Visualizing Global Correlation in Large-Scale Molecular Biological Data

A.N.M. Imroz Choudhury* Kristin Potter* Theresa-Marie Rhyne† Yarden Livnat* Chris R. Johnson* Orly Alter*

Scientific Computing and Imaging Institute, University of Utah

Detection of novel patterns of correlation in large-scale molecular biological data can hint at the existence of as yet unknown cellular regulatory mechanisms. For example, correlations were observed among the DNA binding of cell cycle transcription factors [5] and the mRNA expression levels of cell cycle-regulated genes [6]. These correlations correspond to a known causal coordination among these processes. Recent experimental results [4] verify a computationally predicted mechanism of regulation [3] correlating genome-wide binding of replication initiation proteins [7] with mRNA expression during the cell cycle. This has demonstrated for the first time that mathematical modeling of DNA microarray data can be used, beyond classification of genes and cellular samples, to correctly predict previously unknown global modes of regulation [2].

In this work, we propose a visualization approach that facilitates exploration and identification of patterns of correlation in biological data. Our method provides a global view of temporal relationships between biological variables and local views of underlying data. This approach empowers researchers to discover global patterns and possible regulatory mechanisms while supporting visual verification of data quality and maintaining confidence in the visualization.

Overview. Our goal is to visualize patterns of protein binding over time based on gene expression. This poses the challenge of representing three-dimensional data along with various measures of correlations in a concise and intuitive display. To address this challenge we employ a two-level display: the global level depicts relationships between proteins and time while the local level depicts relationships between gene expression and protein binding levels.

Local View. The role of the local view is to provide a direct view of the data augmented with a concise representation of statistical measures. Each local view focuses on a single protein's activity at a specific time and consists of a scatter plot where each point represents the level of protein binding adjacent to and gene expression level of a given gene. To show possible (anti-)correlation between the levels of protein binding and gene expression, we depict the data's principal components using an ellipse. An elongated ellipse indicates a high correlation while its orientation indicates positive or negative correlation. The ellipse provides a concise view of the correlation and the associated uncertainty for the data, allowing researchers to determine their own level of confidence in the data.

Global View. To depict the temporal aspect of protein activity we tile several local views, one per protein per time point, into a table. Each row represents a particular protein and each column a time point. The table forms a gallery of images (Figure 2). While the ellipses provide important local information, they are not well-suited to detecting global patterns because it is hard to visually integrate and comprehend a four dimensional space (two ellipse axes spanning protein binding level × mRNA expression × proteins × time). We reduce this complexity by computing a single statistical correlation value in each local view and displaying it as the view's background color. The background color and ellipse are orthogonal visual representations that do not interfere with each other. In this way we can depict correlations for both local and global views within the same visual space.

Our work applies concepts from color theory to develop color maps [1]. Figure 1 shows two traditional colormaps used by biologists, with fully saturated colors showing non-uniform luminance. The red-green colormap is challenging for colorblind viewers to examine. By contrast, we have designed several colormaps (Figure 2) to address these concerns. We use muted color tones with complementary hues that retain strong contrast between extremes and are optimized to contrast with our scatter plots. Figure 2, bottom left, includes a color analysis using Adobe's Kuler tool (http://kuler.adobe.com).

The combination of the image gallery and the background colors (Figure 2), highlights temporal patterns of correlation, which may indicate biologically significant patterns. The top nine rows display the temporal activity of the cell cycle transcription factors [5] and its correlation with the global gene expression during the cell cycle that these transcription factors are known to regulate [6]. Each protein exhibits regions of strong positive correlation, which move forward in time from protein to protein. In contrast, the bottom four proteins, i.e., the replication initiation proteins, computationally predicted to exhibit a particular anti-correlation, show strong negative correlation during the predicted phase. We note that these genes were selected on the basis of data quality alone and were not limited to those that are classified as cell cycle-regulated. This suggests that the underlying cellular coordination spans the whole yeast genome.

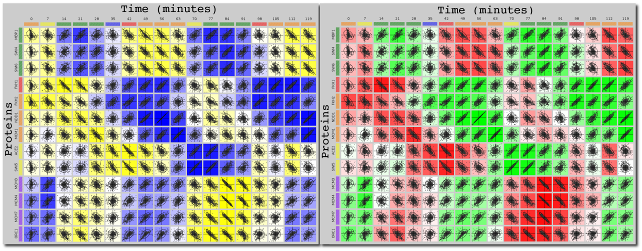

Figure 1: Our visualization scheme with traditional colormaps used in biological data display.

Acknowledgements

This research was supported by grants from the National Human Genome Research Institute R01 Grant HG-004302, NSF CAREER Award DMS-0847173 (to OA) as well as NSF, DOE, Award No. KUS-C1-016-04, made by King Abdullah University of Science and Technology (KAUST), and the NIH/NCRR Center for Integrative Biomedical Computing, 2P41 RR0112553-12 (to CRJ).

References

[1] J. Albers. *The Interaction of Color*. Yale University Press, 1975.

[2] O. Alter. Discovery of principles of nature from mathematical modeling of DNA microarray data. *PNAS*, 103:16063, 2006.

[3] O. Alter and G. H. Golub. Integrative analysis of genome-scale data by using pseudoinverse projection predicts novel correlation between DNA replication and RNA transcription. *PNAS*, 101:16577, 2004.

