BicOverlapper

User's Guide

Version 1.4

September 4, 2009

VNiVERSiDAD
D SALAMANCA

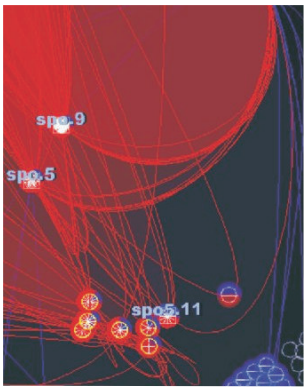
VNiVERSiDAD
D SALAMANCA
Facvltad D Ciencias

Rodrigo Santamaría, Roberto Therón and Luis Quintales
Department of Computers and Automation
Faculty of Sciences
University of Salamanca (Spain)

Contents

List of Figures

List of Tables



Chapter 1
Install

1.1 Requirements

BicOverlapper has been tested for the platforms in Table 1.1. Developed in Java, it is expected to work under other Operating Systems that support the required JRE version and with enough disk and RAM.

Table 1.1. *Tested Platforms for BicOverlapper 1.4*

Operating System	Memory	RAM	JRE
Windows XP	50MB	1GB	1.6.0_01
Windows Vista	50MB	1GB	1.6.0_01
Ubuntu 8.0 (Debian)	50MB	1GB	1.6.0_01

This version of BicOverlapper integrates biclustering analysis of microarrays and requests for gene information, by means of the statistical package R and BioConductor. Therefore, in order to load and analyze gene expression matrices, R 2.8 or higher must be installed, with the following requirements:

- *R_HOME* must be set to the R folder containing the folders bin, doc, library, etc. (for example: *C:/Program Files/R/R-2.8.1*)
- The package *rJava* must be installed to allow R-Java communication. It must be installed in the *R_HOME/library* folder and not anywhere else. This is specially important in systems like Windows Vista or Linux with rigid user accounts, where libraries are installed in specific user folders. In R, you can change the folder to which libraries are installed with the *lib* parameter on *install.packages()* method or check the default library paths with *.libPaths()* (note that you must also have writing permissions for *R_HOME/library*). Or you can run R as administrator to install the packages (for example, in Windows Vista, right-click on the R icon and select *run as administrator*). If *rJava*

is not properly installed in *R_HOME/library*, the program will exit with the following error:
Creating Rengine (with arguments)
Cannot find JRI native library!
java.lang.UnsatisfiedLinkError: no jri in java.library.path

- *R.dll* must be in the path. If it isn't, you will get an error similar to this one:
Creating Rengine (with arguments)
Cannot find JRI native library!
java.lang.UnsatisfiedLinkError:
C:/Program Files/R/R-2.4.0/library/rJava/jri/jri.dll: Can't find dependent libraries

Finally, BicOverlapper should be able to install additional required R/BioConductor packages on the fly, but it is recommended to install them off-line, following the same requirements that the *rJava* package above. The required packages are:

- *biclust*: This package is used to perform biclustering analysis of gene expression matrices. If it is not properly installed, the analysis menu will be disabled or not working.

- *GO.db, annotate*: These packages are used in order to manipulate GO annotations for the genes in the expression matrix, providing that they were proper IDs for the corresponding microarray database package (see more about how to specify microarray annotation packages in section 2.1). If it is not properly installed, GO details and GO significant tests won't work.

- *GOstats*: This package is used to perform hypergeometric GO significance tests. If not properly installed, these kind of tests will be disabled or not working.

- *annotation packages*: In the case of microarray platforms with an available BioConductor annotation package, if it is properly installed and specified (see section 2.1)), gene annotations and GO tests will be available. For example, if the expression matrix comes from Affymetrix Yeast Genome S98 chips, then *ygs98.db* package should be installed.

For any issue, please contact with rodri@usal.es, providing the details of your problem running BicOverlapper, along with your platform characteristics.

1.2 Installation and run

To obtain the framework, go to *vis.usal.es/bicoverlapper* and take the latest version in the *downloads* section. Please note that alpha versions can still have some bugs. After downloading the package, please follow these steps:

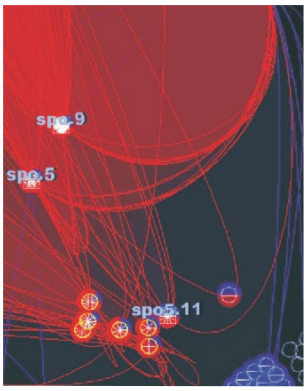
1. Unzip the downloaded file wherever you want it to be installed

2. Open the unzipped folder

3. Depending on the OS:

   - *Windows:* Double click on file *bicoverlapper.bat* or write in the command line *bicoverlapper.bat*
   - *Linux:*
     - Check that *bicoverlapper.bat* has execution privileges. If not, give them with *chmod +x bicoverlapper.bat*.
     - Run the application with *./bicoverlapper*.

If everything is ok, BicOverlapper should start right now. If it does not start or closes suddenly after loading a expression matrix, it is possibly a problem of requirements (see section 1.1)



Chapter 2
Files

2.1 File Formats

Three different kind of files can be loaded by BicOverlapper:

- *Microarray*: Gene expression data with expression levels, sample and condition names, species and platform (extension *.txt*)
- *TRN*: Transcription Regulatory Network with information about what is known about how gene transcription works for a determinate organism (extension *.xml* or *.gml*
- *Biclustering Results*: the base of BicOverlapper, the set of biclusters returned by one or more biclustering techniques (extension *.bic*)

The three files can be opened for the same analysis: the more information we have, the deeper understanding of the problem we will get. Microarray data and bicluster data have delimited text formats as described in Tables 2.1 and 2.2, respectively.

The **microarray format** has, in rows, the expression level for each gene/probe set and the species and chip name. The first column contains the gene or probe set names, and the first row the condition (array) names. The first column of the first row contains information of the organism name and the array platform, separated by a bar (organism/platform). The name of the organism should follow the standard species name, just as NCBI stores it (for example, for human it should be *Homo sapiens*). It the microarray platform has a corresponding BioConductor annotation package, such as *hgu95av2.db* for the platform Affymetrix Human Genome U95, then the name of the package is the string that must be in this field. If there is no annotation package for it, you can use Entrez gene symbols or IDs to name genes; in this case use *GeneName* or *GeneID*, respectively, and BicOverlapper will search for annotations directly on Entrez Gene and QuickGO[1]

[1]Note that this method is slower and sometimes less reliable than the use of BioConductor packages. Also, the GO significance tests are not available if there is no information about the

Table 2.1. *Microarray Data File Format*

Organism_name/platform	colName1	colName2	...	colNameM
rowName1	expr11	expr12	...	expr1M
rowName2	expr21	expr22	...	expr2M
...		...		
rowNameN	exprN1	exprN2	...	exprNM

Table 2.2. *Bicluster File Format*

Number_of_biclusters
bicluster_set_1
#rows_bic1.1 #columns_bic1.1
row1 row2 ... rowN
col1 col2 ... colN
#rows_bic1.2 #columns_bic1.2
row1 row2 ... rowN
col1 col2 ... colN
...
bicluster_set_2
#rows_bic2.1 #columns_bic2.1
row1 row2 ... rowN
col1 col2 ... colN
#rows_bic2.2 #columns_bic2.2
row1 row2 ... rowN
col1 col2 ... colN
...

Biclusters files, although simple text files, have extension *.bic* to distinguish them from microarray files (*.txt*). Note also that microarray format uses tabs as delimiters, while **bicluster format** uses blank spaces (this is done to fit with bicluster files generated by *biclust* R package [1] and BicAT [2]). You can have several bicluster sets in a single bicluster file, each one must be preceded by a one-line description of the bicluster set. In order to correctly link microarray and bicluster visualizations, the names of rows and columns must be exactly the same in the microarray and bicluster files.

Transcription networks follow SynTReN [3] XML format for graphs. The minimal information necessary for this format is described in fig. 2.1. We are working to support more extended formats as GML or GraphML.

Please be strict with the formats. Even small errors could make the program not to load the file correctly, or to fail after opening some of the visualizations.

chip.

$< GeneNetwork >$
$< Nodes >$
$< Nodeid = "0" name = "name0" type = "type0" >$
$< /Node >$
$< Nodeid = "1" name = "name1" type = "type1" >$
$< /Node >$
...

$< Nodeid = "n" name = "namen" type = "typen" >$
$< /Node >$
$< /Nodes >$
$< Edges >$
$< Edge >$
$< from > 0 < /from >$
$< to > 1 < /to >$
$< /Edge >$
...
$< Edge >$
$< from > i < /from >$
$< to > j < /to >$
$< /Edge >$
$< /Edges >$
$< /GeneNetwork >$

Figure 2.1. *GML format*

Some things to have in mind:

- Do not leave blank lines at the start, middle or end of the documents.
- Separation between fields are blank spaces in Biclustering Results and tabs in Microarray Data.
- Do not use blank spaces or delimiters as names of samples or conditions. Use "_" or " − " instead.
- Do not duplicate gene or conditions names. Use the same names for genes and conditions in all the files.
- Use points (.) as decimal separators, not commas (,)
- Remember that the first line of Microarray Data has as first field a name for the organism/kind of row for the microarray data, and then the condition names.

2.2 Minimal guide to obtain biclustering results

At this stage, maybe you are wondering how to make a biclustering analysis of your microarrays; and then, how to generate the required files for BicOverlapper.

2.2.1 Outside BicOverlapper

If you deal with Affymetrix chips and all you have are the .CEL and .CDF files, a typical way of getting biclustering results is to start with the same steps that in a *Differential Gene Expression Analysis* by means of R-Bioconductor (see, for example, [4]). Once you get the gene level expression matrix, after making the quality checks, background correction, normalization, etc.; and instead or in addition to making the traditional differential analysis, you can use the R package *biclust* [1], that provides several biclustering algorithms you can use to analyze your expression matrix (*exprs(ExpressionSet)*). Finally, you can write to text files both the processed gene expression matrix and the biclustering results with the methods *write.exprs()* and *writeBiclusterResults()*. The file written by *writeBiclusterResults()* must be edited to add a first line with the number of biclusters in the results. The file written by *write.exprs()* must be edited to specify the organism name and kind of rows in the first row, first column, as explained above.

