

GENE EXPRESSION

BiVisu: Software Tool for Bicluster Detection and Visualization

K.O. Cheng*, N.F. Law, W.C. Siu and T.H. Lau

Centre for Signal Processing, Department of Electronic and Information Engineering, The Hong Kong Polytechnic University, Hong Kong.

Associate Editor: Dr. Olga Troyanskaya

ABSTRACT

Summary: BiVisu is an open-source software tool for detecting and visualizing biclusters embedded in a gene expression matrix. Through the use of appropriate coherence relations, BiVisu can detect constant, constant-row, constant-column, additive-related as well as multiplicative-related biclusters. The biclustering results are then visualized under a two-dimensional setting for easy inspection. In particular, parallel coordinate (PC) plots for each bicluster are displayed, from which objective and subjective cluster quality evaluation can be performed.

Availability: BiVisu has been developed in Matlab and is available at <http://www.eie.polyu.edu.hk/~nflaw/Biclustering/>.

Contact: ennflaw@polyu.edu.hk

1 INTRODUCTION

Biclustering aims at detecting co-regulated genes in a gene expression matrix obtained from microarray experiments. Unlike classical clustering methods which partition data based on whole set of genes or conditions, biclustering groups a subset of genes (rows) over a subset of conditions (columns). Hence, genes which are co-regulated under certain biological processes can be identified. It is well-known that the biclustering process is NP-complete, (Madeira and Oliveira, 2004) so an efficient but reliable biclustering tool is highly desirable. Several efficient approaches (Cheng and Church, 2000; Wang *et al.*, 2002; Yoon *et al.*, 2005) have been proposed for detecting additive-related biclusters. These approaches can be applied for multiplicative models if the log values of expression data are processed. However, results deteriorate when noise is present. Besides biclusters detection, an efficient way for visualizing the high-dimensional biclusters is often desired. In order to integrate bicluster detection and visualization, a novel biclustering algorithm with the use of the parallel coordinate (PC) plots is implemented in the BiVisu software tool.

2 METHODS AND IMPLEMENTATIONS

The biclustering software, BiVisu, is based on using the parallel coordinate (PC) plots in the representation of a data matrix (Inselberg and Dimsdale, 1990; Wegman, 1990). In the PC plot, all axes are drawn in parallel to each other in a plane. The main problem is that biclusters are hidden in the PC plot if the axes are not arranged appropriately. However, once axes are arranged properly, biclusters can be visualized. We have proposed a split-and-merge algorithm for bicluster detection. Details of the algorithm are given in our article which was recently submitted to IEEE/ACM

Transactions on Computational Biology and Bioinformatics. An overview of the algorithm is provided in the webpage.

In BiVisu, row clustering is first performed by comparing every two-columns and potential row clusters in each column-pair are identified. These row clusters are then intersected to identify column pairs that can be merged together to form big biclusters. Note that the intersection process would not be performed if there is a significant drop in the number of rows after merging certain columns onto the current biclusters. The type of biclusters found is determined by how columns are compared. If the column pair is compared by calculating their differences in expression levels, additive-related biclusters are found. If the column pair is compared by calculating the ratio of their expression levels, multiplicative-related biclusters are found.

BiVisu has been developed in Matlab. Besides bicluster detection, BiVisu also provides functions for pre-processing, filtering and bicluster analysis. Details are given below,

- *Pre-processing:* the user can decide whether a logarithm function is required for the input data. Logarithm is necessary for detecting multiplicative-related biclusters using conditions set for the additive model.
- *Biclustering:* this performs bicluster detection. The user selects the type of biclusters to be detected, either an additive model or a multiplicative model. For both models, there are two mandatory and two optional parameters to be specified. The two mandatory parameters are the noise threshold and the minimum percentage of rows in a bicluster. The noise threshold controls the coherence of genes over the subset of conditions in biclusters. If the difference between a particular expression level and the cluster centroid is smaller than the noise threshold, it is considered to be in the cluster. The minimum percentage of rows in a bicluster acts as a stopping criterion for growing a bicluster as well as a pre-set requirement for a valid bicluster. The two optional parameters specify another two requirements for a valid bicluster, namely the minimum number of conditions (columns) in a bicluster and the maximum percentage of overlapping allowed among the detected biclusters. The latter sets an upper bound for the maximum percentage of rows and columns that can be overlapped between two biclusters. Appropriate settings can avoid unnecessary processing and reduce the processing time.
- *Filtering:* Biclusters can be filtered subject to criteria: the minimum number of rows, minimum number of columns, maximum number of biclusters and maximum percentage of overlapping allowed. The filtering can refine the result without re-performing the expensive biclustering process.
- *Bicluster analysis:* The main tool for visualizing biclusters is PC plots. BiVisu allows navigation of biclusters one by one. Apart from raw expression values, options of displaying the difference matrix and ratio matrix are available for additive

*To whom correspondence should be addressed.

