Integrating Data Clustering and Visualization for the Analysis of 3D Gene Expression Data

Oliver Rübel, Gunther H. Weber, Min-Yu Huang, E. Wes Bethel, Mark D. Biggin, Charless C. Fowlkes, Cris L. Luengo Hendriks, Soile V.E. Keränen, Michael B. Eisen, David W. Knowles, Jitendra Malik, Hans Hagen, and Bernd Hamann

Abstract—The recent development of methods for extracting precise measurements of spatial gene expression patterns from three-dimensional (3D) image data opens the way for new analyses of the complex gene regulatory networks controlling animal development. We present an integrated visualization and analysis framework that supports user-guided data clustering to aid exploration of these new complex data sets. The interplay of data visualization and clustering-based data classification leads to improved visualization and enables a more detailed analysis than previously possible. We discuss 1) the integration of data clustering and visualization into one framework, 2) the application of data clustering to 3D gene expression data, 3) the evaluation of the number of clusters k in the context of 3D gene expression clustering, and 4) the improvement of overall analysis quality via dedicated postprocessing of clustering results based on visualization. We discuss the use of this framework to objectively define spatial pattern boundaries and temporal profiles of genes and to analyze how mRNA patterns are controlled by their regulatory transcription factors.

Index Terms—Bioinformatics visualization, multimodal visualization, integrating Infovis/Scivis, visual data mining, three-dimensional gene expression, data clustering, cluster visualization, gene expression pattern, temporal expression variation, gene regulation, spatial expression pattern.

1 Introduction

Understanding the control of embryo development is a fundamental question in biology. A cell's unique fate is determined by specific combinations of developmental regulatory factors that form part of complex genetic regulatory networks ultimately coordinating the expression of all genes. As a result, the developing embryo exhibits an extraordinarily complex set of spatial and temporal gene expression patterns. The basic structure of the genetic regulatory network is defined by the genome sequence. However, we currently cannot adequately decipher this information or correctly predict how patterns of gene expression evolve.

The *Berkeley Drosophila Transcription Network Project* (BDNTP) is generating multiple complementary data sets to address these challenges using the early *Drosophila* developmental regulatory network as a model. These data sets include in vitro and in vivo DNA binding data for key transcriptional regulators and, of particular relevance to this work, three-dimensional (3D) gene expression data that describes the spatial output of the network at cellular resolution for multiple time points [1], [2].

A large variety of questions can be addressed using these new 3D gene expression data sets [2], [3]. For some analyses, such as logic-based network models, it is helpful to have an objective description of the pattern of a gene at a particular time point, i.e., to define which cells do or do not express a gene. Analysis of the temporal dynamics of gene expression, i.e., how patterns change over time, is essential for gaining a deeper understanding of complex network interrelationships. Knowledge of the input and output of a network, i.e., the response of the gene expression network at time $t = t_{i+1}$ to the input of the expression levels of regulators at time $t = t_i$, is paramount to identifying regulatory interactions.

- *O. Rübel is with the Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Mail Stop 50F-1650, Berkeley, CA 94720, and also with IRTG 1131, University of Kaiserslautern, 67653 Kaiserslautern, Germany, and the Institute for Data Analysis and Visualization, University of California, Davis, One Shields Avenue, Davis, CA 95616. E-mail: ORuebel@lbl.gov.*
- *G.H. Weber is with the Computational Research Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Mail Stop 50F-1650, Berkeley, CA 94720 and the Department of Computer Science, University of California, Davis, One Shields Avenue, Davis, CA 95616-8562. E-mail: GHWeber@lbl.gov.*
- *M.-Y. Huang is with the Institute for Data Analysis and Visualization and the Department of Computer Science, University of California, Davis, One Shields Avenue, Davis, CA 95616. E-mail: myhuang@ucdavis.edu.*
- *E.W. Bethel is with the Computational Research Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Mail Stop 50F-1650, Berkeley, CA 94720. E-mail: EWBethel@lbl.gov.*
- *M.D. Biggin and S.V.E. Keränen are with the Genomics Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Mail Stop 84R0171, Berkeley, CA 94720. E-mail: {MDBiggin, SVEKeranen}@lbl.gov.*
- *C.C. Fowlkes is with the Donald Bren School of Information and Computer Sciences, University of California, Irvine, CA 92697-3425. E-mail: fowlkes@ics.uci.edu.*
- *C.L. Luengo Hendriks is with the Centre for Image Analysis, Swedish University of Agricultural Sciences, Box 337, SE-751 05 Uppsala, Sweden. E-mail: cris@cb.uu.se.*
- *M.B. Eisen is with the Howard Hughes Medical Institute and the Department of Molecular and Cell Biology, University of California, Berkeley, Stanley Hall 304B, Berkeley, CA 94720-3220. E-mail: MBEisen@lbl.gov.*
- *D.W. Knowles is with the Life Sciences Division, Lawrence Berkeley National Laboratory, 1 Cyclotron Road, Mail Stop 50F-1650, Berkeley, CA 94720. E-mail: DWKnowles@lbl.gov.*
- *J. Malik is with the Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94720-1776. E-mail: malik@eecs.berkeley.edu.*
- *H. Hagen is with the University of Kaiserslautern, 67653 Kaiserslautern, Germany. E-mail: hagen@informatik.uni-kl.de.*
- *B. Hamann is with the Institute for Data Analysis and Visualization, Department of Computer Science, University of California, Davis, One Shields Avenue, Davis, CA 95616-8562. E-mail: hamann@cs.ucdavis.edu.*

Manuscript received 17 Aug. 2007; revised 23 Dec. 2007; accepted 7 May 2008; published online 21 May 2008.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBB-2007-08-0107.
Digital Object Identifier no. 10.1109/TCBB.2008.49.

1545-5963/10/$26.00 © 2010 IEEE Published by the IEEE CS, CI, and EMB Societies & the ACM

To address these and other challenges, we need a flexible visualization tool that allows for interactive exploration of the data. Since *Drosophila melanogaster* has been used as a model for genetic research for decades, there exists a large accumulated body of knowledge about it. A tool designed for the analysis of 3D gene expression data must therefore allow researchers to incorporate this existing knowledge in the analysis, for example, by providing ways to modify analysis results and, thus, the visualization accordingly. The tool must also capture the biological context of the embryo and allow different subsets of the data (cells or gene expression patterns) to be examined.

While visualization is a powerful approach to gain deeper insights into such complex data sets, it is limited in this case because the intricate and often subtle nature of 3D gene expression data makes visual detection of all existing features very difficult. For example, a typical feature of interest would be various groups of cells behaving similarly with respect to the expression of several genes. A human's eye and mind, however, cannot readily compute relative concentrations of gene products. Data clustering has already proven to be very powerful at revealing details from conceptually simpler forms of expression data, such as that from microarray experiments, that are not easily detected visually in raw data. Appropriately defining clustering parameters such as the number of clusters, as well as validation and interpretation of clustering results, is a nontrivial endeavor. To overcome these difficulties in both visual analysis and data clustering, we have adapted data clustering for 3D gene expression analysis by integrating it into PointCloudXplore (PCX). PCX is a visualization tool that features linked physical and information visualization views specifically developed for visualization of 3D gene expression data [4], [5].

