BIOINFORMATICS APPLICATIONS NOTE Vol. 24 no. 9 2008, pages 1212–1213
doi:10.1093/bioinformatics/btn076

*Data and text mining*

# BicOverlapper: A tool for bicluster visualization

Rodrigo Santamaría*,†, Roberto Therón† and Luis Quintales

Departamento de Informática y Automática, Pz. de Los Caídos S/N, 37008 Salamanca, Spain

Received on December 17, 2007; revised on February 9, 2008; accepted on February 24, 2008

Advance Access publication March 5, 2008

Associate Editor: John Quackenbush

**ABSTRACT**

**Summary:** BicOverlapper is a tool to visualize biclusters from gene-expression matrices in a way that helps to compare biclustering methods, to unravel trends and to highlight relevant genes and conditions. A visual approach can complement biological and statistical analysis and reduce the time spent by specialists interpreting the results of biclustering algorithms. The technique is based on a force-directed graph where biclusters are represented as flexible overlapped groups of genes and conditions.

**Availability:** The BicOverlapper software and supplementary material are available at http://vis.usal.es/bicoverlapper

**Contact:** rodri@usal.es

## 1 INTRODUCTION

Biclustering is a unsupervised learning technique which over the last few years has been widely used in microarray analysis, outperforming traditional clustering. While clustering techniques group genes similarly expressed under all conditions or viceversa (clusters), biclustering techniques group them under a certain *subgroup* of conditions (groups of both genes and conditions are called biclusters). A gene or condition can be in more than one bicluster at the same time (overlapping), while in clustering a gene or condition is usually assigned to a unique cluster. A complete survey of biclustering algorithms can be found in Madeira and Oliveira (2004).

Biclusters are more flexible and fit biological behavior better than clusters, but their special characteristics (overlapping and grouping of genes and conditions) make it difficult to apply cluster visualizations to biclusters. While some cluster visualization techniques can be adapted to the representation of single biclusters [for example, heatmaps or parallel coordinates as in Barkow *et al*. (2006) and Cheng *et al*. (2007)], the simultaneous visualization of biclusters from one or more biclustering methods is a less explored field. Biclustering outputs range from one to thousands of biclusters that must be individually inspected, a slow task, or filtered using statistical methods or biological knowledge.

Even with these filters, it is difficult to show the selected biclusters in a single view because of overlapping. For example, Grothaus *et al*. (2006) visualize various biclusters in a heatmap, but the method needs replication of rows and columns because of the geometrical limitations of heatmaps. This replication of rows and columns increases the space needed for visualization and could lead to confusion. BicOverlapper is based on a novel visualization technique that simultaneously displays different biclusters, addressing the problem of bicluster overlapping.

## 2 METHODS

Our tool is based on a graph where nodes represent genes or conditions, and edges join nodes that are grouped by one or more biclusters (Fig. 1A). Therefore, each bicluster is represented as an undirected complete subgraph.

To avoid edge cluttering, edges are not drawn and, instead, each bicluster is wrapped in a rounded shape (hull) built by splines that take the positions of the outermost nodes in the bicluster as anchor points. Unlike other zone graph visualizations such as those of Perer and Shneiderman (2006), a node can be in more than one zone, reflecting overlapping between biclusters, which can usually affect more than a node. Hulls are drawn with a transparent color, so intersecting areas become more opaque and easily distinguishable.

The nodes are positioned following a force-directed layout (Fruchterman and Reinhold, 1991). In this model, each pair of nodes can be affected by up to two forces. If the nodes are connected, a spring force acts to keep them close. Additionally, an expansion force repels every pair of nodes, whether connected or not. This way, nodes in the same biclusters tend to be close, while nodes in different biclusters are separated.

Apart from node positions, additional information is given by node representation, by means of glyphs. A glyph is a graphical object designed to convey multiple data values (Ware, 1999). The geometrical properties of the glyph represent different dimensions. Shape distinguishes between genes (circles) and conditions (squares). Pie charts with as many sectors as biclusters in which the node is grouped are used to convey the degree of intersection.

In order to foster knowledge discovery, the visualization is not a static image, but a user-driven representation that can be manipulated in a number of ways. Besides controlling node representation, the user can change force parameters, drag and fix node positions, search for gene or condition names, visualize or hide nodes, edges and hulls, highlight the nodes connected to a particular node, navigate through the graph without losing overview and export images to different formats.

## 3 RESULTS

For demonstration purposes, the tool has been applied to the biclustering results of the Order Preserving SubMatrix search algorithm (OPSM) (Ben-Dor *et al*., 2003) and Bimax (Prelic *et al*., 2006) in the analysis of a microarray data matrix containing two types of Diffuse Large B-Cell Lymphomas

*To whom correspondence should be addressed.
†The first two authors should be reported as joint first authors.

 

