Interactive Poster: Visualization of Gene Combinations

Christian Tominski* Clemens Holzhüter Andrea Unger† Heidrun Schumann‡

Institute for Computer Science
University of Rostock

ABSTRACT

The analysis of microarray data is a key to understanding the influence and role of genes. Visualization is one way to support analysts in finding potentially important genes. However, most visualization tools focus on representing single genes, important gene combinations are not always easy to spot and compare with existing approaches.

What we propose here is the novel idea of making gene combinations visually explicit, and thus making them easier to grasp and understand. We describe a tool that integrates several visual and interaction methods to help users gain insight into their microarray data. The tool further provides an interface to plug in analytical methods, which are important to filter relevant gene combinations out of the massive number of theoretically possible ones.

Keywords: Microarray, visualization, gene combination.

1 INTRODUCTION

In recent years, microarray analysis has paved the way to understanding the interplay of genes. Visual methods play an important role in the process of gaining insight from microarray experiments. Today, a variety of useful tools is available, which mostly help in understanding the impact of single genes. However, still much knowledge and experience is required to spot not only important genes, but important combinations of genes. The situation worsens if the analyst has to assess similarities between different gene combinations.

In this work, we describe concept and implementation of a tool that aims at explicit visualization of gene combinations. For that purpose, we make a switch from data items in form of single genes to data items that accommodate gene combinations. Apparently, since theoretically any combination of genes can carry important information, such a switch implies a massive increase in the volume of data to consider for the analysis. We will briefly describe how plugable filters can help to cope with these vast data volumes. The main contribution of this work is an interactive tool that integrates several visual concepts to visualize combinations of genes. We extend the classic heatmap approach in order to emphasize on the representation of gene combinations. The representation is enhanced with additional visual clues to support comprehension of dis/similarities between different combinations of genes.

2 CONCEPT & IMPLEMENTATION

Let us now be more specific in terms of how combinations of genes can be analyzed and visualized.

2.1 Approach Outline

Microarray data contain information on the expression of genes G for several samples S (e.g., time steps), which can be formalized as a function $exp_g : G \times S \rightarrow \mathbb{R}$. Classic heatmap visualizations represent exactly that relationship between genes/samples and corresponding expression. In other words, the visualization focusses on genes themselves. What we pursue is the visualization of gene combinations. This will allow us to assess not only the relationships between single genes, but also the interplay of combinations of genes. A gene combination is a set $GC \in \mathfrak{P}(G)$. In order to represent the expression of gene combinations, we need to aggregate the expression of those genes participating in a gene combination, which can be modeled as a function $exp_{gc} : GC \times S \rightarrow \mathbb{R}$. That is, the aggregated expression value determines the level of regulation of an entire gene combination. Usually, we want to use the average, but it is also possible to consider other aggregates if needed for a particular application.

It is obvious that the step from genes to gene combinations increases the data volume by magnitudes. And indeed, we are not able to represent that huge data volume. However, since we are not interested in all theoretically possible gene combinations, but only in biologically interesting and relevant ones, we can apply the following three-step approach to achieve our goal: 1) Generate gene combinations; 2) on the fly filter out biologically less relevant combinations; 3) visualize only gene combinations that passed step 2).

Step 1) is relatively straight-forward to implement as a serial permutation generator. Explanations on step 2) and 3) will be given in the next paragraphs.

2.2 The Need for Filters

In order to practically visualize gene combinations, it is mandatory to filter out biologically less relevant gene combinations. Commonly, expert knowledge is required to assess which combinations are relevant and which are not. Therefore, we provide two basic options to drive the filtering process.

The first option is to restrict the number of genes to consider for the analysis. That is, the (expert) analyst selects (out of the many genes in a data set) only those that are relevant with respect to the task at hand. This results in significant, though coarse reduction of the data volume.

Secondly, in order to fine-tune the filtering, analytical methods are applied [2]. This step further crystallizes possibly relevant gene combinations. Parameters to control this analysis step are exported to the user interface. Hence, it is easily possible to steer the filtering process interactively.