Sections 2 and 3 present essential biological background necessary for understanding this work. After describing our integrated system in detail in Section 4 and evaluating the question of how to choose the number of clusters k in Section 5, we discuss, using a few example cases, how our integrated data clustering and visualization tool can be used in practice to address three relevant questions: 1) How can we usefully divide cells into distinct components of a gene's expression pattern (Section 6)? 2) What is the temporal variation of a gene expression pattern (Section 7)? 3) What components of a gene's expression pattern are related to the expression patterns of the regulatory factors that control it (Section 8)? In Section 9, we present our conclusions and describe future plans.

2 Background

All cells of living organisms contain *DNA*, which encodes the genetic information of the organism. *Genes* are functional subsequences of the DNA. Most genes code for the amino acid sequences of proteins and additional *cis*-regulatory elements that help to determine in which cells the gene's product will be expressed. An important class of protein coding genes are developmental regulatory *transcription factors* that function by binding to *cis-regulatory sequences* in many genes and direct their patterns of gene expression. Complex *genetic regulatory networks* are built up where cascades of differently expressed transcription factors ultimately regulate all genes' expression. These networks guide the development of all living organisms. The characteristic spatial and temporal patterns of regulatory transcription factors define the body plan of the developing animal (see Fig. 1).

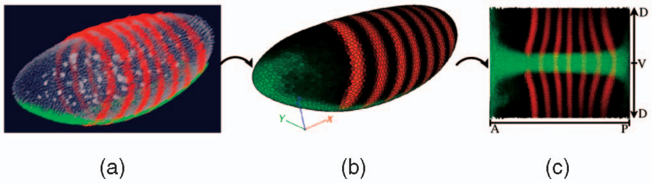


Fig. 1. (a) Three-dimensional images, each containing a whole embryo, are transformed into PointCloud files containing information about cell positions and the expression of the measured genes. Our visualization tool, PCX, (b) uses a 3D physical model to visualize the embryo. (c) To provide an overview of all cells in PCX, the embryo is projected onto a rectangular plane using cylindrical projection along with annotations indicating the anterior (A), posterior (P), dorsal (D), and ventral (V) orientation of the embryo. Here, the expression pattern of the gene *even skipped (eve)* is shown in red, and the pattern of *snail (sna)* is shown in green.

To provide a quantitative description of these patterns of gene expression in the early *Drosophila* embryo, the BDTNP has developed a data processing pipeline for extracting precise measurements of spatial gene expression patterns in 3D space. *Drosophila* embryos are first fluorescently stained and imaged using two-photon microscopy (see Fig. 1a). Each image is segmented to extract information such as nuclear positions and volumes, as well as expression values in the neighborhood of each nucleus for the chosen genes [2]. The resulting *PointCloud* file contains information about either the protein or mRNA expression of the genes. It is not practical to obtain the expression of more than a few genes in a single embryo, due to the limited number of different distinguishable fluorophores as well as the difficulty in adding multiple labels to embryos.

To allow relationships between multiple transcription factors and their target genes to be compared in a common coordinate framework, PointClouds are registered into a *Virtual Embryo* using both morphology and a common reference gene to determine cell correspondences [6], [7]. Because the spatial patterns of the genes change rapidly during stage 5, we stage the embryos based on invagination of cell membranes and group the PointClouds into six temporal cohorts [2].

For temporal comparisons, different cohorts are matched using the cellular flow fields that predict the positions of individual cells at each time point [3], [7]. This method enables us to follow gene expression levels within a particular cell over time using only data measured in fixed embryos. Hence, each cell in the Virtual Embryo contains gene expression levels for each of the six time steps. This cellular-level link between embryos of different ages makes it possible to study the development of gene expression patterns over time, as well as to use an mRNA expression pattern as an approximate substitute for a later protein expression pattern, when suitable protein data is not yet available [7].

Fig. 1b shows a 3D representation of a Virtual Embryo with an average expression pattern using the BDTNP's visualization tool PCX. To provide an overview of all cells while preserving the relative spatial expression patterns, PCX offers a second physical view in which a cylindrical projection maps all cells onto a rectangular plane (see Fig. 1c). For simplicity, here, we use this *Unrolled View* as our standard physical embryo view. A 3D view can equally be used to view embryos and developmental stages with more complex morphologies than in the early *Drosophila* embryo.

3 Related Work

Data classification is the systematic grouping of data into categories according to some criteria. *Data clustering* is a class of techniques for unsupervised classification of data samples (here, cells) into groups (clusters) of similar behavior. Data clustering provides means for the automatic discovery of data subclasses [8].

In some experimental contexts such as expression microarrays, gene expression data is often represented as a data matrix, where each gene corresponds to a row, and each data sample (cell, microarray, experiment, or condition) corresponds to a column. Each matrix entry describes the expression level of a gene in a specific experiment. In these applications, data clustering has proven very useful to classify expression data matrices and thereby identify characteristic substructures of each matrix.

Gene expression data clustering can roughly be subdivided into three applications: 1) clustering of genes to identify genes of similar function [9], 2) clustering of data samples to identify, for example, different tumor cell types [10], and 3) biclustering, i.e., clustering of genes and data samples at the same time to find subgroups of genes and data samples where highly similar activities are seen for the genes in the subset of data samples [11].

Clustering results are most commonly visualized using scatterplots, plots of statistics, and color table views with columns and/or rows sorted with respect to the clustering. The broad applicability of clustering to gene expression has led to the development of several commercial and publicly available tools for clustering and visualization of gene expression data [12], [13], [14], [15], [16], [17]. However, these tools are limited to what essentially are one-dimensional (1D) analyses of gene expression in homogenized populations of cells. They do not take account of spatial position or the complex relationship of expression across neighboring cells and are consequently not suitable for interactive visualization and exploration of 3D gene expression data produced by the BDTNP.

Validation of clustering results and evaluation of an "optimal" number of clusters k is an important problem in clustering of gene expression data. A survey of computational cluster validation techniques for gene expression data analysis is provided by Handl et al. [18]. Cluster evaluation functions are commonly subdivided in external and internal measures. *External evaluation measures* compare the result of a single clustering with a known set of class labels (the "gold standard" or "ground truth"). For our data a "gold standard" is not known, and consequently, we cannot consider external cluster evaluation functions. *Internal evaluation measures* do not rely on a "gold standard" but evaluate the clustering based on clustering results and the classified data set.

The most common cluster evaluation measures consider the compactness, connectedness, and/or separation of a clustering. Such general measures, however, do not employ any specific characteristic of gene expression data. The *Figure Of Merit* (*FOM*) is an internal measure for gene clustering proposed by Yeung et al. [19] and extended by Datta and Datta [20] that employs explicitly the redundancies and correlations often present in gene expression data. In our application, the level of redundant information is generally low. As a result, FOM and analysis techniques such as the overabundance analysis proposed by Ben-Dor et al. [21] are often not appropriate for our applications but may be interesting when the cells of the embryo are to be classified based on the information of a very large number of genes. To the best of our knowledge, none of these existing cluster quality measures directly employ the fact that genes are expressed in characteristic spatial patterns.