Besides, you can also use BicAT tool [2] to generate biclustering results, and export them to a text file. This file has the same format that BicOverlapper requires, but you must add two lines at the start, the first one with the number of biclusters found and the second one with a text description of the bicluster set.

2.2.2 Inside BicOverlapper

From version 1.4 on, BicOverlapper integrates several biclustering methods for gene expression analysis. These methods correspond to the methods in package *biclust* and use its implementation. Currently, they are Bimax [5], xMotifs [6], Cheng and Church [7] and Turner's Plaid Model [8]. Help about each method is available with the *biclust* package. They can be run inside BicOverlapper from the *Analysis* menu.

All the methods allow to write the results to a file. If no path is selected, a default name is given to the file, and it is stored in the current path. You can also add additional information about the biclustering results with the *Add description* checkbox (see fig. 2.2).

It is usual that biclustering algorithms return a very large number of biclusters. In order to reduce this complexity in results, some authors perform a post-filtering of biclusters, discarding those highly overlapped or excessively large. The methods we present allow a post-filtering of biclusters based on [5], where the user can select the maximum overlap among biclusters (max_o), the maximum number of genes or conditions (max_g and max_c) and the maximum number of biclusters (max_b). The post-filter selects the max_b larger biclusters but with maximum dimension $max_g \times max_c$ that do not overlap with any other more than a max_o%.

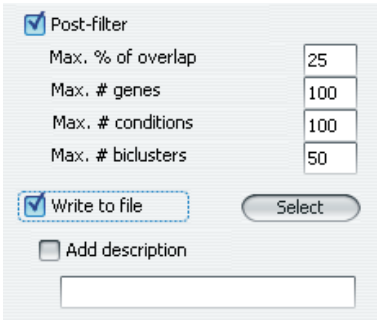


Figure 2.2. *Post-filtering and write biclusters interface.*

Bimax

This algorithm searches for biclusters highly over or under expressed. It performs a binarization of the expression matrix, assigning one to the expression levels above a given threshold and zero to the ones below (for high expression search, the opposite for low expression). Then all the biclusters formed by ones are returned.

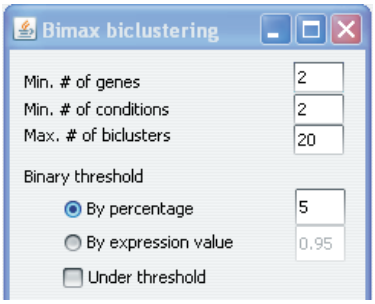


Figure 2.3. *Bimax biclustering interface.*

The first part of the interface relates to the size and number of biclusters: the minimum number of genes/conditions that every bicluster must have, and the maximum number of biclusters to be returned (see fig. 2.3). Binarization is controlled by the *Binary threshold* section of the Bimax biclustering interface. A radio button allows to select if the threshold should be determined by percentage (for example, the higher 5% expression levels will be one) of by value (for example, expression levels above 3.5 will be one). The check button *Under threshold* inverts the binarization to low expression values.

XMotifs

This biclustering algorithms searches for genes with preserved states across several conditions. A state is a range of expression, and a bicluster may include several

genes, in the same or different states. It is considered as a coherent evolution biclustering.

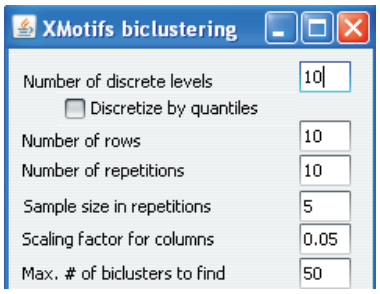


Figure 2.4. *Xmotifs biclustering interface.*

The XMotifs interface allows to determine the *Number of discrete levels*, that will determine the expression value ranges for each state (ideally, two levels should be enough in order to find high and low expression, but it its usually not enough, providing very broad ranges). They may also be determined by quantiles instead of by raw expression. The *Number of rows*, *Number or repetitions*, *Sample size in repetitions* and *Scaling factor for columns* refer to the n_s, n_d, s_d and α parameters in [6].

Cheng and Church

This was the first biclustering algorithm published that was applied to expression matrices. It searches for a bicluster at a time, masking it with random numbers afterwards in order to search for the next different one. Each bicluster is a subset of rows and columns with a high similarity score (the mean squared residue is used, a measure of coherence). A subset of rows and columns is a bicluster if its mean squared residue is below a δ threshold. α is a factor for determining the rows and columns to be deleted in order to speed up the algorithm on a multiple node deletion paradigm. These values are restricted to $\alpha \geq 0$ and $\delta > 1$. See [7] for the details about the algorithm and its parameters.

Figure 2.5. *Cheng and Church biclustering interface.*

Turner's Plaid Model

This method [8, 9] is an improvement of the original Plaid Model algorithm [10]. It tries to fit the matrix to a sum of layers plus a background noise. Each of the layers correspond to a bicluster. The method can be modified to group rows, columns or both. The *Row release* and *Column release* parameters are values in [0-1] (recommended interval in [0.5-0.7]) that are used as thresholds to remove rows or columns in the layers, depending on row/column homogeneity (the lower they are, the more strict is the search).

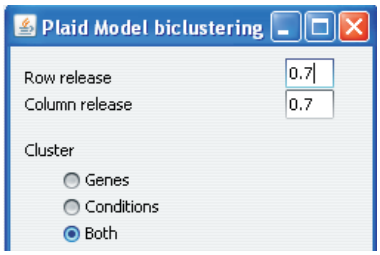
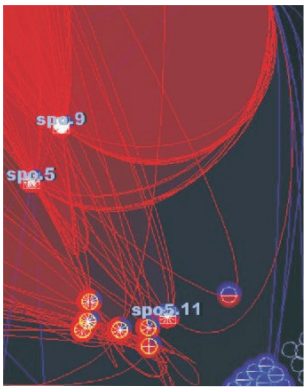


Figure 2.6. *Turner's plaid model biclustering interface.*



Chapter 3
Data Visualizations

Currently, five visualizations of data can be displayed, as outlined in Table 3.1. Here we will briefly refer to the practical use of these visualizations, specially the Overlapper. For more information please refer to the technical report document, available at *http://code.google.com/p/bicluster* or to [11]. Resumed interfaces for each visualization are shown in the tables in the Appendix at the end of this document.

Table 3.1. *Data Visualizations*

Visualization	Data	Brief Description
Heatmap	Microarray	Color coded expression levels for representation of microarray data
Parallel Coordinates	Microarray	Representation of gene profiles as lines
Bubble Map	Biclusters	2D Projection of biclusters
Overlapper	Biclusters	Force-directed graph for Biclusters
Transcription Network	TRN	Force-directed graph for TRNs
Word Cloud	Annotations	Representation of gene annotation occurrences

3.1 Microarray Heatmap

To display the microarray heatmap visualization, microarray data must be previously loaded with *File/Open Microarray*. Once the data are loaded, the option *View/Microarray Heatmap* is unlocked so you can click on it to display a microarray heatmap instance (see fig. 3.1).

The microarray heatmap can be navigated with the mouse to inspect expression levels, that are enlarged when it is over them. In addition, the gene and condition profiles in which the expression level is in are also enlarged, vertically

and horizontally respectively. The rest of expression levels are shrunk to fit the remaining space in the display. Finally, expression levels and gene and condition tags can be clicked. When it is done, the corresponding rows and/or columns are rearranged to the top left of the heatmap, and its enlargement is maintained. More than one row and column can be selected by pressing a key while clicking.

Figure 3.1. *Heatmap visualization of a 150x20 microarray data matrix. Even with a small number of genes, heatmaps usually are very tall and thin.*

3.1.1 Sampled microarray heatmaps

Microarray data matrices can be very large, mainly in the number of rows. The visualization and interaction with large heatmaps is slow, and to avoid this only

sampled matrices are visualized. Only some genes randomly sampled from the original microarray data matrix are visualized, giving an overall impression of how is the dataset. However, when any group of genes is selected in the other visualizations (by parallel coordinates thresholds or by selecting any bicluster), the actual genes are represented and highlighted in the heatmap.

3.2 Parallel Coordinates

To display the parallel coordinates visualization, microarray data must be previously loaded with *File/Open Microarray*. Once the data are loaded, the option *View/Parallel Coordinates* is unlocked so you can click on it to display a parallel coordinates instance (see fig. 3.2).

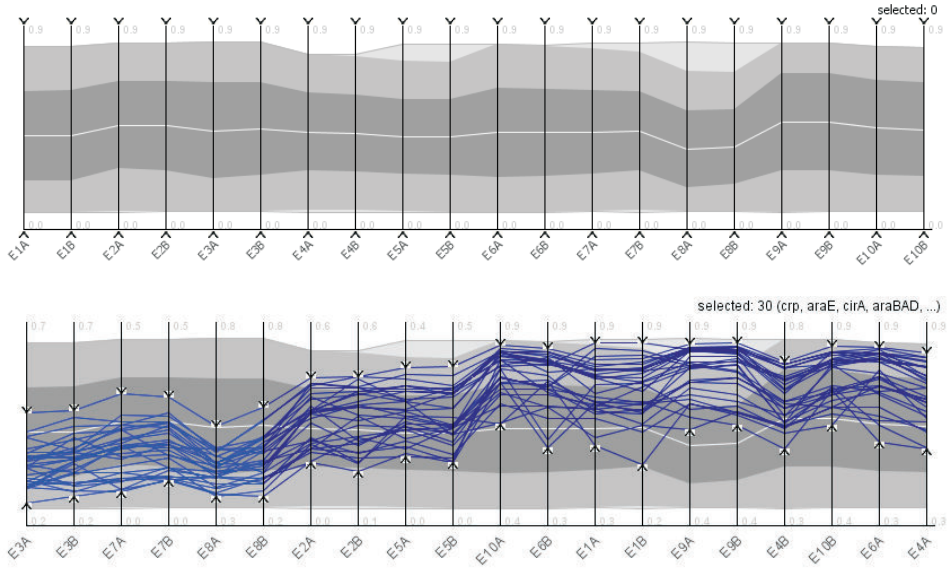


Figure 3.2. *Parallel coordinates overview of an expression matrix (top). Selection of the genes profiles for a bicluster, the section of the gene profile corresponding to conditions in the bicluster is colored brighter at the left (bottom).*

The most important interaction with parallel coordinates is the setting of the scroll handles by dragging them. These scrolls set threshold limits for the expression levels for the corresponding condition. This way, when you set a threshold, gene profiles that fulfill the new threshold criterium are highlighted. In addition, the remaining scrolls are set to the maximum and minimum expression level values of the highlighted profile set. This is done to help to refine an initial threshold criterium. Therefore, when you set a new scroll limit, only the profiles that fulfil this limit and any other imposed limits are highlighted. To broaden a search, if *Ctrl* is pressed while releasing a scroll, the visualization will display all the profiles

that fulfil the threshold criteria for the corresponding coordinate, but ignoring any other restriction in the rest of coordinates.

