Visually Relating Gene Expression and \textit{in vivo} DNA Binding Data

Min-Yu Huang\(^*\), Lester Mackey\(^\dagger\), Soile V. E. Keränen\(^\ddagger\), Gunther H. Weber\(^\S\), Michael I. Jordan\(^\S\), David W. Knowles\(^\S\), Mark D. Biggin\(^\dagger\) and Bernd Hamann\(^*\)

\(^*\)Institute for Data Analysis and Visualization, Department of Computer Science, University of California, Davis, CA 95616, USA
Email: \{myhuang, bmhamann\}@ucdavis.edu

\(^\dagger\)Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA 94720, USA
Email: \{lmackey, jordan\}@cs.berkeley.edu

\(^\ddagger\)Genomics Division, \(^\S\)Life Sciences Division, and \(^\S\)Visualization Group, Lawrence Berkeley National Laboratory, CA 94720, USA
Email: \{svekeranen, ghweber, dvknowles, mdbiggin\}@lbl.gov

\textbf{Abstract}—Gene expression and \textit{in vivo} DNA binding data provide important information for understanding gene regulatory networks: \textit{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 a straightforward approach that makes use of the average expression driven by multiple of \textit{cis}-control regions within a binding strength cohort to visually relate gene expression and \textit{in vivo} DNA binding data. We demonstrate the utility of this tool using examples that explore the network controlling early \textit{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.

\textbf{Keywords}—Interactive Data Exploration, Gene Expression, \textit{in vivo} DNA Binding Data, Visualization.

\section{Introduction}

One important approach to decipher the complex gene regulatory networks that control animal development is to analyze the spatial and temporal expressions patterns of transcription factors and their target genes. 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. The main reason for this diversity is that the expression of genes is selectively activated or deactivated in different cells at different times by transcription factors. In general, there should be some correlation between the spatial and temporal expression patterns of transcription factors and the target genes that they regulate that would help us understand the regulatory interactions within a transcription network.

Another important approach to infer regulatory relationships is to investigate interactions between transcription factors and DNA \textit{in vivo} on a genome-wide scale by chromatin immunoprecipitation followed by either microarray analysis (ChIP-chip) or sequencing (ChIP-seq). Genomic regions, including \textit{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. By examining \textit{in vivo} DNA binding data for multiple transcription factors, we can identify at what levels the transcription factors bind to each CCR. We can also obtain data on the transcriptional output pattern driven by each CCR in an animal or its developing embryo. Since gene expression data are the output of gene transcription networks, there must exist relationships between the expression patterns of transcription factors, \textit{in vivo} DNA binding data, and target CCR expression data. We demonstrate a visualization tool that helps the user visually relate gene expression and \textit{in vivo} DNA binding data to explore these relationships.

Section II summarizes previous related work. Section III describes the data sets, the approach, and the visualization components of our tool. Section IV demonstrates the use of our tool with example data sets. Finally, we present ideas for future research directions and conclude our paper in Section V.

\section{Previous Work}

Previous research efforts developed methods to record spatial and temporal gene expression patterns in several animals [1]–[6]. However, these approaches do not provide quantitative data on gene expression in a whole embryo with the cellular resolution needed for creating a detailed model of transcription networks. To address this deficiency, researchers in the Berkeley Drosophila Transcription Network Project (BDTNP) have developed methods [7], [8] 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 [9] to support quantitative computational analysis. PointCloudXplore, a visualization tool, has also been developed to interactively explore these high-resolution expression data [10] and to analyze relationships between the expression patterns of important transcription factors [11]. MulteeSum [12] is a second visualization tool devoted to VirtualEmbryo data, but in this case it focuses on comparing information between VirtualEmbryos from different Drosophila species. Visualization tools to explore three-dimensional (3D) expressions data sets in other animal systems have also been developed, such as the Allen Brain Atlas viewer [13], which maps color-encoded gene expression onto a 3D representation of a mouse brain.

One standard approach to visualize in vivo DNA binding data is to use a genome browser [14]–[16] to examine binding information and other data, such as DNA sequence and annotations from different gene models, in track views. There are also several tools for integrated analysis of in vivo DNA binding data. For example, CisGenome [17] can perform basic analysis tasks, such as peak detection, false discovery rate computation, motif analysis and so on. However, most of these analysis tools are designed to perform computationally intensive tasks rather than user-interactive analysis, 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. This is a serious limitation as recent studies [18], [19] 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 [20].