Internal cluster quality measures have been used to estimate the number of clusters k in a data set. Estimation of an "optimal" k is usually done by computing a series of clustering results for an increasing number of clusters k. If a clustering algorithm and an internal evaluation measure are adequate for the data to be classified, an "optimal" value of k can often be identified as a "knee" (or elbow) of the resulting performance curve. Tibshirani et al. [22] introduced the *gap statistic*, a statistical procedure that formalizes this heuristic. Milligan and Cooper performed a Monte Carlo evaluation of 30 procedures for determining the number of clusters in a data set [23]. Existing cluster evaluation measures are designed to find "one perfect" k. As we show later in Sections 4.4 and 5, when clustering cells in a 3D gene expression data set, we typically find a series of valid values for k, rather than the one "perfect" k.

To enable visualization of high-dimensional 3D gene expression data, PCX uses the established concept of *linked multiple views* [24]. Henze [25] proposed a system based on multiple views (termed portraits) for exploration of time-varying computational fluid dynamics data sets; advanced queries can be performed by selecting data subsets in these portraits. In the WEAVE system, a combination of Physical Views and Information Visualization Views (or abstract views as we refer to them in this paper) is used for exploration of cardiac simulation and measurement data [26]. Doleisch et al. [27] formalized the concept of using abstract views to define data queries.

It is often useful to interactively select data samples from a visual data representation, an operation generally referred to as *brushing*. A *brush* is an object that defines one specific selection of data samples. In PCX, brushing is used in a variety of views to select groups of cells with respect to associated quantities. To make this concept more intuitive to the biologist users, brushes are referred to in PCX as *cell selectors*, and the operation of brushing is referred to as *cell selection*. Furthermore, cell selectors defined in one view are also highlighted in all other views, greatly aiding identification of further data properties. This process is termed *linking*.

PCX was also inspired by the work of Kosara et al. [28], Piringer et al. [29], and Fua et al. [30], who described several

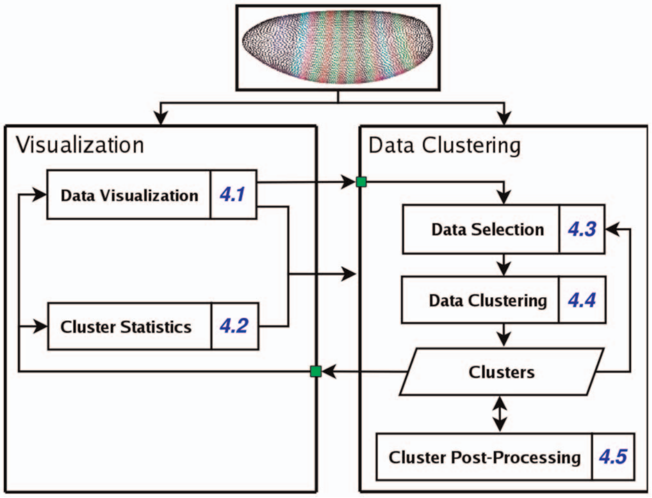


Fig. 2. The data clustering and visualization pipeline. Each box represents a stage of the pipeline and contains the section number where we describe that part of the pipeline in this paper.

important extensions to standard scatterplots and parallel coordinates that are incorporated as abstract views in PCX and also used here [4], [5].

4 Data Clustering and Visualization Pipeline

The PCX processing pipeline consists of two main interconnected components: visualization and data clustering. Visualization provides the ability to explore the data, to determine appropriate parameters for the clustering, to validate and analyze clustering results, and to modify clustering results using several dedicated cluster postprocessing techniques (see Fig. 2). Clustering provides ways for automatic identification of data features by classifying cells into groups (the clusters) based on the similarity of their gene expression profiles. By highlighting clusters in the visualization, analysis and comparison of specific data features becomes possible, leading to a much more focused analysis of the data. Fig. 2's flowchart shows the basic structure of the data clustering and visualization pipeline, as well as the connections between its main components, which are described in detail in the following sections.

4.1 Visualizing 3D Gene Expression Data

As described above, PCX is a visualization tool specifically developed for the analysis of 3D gene expression data [4], [5]. Physical and abstract views are integrated into a common framework using the established concept of brushing and linking. In physical views, color and height are used for visualizing spatial gene expression patterns (see Section 2). In abstract views, physical cell positions are ignored and expression levels for multiple genes are plotted with respect to each other using scatterplots or parallel coordinates.

Selecting cells of interest can be executed in any view in PCX. Depending on the view, different data properties are employed to select cells. User-defined cell selections are then stored and managed in a central cell selector management system. Since all views have access to the same set of cell selectors, features of interest can be defined in any one view and then further analyzed in any other view (as will be shown later in Figs. 6, 13, and 14). The most common way to visualize cell selectors in PCX is to use a consistent color mapping. Depending on the current view, additional functions for highlighting cell selectors are available, such as cell selector bands in two-dimensional (2D) parallel coordinates (see, e.g., Figs. 6 and 14).

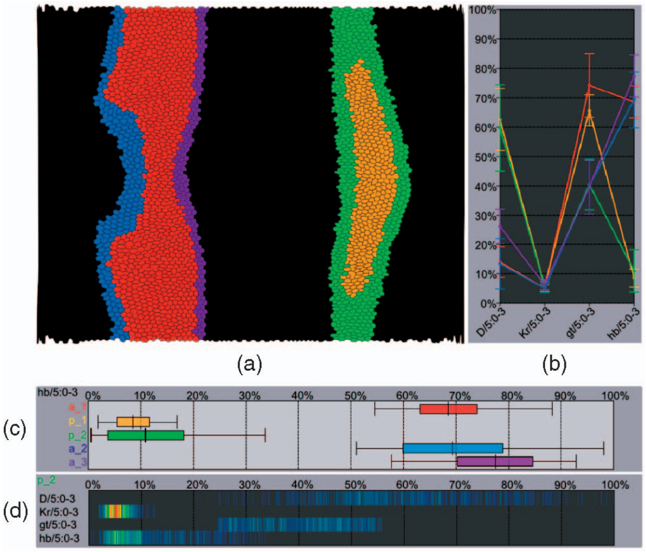


Fig. 3. An analysis of characteristics of the *giant (gt)* expression pattern using cluster statistics. (a) An unrolled view showing the spatial structure defined by five clusters. The red and orange cluster define the centers of the two expression regions of *gt*, and the other clusters define the boundaries. (b) A curve plot showing the average expression profiles of the genes *D*, *Kr*, *gt*, and *hb* in each of the five clusters (x-axis). The y-axis represents the expression level. (c) A box plot comparing the expression of *hb* in the five clusters. The x-axis represents the expression level. (d) A color/transparency histogram comparing the expression of *D*, *Kr*, *gt*, and *hb* for cluster p_2 (green). The x-axis indicates the gene expression level. We use a "heat map" coloring scheme to indicate the number of cells in the cluster having a given expression level: red indicates many cells, while blue indicates few cells.

4.2 Cluster Statistics

Analysis of statistical properties of clusters is essential for both the validation and analysis of clustering results. Cluster properties provided by PCX include the percentage of cells selected by a cluster, as well as the minimum, maximum, average, and standard deviation values for gene expression levels in a cluster. To compare these statistical properties for one gene in multiple clusters or multiple genes in one cluster, PCX provides box plots and multidimensional color/transparency histogram plots. In histogram plots, we use both color and transparency to visualize the number of cells within a cluster that express the gene over a range of expression levels. Average curve plots (with optional error bars showing standard deviation values) aid in simultaneous analysis of multiple clusters in multiple genes. A simple example shown in Fig. 3 illustrates the use of cluster statistics.