[4] L. Omberg *et al.* Global effects of DNA replication and DNA replication origin activity on eukaryotic gene expression. *Mol. Syst. Biol.*, 5(312), 2009.

[5] I. Simon *et al.* Serial regulation of transcriptional regulators in the yeast cell cycle. *Cell*, 106:697, 2001.

[6] P. T. Spellman *et al.* Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol. Biol. Cell*, 9:3273, 1998.

[7] J. J. Wyrick *et al.* Genome-wide distribution of ORC and MCM proteins in *Saccharomyces cerevisiae*: high-resolution mapping of replication origins. *Science*, 294:2357, 2001.

*e-mail:{roni,kpotter,yarden,crj,orly}@sci.utah.edu

†e-mail:theresamarierhyne@gmail.com

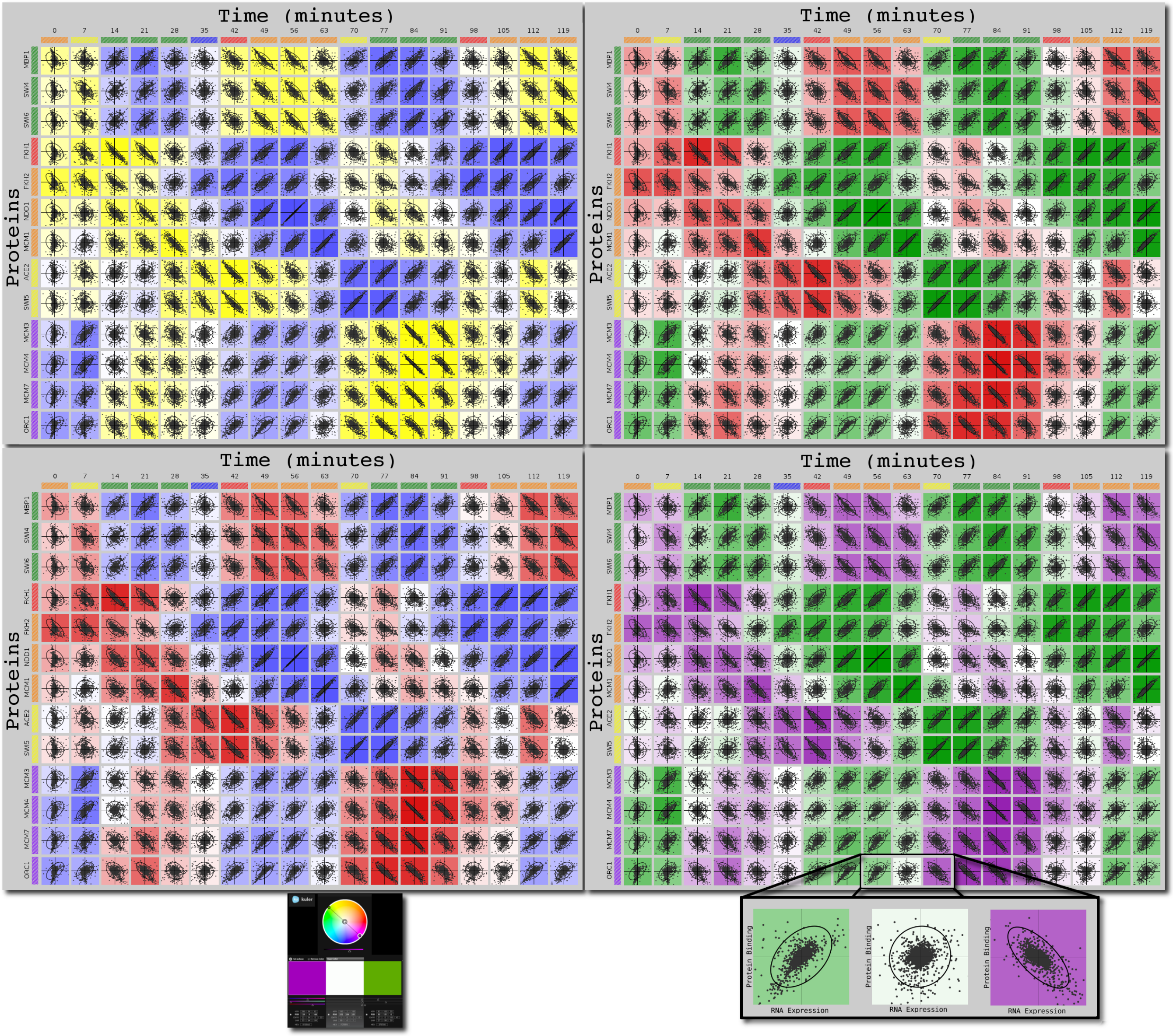

Figure 2: Tiled scatter plot display using redesigned colormaps. At a global scale, coordinated patterns of correlation and anti-correlation are visible. The top nine rows show nine proteins, cell cycle transcription factors, according to their order of activation during cell division in yeast. The visualization confirms the order, showing each protein in a positive correlation with gene expression in a definite temporal order. The bottom four proteins, replication initiation proteins, were predicted to be anti-correlated with a particular cell cycle phase, and indeed, these four proteins show a strong negative correlation during that phase (once in each of two full cell cycle periods). The inset demonstrates the local view by magnifying a selection of plots showing negative, zero, and positive correlations through scatter plot point distributions, data ellipses, and colors. A color analysis using Adobe's Kuler tool (http://kuler.adobe.com/) demonstrating our use of a complementary color scheme is shown in the bottom left.