where a set of similar overlaps is selected. The matrix of overlaps is clustered by Levenshtein distance [23] (on the left). Genes and conditions of a set of similar selected overlaps are represented as a heatmap (18 genes and 9 conditions). Genes of this group are down-regulated under sporulation and diauxic shift samples, which are marked in bold (on the right).*

Since the Plaid model searches for additive coherent evolution biclusters, we observe based on the heatmap visualization the existence of biclusters where their genes are over-expressed under considered conditions such as biclusters 2 or 4 while there are other biclusters where their corresponding heatmaps show genes that are under-expressed with the corresponding active conditions. This observation reflects one of the characteristics of Plaid algorithm.

After clustering the two-dimensional matrix of overlaps, we find a similar group of overlaps (see Figure 9 on the left) which contains genes and conditions from 8 out of the 10 overlapped biclusters (biclusters 1 to 6, 8 and 10). These overlaps contain transcripts that are under-expressed according to sporulation and diauxic shift conditions (see Figure 9, right part). We mention that genes of this similar group with locus tags *YLR441C*, *YML063W,* and *YPL081W* are grouped together. These three genes are protein components of the ribosomal subunits *40S*. This explains why they are grouped together in biclusters by Plaid model. In this case, they serve as validation of the method because there is biological evidence of the relation among genes (components of ribosomal subunits) but in other cases, these identifications could lead to new knowledge.

4. Comparison with Other Tools

We evaluated the performance of VisBicluster relative to two other state-of-the-art tools on a range of tasks, which are BicOverlapper [15] and Furby [16]. We chose these tools for four reasons: the first one is that these techniques focus on how to visualize overlaps between biclusters. The Second one is that in general, the number of tools for biclustering visualization is low. Third, the source codes of most of the tools are not available. So, running these tools as executables or scripts is not a straightforward task. Fourth, the chosen tools are the most developed ones. In fact, they implemented new techniques of bicluster visualization in combination with traditional ones like heatmaps or parallel coordinates. Thus, BicOverlapper uses a *Venn-like* representation to visualize biclusters. Biclusters are depicted as irregular surfaces called *hulls*, and overlaps between biclusters are shown by *intersections* of hulls. Groups of genes and conditions that are either on just one bicluster or on specific overlaps are represented by *glyphs*. A glyph is a pie chart divided into sectors, where the number of sectors represents the number of biclusters to which the genes and conditions belong to while Furby depicts biclusters and their overlaps as a *node-link* graph. Biclusters are the *nodes* of the graph and shared genes and conditions between biclusters are considered as *edges* or *bands*. Each bicluster node is depicted as a heatmap matrix, where rows represent genes and columns represent conditions of the corresponding bicluster. Overlaps between each pair of biclusters are encoded using bands that link the corresponding heatmaps at the position of the shared rows and columns. The Venn-like representation provides a more global view of the biclusters, while the node-link graph provides a more detailed view of the relationships between the biclusters

We evaluated the clarity and simplicity of the visualization methods for quickly identifying target information about biclusters and their overlaps. In order to do that, we used a synthetic dataset from Padilha and Campello repository [26] to perform a user study with 14 participants. For the real dataset, we used the *human lung carcinomas* data [27]from the same repository. This dataset is validated through gene set enrichment and clustering accuracy [26].

The physical setup for the comparison process consisted of an Intel Core 2 Duo laptop computer at a frequency of 2 GHz and 3 Gigabytes of RAM.