4.3 Data Selection

While it is possible to execute the clustering algorithms on an entire data set, a more typical use pattern is to focus clustering on a data subset relevant to a specific line of scientific inquiry. The researcher therefore needs to define which parts of the data are relevant to address the current problem. In this section, we describe the different steps involved in the data selection process, as well as the effects of data selection on the cluster analysis, and describe how spatial information can be incorporated in the data analysis process. In the following two sections, we will then describe clustering of 3D gene expression data and postprocessing of clustering results.

Three-dimensional gene expression data can be described as a matrix where each row represents one cell and each column represents one cell attribute, i.e., the expression of a gene at a specific time point or the x, y, and z positions of the cells in physical space. In order to define which parts of the expression data matrix are relevant, one needs to define 1) which rows (cells) and 2) which columns (gene+time point, x, y, and z) are of interest. Note that this form of data is quite different from that of gene expression microarray matrices, where each row represents a gene, and each column represents expression under a different experimental condition, and spatial relationships are meaningless.

Cells of interest can be defined in PCX by using any cell selector or by using the results of a previous data clustering. Defining cells of interest focuses the analysis on a specific part of the data and also reduces the impact of surrounding noise on the analysis. By explicitly allowing data selection based on cell location, PCX overcomes one of the limitations of clustering methods designed for expression microarray data. By using an earlier clustering to define cell subsets of interest, one can first use PCX to group cells into a smaller number of clusters representing the predominant data features and then refine these clusters again using additional rounds of data clustering. In PCX, data clustering, as well as validation of clustering results, can in this way be performed in a step-by-step iterative process.

Defining which cell attributes are of interest is mandatory prior to clustering in PCX since these attributes define the actual biological context of the cells. To account for the complexities of 3D expression data, a variety of unique cell attribute data selection strategies is supported within PCX. First, genes of interest are generally identified based on visualization of the 3D gene expression data, as well as based on input from other biological experiments such as in vivo protein-DNA binding affinity data. Second, to account for spatial location in the clustering analysis, it is possible to directly use cell coordinates as input to the PCX clustering process. Adding this data enforces creation of spatially separated clusters along the AP (x) and/or the DV (y and z) body axes. Individual weights can be defined for x, y, and z. These weights are then considered in the distance metric (see Section 4.4). However, in most cases, the preferred way to incorporate spatial information in the analysis process is by splitting the newly computed clusters into their main independent spatial components. The main advantage of such a cluster postprocessing technique over including cell coordinates in the clustering process is that cells with similar expression behavior in different parts of the embryo can be identified, and possible clustering artifacts due to the mixing of expression and spatial information can be prevented.

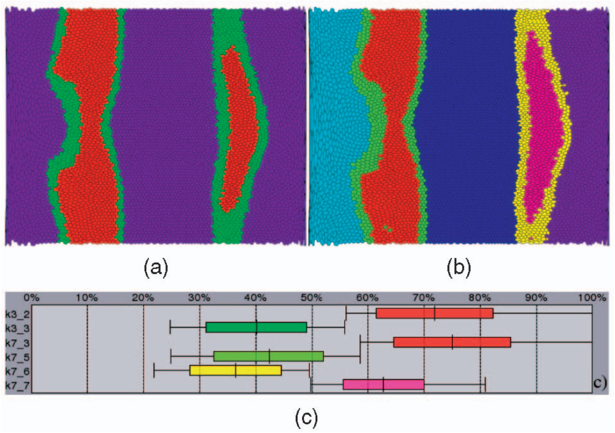


Fig. 4. (a) *Giant (gt)* expression pattern classified using k-means clustering with euclidean distances and $k = 3$. (b) Same, using $k = 7$ and including x cell positions weighted with 0.24 (after normalization). (c) A box plot showing the statistics in *gt* expression (x-axis) for the two main clusters of the result shown in (a) (first two entries on the y-axis) and for the four main clusters of the clustering shown in (b). Including spatial information in the clustering resulted in spatially separated clusters for the main regions of *gt*, as well as in different threshold levels, depending on the physical cluster locations.

We observed an improved quality of analysis results by adding spatial information to the clustering process when classifying the static pattern of a single gene that has a wide spatial distribution. In the example shown in Fig. 4, we classified the pattern of the gene *giant* (*gt*) using k-means clustering with and without using x (AP) cell positions in the clustering process. In the first case, three clusters were created, each selecting cells expressing *gt* at different levels, i.e., low, medium, and high expression (Fig. 4a). By considering x cell positions, we create separate clusters for the different major spatial components of the *gt* expression pattern (see Fig. 4b). In this case, each cluster includes only cells that express *gt* at specific levels, while the minimum and maximum expression levels selected by each cluster also depend on its physical location. In this case, higher threshold levels were created in the anterior, and lower thresholds were created in the posterior region of the embryo (see Fig. 4c). Creation of region-dependent threshold levels is often desirable when analyzing the static pattern of a single gene since each domain of a pattern may be regulated differently, and therefore, different thresholds may be appropriate. For gap genes with spatially distant independent expression domains, such as *gt*, this simple strategy works well, whereas for patterns with shorter interdomain distances, such as *eve*, this strategy fails.

4.4 Clustering 3D Gene Expression Patterns

To implement clustering operations in PCX, we use portions of the open source clustering library "Cluster 3.0" [31]. We have integrated data clustering directly into PCX and created a dedicated GUI that provides access to data clustering and allows management of clustering results. Clustering algorithms currently available in PCX include the most commonly used methods for microarray gene expression data analysis, such as k-means, k-median, and k-medoid clustering, as well as several hierarchical

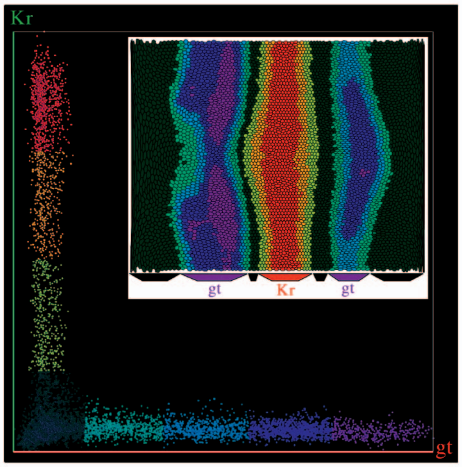


Fig. 5. An example clustering of *giant* (*gt*) and Krüppel (*Kr*) using k-means clustering and euclidean distances with $k = 8$. In the scatterplot, the structure of the clusters is shown in expression space, while the unrolled view reveals spatial structures formed by the different clusters.

clustering algorithms, and self-organizing maps (SOMs) [31], [9], [10], [32]. All these clustering algorithms require an appropriate distance function in order to define similarity between cells. In PCX, we included the most common metrics for defining distances in gene expression space: euclidean distance, city block distance, and several derivatives of the Pearson correlation [31].

Some clustering algorithms require additional parameters such as the number of clusters k to be specified by the user. In the context of 3D gene expression data, there exists in general not a single "perfect" value for k, but we rather find a number of valid values, each representing a different level of detail. This behavior is due to the fact that quantitatively different expression levels of a gene may lead to multiple different outputs of the underlying genetic regulatory network. It is therefore valid to subdivide elongated structures formed in gene expression space into several subclusters.

