Legal Disclaimer

Visually Relating Gene Expression and *in vivo* DNA Binding Data

Min-Yu Huang*, Lester Mackey†, Soile V. E. Keränen‡, Gunther H. Weber¶, Michael I. Jordan†,
David W. Knowles§, Mark D. Biggin‡ and Bernd Hamann*
*University of California, Davis, CA 95616, USA
Email: {myhuang,bhamann}@ucdavis.edu
†University of California, Berkeley, CA 94720, USA
Email: {lmackey,jordan}@cs.berkeley.edu
‡Lawrence Berkeley National Laboratory, CA 94720, USA
Email: {svekeranen,ghweber,dwknowles,mdbiggin}@lbl.gov

***Abstract*—Gene expression and *in vivo* DNA binding data provide important information for understanding gene regulatory networks: *in vivo* DNA binding data indicate genomic regions where transcription factors are bound, and expression data show the output resulting from this binding. Thus, there must be functional relationships between these two types of data. While visualization and data analysis tools exist for each data type alone, there is a lack of tools that can easily explore the relationship between them. We propose an approach that uses the average expression driven by multiple of *cis*-control regions to visually relate gene expression and *in vivo* DNA binding data. We demonstrate the utility of this tool with examples from the network controlling early *Drosophila* development. The results obtained support the idea that the level of occupancy of a transcription factor on DNA strongly determines the degree to which the factor regulates a target gene, and in some cases also controls whether the regulation is positive or negative.**

***Keywords*-Interactive Data Exploration, Gene Expression, *in vivo* DNA Binding Data, Visualization.**

I. Introduction

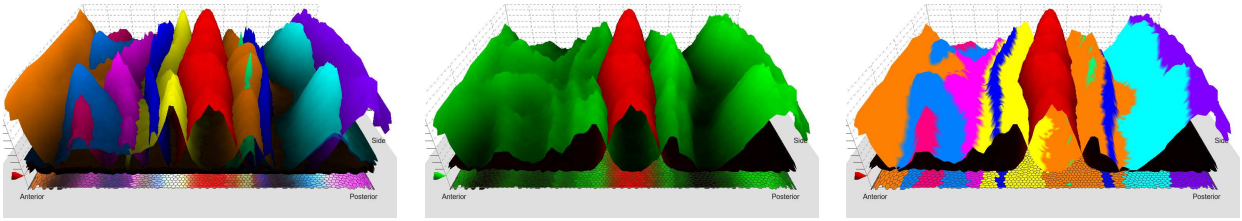
Although most cells in the animal carry identical genetic information in DNA, cells in different tissues and at different stages of development can have very different functions, as the expression of genes is selectively activated or deactivated in different cells at different times by transcription factors. Understanding the complex regulatory networks that control animal development and gene expression requires analysis of the spatial and temporal expressions patterns of transcription factors and their target genes.

One approach is to obtain data on the transcription output pattern driven by each CCR in an animal or its developing embryo using immunohistochemistry or *in situ*-hybridization [1], [2]. Another approach is to infer regulatory relationships between transcription factors and DNA on a genome-wide scale *in vivo* by chromatin immunoprecipitation followed by either microarray analysis (ChIP-chip) or sequencing (ChIP-seq). Genomic regions, including *cis*-control regions (CCRs), that are bound by a specific transcription factor can be identified by these techniques, as can the degree of factor occupancy on each region. CCRs are typically bound by several transcription factors. Since gene expression data are the output of gene transcription networks, there must exist relationships between the expression patterns of transcription factors, *in vivo* DNA binding data, and target CCR expression data. We demonstrate a visualization tool that helps the user visually relate gene expression and *in vivo* DNA binding data to explore these relationships.