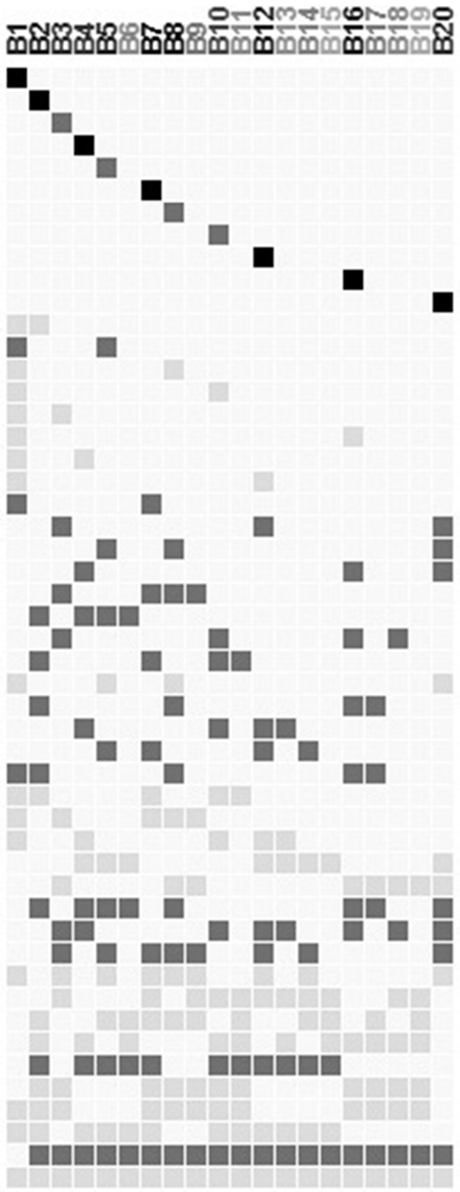


Figure 10. *Bimax bicluster visualization for synthetic data with VisBicluster.*

4.1. Comparison Results from Synthetic Dataset

We chose 14 participants for this study. Six participants (3 male, 3 female) were Ph. D. students with intermediate experience, five participants (3 male, 2 female) were computer science teachers at secondary level school, and three participants (3 male) were senior researchers and practitioners at a computer science faculty. The training session began with a lecture on the relevant functionality of each visualization tool and an introduction to the bicluster notion. After the explanation, participants were invited to work independently with each tool for as long as they wanted, discovering all of its features. During this training phase, participants could ask questions to the instructor. Finally, participants were given a quiz to assess their understanding of the operating principles of each visualization tool and their ability to use them to answer specific tasks correctly.

We used the synthetic dataset generated by Padilha and Campello [26] which is composed of 500 rows and 200 columns. The used biclustering algorithm is Bimax [22]because it has an easy interpretation of biclusters (highly up or down-regulated constant biclusters). It was executed with a binary threshold value equal to 0.5, so only transcription levels that are higher than this threshold are considered. The minimum size of biclusters was set to 2x2, finding 100 biclusters. We focus on the first 20 biclusters. Figure 10 shows the results for VisBicluster while Figure 11 shows a snapshot of the overall visualization result for BicOverlapper and Furby. Inspired by the work of [28], we identified three main categories of tasks to assess, some of which included multiple subtasks [17]:

1. Tasks related to biclusters like find out the total number of overlapped biclusters.
2. Tasks related to overlaps between biclusters. As an example, find out the largest/smallest overlap.
3. Tasks related to bicluster elements (genes and/or conditions) like find out the elements belong to a specific bicluster, that are not integrated into any overlap.

Ten tasks were given to participants, some of which had subtasks. The following list describes the tasks and subtasks that were used in the study:

1. *Task 1*: Find out the total number of overlapped biclusters.
2. *Task 2*: Find out the total number of overlaps.
3. *Task 3*: Find out the list of biclusters that have elements (genes and conditions) not integrated into any overlap.
4. *Task 4*: Analyze overlaps: e g. find out if a certain pair of biclusters or if a certain group of biclusters overlap (i e. have non-empty intersections):

4.1. Identify overlaps between k biclusters (In our test, we ask to find the number of overlaps between two biclusters).

4.2. Identify the biclusters involved in a certain overlap (In our test, the choice of the overlap is up to the participant).

4.3. Identify the biclusters not integrated into any overlap.

5. *Task 5*: Analyze exclusion overlaps: e g. find out if bicluster A does not intersect with bicluster B (In our test, we ask to find out if bicluster 1 has intersections with bicluster 2. If yes, what are the numbers of overlaps between them?).
6. *Task 6*: Find out the largest/smallest overlap (In our test, we ask to find only the largest overlap).
7. *Task 7*: Find out the largest/smallest bicluster (In our test, we ask to find only the largest bicluster).
8. *Task 8*: Analyze and compare bicluster exclusiveness: e g. find out if bicluster A contains more exclusive elements than bicluster B, or more elements shared with 1, 2, 3 or any number of biclusters (In our test, we ask to find out if bicluster 1 has more exclusive elements than bicluster 2).
9. *Task 9*: Analyze elements:

9.1. Find elements that belong to a specific bicluster (Participant's choice).

9.2. Find elements that belong to a specific overlap (Participant's choice).

9.3. Find elements belong to a bicluster that are not integrated into any overlap (Participant's choice).

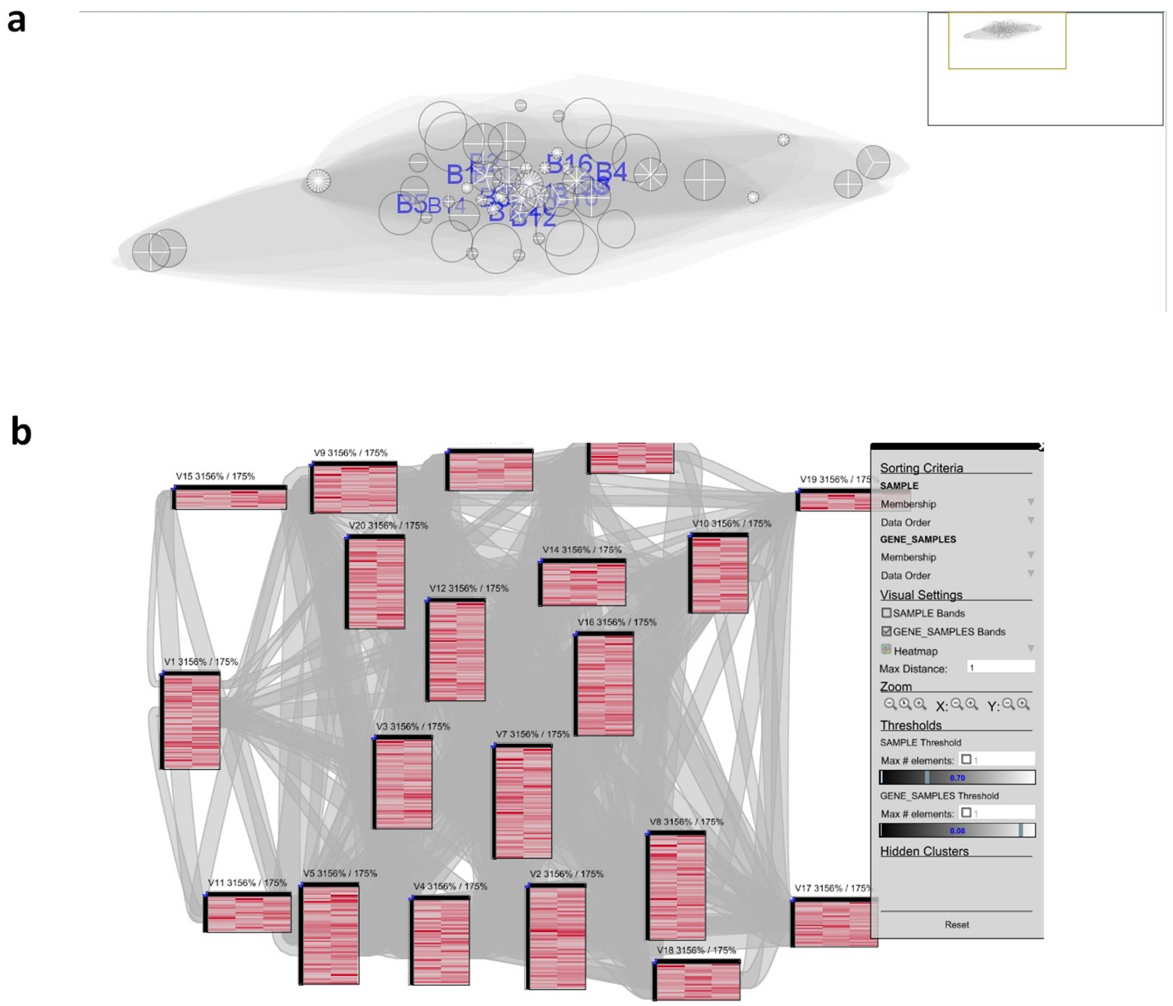
10. *Task 10*: Find elements based on their bicluster memberships: e g. elements in bicluster A and in bicluster B but not in C (In our test, we ask to find out shared elements between biclusters 1 and 2 but are not involved in bicluster 3).

To avoid pre-learned behavior biases from the first tested tools, we randomized the test order for the three tools across all the 14 participants.

The results of the pilot study are summarized in Figure 12.

The average time to answer 8 out of 10 tasks is the lowest one for our tool compared to Furby and BicOverlapper. Furthermore, the global average time of VisBicluster is also the lowest one (1.8s). Our visualization technique is unique in its ability to clearly and concisely display overlaps between biclusters. By using a matrix to represent overlaps, we can sort them in a way that makes them easy to analyze and interpret. This is in contrast to other visualization techniques that use linkage elements to represent overlaps, which can be cluttered and difficult to understand. By avoiding clutter, our visualization makes it easy for users to answer questions about overlaps, such as which biclusters have the most overlap or which genes are involved in the most overlaps.

However, we notice that Furby has the lowest answer time in tasks 1 and 7 (finding the number of biclusters and the largest or smallest one). This can be explained by the fact that Furby's node-link diagram is effective for overview tasks because the nodes that depict biclusters are easily perceived by participants. In contrast, BicOverlapper hulls are very overlapping and extensive, and VisBicluster visualization focuses on the overlaps themselves, making these two approaches less efficient for these tasks. Yet, Furby has the highest answer times for some tasks like tasks 2, 3, 10, and subtask 4.1 since the links between biclusters that represent overlaps are too dense and cluttered, it is impossible to answer these tasks (9 out of 14 participants cannot give answers about these tasks). In contrast, BicOverlapper gives, in general, reasonable results for most of the tasks (global average answer time of 5.5 s). Despite being the only approach that dissolves biclusters in favor of overlaps, BicOverlapper only underperforms Furby and VisBicluster on three tasks (1, 5, and 7), two of which are bicluster-centered. VisBicluster, an overlap-focused biclustering algorithm, keeps bicluster information on a secondary level (as columns in the overlap matrix), addressing this issue.