For example, consider early-stage *giant* (*gt*) and *Krüppel* (*Kr*), which are expressed in spatially nonoverlapping patterns, leading to the formation of an L-shaped scatterplot (see Fig. 5). Even though one could interpret this structure as one cluster—possibly indicating a *NOT* relationship between *gt* and *Kr*—it is also valid to subdivide this structure into, e.g., eight clusters, resulting in one cluster representing the background expression, a three-level description of the pattern of *Kr*, and a four-level description of the *gt* pattern.

The choice of k depends on the level of detail required by the user. Therefore, PCX uses an interactive process to define k based on visualization. The spatial structure formed by the cells selected by clusters, cluster statistics, and standard data visualizations provides a way to decide if the number of clusters should be increased or reduced. Depending on the characteristic spatial patterns of genes, the cells included in a cluster often define some coherent spatial pattern. Thus, the presence of clusters that show high spatial scattering may be an indication that the chosen k was too large. To assist in this evaluation process, we have developed a dedicated cluster quality measure indicating the physical scattering of clustering results along with a function for suggesting a good initial k. These measures will be described in more detail in Section 5. In Section 6, an example is provided where the pattern of *eve* is classified using different values of k.

Like a manually created cell selector, an automatically created cluster defines a subset of cells in the embryo and can therefore be stored and visualized in the same way as cell selectors. Thus, clustering can be used for highlighting data features in physical or abstract views, enabling a much more focused analysis. In the visualization, PCX allows colors to be assigned to clusters either randomly, manually, or according to the average or ranked average expression of a selected gene in each cluster. Using physical views, the spatial pattern defined by a cluster can be analyzed, and abstract views allow for identification of cluster characteristics in gene expression space.

4.5 Cluster Postprocessing

Cluster postprocessing is essential to allow users to modify clustering results with respect to validation results or prior knowledge. There are four ways to postprocess clusters in PCX. Manual correction and cluster filtering are two ways to correct small groups of misclassified cells. Cluster merging and splitting provide means to derive coarser or finer representations based on spatial information from the initial clustering.

Manual correction of clustering results can be performed in any physical view. By drawing on the embryo surface, one can interactively add and erase cells from the selection defined by a cluster. In contrast, filtering provides an automatic way to correct misclassified cells. Because genes are expressed in coherent spatial patterns, outliers in physical space tend to be also outliers in gene expression space. Therefore, we have developed a cluster filtering method that identifies and reassigns misclassified cells to the spatially neighboring cluster that is closest in expression space. First, all spatially independent components of a cluster that consist of less than M cells are identified. To rule out false filtering, a minimum distance in physical space, as well as a maximum error in expression space, can be defined. In the example shown in Fig. 6, it would be possible to exclude the cells shown in green from the filtering process either by increasing the minimum spatial distance or by reducing the maximum allowed error in expression space.

Merging clusters allows coarser representations to be created from an initial finer clustering. Such coarser descriptions often provide a clearer visualization that focuses on the main question being addressed (see, e.g., Section 6). Splitting clusters, on the other hand, provides means to derive a finer representation from clusters based on spatial information. A cluster often consists of several spatially independent components (for example, Fig. 12), which may need to be treated differently in the subsequent analysis. In general, however, one major component of a cluster may be defined by a number of small spatially independent components. PCX uses a modified single-linkage clustering approach to split up such a cluster into a selected (often smaller) number of components.

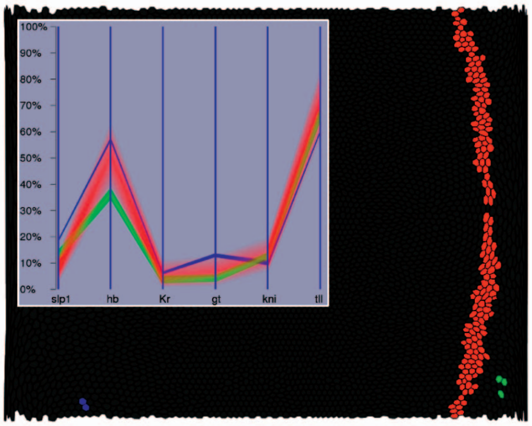


Fig. 6. Filtering applied to an example cluster. The cluster is split into its four independent spatial components (red, blue, and two shown in green). The profiles of these regions in gene expression space are shown in parallel coordinates. Here, the genes *slp1*, *hb*, *Kr*, *gt*, *kni*, and *tll*, which were used to obtain this clustering result, are each represented by one axis, and the percentage of expression is shown in ordinate direction (y-axis). One can see that the blue cells are spatially more distant to the main component of the cluster (red) than the green cells and that they show a higher divergence from the main spatial component of the cluster in gene expression space.

The splitting algorithm works as follows: A cluster is first split into all its spatially independent components. The smallest components are subsequently merged with the spatially closest component. This approach is computationally more efficient and less sensitive to outliers than a classical single-linkage clustering and also guarantees that the independent spatial components of a cluster are preserved while small scattered components can still grow to define major cluster components. An example for cluster splitting will be described later in Section 6.

5 Identifying Good Values for k, the Number of Clusters

Many clustering algorithms, such as k-means, require the user to specify as an input parameter the target number of clusters, k. The quality of clustering results often depends on a proper choice of k. Unless users have a priori knowledge concerning the number of clusters present in the data, it is helpful for the user that the software offers a reasonable initial value for k. Different approaches for finding an "optimal" k have been proposed. Among them, those based on internal cluster measures appear to be more appropriate for our application [18], [23]. Our objective here, as described below, is to provide the user assistance in interactively searching for a good k as opposed to trying to automatically compute the optimal value of k.

Even though internal cluster quality measures (see Section 3) may be useful here, we are not aware of any such measure that takes the specific characteristics of 3D gene expression data into account. Since genes are often expressed in compact spatial patterns, we expect the derived clusters to be spatially compact. The presence of computed clusters with high spatial scattering typically suggest that the value of k was too large. Because we do not use information about physical cell position in the clustering process, spatial compactness is a criterion available as an independent measure for clustering quality. As we will discuss below, spatial cluster scattering can also serve as a measure to indicate a series of adequate values of k. Combining spatial cluster scattering and the clustering error in expression space yields a method to identify a good initial value for k that accurately reflects the structures present in the data but with relatively low spatial scattering.

We propose to use

$$\varepsilon_{\mathrm{scatter}}(k) = \frac{\sum_{i=1}^{k} R_1(i)}{\sum_{i=1}^{k} R_\infty(i)} \tag{1}$$

as an objective measure for the relative spatial scattering of a clustering result. $R_s(i)$ (with $s > 0$) is the number of spatially independent components of cluster i consisting of at most s cells. $R_1(i)$ thus defines all single cell regions in cluster i. $R_\infty(i)$ is the total number of spatially independent regions in cluster i. $\varepsilon_{\mathrm{scatter}} \in [0, 1]$ is independent of the clustering algorithm, usually has discontinuities, and shows a larger variation for smaller values of k than for large values of k. Local minima of $\varepsilon_{\mathrm{scatter}}$ indicate values of k for which clusters are relatively compact and thus indicate a series of appropriate values of k. In the context of 3D gene expression data, clustering errors introduced by single cells isolated in physical space are quite common, and our choice for $\varepsilon_{\mathrm{scatter}}$ performs well. An alternative approach might work better when these cluster outliers consist of small groups of cells. One approach might be a less sensitive weighted cascade measure that also accounts for larger regions as potential scatter, such as

$$\varepsilon_{\mathrm{sc}}(k) = \frac{\sum_{s=1}^{p}(\frac{1}{s} * \sum_{i=1}^{k} R_s(i))}{\sum_{s=1}^{p}(\frac{1}{s} * \sum_{i=1}^{k} R_\infty(i))},$$

with $p > 0$ being much smaller than the number of cells.