In the case of biclusters, the profiles of the genes in the bicluster are selected. The conditions in the biclusters are placed first (at the left), and the portions of the gene profiles corresponding to conditions in the bicluster are colored with a brighter hue (see fig. 3.2, bottom).

3.2.1 Sampled parallel coordinates

As in the case of heatmaps, large parallel coordinates visualization and interaction is slow, so the whole set of lines is not displayed. Instead, polygons representing 1-fold to 4-fold areas are drawn, in a grey scale (see fig. 3.2, top). Also, lines representing the average value and the maximum and minimum values are drawn. If we select more than 200 gene profiles by setting parallel coordinate thresholds, a blue polygon formed by the maximum and minimum expression values of the selected gene profiles is drawn. If we select less than 200 genes, the actual gene profile lines are drawn.

3.3 TRN Graph

To display the TRN Graph visualization, TRN data must be previously loaded with *File/Open TRN*. Once the data are loaded, the option *View/TRN* is unlocked so you can click on it to display a parallel coordinates instance (see Fig 6.5).

The TRN graph initially reconfigures by directed forces, then becoming static. The graph can be navigated by clicking in the background and dragging the whole visualization. By clicking with the right button in the background, the visualization is centered and fitted to the window space. By using the mouse wheel, we can zoom in or out.

When nodes are hovered over with the mouse, they are highlighted. If they are clicked, they keep highlighted until clicked again (we will say that they are selected). To select more than a node simultaneously, press a key while clicking a new node. Nodes can also be dragged to modify their position.

Finally, a text box at the right-bottom of the visualization allows searches by text, so nodes with names that starts with the searched text are highlighted (the number of matched nodes appears next to the search box).

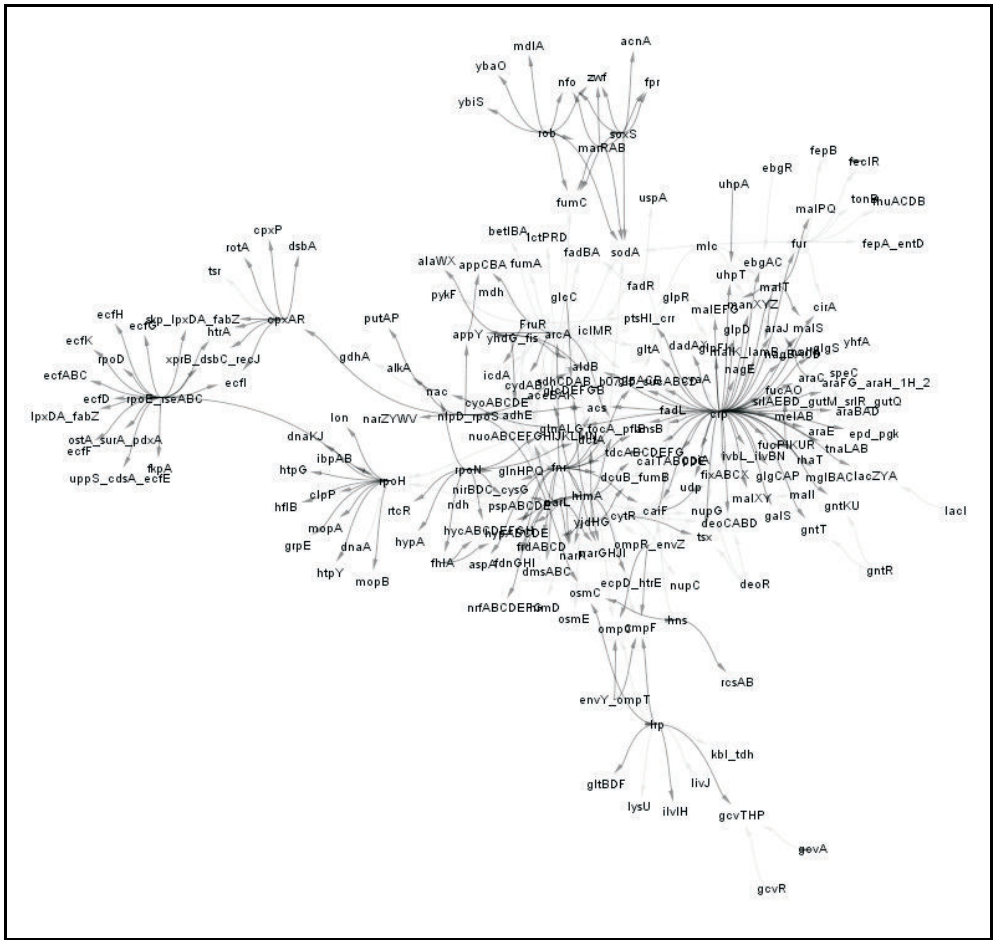


Figure 3.3. *Transcription Network visualization, representing 200 genes for* E. coli*. These networks have "small-network" configurations, which roughly means that there are hub nodes with high connectivity and you can reach any node from any other node passing by a low number of intermediate nodes.*

3.4 Word Cloud

To display the Word Clouds visualization, a gene expression matrix must be previously loaded with *File/Open Microarray*. The matrix should have Entrez Gene IDs or symbols or, preferably, gene identifiers corresponding to a BioConductor annotation package already installed (see section 2.1). Once the matrix is loaded, the option *View/Word Cloud* is unlocked so you can click on it to display a word cloud. A word cloud displays the gene annotations of a selection of genes that must be previously done by means of another visualization. You can have several Word Clouds open at the same time to visualize different annotations.

The Word Cloud displays gene annotations, either descriptions or GO terms, for the genes selected by means of the other visualizations. The gene annotations are retrieved in real time from the installed BioConductor packages or from the Entrez Gene and QuickGO databases. The terms or description are displayed alphabetically sorted, with a size proportional to the number of times that the term or description appear in the selected genes (the number of times is displayed in brackets following the word, see fig. 3.4). Both types of annotations can be split into single words or di-words[2]. GO terms may be filtered, so only biological processes (bp), molecular functions (mf) or cellular processes are displayed.

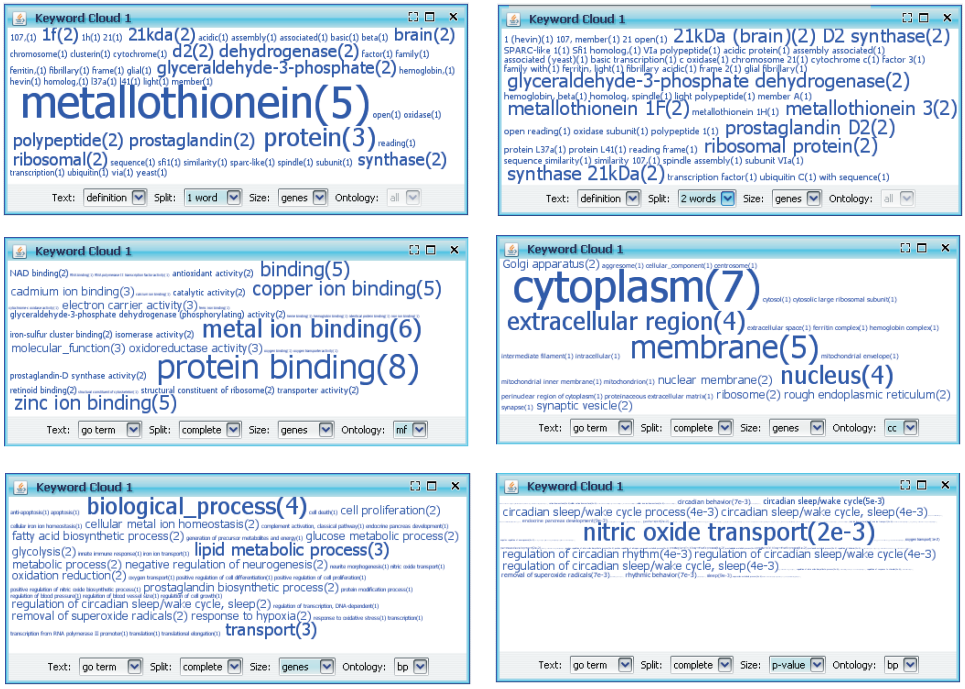


Figure 3.4. *Word clouds for ten highly expressed genes in human brain. Top: words appearing in gene descriptions, one by one (left) and in pairs (right); it can be changed with the "Split" combo box. Center: GO terms annotated, molecular functions (left) and cellular components (right); it can be changed with the "Ontology" combo box. Bottom: biological processes by number of occurrences (left) and p-values (right); it can be changed with the "Size" combo box.*

Finally, you can select that the size of the GO terms will depend on the statistical significance of the term instead of on the number of occurrences. In this case, the size is determined by the inverse of the p-value calculated by means of a hypergeometric tests. This value is the probability that a GO term will appear n times in m genes randomly selected, being n the number of genes, among the

[2] A di-word is a combination of two words

selected ones, for which the GO term is actually annotated. Therefore, the lower the value, the more probable is that the GO term is not grouped by chance in the selected genes.

3.4.1 Word Cloud Performance

Note that the Word Cloud needs to search for gene annotations in BioConductor annotation packages or via programmatic access to Entrez Gene and QuickGO. This can take some time, so it is advisable not to select a large number of genes. The search in BioConductor retrieves about 20 genes per second, while the search in the web retrieves around 1 gene per second.