Building on the previous success of our 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 stage-5 Drosophila melanogaster embryos: mRNA expression data of fifteen transcription factors, mRNA expression data of the ninety-five CCRs, and in vivo DNA binding data of twenty-one transcription factors at ninety-five CCRs. Expression data are in the form of a VirtualEmbryo [9], which provides measured expression on a per-cell basis. In the blastoderm stage (stage-5), the embryo consists of about six thousand nuclei, and the Virtual Embryo specifies an individual expression value for each cell. Data for the Virtual Embryo comprise six 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

Figure 1. The expression surface of Kräppel (KR) and that of the average of a cohort of strongly bound CCRs (rank 1–10) shown in Unrolled View [10]. (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 map of individual CCRs on the average expression surface of the cohort. This map indicates which is the major contributing CCR in each cell using that CCR’s colors. 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.
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) and Figure 3, 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). It is not easy to see whether KR is a repressor for these CCRs, even though molecular genetic data suggest that it is. However, when looking at the average pattern of all of the CCRs shown in Figure 1(a), it becomes much more apparent that KR represses these CCRs (Figure 1(b)). Hence, the use of an average pattern derived from multiple CCRs makes it possible to better understand a transcription factor’s role.

In order to understand whether the level at which a transcription factor occupies a CCR is important in how transcription factor affects transcriptional 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 to 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)). The user can then compare the resulting averaged CCR patterns along with the transcription factor’s pattern in our visualization tool and deduce which CCRs contribute most to each part of the averaged pattern.

C. Visualization Components

This section describes the major visualization components we have developed for our integrated framework and exploration tool.

The in vivo DNA Binding Table: 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 transcription factor name initiates a sorting process for the CCR names based on that transcription factor’s ChIP scores. Double-clicking on the table label causes the display of the corresponding expression in Unrolled View [10], as shown in Figure 3.

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 this CCR. Due to the size of the table (95 by 21), it is not easy to view the entire table at once. Hence, we also provide a bird’s eye view (in Figure 2(b)) of the table to help
the user see the distribution of binding strengths among transcription factors and CCRs. In this bird's eye view, each table cell is represented as a small color square. The user can obtain detailed information about a cell in the tooltip by moving the mouse cursor over that cell. The user can also change the order of columns freely by dragging transcription factor names to their new positions. This helps the user cluster transcription factors of interest together in the table to trace and compare the change of binding strengths among transcription factors and CCRs. The user can also manually select transcription factors and CCRs of interest by checking the checkbox in each table cell to compute the average expression of the transcription factor at the same time.

The MultiView Window: 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. Figure 3 shows an example of the MultiView window. In this example, we compare KR and individual patterns of its more strongly bound CCRs. Later figures (Figure 4–Figure 5) shown were exported from the MultiView window.

IV. Case Study

Early Drosophila embryo development is coordinated by several groups of transcription factors that generate different types of expression patterns. Along the anterior-posterior (A-P) body axis, pattern formation is initiated by the A-P early factors, and subsequently refined by the A-P pair-rule factors. Along the dorsal-ventral (D-V) axis, a separate group of D-V factors regulate patterning. We demonstrate the visualization results for several A-P and D-V transcription factors in the following sub-sections.

In all images, a cylindrical projection is used to map the entire embryo blastoderm to a plane (an Unrolled View [10]), and the expression is shown as a height field, where the height represents the expression value. Additionally, the locations with higher expression values are shown as brighter areas while locations with lower expression values are shown dimmer on the surface. All transcription factors expression patterns are shown in red while the average expression patterns are shown in green.

A. Krüppel and Giant

Krüppel (KR) and Giant (GT) are two gap transcription factors that regulate early A-P patterning. KR is expressed as a stripe around the middle of the embryo. Figure 4 shows the KR expression surface along with the average patterns of 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 4(a)) show pronounced A-P patterns, while weakly bound CCRs (Figure 4(c)) show D-V patterns. If we examine the average patterns from strongly bound CCRs to weakly bound CCRs (although we only put three representative images here), these patterns 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 4(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 4(c)) shows no correlation with that of KR and thus likely low levels of KR binding have no function in controlling these CCRs.

Figure 5 shows the GT expression surface along with the averaged expression patterns of CCR cohorts to which GT shows strong, medium, and weak binding. Like KR, Figure 5(b) and Figure 5(c) suggest that GT may enhance some moderately bound CCRs and have no function at weakly bound CCRs, whereas Figure 5(a) suggests that high levels of GT binding likely repress transcription.