II. Previous Work

Previous research efforts developed methods to record spatial and temporal gene expression patterns in several animals [1], [3]–[6]. However, to create a detailed model of transcription networks, cellular resolution quantitative data on gene expression in a whole embryo is needed. To address this deficiency, researchers in the Berkeley Drosophila Transcription Network Project (BDTNP) have developed methods to measure gene expression over an entire embryo blastoderm at cellular resolution based on fluorescence microscopy. After collecting expression data for different genes within different time cohorts from hundreds of embryos, a model VirtualEmbryo was constructed using registration techniques [2] to support quantitative computational analysis. PointCloudXplore, a visualization tool, was developed to interactively explore and analyze these high-resolution expression data [7]. MulteeSum [8] is a second visualization tool devoted to VirtualEmbryo data, developed for comparing VirtualEmbryos from different *Drosophila* species. Visualization tools to explore three-dimensional (3D) expression data sets also exist for other animal systems. For example, the Allen Brain Atlas viewer [9] maps color-encoded gene expression onto a 3D representation of a mouse brain.

A genome browser is often used for co-visualizing *in vivo* DNA binding and other data, such as DNA sequence and annotations from different gene models, in track views [10], [11]. There are also several tools for integrated analysis of *in vivo* DNA binding data that can perform basic analysis tasks, such as peak detection, false discovery rate computation, motif analysis and so on [12]. However, most of these analysis tools are designed for computationally intensive



(a) (b) (c)

Figure 1. The expression surface of *Krüppel* (KR) and that of the average of a cohort of strongly bound CCRs (rank 1–10) shown in Unrolled View. In this view, a cylindrical projection is used to map the entire embryo blastoderm to a plane and the expression is shown as a height field, where the height represents the expression value. (a) shows KR in red and individual CCRs' patterns in other colors. (b) shows KR and the averaged expression surface of this cohort of CCRs in green. The average expression surface makes it easier to observe KR's repressing role. (c) shows the average expression surface of the cohort with strongest contributing CCR's color in each cell. One can see that each color region maps to its corresponding CCR peak shown in (a). The user can switch among these views freely to explore the data.

tasks rather than user-interactive analyses, or they analyze data for only one transcription factor at a time. They do not allow quantitative comparison of results for many factors at once directly within the tool, which is a serious limitation as recent studies [13] show that many transcription factors bind quantitatively to highly overlapping sets of thousands of genomic regions *in vivo*. Regions occupied at high levels by transcription factors are quite different in character from those that are more poorly bound, with only the more highly bound regions being functional CCRs. To address this limitation, we previously developed a visualization framework that combines a genome browser, a correlation table, scatter plots, and parallel coordinates via brushing-and-linking, to support quantitative analysis and exploration of data for many transcription factors at once [14]. Building on our previous tool to analyze *in vivo* DNA binding data, we have now established a tool that integrates features of this tool with PointCloudXplore to make use of the high-resolution VirtualEmbryo gene expression data sets. We demonstrate this novel integration and show its unique capability to relate gene expression and *in vivo* DNA binding data.

III. System Design

We describe the data sets, the approach, and the visualization components in our tool in this section.

A. Data Sets

We consider three types of the data sets obtained from early *Drosophila melanogaster* embryos: mRNA expression data of 15 transcription factors, mRNA expression data for 95 CCRs, and *in vivo* DNA binding data of 21 transcription factors at the 95 CCRs. Expression data are in the form of a VirtualEmbryo [2], which provides measured expression on a per-cell basis. In the late blastoderm stage (stage-5), the embryo consists of $\sim$6000 nuclei, and the VirtualEmbryo specifies an individual expression value for each blastoderm cell. The VirtualEmbryo data used here comprise of 6 time cohorts in stage-5. Expression values are normalized between zero and one for each transcription factor and for each CCR respectively. The *in vivo* DNA binding data are also normalized between zero and one for each transcription factor. All data below 1% false discovery rate (FDR) are set to zero.