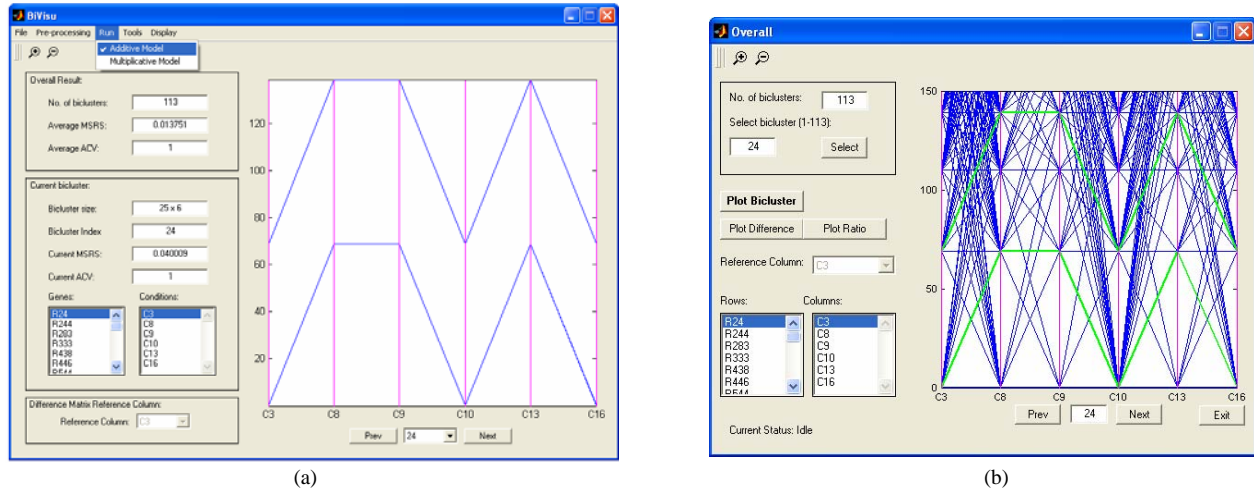


Fig. 1. (a) The main window of the BiVisu software. Core functions are accessible from the menu bar. The left panel shows information about the biclustering results while the right panel shows the PC plot of the currently selected bicluster. (b) A separate window showing the PC plot of all genes over the conditions of the currently selected bicluster. The genes in the selected bicluster are in green color while genes not in the selected bicluster are in blue color.

models and multiplicative models respectively so that the coherence of genes within each bicluster can be judged in a subjective manner. The genes inside and/or outside a bicluster can be drawn for comparison purpose. Besides PC plots, one of the common visualization tools called heat maps is included too. Mean square residue score (MSRS) (Cheng and Church, 2000) and average correlation value (ACV) (Teng and Chan, 2006) are provided as objective measurements of coherence. Other available information includes bicluster size, names of genes and conditions. The detected biclusters and all the summary information can be saved as text files.

3 RESULTS

The BiVisu program is applied on yeast *Saccharomyces cerevisiae* cell cycle dataset (Cheng and Church, 2002), which consists of 2884 genes and 17 conditions, for demonstration. It runs on Matlab 6.5 in a computer with Intel Pentium 4 2.4GHz CPU. Fig 1 (a) shows the graphical user interface of the BiVisu program. The main functions such as pre-processing, model selection, filtering and display options can be accessed from the menu bar. Statistical information regarding the individual bicluster and overall results are given in the left panel. The PC plot of the selected bicluster is shown at the right panel. Below the PC plot is the navigation interface for individual bicluster. A separate PC plot showing genes in the selected bicluster and the outside genes is illustrated in Fig. 1 (b). This diagram helps comparing the expression levels of the genes in the current bicluster and those of the other genes. For an additive model, the minimum percentage of rows in a bicluster, noise threshold, minimum number of columns and maximum percentage of overlapping allowed are 0.6, 4, 6 and 80 respectively. There are 113 biclusters detected with an average MSRS of 0.013751 and an average ACV of 1. These values and the PC plots of biclusters show high homogeneity within each detected biclusters. For a multiplicative model, 97 biclusters are found with an average MSRS (on log values) of 0.003613 and an average ACV of 0.9996 using the same settings as in the additive model except that the noise threshold is 0.025. The processing time for

additive model and multiplicative model are about 70 and 80 sec respectively.

4 CONCLUSION

An open-source software tool known as BiVisu for detecting and visualizing biclusters from a gene expression matrix is described. The program is applicable to additive models as well as multiplicative models. The parallel coordinate (PC) plot is used as a visualization tool for each detected bicluster. Together with the mean square residue score and average correlation value, subjective and objective judgment of bicluster homogeneity can be achieved. Statistical information and the PC plot of each detected bicluster are provided in a panel to facilitate further analysis. The effectiveness of the BiVisu has been demonstrated using a yeast dataset.

ACKNOWLEDGEMENTS

This work is supported by the Centre for Signal Processing, Department of Electronic and Information Engineering, the Hong Kong Polytechnic University. K.O. Cheng acknowledges the research studentships provided by the University.

Conflict of Interest: none declared.

REFERENCES

Cheng, Y. and Church, G.M. (2000) Biclustering of Expression Data. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 93 – 103.
 Inselberg, A. and Dimsdale, B. (1990) Parallel Coordinates: A Tool for Visualizing Multidimensional Geometry. *Proc. Of Visualization*, 361-378.
 Madeira, S.C. and Oliveira, A.L. (2004) Biclustering Algorithms for Biological Data Analysis: A Survey. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 1, 24 – 45.
 Teng, L. and Chan, L.W. (2006) Biclustering Gene Expression Profiles by Alternately Sorting with Weighted Correlated Coefficient. *Proc. IEEE Int. Workshop Machine Learning Signal Processing*, 289 – 294.
 Wang, H. et al. (2002) Clustering by Pattern Similarity in Large Data Sets. *Proc. ACM SIGMOD Int. Conf. Management of Data*, 394 – 405.
 Wegman, E.J. (1990) Hyperdimensional Data Analysis Using Parallel Coordinates. *J. American Statistical Association*, 85, 664-675.
 Yoon, S. et al. (2005) Discovering Coherent Biclusters from Gene Expression Data Using Zero-Suppressed Binary Decision Diagrams. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, 2, 339 – 354.