Sorting Criteria
SAMPLE
Membership
Data Order
GENE_SAMPLES
Membership
Data Order
Visual Settings
□ SAMPLE Bands
☑ GENE_SAMPLES Bands
Heatmap
Max Distance: 1
Zoom
X: Y:
Thresholds
SAMPLE Threshold
Max # elements: □ 1
0.70
GENE_SAMPLES Threshold
Max # elements: □ 1
0.08
Hidden Clusters
Reset

Figure 11. *Bimax bicluster visualization for synthetic data with (a) BicOverlapper and (b) Furby.*

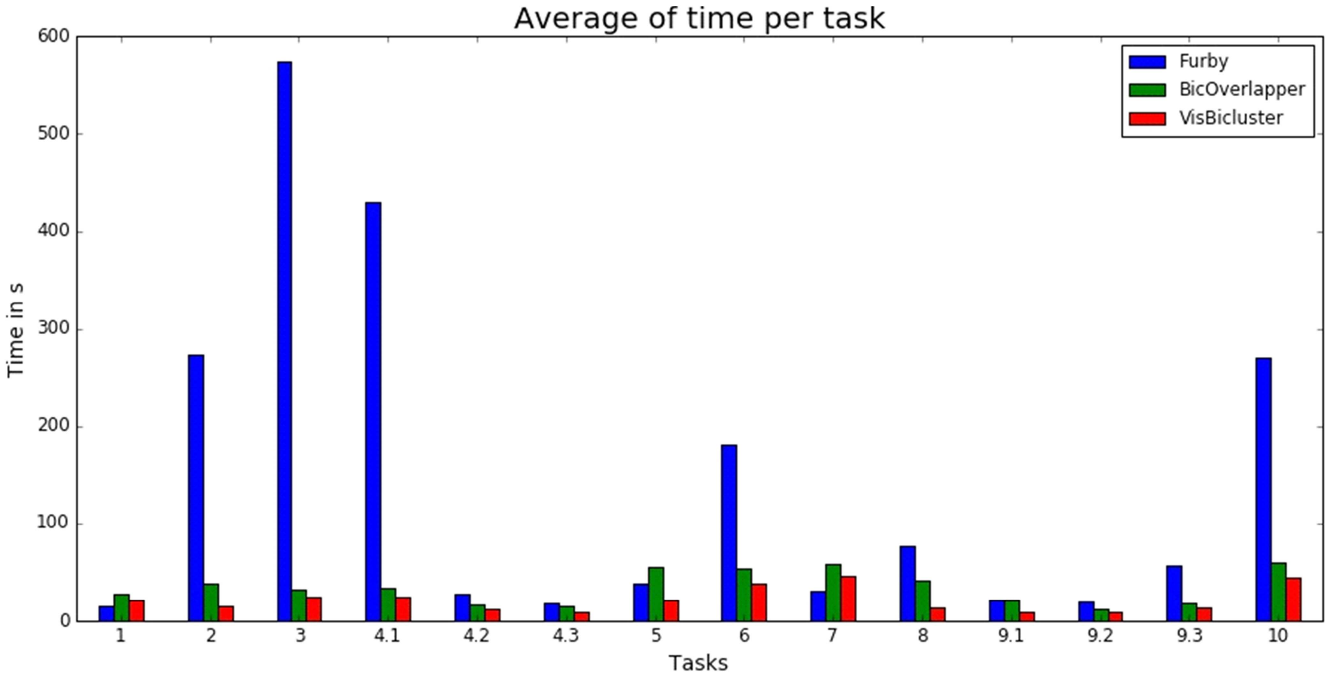


Figure 12. *Answering times of the tools Furby, BicOverlapper, and VisBiclusters for the proposed tasks.*

4.2. Comparison Results from Real Dataset

In order to evaluate our visualization method biologically, we conducted a real case study based on data from the same repository used in the synthetic comparison subsection. From this repository [26], we chose the *human lung carcinomas* dataset [27] from a collection of 35 cancer datasets. The dataset contains the gene expression values of 12600 carcinomas transcripts for 203 snap-frozen lung tumors and normal lung samples that have been analyzed using Bimax algorithm [22]. Because of the exhaustiveness of Bimax and to show a reasonable number of biclusters with reasonable sizes as a result, we fixed the binary threshold parameter to a high value (2000), so only transcription levels that are higher than this threshold are considered. The minimum size of biclusters was set to 2x2, finding 100 biclusters. We consider the first 20 ones. Figure 13 shows the visualization result as a two-dimensional matrix based on VisBicluster while Figure 14 shows visualization results for BicOverlapper and Furby.

From the visualization of Figure 13, we find easily that 8 out of the 20 biclusters with exclusive elements (genes and/or conditions not integrated into any overlap) since they have rows in the matrix of overlaps with only one filled cell while the remaining overlaps are between at least two biclusters. Based on the used color scale, bicluster 1 is the largest one (73 genes and 2 conditions). Also, the overlaps with the high number of genes and conditions are overlaps 20, 38, 49, and overlaps from 52 to 55. Among them, we focus on overlap 55, which therefore can be an interesting candidate for a detailed inspection. By selecting rows of this big overlap and visualizing its corresponding genes and conditions as heatmaps and based on the original paper describing the analysis of this *lung cancer* data [27], we depict that these 35 genes are activated under conditions from all the biclusters. The interesting knowledge that can be deduced from these genes is that they are activated under the two conditions of bicluster 1 (*AD230* and *AD 252*) which are classified in one of the six defined clusters generated from the clustering of the adenocarcinoma human lung tumors [27] using two clustering algorithm; hierarchical [1] and probabilistic model-based clustering [29]. These two conditions are in the class of *proliferation-related* gene expression (C1). The list of defined clusters includes also *colon metastases* cluster (CM), *neuroendocrin*e gene expression cluster (C2), *ornithine decarboxylase 1* and *surfactant* gene expression cluster (C3), *Type II pneumocyte* gene expression cluster (C4) and *normal lung* cluster (NL). We mention that tumors in the C1 cluster express high levels of genes associated with cell division and proliferation which are also expressed under other samples such as squamous cell lung carcinoma (SQ) and SCLC [27].