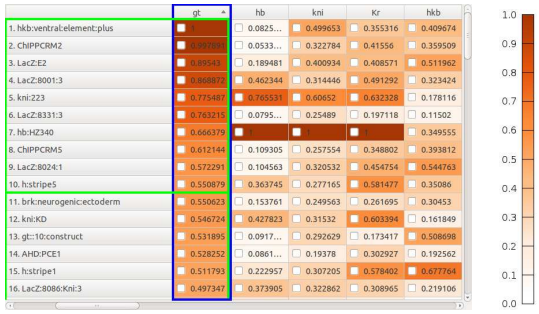
B. Approach

Determining the function of a transcription factor by comparing its expression pattern with that of an individual CCR is challenging. Consider Figure 1(a), where we observe the expression surfaces of *Krüppel* (KR, shown in red) and a number of its target CCRs (shown in, for example, magenta, blue, or cyan). When looking at the average pattern of all of the CCRs shown in Figure 1(a), it is apparent that KR likely represses these CCRs (Figure 1(b)). Hence, using an average pattern derived from multiple CCRs makes it easier to understand a transcription factor's role.

To test if the level at which a transcription factor occupies a CCR is important for how transcription factor affects expression output, we sort CCRs based on transcription factor ChIP-chip scores, with lower ChIP scores being ranked lower. We compute the average CCR expression pattern for every group of n CCRs down the rank list, where n is specified by the user and might typically be around ten. During the averaging process, we also create a CCR map by recording which CCR has the maximum expression value in each cell and this information can be displayed on the average expression surface using color (Figure 1(c)).

C. Visualization Components

The *in vivo* DNA binding table is the central graphical user interface (GUI) and the starting point of our tool. The user uses this GUI to load all expression and *in vivo* DNA binding data. Figure 2 shows an example. The column labels show the names of transcription factors and the row labels indicate the names of each CCR. Each table cell shows the normalized ChIP score of its corresponding transcription factor at its corresponding CCR. Clicking on a column label initiates a sorting process for the CCR names based on that transcription factor's ChIP scores. Double-clicking



(a) (b)

Figure 2. An example of the *in vivo* DNA binding table. (a) shows the ordinary interface. Table rows (CCRs) are sorted by the ChIP-chip scores of the transcription factor *Giant* (GT, surrounded by the blue box). We compute averaged CCR patterns for every group of n CCRs ($n = 10$ in this example) down the CCR list (surrounded by green boxes). (b) the color map for the table. The background color of each table cell is mapped by its normalized ChIP-chip score.

on the table label causes the display of the corresponding expression pattern in Unrolled View.

We color-map each table cell's background based on its score. This color-mapping GUI helps the user discover the binding strengths of different transcription factors at different CCRs. For example, in Figure 2(a), *Giant* (GT, the first column surrounded by the blue box) shows a very strong binding at ChIPPCRM2 (the second row), while *Hunchback* (HB, the second column) only has a weak binding to it.

The MultiView window consists of a grid of images that share the same view point to allow the user to compare multiple expression patterns easily. Each sub-window can show an average CCR expression pattern for the transcription factor in the results. The user can also choose to display the transcription factor and/or the individual CCR expression pattern in the same sub-window if they are available.

IV. Case Study

Early *Drosophila* embryo development is coordinated by two groups of transcription factors that control patterning along the anterior-posterior (A-P) and dorsal-ventral (D-V) body axes, respectively. We have used our tool to explore the activity of several A-P and D-V transcription factors.

The A-P patterning transcription factor *Krüppel* (KR) is expressed as a stripe around the middle of the embryo. Figure 3 shows the KR expression surface along with the average patterns of three example cohorts of CCRs to which KR binds to at strong, medium, and weak levels. It can be seen that the CCRs that are strongly bound by KR (Figure 3(a)) show pronounced A-P patterns, while weakly bound CCRs (Figure 3(c)) show D-V patterns. Comparison between the average patterns of strongly bound CCRs and weakly bound CCRs suggest that KR represses the strongly bound CCRs in the middle of the embryo, as there exists a pronounced anti-correlation between KR expression and that of the CCR cohort. The average expression patterns of moderately bound CCR cohorts (Figure 3(b)) have a more complex relationship with KR, which could indicate that this transcription factor may activate some moderately bound CCRs. The expression of weakly bound CCR cohorts (Figure 3(c)) shows no obvious correlation with KR expression and thus the low levels of KR binding have likely no effect in controlling these CCRs.