3.5 Bubble Map

To display the Bubble Map visualization, bicluster data must be previously loaded with *File/Open Biclustering Results*. Once the data are loaded, the option *View/Bubble Map* is unlocked so you can click on it to display a bubble map instance.

The Bubble map draws biclusters as colored circles (bubble) in a static display (see Fig 3.5). The display can be navigated by clicking in the background and dragging the whole visualization. By clicking with the right button in the background, the visualization is centered and fitted to the window space. By using the mouse wheel, we can zoom in or out.

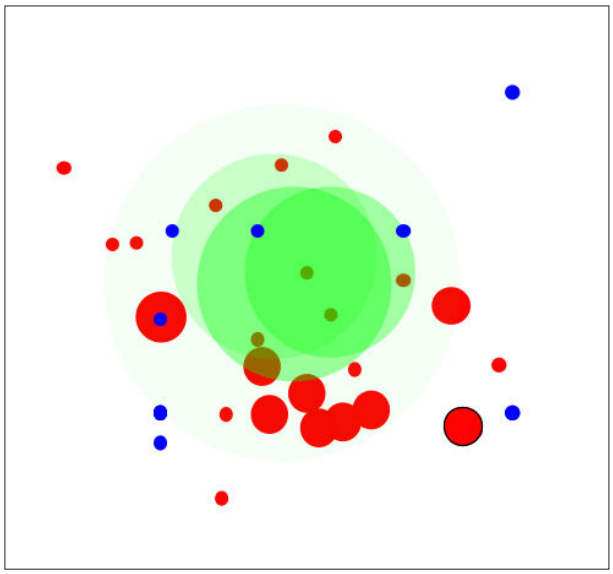


Figure 3.5. *Bubble map visualization. Biclusters from three biclustering methods are drawn in different colors. 2D projection and transparencies reduce cluttering.*

When bubbles are hovered by the mouse, they are highlighted. If they are clicked, they keep highlighted (we will say that they are selected). To select more than a node simultaneously, press a key while clicking a new node. You can also select all the genes and conditions inside an area by drawing a line surrounding it.

Finally, a text box at the right-bottom of the visualization allows searches by text, so bubbles that correspond to biclusters with genes whose name starts with the searched text are highlighted (the number of matched bubbles appears next to the search box).

3.6 Overlapper

To display the Overlapper visualization, Bicluster data must be previously loaded with *File/Open Biclustering Results*. Once the data are loaded, the option *View/Overlapper* is unlocked so you can click on it to display an overlapper instance.

The Overlapper draws biclusters as colored rounded shapes (hulls) in a dynamic display (see fig. 3.6). Each hull contains circles (genes) and squares (conditions) grouped by the corresponding bicluster. Hulls overlap if they share some of the genes or conditions. Hulls and nodes are in continuous movement until they reach an optimal configuration and visualization becomes static or almost static.

The display can be navigated by the top-right overview. This overview shows the whole display, with the portion displayed highlighted by an orange square. By clicking on this square and dragging it, the graph can be navigated.

When hulls are hovered over with the mouse, they are highlighted. When a node (gene or condition) is hovered by the mouse, it is highlighted, along with all the nodes that are grouped with it in at least one of the biclusters. Nodes and hulls can be selected by clicking on them when they are highlighted. To select a hull implies the selection of all the nodes in it. More than one node or hull can be selected by pressing a key while clicking them.

At the bottom of the display, there is a toolbar with several interaction options. When hovered, the buttons show a label with a brief explanation of their function. Table 6 in the Appendix explains them in more detail.

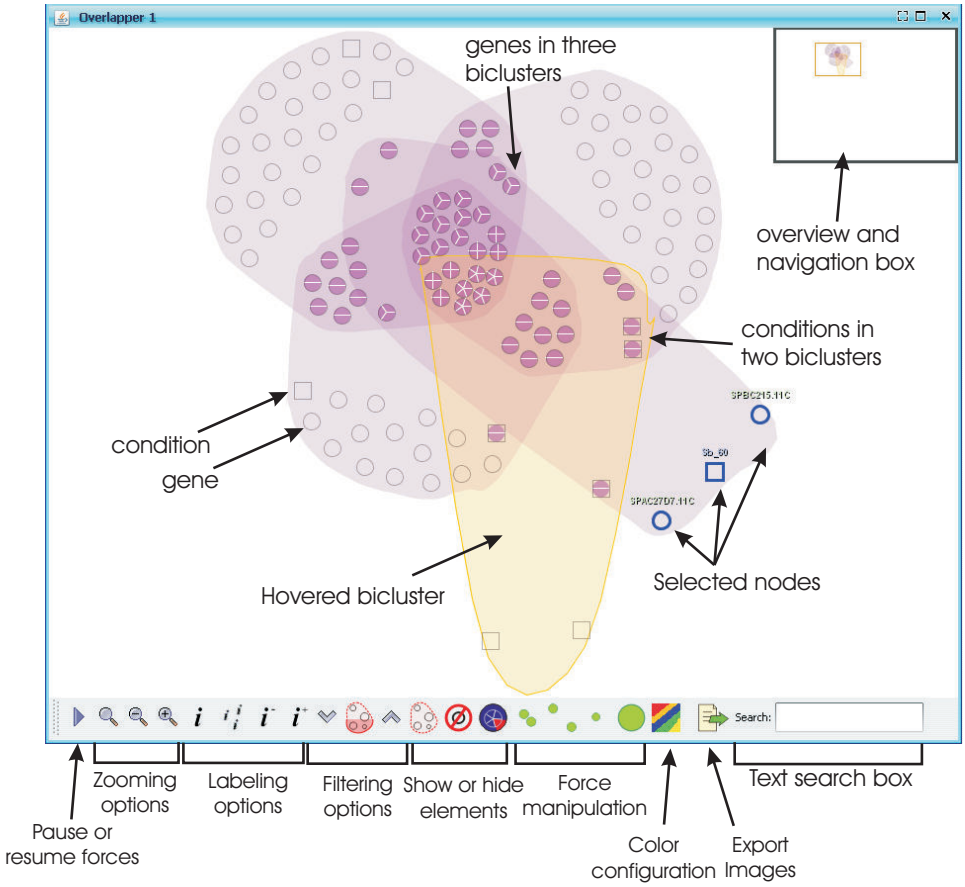
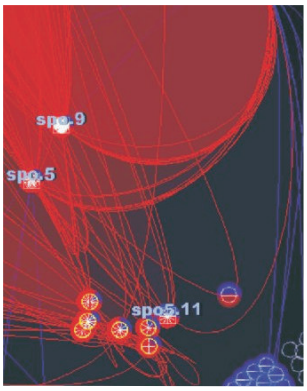


Figure 3.6. *An example of the Overlapper visualization for five biclusters. Visualization elements and options detailed in the figure.*



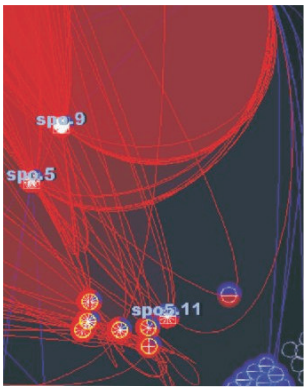
Chapter 4

Data Interaction

Data analysis is performed by interaction with the data visualizations described above. The visualizations are inspected, then we select or search for some interesting items (genes, conditions or both), visualizations are reconfigured to focus on those items, and the loop starts again up to obtaining enough insight in the data. Data interaction is therefore focused in three actions:

- *Select*: items are usually selected by clicking on them, reconfiguring the visualization to highlight them. Selected items change their color to a different, identifiable one (by default, blue). Selected items can be exported to a text file with the menu option *File/Export Selection*. The selection history is recorded, so you can use ctrl-Z and ctrl-Y to go back and forward in selections.
- *Hover*: items are hovered by moving the cursor over them, being highlighted in the visualization, by changing the items's color to a brighter one (by default yellow).
- *Search*: one or more items are actively searched for, and they are highlighted if found. BicOverlapper allows textual searches for genes in some of the visualizations (TRN, Bubble map and Overlapper). The search for genes in the Bubble map highlights the biclusters that contain the gene. In addition, the option in the menu *Analysis/Search* allows searches by any textual fields, usually gene and condition names, but also gene annotations, etc.

In the appendix 6.5 you can find several tables detailing the interaction with each visualization.



Chapter 5
Configuration

5.1 Color configuration

All the visualizations allow the configuration of colors used for main drawing elements. Some of the colors are also linked for most of the visualizations:

- *Selection*: color used to highlight selected entities (genes, profiles or biclusters). Entities are usually selected by clicking on them. Default selection color is blue.
- *Hover*: color used to highlight hovered entities, which is usually done by placing the cursor over them. Default hover color is yellow.
- *Search*: some of the visualization have text boxes to do searches by name of genes or conditions. All entities that match to these searches are highlighted by this color. Default search color is magenta
- *Bicluster sets*: there are three colors defined to draw biclusters on different bicluster sets, shared by the bicluster visualizations. Default colors for biclusters are purple, orange and spindrift blue.

The color configuration panel can be activated with the interactions shown at Table 5.1. All the visualizations have similar interfaces for color configuration, in fig. 5.1 is shown the color configuration panel for Overlapper.