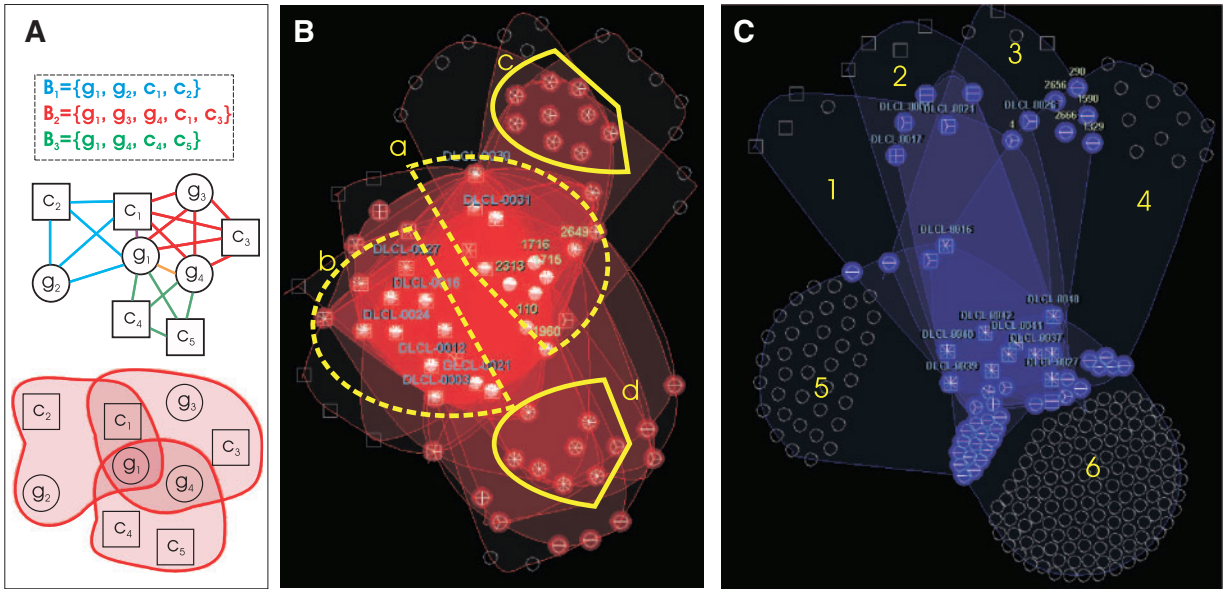


**Fig. 1.** **A**) Three simple biclusters and their representation. $c_1$ and $g_1$ appear on both $B_1$ and $B_2$ so the edge between them is shorter (a) and they appear in the intersecting area in the final visualization (b). Intersecting areas are more opaque to highlight overlapping. **B**) Representation of the 50 biggest Bimax biclusters. A central group of some genes and several conditions appear at the center, slightly separated in groups a) and b). High overlapping of biclusters, conveyed as thicker areas is explained by the exhaustiveness of Bimax biclustering. Some genes, less overlapped than the central group but still closely related, are formed in c) and d). The names of some relevant genes and conditions have been highlighted. **C**) OPSM bicluster visualization. Biclusters grouping mainly conditions (1, 2, 3) or genes (4, 5, 6) are easily identified, revealing asymmetry in OPSM method. The relaxed condition of order preservation searched by OPSM produces very large biclusters in some cases (5, 6). Conditions grouped in all the biclusters of OPSM (*DLCL0027*, *DLCL0036*, *DLCL0039-0042* and *DLCL0048*) have a strong influence in order preserving of gene expression levels. Most of them correspond to activated B-like lymphomas.

(DLCL), previously identified by gene-expression profiling (Alizadeh *et al.*, 2000). For Bimax results (Fig 1B), high connectivity of the nodes demonstrates the exhaustiveness of Bimax. A central group of genes and conditions (a, b) with over-expressed levels is easily identifiable. This group is present in almost all Bimax biclusters. Other groups of genes usually biclustered, but less frequently, also appear (c, d).

OPSM biclusters (Fig. 1C) agree in the importance of some DLCL lymphomas, all conditions classified as *activated B-like* lymphomas by Alizadeh *et al.* (2000). Regarding the grouping criteria of OPSM, these conditions are very interesting because they are able to keep an order in expression levels over a high number of genes (those in biclusters 5 and 6).

## 4 CONCLUSION

We present a novel visualization technique that allows the simultaneous representation and interaction with biclusters, gaining insight into overall biclustering results. The overlap between biclusters is visualized by means of intersecting hulls, thus solving one of the most serious problems with bicluster visualization. The use of glyphs on gene and conditions nodes improves our understanding of instances of overlapping when the representation becomes complex. The effectiveness of BicOverlapper has been demonstrated using a lymphoma dataset, extracting actual biological features through the interaction with the tool without wasting time inspecting biclusters individually. Following these promising results, the tool is currently being upgraded with new linked visualizations within a visualization framework and by means of improvements in the graph layout algorithm.

## ACKNOWLEDGEMENTS

This work was supported by the MCyT of Spain (project TIN2006-06313) and by a grant from the Junta de Castilla y León.

*Conflict of Interest*: none declared.

## REFERENCES

Alizadeh,A.A. *et al.* (2000) Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, **403**, 503–511.

Barkow,S. *et al.* (2006) Bicat: a biclustering analysis toolbox. *Bioinformatics*, **22**, 1282–1283.

Ben-Dor,A. *et al.* (2003) Discovering local structure in gene expression data: the order-preserving submatrix problem. *J. Comput. Biol.*, **10**, 373–384.

Cheng,K.O. *et al.* (2007) Bivisu: software tool for bicluster detection and visualization. *Bioinformatics*, **23**, 2342–2344.

Fruchterman,T.M.J. and Reinhold,E.M. (1991) Graph drawing by force-directed placement. *Softw. – Pract. Exper.*, **21**, 1129–1164.

Grothaus,G.A. *et al.* (2006) Automatic layout and visualization of biclusters. *Algorithms Mol. Biol.*, **1**, doi: 10.1186/1748-7188-1-15.

Madeira,S. and Oliveira,A. (2004) Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Trans. Comp. Biol. Bioinf.*, **1**, 24–45.

Perer,A. and Shneiderman,B. (2006) Balancing systematic and flexible exploration of social networks. *IEEE Trans. Vis. Comp. Graphics*, **12**, 693–700.

Prelic,A. *et al.* (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, **22**, 1122–1129.

Ware,C. (1999) *Information Visualization: Perception for Design*. Morgan Kaufmann, San Diego, California.