In conclusion, we notice the importance of bicluster 1 which is the largest bicluster and it is integrated into the largest overlap (number 55). Also, his conditions are interesting in human lung tumors analysis. This would suggest that this bicluster can demonstrate the ability of gene expression analysis to discriminate primary lung adenocarcinomas from metastases of extra-pulmonary origin [27].

Focusing on the visualization of Figure 14, we mention that with BicOverlapper visualization (see Figure 14a), with pie charts usage, inferring the most interesting overlaps is a little bit straightforward but deducing information about biclusters, either their names is not a simple task. This difficulty can be explained by the exhaustiveness of Bimax (high rate of overlaps). For Furby visualization (see Figure 14b), with the node-link diagram representation, biclusters are easily perceived but it's not the same case with overlaps since bands coding them become quickly too cluttered with either a medium rate of overlapping.

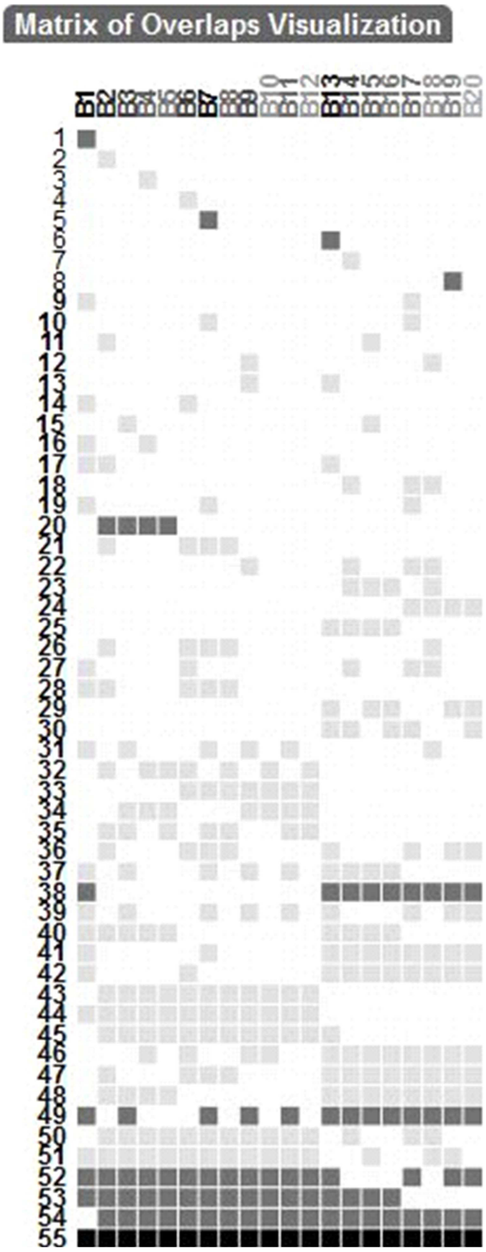


Figure 13. Bimax bicluster visualization for the human lung carcinomas dataset with VisBicluster. Overlap 55 is the most important one based on the defined color scale.