Figure 4 shows the expression surface of another A-P patterning factor, *Giant* (GT), along with the averaged expression patterns of CCR cohorts to which GT shows strong, medium, and weak binding. Like KR, GT may activate some moderately bound CCRs (Figure 4(b)) and have no effect at weakly bound CCRs (Figure 4(c)), whereas high levels of GT binding likely repress transcription (Figure 4(a)).

We have also examined the relationship between DNA occupancy levels and CCR output expression patterns for the A-P regulators BCD and CAD and those of the D-V regulators SNA and TWI, obtaining broadly similar results to those seen for KR and GT (unpublished data).

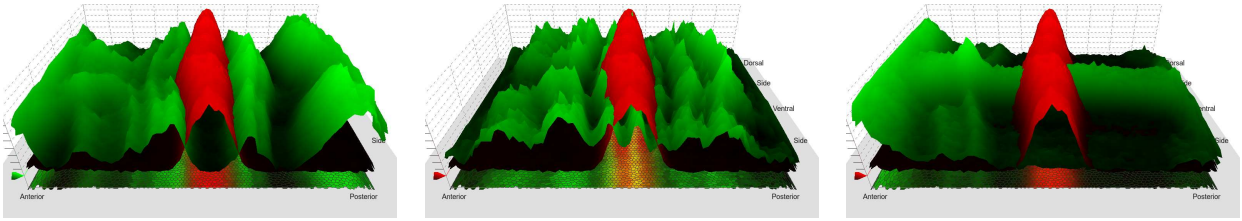
The above examples provide new evidence that the level of factor occupancy on a CCR, as measured by ChIP assay, is an important determinant in how or whether the transcription factor regulates the CCR. CCRs that are more highly bound tend to be significantly regulated by the factor. CCRs that are occupied at lower levels tend to either not be regulated by the factor, regulated to a smaller degree, or, in the case of SNA, GT and KR, perhaps regulated in a different direction (*i.e.*, activated instead of repressed). This result supports earlier biochemical and genetic evidence that transcription factors show a quantitative continuum of binding and function *in vivo* [13] and illustrates the importance of quantitative analyses of both *in vivo* DNA binding and gene expression.

V. Conclusions

We have introduced an effective approach to visually relate gene expression and *in vivo* DNA binding data: for each transcription factor, CCRs are grouped by the level of *in vivo* DNA occupancy of the transcription factor, and for each cohort, the expression patterns are then averaged to allow visual comparison with the expression pattern of the transcription factor. Our tool allows the user to visually explore the relationships among transcription factors and CCRs easily based on *in vivo* DNA binding data. We have provided several examples that illustrate the strength of this method to visualize integrated data.

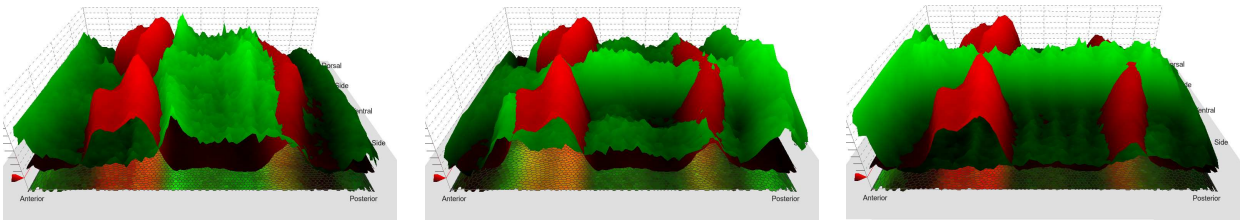
Acknowledgments

This work was supported by the National Institutes of Health through grant GM704403. Work conducted at Lawrence Berkeley National Laboratory (LBNL) is performed under Department of Energy contract DE-AC02-05CH11231.