The result of the described filtering procedure is a set of potentially significant gene combinations, which are passed to the visualization step. Indisputably, the term "biologically significant" depends on many factors. Therefore, we designed the filtering procedure as a modular component, which allows for integration of task and data specific filter implementations.

2.3 Visualization of Gene Combinations

The first question to ask when visualizing abstract data – in our case gene combinations extracted from microarray data – is how to layout information on screen. Since heatmaps are widely accepted among biologists, we have decided to extend that approach for gene combinations. Heatmaps are based on a matrix-like display and commonly use a red-black-green color scale to visualize gene expression (red: up-regulation; black: no regulation; green: down-regulation). We extend that approach as follows.

*e-mail: ct@informatik.uni-rostock.de
†e-mail: aunger@informatik.uni-rostock.de
‡e-mail: schumann@informatik.uni-rostock.de

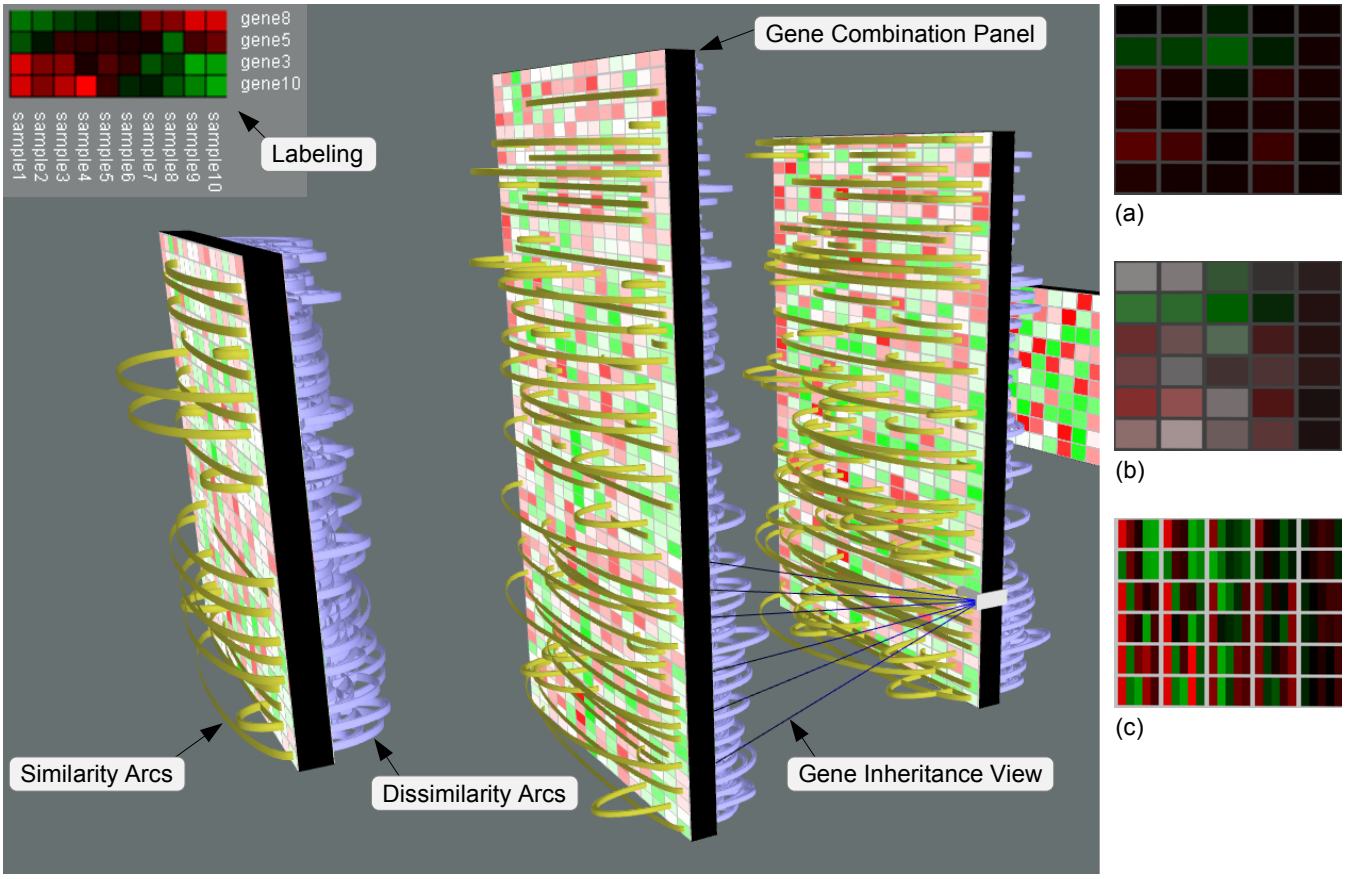

Figure 1: ViGeCo – Visualization of Gene Combinations.

First, we generate multiple panels – one for each possible size of gene combinations. That is, gene combinations consisting of two genes are represented by one panel, combinations of three genes by another panel, and so forth. These panels are arranged in a 3D presentation space as shown in Fig. 1.

Secondly, we need a way to represent aggregated data values. These values indicate the regulation of a gene combination as a whole. The problem is that the classic coloring scheme of heatmaps (see Fig. 1 (a)) can not be applied as is. This becomes clear when we consider a gene combination that consists of one up-regulated and one down-regulated gene. The aggregated expression value would be represented by a color close to black. The analyst is likely to interpret this as a gene combination with no regulation, which could be a wrong conclusion depending on the task at hand. Therefore, we can optionally use the brightness channel of colors to encode the average difference in the expression value of genes. This reduces the chance of misinterpretation, because darker colors are only created for gene combinations that are really not or less regulated (see Fig. 1 (b)). Alternatively, users can switch to a "small multiples" representations, for which each cell of a gene combination panel is subdivided to accommodate a color representation of the original (non-aggregated) expression values (see Fig. 1 (c)).