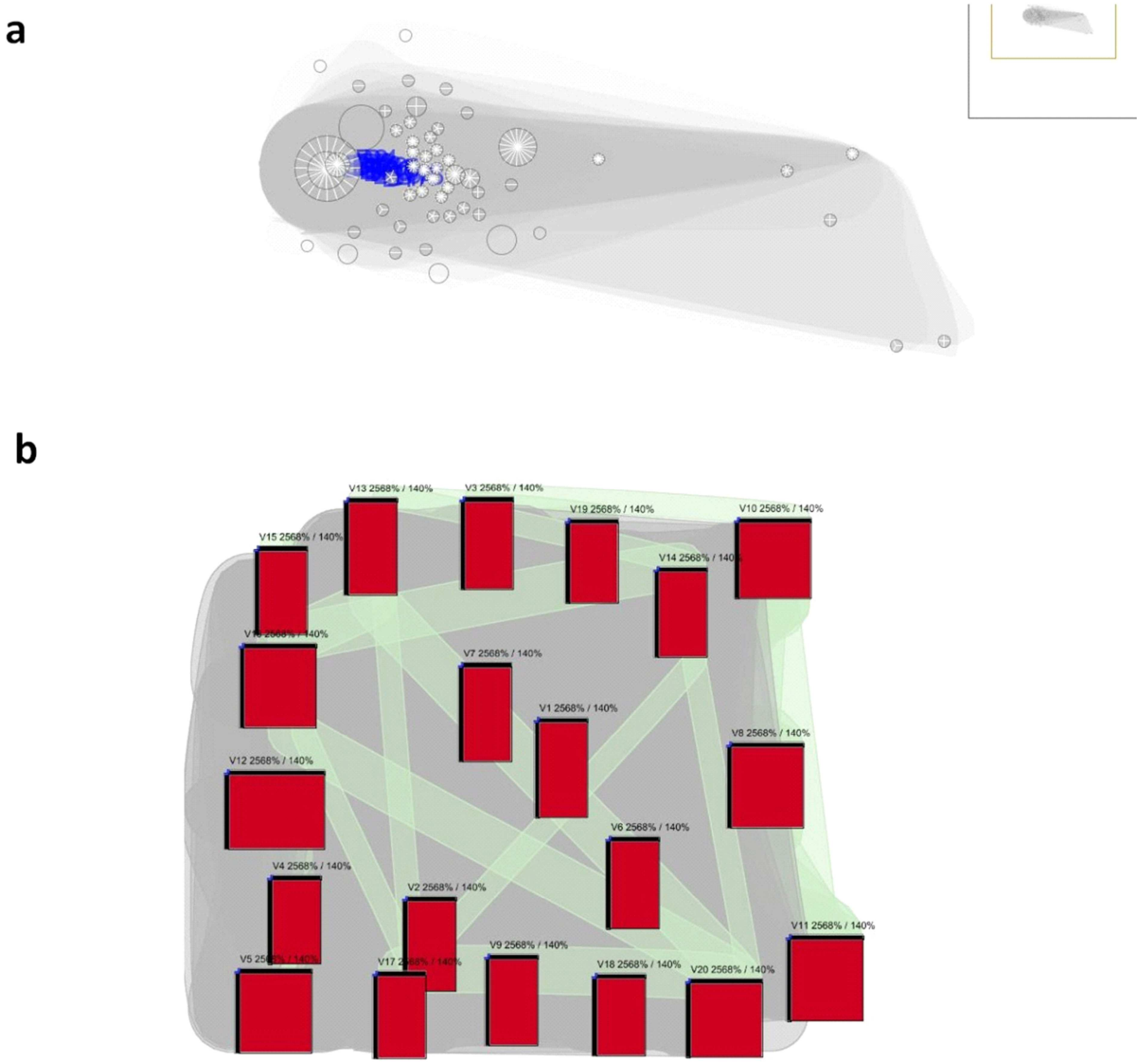


Figure 14. *Bimax bicluster visualization for the human lung carcinomas dataset with (a) BicOverlapper and (b) Furby.*

5. VisBicluster and the Visualization Principles

VisBicluster is designed following the rules of information visualization (InfoVis), especially the Gestalt laws and Mackinlay visual variables for overlaps perception [30, 31] and also, the visualization mantra for the visualization process [32, 33]. Our technique is based essentially on *similarity* and *connectedness* principles as well as *position* and *color saturation* variables. Table 1 resumes the main principles observed in VisBicluster.

Table 1. *VisBicluster visualization objectives and InfoVis principles.*

Objective	Visbicluster	Principle
Represent biclusters	Text, color	Similarity, position, color saturation
Represent overlaps	Cells, color	Similarity, position, color saturation, shape
Represent elements	Text	Area, containment
Distinguish biclusters and elements	Text	Area, containment
Dissolve ambiguities	Hovering, color	Shape, color hue
Overall display	Matrix, overview	Overview first
Reduce cluttering and navigation	Interaction	Focus+context
Analyze and filter data	Filtering options	Analyze further
Textual information	Text	Details on demand

6. Data Communication and Retrieval

Following the multiple-linked views philosophy [33], the visualization techniques implemented by VisBicluster are interconnected so the interaction with one of them affects the rest. A communication layer translates the items selected on a visualization technique to the related items on other visualizations (see Figure 15). Despite their nature, all the data sources share two entities, genes, and conditions that are used to perform the translations. For example, the selection of a cell in the two-dimensional matrix of overlaps leads to the selection of its genes and conditions as a heatmap or to the biclusters it belongs to.

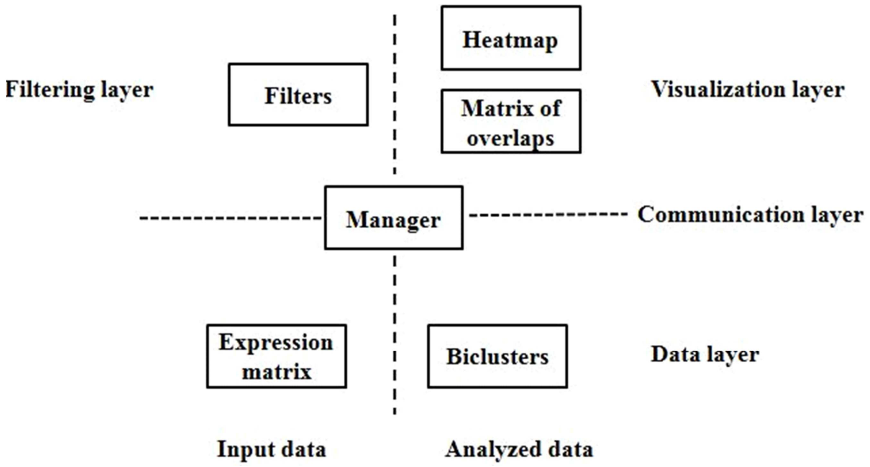


Figure 15. *Layer and data schema of VisBicluster.*

Another interesting aspect to discuss is how data of different natures are retrieved by VisBicluster. The straightforward method to obtain data is to let the user provide them. In fact, it is important to give the user control about the nature of data especially in the case of expression data and also in the case of analyzed data. Therefore, users must manually load any expression matrix they want to analyze. The format is kept simple so the user's data can be easily translated from other formats. We allow the load of bicluster results from sources generated with *biclust* R package [34].