To evaluate the clustering error in expression space, we use $\varepsilon_{\mathrm{exp}}$, the average distance, in expression space, of a cell to the center of the cluster it belongs to:

$$\varepsilon_{\mathrm{exp}}(k) = \frac{1}{n}\sum_{i=1}^{n} dist(center(c_i), c_i), \tag{2}$$

where n is the number of clustered cells, c_i is the ith cell, $center(c_i)$ is the center of the cluster to which c_i belongs, and $dist(\cdot,\cdot)$ is the distance operator used in the clustering process.

We compute $\varepsilon_{\mathrm{scatter}}$ and $\varepsilon_{\mathrm{exp}}$ for $2 \leq k \leq m$, with m being the first value where $\varepsilon_{\mathrm{scatter}} > 50$ percent. If the pattern of only one gene is used in the clustering, we use $\varepsilon_{\mathrm{scatter}} > 60$ percent as the termination criterion instead because variations in the background expression have a stronger impact on the cluster analysis and because more complex structures are possible when multiple genes with spatially overlapping patterns are clustered. By using these thresholds for $\varepsilon_{\mathrm{scatter}}$, we ensure that we iterate over all potentially useful clusterings and do not terminate prematurely. We use $k = 2$ as the starting point because it represents the first potentially useful clustering. Furthermore, considering the relatively large value of $\varepsilon_{\mathrm{exp}}(1)$, starting at $k = 1$ would result in a suggested value for k that is too small.

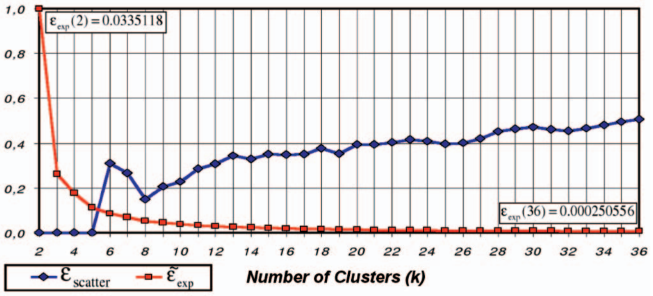


Fig. 7. Cluster evaluation functions $\tilde{\varepsilon}_{\text{exp}}$ (red) and $\varepsilon_{\text{scatter}}$ (blue) for the clustering of *gt* and *Kr*, with $w = 5$ and $m = 36$. The suggested k is eight, as shown in Fig. 5.

To identify a value w for k for which the error in expression space is sufficiently low to well characterize the data, we identify the first k for which the decrease in ε_{exp} is lower than the average decrease $\bar{\Delta}_{\varepsilon_{\text{exp}}} = \frac{\varepsilon_{\text{exp}}(2)-\varepsilon_{\text{exp}}(m)}{m-2}$. Alternatively w could also be defined as the k that corresponds to the point of the ε_{exp} evaluation curve that is furthest from the line defined by $\varepsilon_{\text{exp}}(2)$ and $\varepsilon_{\text{exp}}(m)$ [33]. While the first approach tries to find the k for which the expression error has sufficiently decreased, the second approach tries to identify the so-called "knee" of the ε_{exp} evaluation curve. Both methods depend on m, but this dependency is well behaved, i.e., with increasing m, the suggested w changes slowly and continuously. During the research and development of this work, both methods seem to work equally well.

We use the following algorithm to identify a good initial $k > w$ that also results in a relatively low physical scattering:

```
$k = w + 1$,
$l = k$,
for $i \leftarrow l$ to $m$
do
    if $(\varepsilon_{\text{scatter}}(i) < \varepsilon_{\text{scatter}}(l) + t)$
    then
        $k = i$
        if $(\varepsilon_{\text{scatter}}(i) < \varepsilon_{\text{scatter}}(l))$
            then $l = i$.
```

Initially, k is set to $w + 1$, which is the lowest value that results in a sufficiently low expression error. Then, the algorithm tries to optimize the expression error, as well as the physical scattering, by searching for a $k > w$ that also results in a relatively low physical scattering. Here, we use a threshold of $t = 4$ percent—determined through empirical testing—to restrict the maximal allowed increase in $\varepsilon_{\text{scatter}}$ with respect to l, i.e., the k with the lowest relative physical scattering visited so far. Since ε_{exp} decreases with increasing values of k, the error in expression space for the suggested k is guaranteed to be smaller than $\varepsilon_{\text{exp}}(w)$.

Alternatively, one can also view the problem of finding a good initial k as an optimization problem by looking for the k that minimizes $\varepsilon_{\text{total}}(k) = |\breve{\varepsilon}_{\text{exp}}(k) - \frac{1}{w}\breve{\varepsilon}_{\text{scatter}}(k)|$, where both ε_{exp} and $\varepsilon_{\text{scatter}}$ are normalized. Conceptually, the first approach is more intuitive, does not require normalization of the evaluation functions, and will always suggest a minimum of $\varepsilon_{\text{scatter}}$ if an adequate local minimum exists. Using $\varepsilon_{\text{total}}$ for finding a good initial k has the advantage that it does not rely on a threshold t. Furthermore, it may result in a more reliable suggestion in cases where $\varepsilon_{\text{scatter}}$ is degenerate since $\varepsilon_{\text{total}}$ does not directly rely on the notion that the physical scattering increases with increasing values of k. In practice, both approaches have shown to be useful.

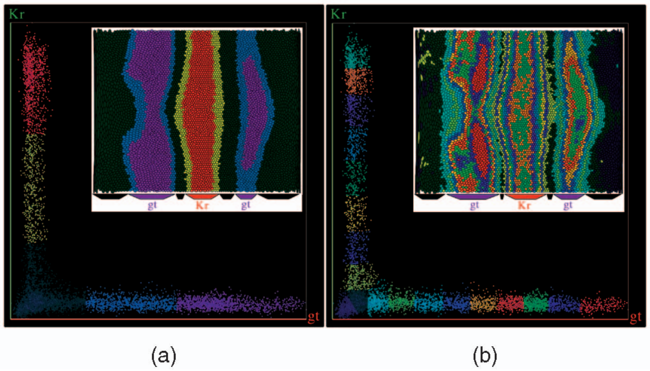


Fig. 8. The patterns of *gt* and *Kr* are classified using k-means clustering, as in Fig. 5, but with (a) $k = 5$ and (b) $k = 19$. One can see that the suggested $k = 8$ provides a compromise between a high-level description, as shown in (a), and a detailed description, as shown in (b).

Beginning with an initial suggested value of k, the user can then determine the "best" k based on the information from the cluster evaluation and previews of the different evaluated clustering results using an Unrolled View. Even though the initial suggested value of k may not always be optimal, our testing has revealed that there is value in providing a "reasonable" value or range of values for k.