Table 5.1. *Color configuration activation*

Heatmap	Right-click in the background
Parallel Coordinates	Right-click in the background
TRN Graph	Right-click in the bottom bar
Bubble Map	Right-click in the bottom bar
Overlapper	Click on "color configuration" button

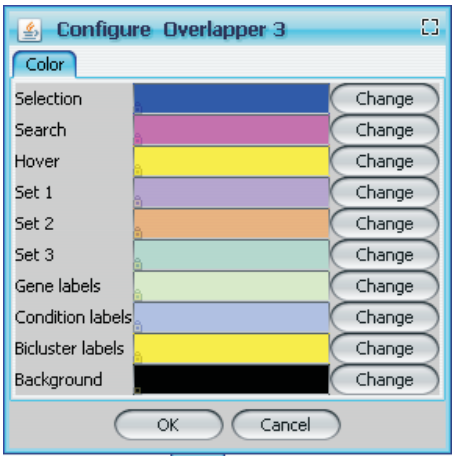
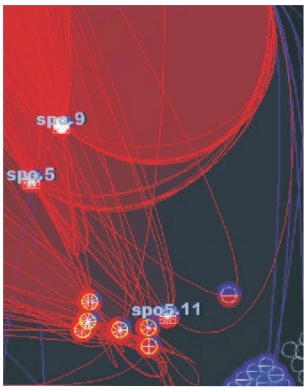


Figure 5.1. *For each configurable item, the item's name, its current color and a button to change it are displayed. "*Selection*", "*Search*", "*Hover*", "*Set 1*", "*Set 2*" and "*Set 3*" colors are updated for all the visualizations that use them.*

5.2 Projects

Your project can be saved, keeping the current opened files and visualizations, and the current selection.

- *File/Save Project*: saves the current data, display information (size and location of each window) and the current selection, into a xml file.
- *File/Load Project*: loads the data and displays the windows of a selected project.
- *File/Load Last Project*: loads the data and displays the windows of the last saved project.
- *File/Load Recent Project*: loads the data and displays the windows of one of the four last saved projects.



Chapter 6
Usage: A practical case

When you start BicOverlapper, an empty window with three menu options pops up:

- *File*: different options to load files
- *View*: different options to visualize information from the files loaded
- *Help*: It contains details related to the BicOverlapper project, authorship and contact information

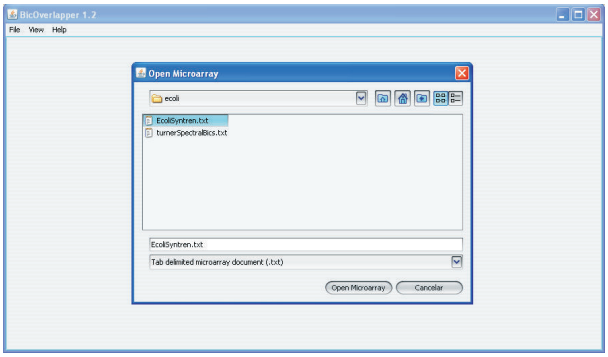


Figure 6.1. *Microarray data file open dialog.*

6.1 Microarray Data Analysis

We are going to analyze the data from a microarray data experiment synthetically generated by SynTReN [3] from E. coli known transcription regulations. We will center on 200 genes and 10 experiments, each one replicated. Although simple,

this example will help as first approach to the framework and provide an excellent opportunity experiment its potential.

The first thing we are going to do is to load the expression levels of our microarray data. To achieve that, we go to *File/Open Microarray* and choose, in the installation folder, *data/ecoli/ EcoliSyntren.txt* (see fig. 6.1). In a moment, a message tells us that the file is loaded. Note that, for larger microarray data (thousands of genes per hundreds of conditions), the load process would need some seconds to load.

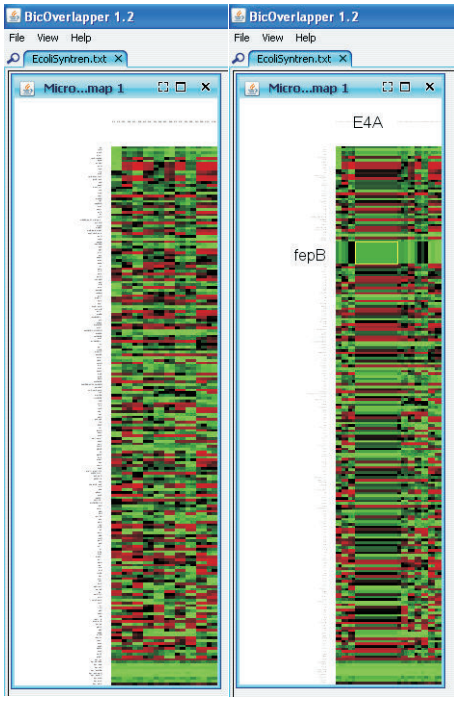


Figure 6.2. *At the left, the microarray heatmap as it loads, at the right, microarray heatmap with distortion.*

Now, if we visit the *View* menu, two options are unlocked: *Microarray Heatmap* and *Parallel Coordinates*. Lets click on *Microarray Heatmap*. A new window pops up with a typical microarray data heatmap (see fig. 6.2). This visualization by itself allows to watch at the whole raw microarray: you can navigate through the heatmap, where rows are genes and columns are experiments. Expression levels are in a blue-white-red scale, where blue colors are for low expression levels and reds for high expression levels, becoming paler is they are close to the mean. By navigating through the matrix, selected expression levels are distorted, amplifying the corresponding gene and condition profiles, along with their gene and condition names. We can observe that the genes at the bottom of the heatmap, are bright green. These are 10 artificial genes that we have added with low expressions through all

the experiments.

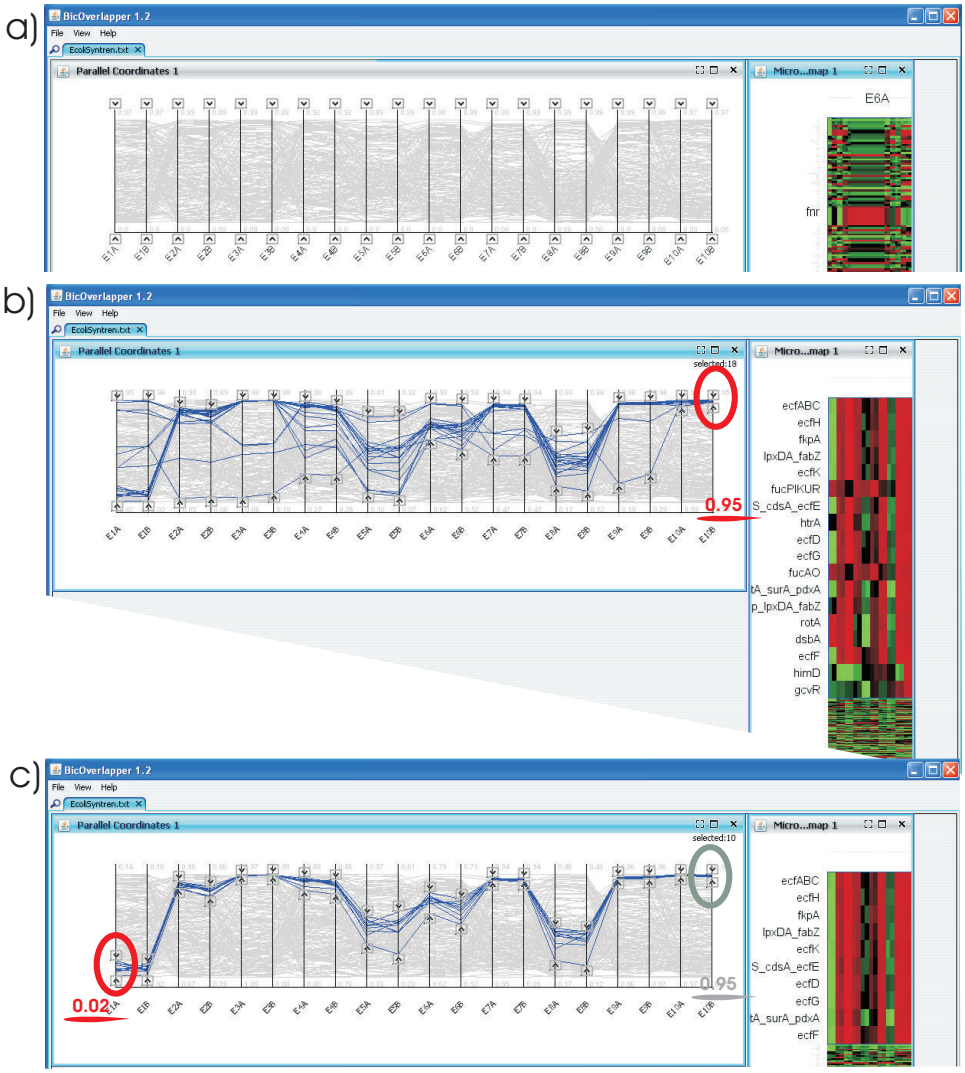


Figure 6.3. *a) Parallel coordinates as they are loaded. b) Parallel coordinates after the selection of genes with the highest expression levels for experiment* E10B*, patterns are revealed in parallel coordinates and microarray heatmap. c) Refinement of the genes that have also low expression levels for experiment* E1A*.*

Now, select *Parallel Coordinates* in the *View* menu. A new window pops up with the parallel coordinates, an entangled mall of lines that flow through 20 parallel vertical lines, each one with the name of one of the experiments in our data (see Fig 6.3). Each of the polygonal lines correspond to a gene profile. Place the

window where you feel more comfortable with it, allowing to see the heatmap at the same time.

Without further interaction, this is a noisy view like the heatmap, since all the gene profiles are being displayed. But now we can filter gene profiles by using the scrolls at the ends of the vertical axis. For example, take the bottom scroll of the last condition (*E10B*) and put it at around 0.95. You will get approximately 20 profiles, those of the genes that have an expression level over 0.95 for condition *E10B*. Also, heatmap rows are rearranged and distorted to show the same gene profiles.

With this simple movement, some patterns can be seen: E10A and E10B show a strong correlation, shared in part with *E9A, E9B, E7A, E7B* and inverted in *E1A* and *E1B*. Now take *E1A* upper scroll and set it down to 0.20. A narrower overall profile for around 10 genes appears.

You can see that the other scrolls have automatically been set to the uppermost and lowermost expression levels of the selected genes for each condition. If you try to drag back the *E1A* upper scroll to a higher level, nothing happens. Dragging a scroll adds or removes the profiles that, *while in the limits of all* other coordinates, are also in the new limits set for the coordinate. If we want to take all profiles that are just within the limits of the dragged coordinate, press *Ctrl* while dragging the scroll.

Take a time to play with parallel coordinates. You will see other overall relationships between genes that can lead to draw preliminary conclusions about the microarray. When you are done, select the initial group of gene profiles with high expression levels in *E10B*.