7. Conclusion

Biclustering is a powerful unsupervised learning technique that can be used to identify patterns in gene expression data. However, the interpretation of biclustering results can be challenging, especially when there are a large number of overlapping biclusters. To address this challenge, we have developed VisBicluster, an interactive visualization tool that allows analysts to explore and analyze biclustering results. VisBicluster represents biclusters and their corresponding overlaps as a two-dimensional matrix, which provides an overview of the overall relationships between all biclusters. VisBicluster comes with an easy-to-use web interface that allows analysts to investigate individual biclusters in detail [17].

The developed visualization technique allows visualizing interesting number of biclusters together within a single representation, which fulfills the main characteristics of biclustering (i e., bi-dimensionality and overlapping). The visualization prioritizes *overlaps* that are displayed in a sorted way (as a matrix) instead of as linkage elements subsidiary to main visual elements (biclusters).

References

[1] M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein, Cluster analysis and display of genome-wide expression patterns, Proceedings of the National Academy of Sciences of the United States of America. 95 (1998) 14863–14868. doi: 10.1073/pnas.95.25.14863.

[2] R. R. Sokal, C. D. Michener, A statistical method for evaluating systematic relationships, Univ. Kansas, Sci. Bull. 38 (1958) 1409–1438. https://ci.nii.ac.jp/naid/10004143217/.

[3] J. A. Hartigan, M. A. Wong, Algorithm AS 136: A K-Means Clustering Algorithm, 1979. http://www.labri.fr/perso/bpinaud/userfiles/downloads/hartigan_1979_kmeans.pdf (accessed July 6, 2019).

[4] Y. Cheng, G. M. Church, Biclustering of expression data., Proceedings. International Conference on Intelligent Systems for Molecular Biology. 8 (2000) 93–103. http://www.ncbi.nlm.nih.gov/pubmed/10977070 (accessed April 4, 2017).

[5] S. C. Madeira, A. L. Oliveira, Biclustering algorithms for biological data analysis: a survey, IEEE/ACM Trans. Comput. Biol. Bioinforma. 1 (2004) 24–45. doi: 10.1109/TCBB.2004.2.

[6] C. North, Information Visualization, in: Handbook of Human Factors and Ergonomics, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2006: pp. 1222–1245. doi: 10.1002/0470048204.ch46.

[7] C. Ware, Information visualization: perception for design, Morgan Kaufman, 2004. https://dokumen.tips/documents/information-visualization-perception-for-design-2nd-edition.html (accessed July 13, 2019).

[8] B. Pontes, R. Giráldez, J. S. Aguilar-Ruiz, Biclustering on expression data: A review, Journal of Biomedical Informatics. 57 (2015) 163–180. doi: 10.1016/j.jbi.2015.06.028.

[9] H. Aouabed, M. Elloumi, R. Santamaría, An evaluation study of biclusters visualization techniques of gene expression data, Journal of Integrative Bioinformatics. 18 (2021). doi: 10.1515/JIB-2021-0019/MACHINEREADABLECITATION/RIS.

[10] H. Aouabed, R. Santamaria, M. Elloumi, Visualizing biclustering results on gene expression data: A survey, ACM International Conference Proceeding Series. (2021) 170–179. doi: 10.1145/3473258.3473284.

[11] D. Gonçalves, R. S. Costa, R. Henriques, Context-situated visualization of biclusters to aid decisions: going beyond subspaces with parallel coordinates, ACM International Conference Proceeding Series. (2022). doi: 10.1145/3531073.3531124.

[12] N. K. Verma, T. Sharma, S. Dixit, P. Agrawal, S. Sengupta, V. Singh, BIDEAL: A Toolbox for Bicluster Analysis—Generation, Visualization and Validation, SN Computer Science. 2 (2021). doi: 10.1007/S42979-020-00411-9.

[13] M. Sözdinler, A Review of Visualization Methods and Tools for the Biclustering, International Journal of Innovative Science and Research Technology. 6 (2021). www.ijisrt.com (accessed June 5, 2023).

[14] H. Aouabed, R. Santamaría, M. Elloumi, Suitable Overlapping Set Visualization Techniques and Their Application to Visualize Biclustering Results on Gene Expression Data, in: Springer, Cham, 2018: pp. 191–201. doi: 10.1007/978-3-319-99133-7_16.

[15] R. Santamaría, R. Therón, L. Quintales, BicOverlapper 2.0: visual analysis for gene expression, Bioinformatics. 30 (2014) 1785. doi: 10.1093/BIOINFORMATICS/BTU120.