(a) (b) (c)

Figure 3. The KR expression surface (red) and the averaged expression patterns of cohorts of CCRs ranked by the level of KR binding (green) shown in an unrolled view (see Figure 1). (a): strongly bound CCRs (rank 1–10); (b): moderately bound CCRs (rank 31–40); and (c): weakly bound CCRs (rank 81–90). These images suggest that KR represses strongly bound CCRs, may activate some moderately bound CCRs, and has no strong effect at weakly bound CCRs.



(a) (b) (c)

Figure 4. The GT expression surface (red) and the averaged expression patterns of cohorts of CCRs ranked by the level of GT binding (green) shown in an unrolled view (see Figure 1). (a): strongly bound CCRs (rank 1–10); (b): moderately bound CCRs (rank 41–50); and (c): weakly bound CCRs (rank 81–90). The images suggest that GT, like KR, represses strongly bound CCRs, may enhance some moderately bound CCRs, and has no strong effect at weakly bound CCRs.

REFERENCES

[1] P. Tomancak *et al.*, "Global analysis of patterns of gene expression during *Drosophila* embryogenesis," *Genome Biology*, vol. 8, no. 7, p. R145, 2007.

[2] C. C. Fowlkes *et al.*, "A quantitative spatiotemporal atlas of gene expression in the *Drosophila* blastoderm," *Cell*, vol. 133, no. 2, pp. 364–374, April 2008.

[3] E. Myasnikova, A. Samsonova, K. Kozlov, M. Samsonova, and J. Reinitz, "Registration of the expression patterns of *Drosophila* segmentation genes by two independent methods," *Bioinformatics*, vol. 17, no. 1, pp. 3–12, 2001.

[4] O. Tassy, F. Daian, C. Hudson, V. Bertrand, and P. Lemaire, "A quantitative approach to the study of cell shapes and interactions during early chordate embryogenesis," *Current Biology*, vol. 16, no. 4, pp. 345–358, 2006.

[5] A. Visel, C. Thaller, and G. Eichele, "Genepaint.org: an atlas of gene expression patterns in the mouse embryo," *Nucleic Acids Research*, vol. 32, no. suppl 1, pp. D552–D556, 2004.

[6] E. S. Lein *et al.*, "Genome-wide atlas of gene expression in the adult mouse brain," *Nature*, vol. 445, pp. 168–176, 2007.

[7] G. H. Weber *et al.*, "Visual exploration of three-dimensional gene expression using physical views and linked abstract views," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 6, no. 2, pp. 296–309, April 2009.

[8] M. Meyer, T. Munzner, A. DePace, and H. Pfister, "MulteeSum: A tool for comparative spatial and temporal gene expression data," *IEEE Transactions on Visualization and Computer Graphics*, vol. 16, pp. 908–917, 2010.

[9] C. Lau *et al.*, "Exploration and visualization of gene expression with neuroanatomy in the adult mouse brain," *BMC Bioinformatics*, vol. 9, no. 1, p. 153, 2008.

[10] W. J. Kent *et al.*, "The human genome browser at UCSC," *Genome Research*, vol. 12, no. 6, pp. 996–1006, June 2002.

[11] J. W. Nicol, G. A. Helt, J. Steven G. Blanchard, A. Raja, and A. E. Loraine, "The integrated genome browser: free software for distribution and exploration of genome-scale datasets," *Bioinformatics*, vol. 25, no. 20, pp. 2730–2731, October 2009.

[12] H. Ji *et al.*, "An integrated software system for analyzing ChIP-chip and ChIP-seq data," *Nature Biotechnology*, vol. 26, no. 11, pp. 1293–1300, November 2008.

[13] S. MacArthur *et al.*, "Developmental roles of 21 *Drosophila* transcription factors are determined by quantitative differences in binding to an overlapping set of thousands of genomic regions," *Genome Biology*, vol. 10, no. 7, July 2009.

[14] M.-Y. Huang, G. H. Weber, X.-Y. Li, M. D. Biggin, and B. Hamann, "Quantitative visualization of ChIP-chip data by using linked views," in *Proceedings IEEE BIBM 2010 Workshops, Workshop on Integrative Data Analysis in Systems Biology*, December 2010, pp. 195–200.