6.2 Word Cloud Analysis

Following the above discussed example, we can now identify which annotations correspond to the 10 genes selected in fig. 6.3c. We can open two word clouds in the *View/Word Cloud* menu option. Initially, if there is any selection (like in this case), the program searches for annotations for the selected genes. It will take some seconds and then the Word Cloud will display the words with a size proportional to their frequency in the genes. Then, we can play with the combo boxes at the bottom of the Word Clouds in order to select the type of gene annotations, the split of words or, in the case of GO terms, the kind of ontology or the use of p-values to determine size[3]. Fig. 6.4 illustrates the di-word frequency in gene descriptions and the occurrences in cellular component GO terms. We can quickly see that 3 of the 10 genes are related to *membrane*, and other close annotations (*plasma membrane*, *cell outer membrane*).

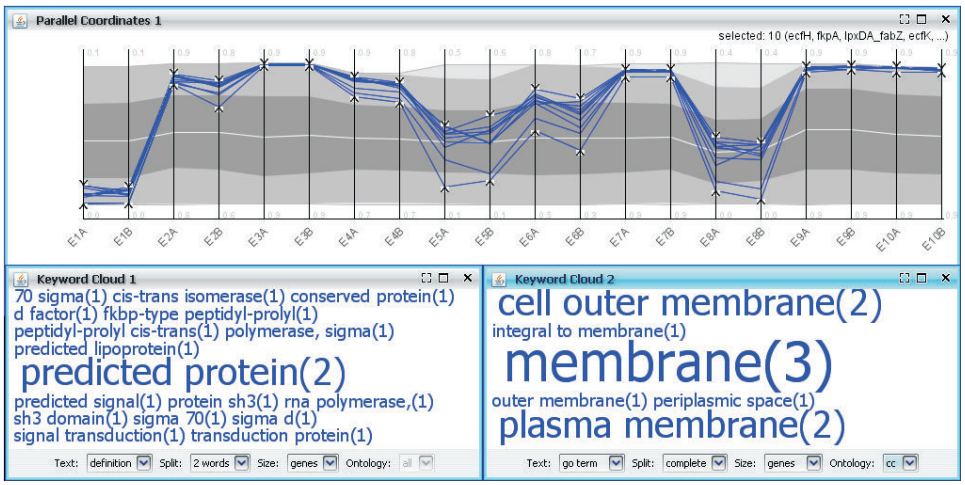


Figure 6.4. *Word clouds for the selection in parallel coordinates. Two word clouds are displayed for the annotations selected with parallel coordinates in fig. 6.3c. The left one shows frequent combinations or two words in gene descriptions. The right one shows the complete GO cellular component terms for these genes, revealing a relationship with the membrane.*

[3]Note that this latter option is only available for expression matrices with row identifiers related to a BioConductor annotation package

6.3 TRN Data Analysis

Now that we are used with our microarray, let's load available transcriptional knowledge for the E. coli, as a TRN data. Go to the *File* menu and click on *Open TRN* and select *Ecoli190.xml*, in the same folder. After the file is loaded, see that the name of the working desktop has changed to *EcoliSyntren.txt|Ecoli190.xml*. By having separate files for microarray, TRN and biclusters, you can combine the microarray data with different bicluster results or TRNs.

Now, in the *View* menu, *Transcription Network* is unlocked. Note that, for a correct analysis and to avoid program failures, the TRN loaded must have as node names the same gene names that we have in the microarray (additionally, other nodes can be in the TRN). When clicking on *Transcription Network*, a third window pops up, with an entangled network that quickly disentangles thanks to force directed layout, letting to see transcription relationships (see fig. 6.5).

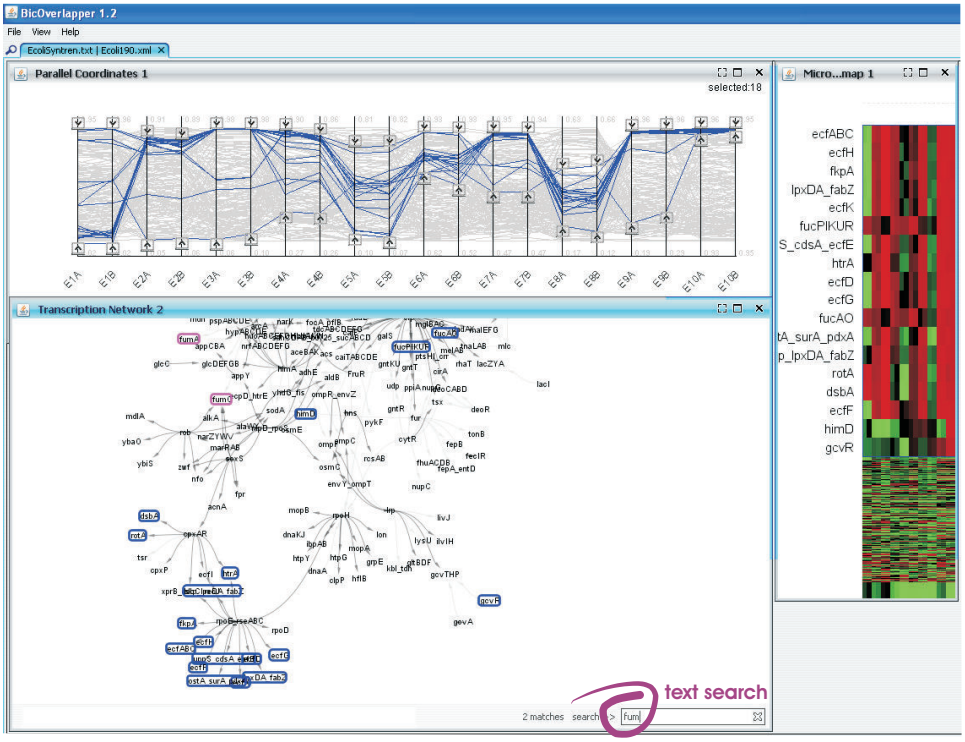


Figure 6.5. *In pink, two genes searched, those which start with* fum*. The blue genes are the ones selected by parallel coordinates, most of them in the same branch of the network.*

Nodes have gene names, and dark grey (activation) or light grey (inhibition) edges connect transcriptionally regulated genes. TRNs have a small-network configuration, with hub genes regulating a high number of other genes. In the TRN,

we can move node positions (by dragging nodes), zoom in and out (with the mouse wheel), navigate the graph (by clicking in the background and dragging), fit the graph in the screen (right-click on the background), search for gene nodes with the bottom-right text box and select nodes by clicking them.

Some gene nodes are highlighted by a blue stroke, corresponding to the gene profiles selected in the parallel coordinates. If we have selected the gene profiles high on *E10B*, we can see that the corresponding genes are mainly in the *rpoE_rseABC* family, but there are other genes, not so close in the TRN, with these high profiles, as *fucAO* or *fucPIKUR*.

Let's search for *lrp*. A pink box highlight this gene. Select it by clicking. See that parallel coordinates and heatmap now show *lrp* expression levels. Now select *gltBDF*, near to *lrp*. Now, by maintaining pressed a key while clicking, select *lrp* in addition to *gltBDF*, and also select *ilvIH*. You can see that their expression profiles are very similar. Now, while pressing a key, select also *lysU*, *livJ* and *kbl_tdh*. Their behavior is just the inverse to *lrp*, as can be seen by the inhibition edges in the TRN.

6.4 Bicluster Analysis with Bubble map

Bicluster analysis is the main focus of BicOverlapper, so let's take some time with it. Load the bicluster data file *turnerSpectralBics.txt* with *File/Open Biclustering Results*. This file has the biclusters found by Turner Plaid Model and Spectral Biclustering in our E. coli synthetic microarray. Note that gene and condition names in the microarray file must coincide with gene and condition names in the bicluster file.

Now, select *View/Bubble Map* to show a Bubble map visualization of biclusters. This is a 2D-projection of biclusters depending on the genes and conditions grouped by them. If two biclusters are near, or partially overlapped, they are likely to share genes or conditions. Each bicluster is a transparent circle, different colors identifying different biclustering methods. The size conveys the number of expression levels grouped by the bicluster.

As in the TRN, you can zoom in and out, fit to the screen, navigate the display, search for gene or condition names (highlighting the biclusters that contain such gene or condition) and select single biclusters (see Fig 6.6).

Let's select *Turner2*, the biggest purple bicluster of the two at the left bottom of the display (biclusters are named with the bicluster set name plus an number). Click on it. Parallel Coordinates, TRNs and Heatmap reconfigure to highlight the genes and conditions in *Turner2*. The Parallel Coordinates axes have been rearranged to show at the left (in a brighter color) the conditions grouped by *Turner2*. The same happened in the microarray, distorting the corresponding rows and columns. TRN highlighted all the genes in the bicluster, most of them in the same families (*rpoE_rseABC* and *cpxAR*), but not all. For example, *uhpT* is far away and has been grouped with the rest. You can now select *uhpT* on the TRN (or any other gene of your interest), to further inspect its profile, see if it is in some other biclusters, etc. Take your time to play with other genes and biclusters and