[16] M. Streit, S. Gratzl, M. Gillhofer, A. Mayr, A. Mitterecker, S. Hochreiter, Furby: fuzzy force-directed bicluster visualization., BMC Bioinformatics. 15 Suppl 6 (2014) S4. doi: 10.1186/1471-2105-15-S6-S4.

[17] H. Aouabed, R. Santamaria, M. Elloumi, VisBicluster: A Matrix-Based Bicluster Visualization of Expression Data, J. Comput. Biol. (2020) cmb.2019.0385. doi: 10.1089/cmb.2019.0385.

[18] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, H. Pfister, UpSet: Visualization of intersecting sets, IEEE Trans. Vis. Comput. Graph. 20 (2014) 1983–1992. doi: 10.1109/TVCG.2014.2346248.

[19] M. E. Baron, A Note on the Historical Development of Logic Diagrams: Leibniz, Euler and Venn, Math. Gaz. 53 (1969) 113. doi: 10.2307/3614533.

[20] R. Santamaría, R. Therón, L. Quintales, A visual analytics approach for understanding biclustering results from microarray data, BMC Bioinformatics. 9 (2008) 247. doi: 10.1186/1471-2105-9-247.

[21] S. Barkow, S. Bleuler, A. Prelić, P. Zimmermann, E. Zitzler, BicAT: A biclustering analysis toolbox, Bioinformatics. 22 (2006) 1282–1283. doi: 10.1093/bioinformatics/btl099.

[22] A. Prelić, S. Bleuler, P. Zimmermann, A. Wille, P. Bühlmann, W. Gruissem, L. Hennig, L. Thiele, E. Zitzler, A systematic comparison and evaluation of biclustering methods for gene expression data, Bioinformatics. 22 (2006) 1122–1129. doi: 10.1093/bioinformatics/btl060.

[23] V. I. Levenshtein, Binary Codes Capable of Correcting Deletions, Insertions and Reversals, Sov. Phys. Dokl. Vol. 10, p.707. 10 (1966) 707. http://adsabs.harvard.edu/abs/1966SPhD...10..707L.

[24] R. Santamaria, Visual analysis of gene expression data by means of biclustering, University of Salamanca, Spain, 2009.

[25] L. Lazzeroni, A. Owen, Plaid Models for Gene Expression Data, CEUR Workshop Proc. 1542 (2000) 33–36. doi: 10.1017/CBO9781107415324.004.

[26] V. A. Padilha, R. J. G. B. Campello, A systematic comparative evaluation of biclustering techniques, BMC Bioinformatics. 18 (2017) 55. doi: 10.1186/s12859-017-1487-1.

[27] A. Bhattacharjee, W. G. Richards, J. Staunton, C. Li, S. Monti, P. Vasa, C. Ladd, J. Beheshti, R. Bueno, M. Gillette, M. Loda, G. Weber, E. J. Mark, E. S. Lander, W. Wong, B. E. Johnson, T. R. Golub, D. J. Sugarbaker, M. Meyerson, Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma subclasses, Proceedings of the National Academy of Sciences of the United States of America. 98 (2001) 13790–13795. doi: 10.1073/pnas.191502998.

[28] B. Alsallakh, L. Micallef, W. Aigner, H. Hauser, S. Miksch, P. Rodgers, Visualizing Sets and Set-typed Data: State-of-the-Art and Future Challenges, Eurographics Conference on Visualization (EuroVis)– State of The Art Reports. (2014) 1–21. doi: 10.2312/eurovisstar.20141170.

[29] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, Knowledge Discovery and Data Mining: Towards a Unifying Framework, 1996. www.aaai.org (accessed January 4, 2020).

[30] D. Chang, L. Dooley, J. E. Tuovinen, Gestalt theory in visual screen design: a new look at an old subject, in: Seventh World Conference on Computers in Education, 2002. https://www.semanticscholar.org/paper/Gestalt-theory-in-visual-screen-design%3A-a-new-look-Chang-Dooley/41ca82e97d5ad678c9578d6a18d4600b708277d2 (accessed November 17, 2019).

[31] J. Mackinlay, Applying a theory of graphical presentation to the graphic design of user interfaces, in: Proceedings of the 1st Annual ACM SIGGRAPH Symposium on User Interface Software and Technology, UIST 1988, Association for Computing Machinery, Inc, 1988: pp. 179–189. doi: 10.1145/62402.62431.

[32] B. Shneiderman, The eyes have it: A task by data type taxonomy for information visualizations, Proceedings IEEE Symposium on Visual Languages. (1996) 336--343. http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.224.3197 (accessed November 16, 2019).

[33] D. Keim, K. Jörn, G. Ellis, M. Florian, Mastering the information age: solving problems with visual analytics, Eurographics Association, 2010.

[34] S. Kaiser, R. Santamaria, T. Khamiakova, M. Sill, R. Theron, L. Quintales, F. Leisch, E. De, T. Maintainer, biclust: BiCluster Algorithms. R package version 1.0.2., (2013). https://cran.r-project.org/web/packages/biclust/biclust.pdf (accessed April 22, 2017).