BicOverlapper 2.0: Visual Analysis for Gene Expression

Rodrigo Santamaria 1,2; Roberto Therón 1 and Luis Quintales 1,2
1 Department of Computer Science, Pz. San Bartolomé 1-8, 37008 Salamanca, Spain.
2 Instituto de Biología Funcional y Genómica. Zacarias González 2, 37007 Salamanca, Spain.

1 INTRODUCTION

BicOverlapper 1.0 (Santamaria et al., 2008) focused on the visualization of complex gene expression analysis results coming from biclustering algorithms. Based on Venn-like diagrams and overlapping visualization layers, it successfully conveyed biclusters. The visualization layer is developed in Java and communicates with the analysis layer via rJava (Urbanek, 2007). This layer contains several methods for bioinformatics analysis (Davis and Meltzer, 2007; Kauffman et al., 2009), although it requires high bandwidth and not all of the experiments are supported. Data analysis includes:

- Gene Set Enrichment analysis is also implemented via GSEA (Oron and Gentleman, 2008). Enriched gene sets are visualized as overlapping groups.
- Biclustering, as in the previous version, is computed with biclust (Kaiser et al., 2013) package. The ISA algorithm is now also available by the isa2 package.
- Correlation networks. This is a simple yet powerful method to find groups. Genes with low overall expression variation are filtered out and the rest are linked if they have a profile distance below some standard deviations. The resulting network is visualized as a force-directed layout, where nodes can be colored by the expression under selected conditions.

The visualization layer is developed in Java and communicates with the analysis layer via rJava (Urbanek, 2007). This layer contains several visualization techniques, with implementations based onPrefuse (Heer et al., 2005) (networks, scatter plots), Processing (Reas and Fry, 2007) (overlap, heatmap) and plain Java (parallel coordinates, word clouds).

2 APPROACH

During the design of BicOverlapper 2.0, we focused on a high level of interaction and a visual analytics (Thomas and Cook, 2005) approach. Another important design principle was the simplification of installation and interfaces. Finally, following the original ‘overlapping’ philosophy, we designed linked visualizations and an agglomerative use of standard numerical analyses. For example, differential expression analysis compares two experimental conditions, but BicOverlapper 2.0 allows to compare several combination of experimental conditions at once and then visualize the relationships between the differentially expressed groups.

3 METHODS

The tool is implemented as two interconnected layers: visualization and analysis. The analysis layer is R/Bioconductor dependent, using several packages and ad hoc scripts. Finding pairs of overlapping expression groups is supported by its corresponding packages (Davis and Meltzer, 2007; Kauffmann et al., 2009), although it requires high bandwidth and not all of the experiments are supported. Data analysis includes:

- Differential expression with limma (Smyth, 2005). In addition to one-to-one comparisons, BicOverlapper allows to perform multiple comparisons at once, visualized as intersecting differentially expressed groups. This way, analysis time is reduced and the differences between the comparisons can be inspected.
- Gene Set Enrichment analysis is also implemented via GSEA (Oron and Gentleman, 2008). Enriched gene sets are visualized as overlapping groups.
- Biclustering, as in the previous version, is computed with biclust (Kaiser et al., 2013) package. The ISA algorithm is now also available by the isa2 package.
- Correlation networks. This is a simple yet powerful method to find groups. Genes with low overall expression variation are filtered out and the rest are linked if they have a profile distance below some standard deviations. The resulting network is visualized as a force-directed layout, where nodes can be colored by the expression under selected conditions.

The visualization layer is developed in Java and communicates with the analysis layer via rJava (Urbanek, 2007). This layer contains several visualization techniques, with implementations based on Prefuse (Heer et al., 2005) (networks, scatter plots), Processing (Reas and Fry, 2007) (overlap, heatmap) and plain Java (parallel coordinates, word clouds).

4 RESULTS

In order to involve biology specialists on bioinformatics analyses, we need simpler and highly interactive tools. For example, Fig. 1 was generated only by clicking two menu options and selecting one visual item and gene/condition labels, on a process that takes no more than five minutes (see Supplementary video 2). Underneath, this requires the seamless connection of different steps: expression data loading, computation of distribution statistics, three differential expression analyses (for up and down regulation), gene annotation retrieval and the visualization of four interactive representations.

Fig. I provides a considerable amount of information about the experiment. First, parallel coordinates (Inselberg, 2009) indicate with boxplots that the data are normalized, although probably skewed towards up-regulation. Second, differential expression

*to whom correspondence should be addressed
1 https://www.princeton.edu/~abarysh/treeview/
2 http://vis.usal.es/bicoverlapper2/docs/tour.mp4
Fig. 1. Yeast gene expression profile along three cell cycles, from experiment GSE3431 (Tu et al., 2005). Each cell cycle is divided into three time intervals (early, mid and late). Differential expression for every combination of such intervals is computed and visualized as overlapping groups. 36 genes high regulated at early and mid intervals have been selected (intersection between 'early vs late' and 'mid vs late' groups at the bottom, left). Their expression profiles are shown in parallel coordinates and heatmap visualizations. Finally the functional annotations, stacked by term, are shown as a word cloud, indicating for example that 9 out of the 36 genes are related to metabolic and oxidation-reduction processes.

5 CONCLUSION
BicOverlapper is a simple to use, highly visual and interactive tool for gene expression analysis. Easily and without programming knowledge, the user can have an overall view of several expression aspects, from raw data to analysis results and functional annotations. This may significantly reduce the analysis time and improve the analytical discourse with the data. For the future, we are working on the support of high-throughput data, especially RNA-Seq, and a comprehensive report and image generation.

Funding This work was funded by grants FI2010-16234 (MICINN), BFU2011-28804 and Consolider-Ingenio CSD2007-00015 (MINECO)

REFERENCES