Using the gene combination panels it is possible to get an overview and to spot gene combinations that are active for all samples (saturated red or green color). The panels are also useful to find samples for which certain gene combinations exhibit similar behavior; this is expressed by similar colors. In order to further facilitate the task of finding similarities, we provide additional arc displays on demand. Arc displays are helpful in making relations between visual elements more explicit (see [4] or [3]). As such, arcs help us to make dis/similarities of the regulation of gene combinations more clear to the analyst. We provide two kinds of arcs: yellow similarity arcs and blue dissimilarity arcs (see Fig. 1). The latter arcs were requested by collaborating biologists, because for them it is also interesting to see "negative" similarity.

The users of our prototype also requested a view to see which genes contribute to which combination. This is achieved by labeling and an additional visual cue called gene inheritance view. This view links gene combinations that share common genes. (see Fig. 1).

3 Summary

We presented a novel approach to visualize microarray data. The novelty of the concept lies in the emphasis of gene combinations. We presented several extensions over the classic heatmap approach. Our concept has been implemented as a component for the microarray analysis framework Mayday [1]. The prototype is highly interactive in terms of adjusting the visualization as well as the provided analysis methods.

References

[1] Dietzsch, Gehlenborg, and Nieselt. Mayday - a Microarray Data Analysis Workbench. *Bioinformatics*, 22(8), 2006.

[2] Drăghici. *Data Analysis Tools for DNA Microarrays*. Chapman & Hall/CRC, 2003.

[3] Neumann, Schlechtweg, and Carpendale. ArcTrees: Visualizing Relations in Hierarchical Data. In *Proc. EuroVis*, 2005.

[4] Wattenberg. Arc Diagrams: Visualizing Structure in Strings. In *Proc. InfoVis*, 2002.