Enriched Heatmaps for Visualizing Uncertainty in Microarray Data

Clemens Holzhüter [*]   Hans-Jörg Schulz [†]   Heidrun Schumann [‡]

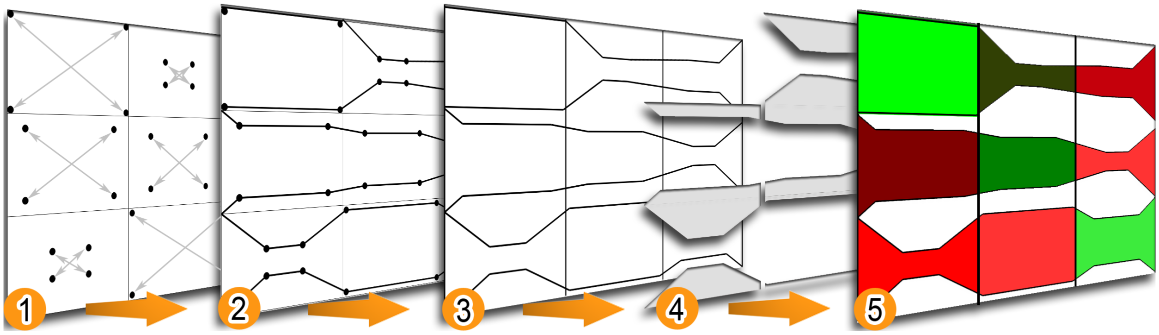

Figure 1: *The steps of generating an "uncertainty grid" and applying it to a standard heatmap. The more of a cell is masked out, the more uncertain the corresponding data value is.*

1 Motivation

Since their invention in the 1990's, DNA microarrays have advanced into a well-established, high-throughput experimentation technique for determining gene regulation. The data resulting from microarray experiments are image scans of physical microarrays. During a multi-step preparation process, a signal intensity for each gene of each sample is extracted, corrected, and aggregated into a numerical value for its regulation. In most cases, these regulation values are then analyzed visually using heatmaps.

As with all experimentation, the resulting data contain uncertainties, such as errors of measurement. A *Quality Control Score* (QCS) can be used to estimate the quality of each digitized intensity signal of a microarray experiment [Wang et al. 2001]. It is a good indicator for the reliability of the data as it combines a number of individual measures, such as the size and intensity of the spots on the microarray, local background variability, and signal-to-noise ratio. The QCS captures the variation of the intensities prior to data correction in a value range from 0 (lowest quality) to 1 (highest quality). These values are then used in later processing steps to exclude measurements that are doubtful by applying a threshold filter. Significance analysis and normalization can further be used to enhance reliability and comparability of the data, respectively.

However, besides potentially influencing the data correction process, the QCS is not visually communicated along with the preprocessed data and sometimes not even kept and stored – despite it being an important aspect of the data and its provenance. It is known that measurement errors can carry valuable information for a subsequent microarray analysis [Weng et al. 2006]. Hence, it is the aim of this work in progress to make the implicit uncertainty of the data explicitly available. In this poster, we explore an approach to enrich standard heatmap visualizations with the information about the uncertainty expressed by the QCS.

2 Our Approach

Heatmaps for microarray visualization represent the measured and processed signal intensities in a matrix. Each cell of the matrix represents one gene of a given sample and the corresponding data value is mapped onto a red/green color scale, with red indicating up-regulation and green indicating down-regulation. To enhance heatmaps with uncertainty information we need to map the QCS values to visual attributes that can be incorporated into the heatmap without hampering its perception. Mapping uncertainty to free visual attributes such as brightness and saturation bears the risk of misinterpretation of a cell's color, while other attributes such as orientation or the cell size potentially interfere with neighboring cells. Our approach proposes a deformation of the heatmap's grid lines that equates a mapping of the QCS values onto the size of an additional geometric shape within each cell. This does not interfere with the perception of a cell's color or its frame within the matrix.