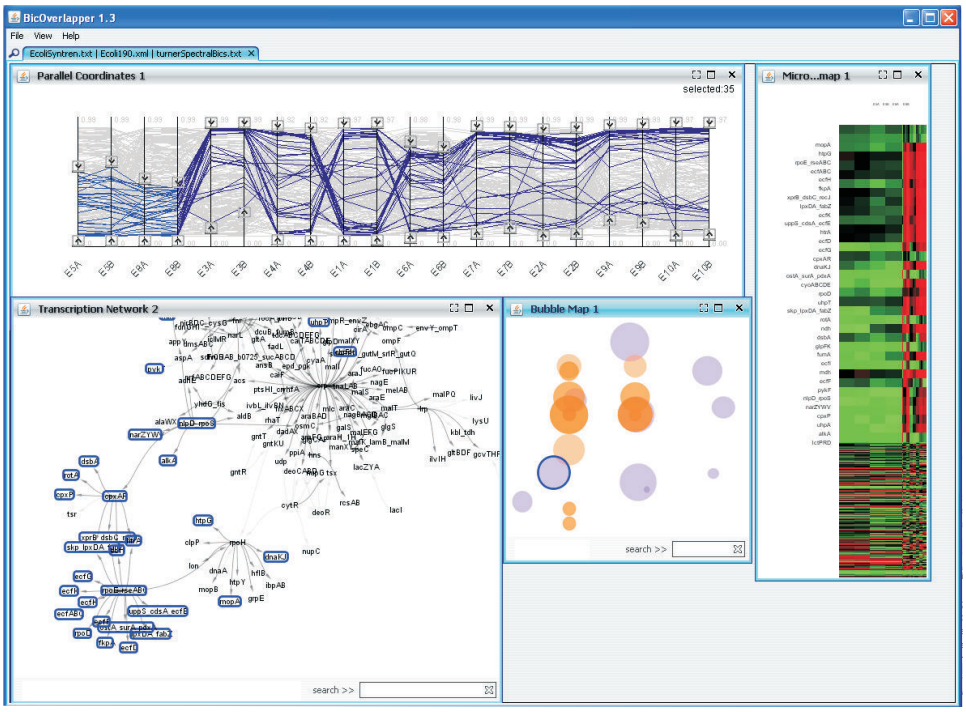


Figure 6.6. *Bicluster* Turner2 *has been selected in the bubble map (int the top-right window, the bubble wrapped in blue). Profiles for the genes grouped by* Turner2 *are selected in parallel coordinates (top left), with axes reordered to show the conditions in the bicluster at the left, and the segment line brighter. Heatmap (right) has also reordered and distorted the rows and columns corresponding to genes and conditions in the bicluster. Finally, the TRN graph (bottom left) highlights nodes corresponding to genes in the bicluster. Strong correlations for some of the genes in the bicluster for conditions not in the bicluster are revealed by the parallel coordinates. According to the TRN graph, most of the genes in the bicluster are in the same two bottom left regulation groups, but some of them are not directly related.*

try to find some relevant patterns.

6.5 Bicluster analysis with Overlapper

Although Bubble map helps to inspect biclusters, and see their biological relevance in the other views, 2D-projection is not good enough to reflect the degree of coincidence and interrelationship between biclusters. For this, the Overlapper is used. Select *View/Overlapper*. A new window pops up. This is a huge one, so maybe you have to close or reorder the other windows (see fig. 6.7). The visualization needs a couple of minutes to understand, but once done it's very useful. Please

read section 3.6 for a full description of the visualization and interaction.

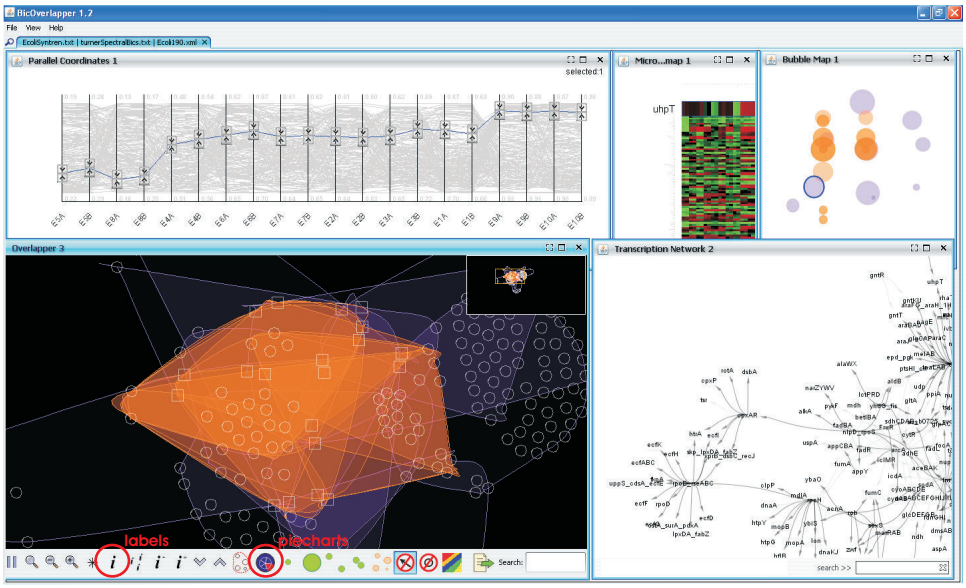


Figure 6.7. *Overlapper and the other visualizations displayed altogether.*

First thing we can appreciate is that biclusters are much more overlapped than what the bubble map showed, being Spectral biclusters (orange hulls), much more overlapped than Turner biclusters (purple ones). Also, both biclustering methods coincide in some of the genes and conditions related. For a more detailed view, click on the piechart button (the one with the radial, blue and red, circle). Now piecharts are drawn for genes and conditions that are in more than one bicluster.

This way we highlight sets of genes grouped together twice or thrice for Turner biclustering (groups of purple, two-three sector piecharts), or highly relevant genes and conditions that appear in several biclusters of both methods (piecharts with lots of sectors, both purple and orange). You can show or hide labels on the nodes with the *"i"* button, and search for node (gene or condition) names.

See conditions as *E4A* and *E4B* or *E1A* and *E1B* (if you don't find them you can search for them with the textual search box at the right-bottom). Observe how they are positioned by pairs. This is logical, because in this case the experimental conditions are experiments replicated twice, so biclustering methods tend to group together such pairs. Search for some relevant genes, as *aspA* and *arcA*, or *lysU*, *lacl* and *livJ*.Lets select, by clicking and maintaining a key pressed, *lysU, lacl, livJ, E1A* and *E1B* (see Fig 6.8). You can also draw a line around to select them all at once. Patterns in parallel coordinates and heatmap reveal that the three genes have a strong correlation, under *E1* conditions but also under the rest. They are mainly grouped by Spectral biclustering (orange), as we can see in the bubble map and in the predominance of orange piechart sectors. Finally, *livJ* and *lysU* are related in the TRN, but *lacl* is very far. However, Spectral biclustering groups the

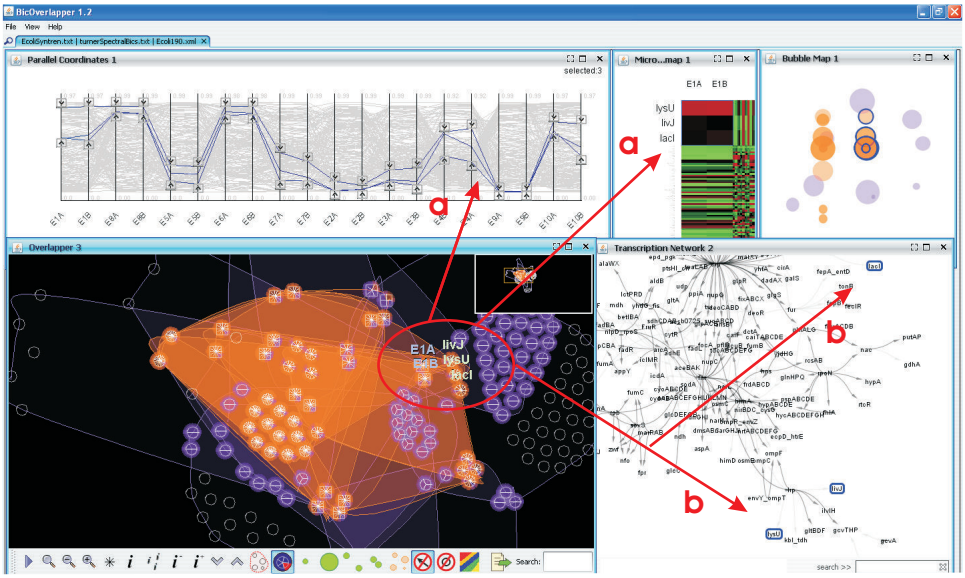


Figure 6.8. *Overlapper focus our attention in 'supergroups' of genes and conditions (genes and conditions highly grouped by one or more biclustering algorithms). Here we have selected five of it (three genes,* lysU, lacI *and* livJ *and two conditions,* E1A *and* E1B*). See how very correlated gene patterns appear for the three genes along most of the conditions (a), and that* lacI *has no direct relation with the other two genes in the TRN (b).*

three together in various biclusters. Have we discovered an unknown relationship? These are the kind of things that are expected to be discovered by the use of the tool: to have a broad view of the problem and to be capable of focusing in certain details, with the help of different views, specially the Overlapper, that helps us to aim efforts in highly grouped genes and conditions.

The number of biclusters and their overlaps may be high for some biclustering algorithms. You can simplify the display in different ways. First, you can hide hulls if they clutter de visualization (*Hide hulls* button or key h), or hide just their contours (key x). Second, you can remove from the visualization the nodes with low degree of overlap with the threshold filter (the buttons marked with "$\vee$" and "$\wedge$"). We can press repeatedly the "$\wedge$" button to do the filtering more strict. With this filtering, we can "cut branches and get the trunk" of biclustering analysis. However, you must be careful if more than one biclustering method is in the analysis, because if one method returns much more biclusters than the others, it surely will be favored by this threshold. Third, you can display just the dual nodes instead of every gene and condition node. The dual nodes resume all the nodes that are together in every bicluster, so they represent tight relationships among genes and/or conditions.

For example, if we click three times the "$\wedge$" button, we reduce the display to the genes and conditions that are at least grouped by three biclusters. Now,

some genes and conditions grouped together in three or more biclusters are located nearly, thus forming a kind of *"superbiclusters"* (see fig. 6.9, left). This is even more clear if we switch to dual nodes, by pressing the u key (fig. 6.9, right). Nodes now represent groups of genes and conditions grouped exactly by the same biclusters (its size represents the number of genes and conditions, the sectors, the number of biclusters they are in together). In both cases, the hull contours have been removed by pressing the x key.