Fig. 7 shows the cluster evaluation functions for the clustering of *gt* and *Kr*. To provide an overview of both functions in one plot, we show $\varepsilon_{\text{scatter}}$ along with $\tilde{\varepsilon}_{\text{exp}}(k) = \frac{\varepsilon_{\text{exp}}(k)}{\varepsilon_{\text{exp}}(2)}$. The suggested k is eight, which is also a strong local minimum of $\varepsilon_{\text{scatter}}$. The corresponding clustering result for $k = 8$ is shown in Fig. 5. Fig. 8 shows two additional example classifications of *gt* and *Kr* using $k = 5$ and $k = 19$. $k = 5$ is the highest level for which $\varepsilon_{\text{scatter}} = 0$, and $k = 19$ is a local minimum of $\varepsilon_{\text{scatter}}$ ($\varepsilon_{\text{scatter}}(19) \approx 35.29$ percent) close to the middle of the range. Here, we see that the suggested level of $k = 8$ provides a good compromise between high-level and low-level descriptions of the patterns. The value of k that is best suited to investigate a biological question depends to a large degree on user requirements. More example usages of $\varepsilon_{\text{scatter}}$ and ε_{exp} are provided in Sections 6 and 7.

ε_{exp} and $\varepsilon_{\text{scatter}}$ are global cluster quality measurement functions in the sense that the clustering quality is evaluated based upon the entire data set (in this case, all classified cells). Global error measures might not be appropriate if the user performs a clustering of a larger number of cells but is interested only in a small subset of clusters defining some local feature of interest.

6 Single-Pattern Analysis

Genes are frequently expressed in complex patterns that show a wide range of quantitative changes in expression across the cells of an embryo. Although for some analyses, the data are best left unclassified in this form—simply using the expression values in all cells—it can also be revealing to

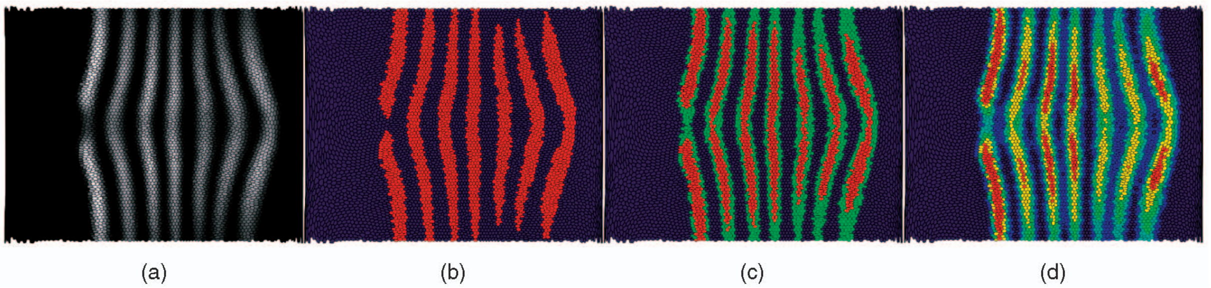


Fig. 9. (a) The expression pattern of *eve* at stage 5 : 9-25 percent. Classification of *eve* with (b) $k = 2$, (c) $k = 3$, and (d) $k = 6$. While the $k = 2$ clustering produced a threshold that was too high, erasing too many cells from the pattern, the $k = 3$ clustering was better able to identify the seven stripes of the *eve* expression pattern. The $k = 6$ clustering identified additional characteristic variations within the stripes along the DV axis, as well as an additional cluster that selects some interstripe cells showing some higher expression of *eve*.

divide a single pattern into one or more distinct regions. For example, on/off descriptions of expression have been useful in logical models of gene networks [34], [35].

However, discretizing a gene pattern via manual thresholding can be problematic—it may be very time consuming, and the choice of thresholds is arbitrary and not fully data dependent. To address this challenge, one can use, for example, k-means clustering and euclidean distances to compute a number of data-dependent thresholds. Each of the k clusters then represents a specific threshold range that can be interpreted as a different confidence level. Different components of a pattern may be regulated by different genes, so different thresholds may be appropriate for different regions. Cluster postprocessing such as splitting clusters into their main spatial components allows different threshold levels to be selected for different components of a gene pattern. Alternatively, as described in Section 4.3, for genes with clearly distinct spatial expression domains, cell positions may be used in the initial clustering to enforce creation of separate clusters for spatially distant components of a pattern. Rather than choosing some arbitrary thresholds, clustering automatically suggests thresholds based on the histogram of the data. The k-means clustering algorithm seeks to minimize the mean squared distance from each data point (cell) to its nearest cluster center. To achieve this goal, the k-means algorithm will create k cluster centers positioned according to the density distribution of the expression values of the selected gene.

Fig. 9 shows three example classifications of the *eve* expression pattern using different numbers of clusters k. While $k = 2$ produces a threshold that is too high and does not capture all parts of each stripe, a clustering with $k = 3$ correctly identifies the seven stripes of the *eve* expression pattern. By increasing the number of clusters, additional details within the stripes along the dorsal-ventral (DV) axis can be seen, as well as an additional cluster selecting cells in the interstripe regions. This complex description illustrates that thinking of a gene as being either on or off is usually too simplistic. The fact that clustering automatically reveals differences along the DV axis demonstrates the usefulness of such analyses. The pair-rule genes such as *eve* are not typically thought of as DV regulators but are consistent with the clustering results, careful quantitation of the levels of *eve* and a similar gene's expression has shown that they indeed show up to twofold changes in expression along the DV axis, suggesting a DV component in pair-rule regulation [3], [2]. Analyzing the actual meaning of these moderate changes requires computational tools such as cluster analysis to provide objective measures of their significance.

Fig. 10a shows the curves of the cluster evaluation functions $\varepsilon_{\text{scatter}}$ and $\tilde{\varepsilon}_{\text{exp}}$. In this case, $\varepsilon_{\text{scatter}}$ is rather smooth and monotonically increasing, indicating that all k with $\varepsilon_{\text{scatter}} > 0$ may result in valid clusterings of the *eve* pattern. This behavior can be explained by the very high signal-to-noise ratio of the *eve* expression data, which was averaged from dozens of embryos. The suggested k is five, which is the largest k for which only one cluster representing a low *eve* expression is created (see Fig. 10b). A clustering with $k = 5$ provides a compromise between a high-level and low-level description of the *eve* expression pattern.

Binarized versions of the *eve* pattern (i.e., on/off descriptions) can be created by merging the different clusters, allowing one to easily compare the different classifications by defining their overlay (see Fig. 11). While $k = 3$ and $k = 6$ result in similar classifications of the seven stripes, the clustering with $k = 2$ misses many cells of the pattern. Thus, first generating multiple clusters and then merging them can provide a more accurate binarization of an expression pattern than an initial $k = 2$.