This approach is inspired by an uncertainty visualization technique for remotely sensed imagery in the geosciences, where errors are introduced by atmospheric corrections, terrain distortions, and variation in the satellite orbit [Bastin et al. 2002]. In this case, the resulting uncertainty is visualized by overlaying multiple grids which are locally distorted by a random perturbation to produce equally probable instances within the bounds of the assumed spatial error. Thus, the underlying grid visualization stays the same. Yet, it is enriched by the multiple overlaid grid lines, with regions where all overlaid grids coincide well being more certain than regions in which they deviate noticeably from one another.

Due to the resemblance of the grid visualization with the matrix-based layout of the heatmaps, we adopt the idea of using the grid lines to adapt microarray heatmaps without losing their original look&feel. While the visualization of the heatmap is done as usual, the grid that separates the matrix cells is altered according to the uncertainty values for each cell. This 5-step-process is illustrated in Figure 1. The deformation is achieved by multiplying the QCS value of a cell with the four vectors that orientate from a cell's center to the different corner points of this cell. This results in four anchor points for the deformed grid: the more certain a cell's value is, the closer these points are to the actual corners of the cell – and the more uncertain a cell's value, the closer they are shifted towards the center of the cell (Step 1). After calculating all anchor points for every cell, the resulting points are connected either horizontally or vertically through piecewise linear interpolation (Step 2), whereas the respective other direction is drawn as straight line as usual. The interpolated line patches and the straight grid lines form together the *uncertainty grid* (Step 3). The resulting whitespace between the deformed gridlines (Step 4) is finally overlaid on the

[*] cholz@informatik.uni-rostock.de
[†] hjschulz@informatik.uni-rostock.de
[‡] schumann@informatik.uni-rostock.de

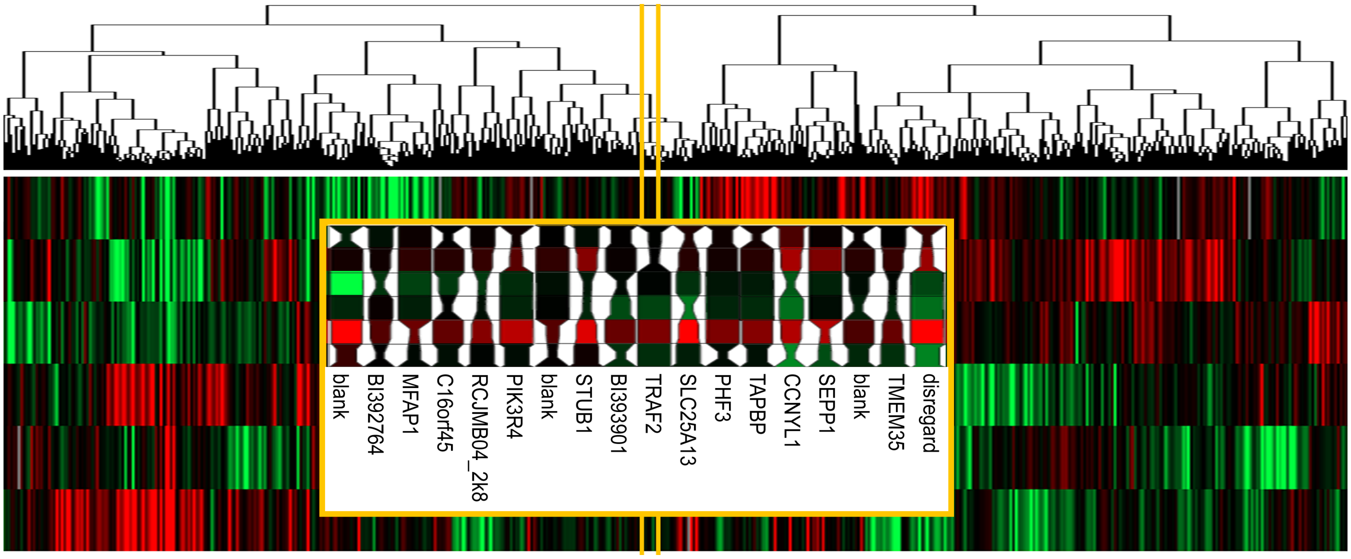

Figure 2: *An overview+detail concept is used together with a lens-like interaction metaphor, to embed our approach in large heatmap visualizations and to provide information about the uncertainty as details-on-demand.*

original heatmap, whereby regions in between grid lines are simply masked-out and filled with a background color (Step 5). This effectively enriches the heatmap visualization with the uncertainty information for each matrix cell.