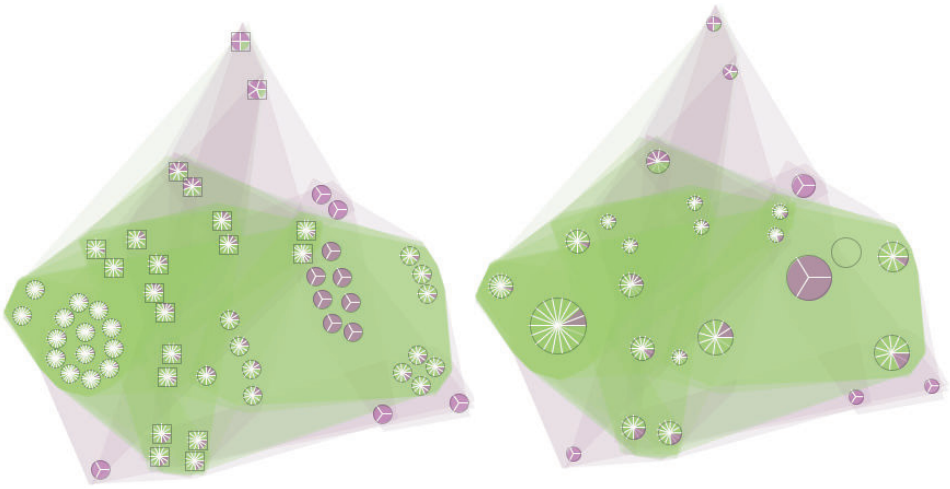


Figure 6.9. *Overlapper filtered by filtering out the nodes grouped less than three times. Left: genes and condition nodes. Right: dual nodes.*

We can also select genes or conditions in other visualizations and see the corresponding biclusters in the Overlapper. For example, search in the TRN for *lrp* gene and select it. Two Turner biclusters have it. In both, it is grouped with *E10* conditions and *fadBA*, a gene not directly related in the TRN (see fig. 6.10).

Now that you know the basics of BicOverlapper, take your time doing analysis tasks similar to the examples given. Of course, this is a synthetic example and any discovery has no true biological meaning. You can try with your own results sets and see what happens. And please if you have any questions, ideas, suggestions or critics, email us, we are eager to improve the tool with your feedback!

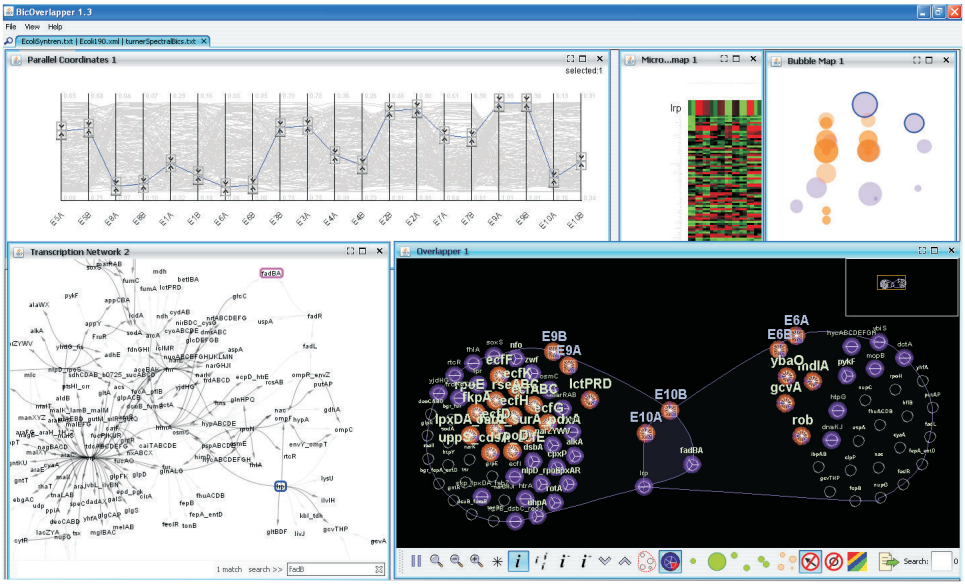


Figure 6.10. *This gene is grouped by two biclusters in Turner Model, thus revealing a bicluster relationship with* fadBA *and* E10 *conditions.*

APPENDIX: INTERFACE TABLES

Table 1. *Heatmap Interface*

Option	Local effects	Other visualizations effects
Navigate over a expression level	Distortion of gene and condition profile	None
Click on expression level	Distortion of gene and condition profile	Selection of gene and condition related
	Reordering of rows and columns	Selection of biclusters that group the gene and condition
Click on gene name	Distortion of gene profile	Selection of the gene or the biclusters grouping that gene
Click on condition name	Distortion of condition profile	Selection of the condition
	Reordering of columns	Selection of the biclusters grouping that condition
		TRN colors nodes with the expr. levels for the condition

Table 2. *Parallel Coordinates Interface*

Option	Local effects	Other visualizations effects
Drag a scroll arrow	Selection of gene profiles with expression profiles in the limits of all the axes Auto-scrolling for the rest of axes	Selection of genes Selection of biclusters containing at least one of these genes
Drag a scroll arrow + Ctrl	Selection of gene profiles with expression profiles in the limits for only this axis	Selection of genes Selection of biclusters containing at least one of these genes
Drag an axis	Horizontal reordering of the axis	None

Table 3. *TRN interface*

Option	Local effects	Other visualizations effects
Drag a node (gene)	Change its position	None
Roll mouse wheel	Zoom in or out	None
Drag the background	Navigate the graph	None
Search text	Highlight genes that contain the text	None
Click a non selected node	Select the node	Selection of genes and biclusters it is in
Click a selected node	De-select the node	De-selection of genes and biclusters it is in
Ctrl+click a non selected node	Select the node in addition to other selections	Selection of genes and biclusters it is in in addition to other selections

Table 4. *Word Cloud interface*

Option	Local effects	Other visualizations effects
Description box	Select the type of annotations	None
Split box	Select the type splits	None
Size box	Select the characteristic used for size	None
Ontology box	In case of GO terms, select the ontology	None

Table 5. *Bubble map interface*

Option	Local effects	Other visualizations effects
Drag a bubble (bicluster)	Change bubble position	None
Roll mouse wheel	Zoom in or out	None
Drag the background	Navigate the map	None
Search text names	Highlight biclusters that group genes/conditions with this text	None
Click a bubble	Select the bubble	Selection of bicluster Selection of genes/conditions in the bicluster

Table 6. *Overlapper interface*

Option	Local effects	Other visualizations effects
Drag a node	Change node position	None
Key pressed + drag a node	Fix node position	None
Drag small square in the top-right view	Navigate the graph	None
Hover over a node	Highlight the node Show the node name Highlight all neighbor nodes	None
Hover over a hull	Highlight the hull	None
Click on a node	Select the node Show the node name	Selection of the gene or condition Selection of the biclusters with the gene or condition
Right-click on a gene node	Retrieve and show gene annotations	None
Click on a hull	Select the hull Select all nodes in the hull and show their names	Selection of the genes and conditions in the hull
Click and drag the mouse out of a node	Select nodes wrapped by the drawn line	Selection of the genes and conditions in the wrapping area
Pause-restore simulation button	Pauses/restores the force simulation	None
Show overview button	Shows-hides top-right overall view	None
Zoom out-in buttons	Zoom out-in the graph (unstable)	None
Show names button	Show-hide node labels	None
Absolute-relative label size button	Make label size change with connection degree	None
Increase-decrease name size buttons	Increase or decrease font size	None
Increase-decrease threshold buttons	Alter by one the number of biclusters a node must be in order to be drawn	None None
Hide-draw zones button	Hide or show hulls	None
Draw node as piecharts button	Hide or show piecharts	None
Increase-decrease cluster size buttons	Alter the spring length	None
Increase-decrease repulsion buttons	Alter gravitational force	None
Change to relative size button	Alter node size to convey connection degree	None
Hide nodes button	Draw or hide nodes	None
Color configuration button	Display a color configuration panel	None
Export image button	Display a save dialog to save visualizations as an image file	None
Search text box	Search for input text in node labels Highlight nodes found If only one node is found, centers on it	None
Hit c key	Draw hulls as polygonal or spline shapes	None
hit t key	Draw bicluster names	None
Hit x key	Show/hide hull contours	None
Hit u key	Draw simple/dual nodes	None
Hit e key	Show/hide edges	None

Bibliography

[1] S. Kaiser and F. Leisch, "A toolbox for bicluster analysis in r," Tech. Rep. 028, Ludwing-Maximilians-Universitat Mnchen (2008).

[2] S. Barkow, S. Bleuer, A. Prelic, P. Zimmermann, and E. Zitzler, "Bicat: a biclustering analysis toolbox," *Bioinformatics* **22**(10), 1282–1283 (2006).

[3] T. V. den Bulcke, K. V. Leemput, B. Naudts, P. van Remortel, H. Ma, A. Verschoren, B. D. Moor, and K. Marchal, "Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms," *BMC Bioinformatics* **7**(43) (2006).

[4] W. G. Alvord, J. A. Roayaei, O. A. Quiones, and K. T. Schneider, "A microarray analysis for differential gene expression in the soybean genome using bioconductor and r," *Briefings in Bioinformatics* **8**(6), 415–431 (2007).

[5] A. Prelic, S. Bleuer, *et al.*, "A systematic comparison and evaluation of biclustering methods for gene expression data," *Bioinformatics* **22**(9), 1122–1129 (2006).

[6] T. M. Murali and S. Kasif, "Extracting conserved gene expression motifs from gene expression data," *Proc. Pacific Symp. Biocomputing* **8**, 77–88 (2003).

[7] Y. Cheng and G. M. Church, "Biclustering of expression data," *Proc. Int'l Conf Intell Syst Mol Biol.* **8**, 93–103 (2000).

[8] H. Turner, T. Bailey, and W. Krzanowski, "Improved biclustering of microarray data demonstrated through systematic performance tests," *Computational Statistics and Data Analysis* **48**, 235–254 (2003).

[9] H. L. Turner, T. C. Bailey, W. J. Krzanowski, and C. A. Hemingway, "Biclustering models for structured microarray data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics* **2**(4), 316–329 (2005).

[10] L. Lazzeroni and A. Owen, "Plaid models for gene expression data," tech. rep., Standford University (2002).

[11] R. Santamaria, L. Quintales, and R. Theron, "A framework to analyze biclustering results on microarray experiments," *Lecture Notes in Computer Science* **4881**, 770–779 (2007).