B. Bicoid versus Caudal

Bicoid (BCD) and Caudal (CAD) are two important A-P early transcription factors that specify the location of anterior and posterior tissues, respectively. BCD expression values are highest in the anterior of the embryo and gradually decrease posteriorly along the A-P axis. Figure 6 shows the BCD expression surface and also the average patterns of three CCR cohorts to which BCD binds at strong, medium, and weak levels, respectively. We see that those CCRs to which BCD binds strongly tend to be expressed at higher
Figure 4. The KR expression surface (red) and the averaged expression patterns of cohorts of CCRs ranked by the level of KR binding (green). (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 function at weakly bound CCRs.

levels in the anterior half of the embryo than in the posterior half, and show little variation in the expression along the D-V axis (Figure 6(a)). In contrast, the CCRs bound most weakly by BCD can be expressed at high levels in both the anterior and posterior halves of the embryo, and show significant changes in expression along the D-V axis (Figure 6(c)). These results are consistent with molecular genetic evidence that BCD generally acts as an activator of transcription and that the more poorly bound CCRs are not significantly regulated by BCD, but by D-V regulators instead [19]. Interestingly, the average expression of the strongly bound CCRs has a dip on the dorsal side at the anterior tip of the embryo where BCD is highest, suggesting some anterior repression.

On the other hand, CAD shows the opposite expression to that of BCD, being expressed at high levels in the posterior of the embryo and at very low levels at the anterior end of the embryo. Figure 7 shows the average patterns of three CCR cohorts that CAD occupies at strong, medium, and weak levels, respectively. Interestingly, these results show that the relationship between CAD occupancy and CCR expression is broadly similar to that of BCD’s in that the average patterns of CCRs that are highly bound by CAD are more strongly expressed where CAD is itself most highly expressed and show little D-V patterning (Figure 7(a)); the CCRs bound at low levels by CAD can be strongly expressed in both the anterior and posterior half of the embryo and show strong D-V patterning (Figure 7(b) and Figure 7(c)). These results are consistent with CAD’s role as a regulator of posterior fate [19].

Figure 5. The GT expression surface (red) and the averaged expression patterns of cohorts of CCRs ranked by the level of GT binding (green). (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 function at weakly bound CCRs.
C. Snail versus Twist

The transcription factors \textit{Snail} (SNA) and \textit{Twist} (TWI) are expressed in similar patterns in the mid-ventral domain of the embryo. The SNA expression surface is shown along with those of the patterns of CCR cohorts bound at different levels by SNA (Figure 8). To allow changes in expression along the D-V axis to be more clearly seen, the view has been rotated 90° compared to previous figures such that the anterior of the embryo is at the bottom. The images suggest that high levels of SNA binding represses CCRs (Figure 8(b)), whereas lower levels of SNA occupancy activate transcription (Figure 8(d)). In contrast, high occupancy by TWI appears to activate CCRs (Figure 9(b)), whereas lower levels have no consistent effect on CCR expression (Figure 9(d)).

D. Discussion

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 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 \textit{in vivo} [18], [19] and illustrates the importance of quantitative...
Figure 8. The SNA expression surface (red) and the averaged patterns of CCR cohorts (green). (a): SNA and strongly bound CCRs (rank 1–10), (b): strongly bound CCRs (rank 1–10), (c): moderately bound CCRs (rank 36–45), and (d): weakly bound CCRs (rank 71–80). The figure suggests that SNA represses CCRs it binds to strongly (see (a) and (b)), but may activate CCRs that it occupies at lower levels (see (d)).

Figure 9. The TWI expression surface (red) and the averaged patterns of CCR cohorts (green). (a): TWI only, (b): strongly bound CCRs (rank 1–10), (c): moderately bound CCRs (rank 41–50), and (d): weakly bound CCRs (rank 81–90). Although TWI’s expression pattern is very similar to SNA’s, their corresponding average CCR patterns are very different. These images suggest that TWI is an enhancer at its strongly bound CCRs, while SNA is a repressor at its strongly bound CCRs.
analyses of both in vivo DNA binding and gene expression.

V. CONCLUSIONS AND POSSIBLE RESEARCH DIRECTIONS

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 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.

There still remain several challenges to be addressed. One is to extend this idea to gene expression patterns that cannot be represented in a plane, such as the more complex expression patterns seen in later stage, differentiated embryos. A second is to develop a way to compare data sets from different stages of development, and/or from different species.

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.

We also thank the members of the Institute for Data Analysis and Visualization (iDAV) at University of California, Davis, and the members of the Berkeley Drosophila Transcription Network Project (BDTNP) and Visualization Group at LBNL.

REFERENCES