3 Discussion and Future Work

To apply our approach to a real-world example, we are using a publicly available data set that can be obtained from the gene expression omnibus database [Ellestad et al. 2006]. This data set provides normalized gene regulations along with the required QCS values for each gene and sample. It features 6 samples with 7,200 genes each. While a complete, overview visualization is still feasible for a data set of this size, the individual matrix cells of the heatmap become too small, to enrich the visualization with our method. Hence, we provide our grid adaptation in a “details-on-demand” fashion: the user can activate the enriched visualization for a selected, enlarged subset using a lens-like interaction metaphor. The resulting view is shown in Figure 2.

First feedback from our partners within the research training group has been positive but needs to be substantiated in a user study. Therefore, our application provides user selectable grid adjustments. One important adjustment concerns the direction of the uncertainty grid. We found that adjusting the grid in horizontal and vertical direction at the same time makes its interpretation much harder. Restricting the adjustment to only one direction aids either in the visualization of how the uncertainty has changed over the course of the experiments, or of how the uncertainty changes from gene to gene. Further adjustments are grouped in variants of shape (trapezoidal and rectangular), line interpolation (cubic spline and linear), fill color (white, gray, gradient fill), and size/spacing of the uncertainty grid. A selection of these variations is illustrated in Figure 3. In future work we will evaluate the variations for their suitability to communicate uncertainties and their progression, while at the same time still allowing a good perception of the original cell color. To evaluate the effectiveness of our enhanced heatmap visualization it is planned to investigate alternative mapping techniques, such as using the saturation of a cell’s color or add textures, and to integrate them into our application.

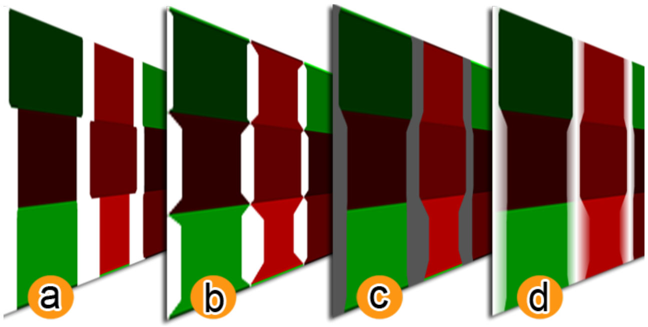

Figure 3: *Variation of the uncertainty grid. Rectangular (a) and trapezoidal shapes (b), gray (c) and gradient fill color (d).*

Acknowledgements

The authors would like to thank Marc Streit for his helpful comments. This work is supported by the interdisciplinary DFG graduate school dIEM oSiRiS.

References

Bastin, L., Fisher, P., and Wood, J. 2002. Visualizing uncertainty in multi-spectral remotely sensed imagery. *Computers & Geosciences 28*, 3, 337–350.

Ellestad, L., Carre, W., Muchow, M., Jenkins, S., Wang, X., Cogburn, L., and Porter, T. 2006. Gene expression profiling during cellular differentiation in the embryonic pituitary gland using cDNA microarrays. *Physiological Genomics 25*, 3, 414–425.

Wang, X., Ghosh, S., and Guo, S. 2001. Quantitative quality control in microarray image processing and data acquisition. *Nucleic Acids Research 29*, 15, e75.

Weng, L., Dai, H., Zhan, Y., He, Y., Stepaniants, S., and Bassett, D. 2006. Rosetta error model for gene expression analysis. *Bioinformatics 22*, 9, 1111–1121.