Heat²map: enriching differential gene expression heatmaps

Anton Zoubarev
azoubare@gmail.com

Domain
Visualization of differential gene expression.

Task
Find differences, similarities, variability between groups of genes or groups of conditions in terms of direction of change in gene expression, its magnitude, its statistical significance, and baseline expression level.

Dataset
The data is a table with rows representing genes and columns representing experimental conditions (or observed factor values). Each table cell has the following attributes:

- magnitude of change in gene expression (baseline vs. condition) (log transformed)
- direction of change (up or down)
- p-value (a measure of statistical significance) (0..1)
- expression level at the baseline

Personal expertise
I have been involved in building and supporting tools (including heatmaps) for biologists and bioinformaticians for about a year.

Proposed solution
A big chunk of this project is actually getting to a good solution. My approach to find one is as follows:

- Firstly (to avoid combinatorial explosion), I will pick a few promising visual encoding channels for each attribute using infovis principles learned in the course.
- Secondly, I will generate families of possible encodings(glyphs) to represent a heatmap cell.
- Thirdly, I will analytically evaluate these choices using criteria such as: expressiveness, accuracy, separability/interference, and others.
- Finally, the best few candidates will be implemented and further evaluated by getting user's feedback.

The proposed solution will use gene expression heatmap structure as a base, but will attempt to encode all 4 attributes inside each cell. It will be a non-interactive visualization.

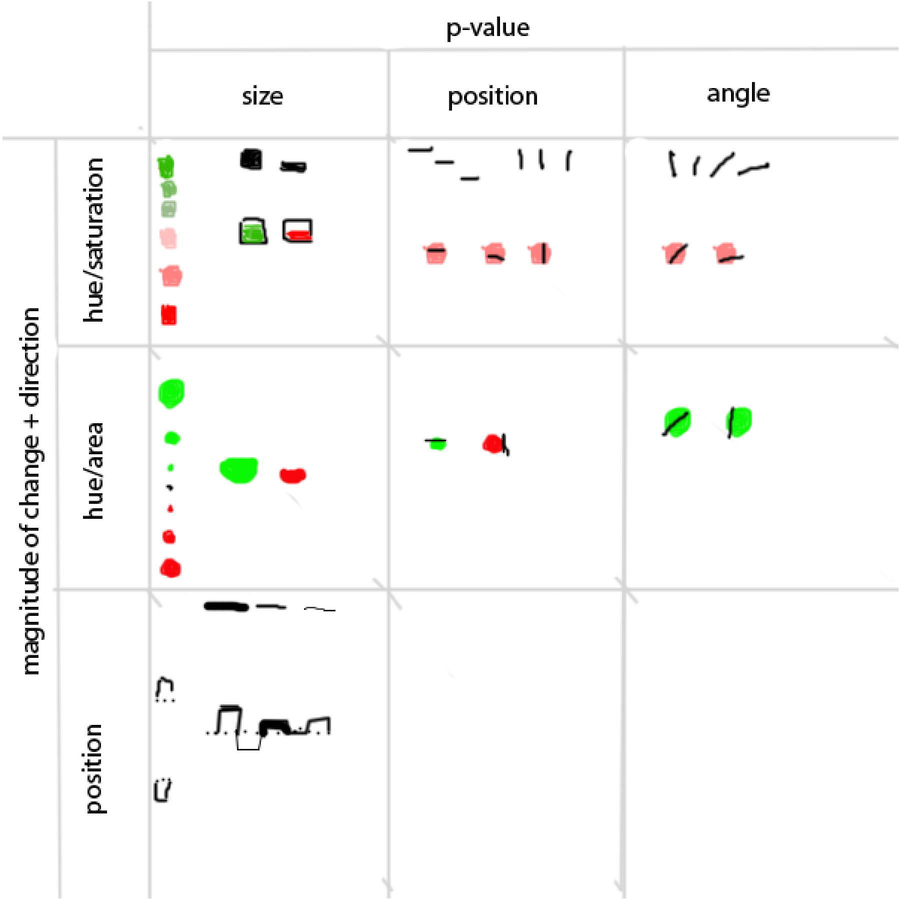
To reduce the search space of possible solutions the following constraints will be used:

- Some attributes are more important to the users. The order of importance (at this moment) is: direction, p-value, fold change, expression at baseline. This will inform the choice of visual channels.
- To allow for larger datasets (200x100), the cell has to be able to compress as much as possible without sudden loss of expressiveness. Say, it has to fit into 8 by 8 pixels box.

- The visualization will be based on a square grid (no distortions).

Note: This list will likely expand or change a little after the interviews with potential users.

Here's a sketch of a subset of possible visualization choices for a heatmap cell based on 3 attributes (direction and magnitude of change, p-value):



Scenario of use

The user will examine static enriched heatmap of clustered/sorted dataset and visually compare rows, columns, or blocks to accomplish the tasks outlined in the 'Task' section.

Implementation

Data : I will extract the data from the gene expression meta-analysis database called Gemma. I will then prepare and format the data using any convenient scripting language. If time permits I will implement hierarchical clustering algorithm otherwise I will use an existing implementation (possibly R) to rearrange the data at this step.

Visualization : The visualization will be implemented in JavaScript using HTML5 canvas.

Milestones

Nov 2	Interviews with users to better understand and constrain the problem.

Nov 8	Families of possible solutions are generated.
Nov 15	Solutions are analyzed/evaluated and best few are chosen for implementation.
Nov 18	Datasets are prepared and converted to the required format.
Nov 22	Best solutions are implemented.
Nov 24	Get users’ feedback and start writing final report.

Previous work

Enriched Heatmaps for Visualizing Uncertainty in Microarray Data by Clemens Holzhüter, Hans-Jörg Schulz, and Heidrun Schumann - poster at Eurographics VCBM 2010