Cluster merging and splitting can also be useful for comparing different gene patterns or for comparing different components of a single gene's pattern. In Fig. 13, for example, the individual clusters shown in Fig. 9d have been merged and then split to obtain one cluster representing each stripe. Fig. 12 shows an example where the cluster that defined the boundary of the stripes, consisting of 296 spatially independent components, is split into its seven

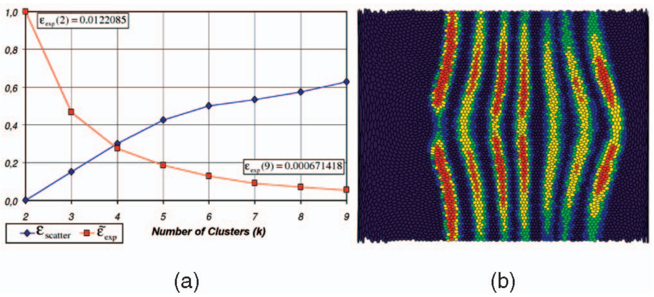


Fig. 10. (a) Cluster evaluation functions $\tilde{\varepsilon}_{\text{exp}}$ (red) and $\varepsilon_{\text{scatter}}$ (blue) for the clustering of the *eve* expression pattern, with $w = 4$ and $m = 9$. (b) The suggested k is five, which is the largest k for which only one cluster representing a low background *eve* expression is created.

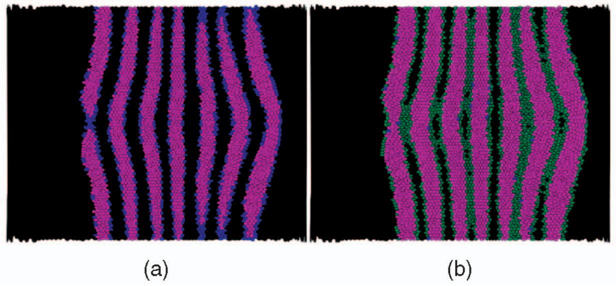


Fig. 11. Comparison of the clustering results shown in Fig. 9. (a) Comparison of $k = 2$ (red) and $k = 3$ (blue) classification of *eve*. (b) Comparison of $k = 3$ (red) and $k = 6$ (blue) classification of *eve*. In (b), the additional interstripe cluster found in the $k = 6$ clustering is shown in dark green. The percentage of cells of the whole embryo selected by the different components are $k2 = 31.31$ percent, $k3 = 42.596$ percent, $k6 = 42.892$ percent, $k3 - k2 = 11.287$ percent, $k6 - k3 = 0.296$ percent, and $k6\_interstripe\_cluster = 21.06$ percent.

main components using the modified single-linkage method described in Section 4.5.

Once derived, we use these individual stripe clusters to highlight the seven *eve* stripes via color in different abstract views. For example, the expression behavior of *gt*, *hb*, and *Kr*—three known transcriptional regulators of *eve*—can be revealingly analyzed within each of the *eve* stripes using a 3D scatterplot (Fig. 13). Here, large differences between stripes are visible, the seven stripes form very distinct point clusters within the scatterplot. This behavior is consistent with current models suggesting that the *eve* expression pattern does not simply consist of seven identical stripes but that many stripes are regulated independently. The available data suggest that *gt*, *hb*, and *Kr* control some stripes, but the scatterplot suggests that these factors have the potential to regulate all stripes by their unique combinations of expression levels. Such plots can be very useful in identifying potential novel regulatory relationships between transcription factors and their targets.

Generally, scatterplots have proven to be a very intuitive and informative gene expression space visualization but are limited due to the fact that only three gene dimensions can be visualized at once. PCX also provides 2D and 3D parallel coordinates to support simultaneous visualization of many more genes [4]. In Fig. 14, the same clusters as in Fig. 13 are shown in a 2D parallel coordinate view of early-stage *hb*, *gt*, *kni*, *Kr*, and *tll*, indicating additional expression differences between the spatial clusters. Because numerical PointCloud data sets are not easily comprehensible, the clustering and cluster manipulation capabilities in PCX provide a reasonably objective method for dividing quantitative spatial expression data into computationally analyzable units.

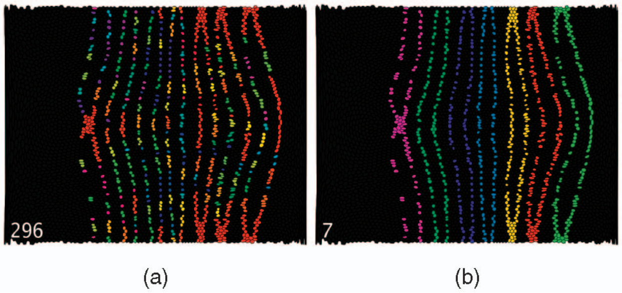


Fig. 12. (a) A cluster consisting of 296 spatially independent components. (b) The same cluster split into its seven main spatial components. Splitting of clusters is essential, e.g., to allow comparison of different main spatial components of a cluster.

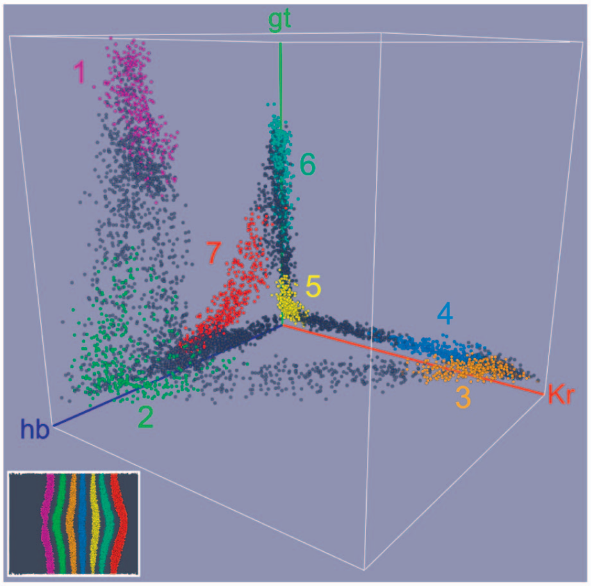


Fig. 13. An unrolled view showing seven clusters, each selecting one stripe of the *eve* expression pattern (bottom left). The same clusters shown in a scatterplot of early-stage *Kr* (red), *gt* (green), and *hb* (blue). The color indicates to which cluster a cell belongs, while cells not selected by any cluster are colored gray. The stripes form characteristic clusters in expression space, indicating a potential relationship between *eve* and the displayed genes.

7 Temporal Variation Analysis

Gene expression patterns are not static but highly dynamic. Understanding the temporal profile of a gene expression pattern is essential if we are to understand the complex relationships between genes. Even though visual inspection of an expression pattern at different time steps provides an impression of the general temporal behavior of a gene, many important features such as groups of cells with a similar temporal expression profile are not easily detected,

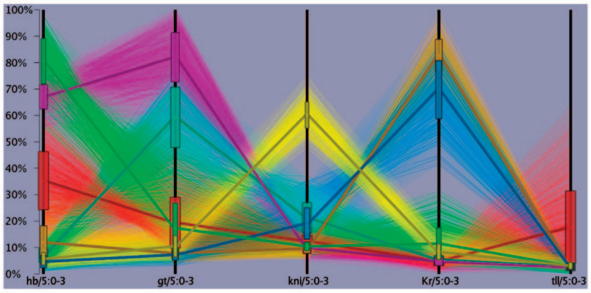


Fig. 14. The same clusters as in Fig. 13 are shown in a 2D parallel coordinate view of early-stage *hb*, *gt*, *kni*, *Kr*, and *tll*. The average expression of the seven clusters in the different genes are shown via additional thicker lines of darker colors, and the associated standard deviations are shown via boxes placed on each parallel axis. Highly transparent color bands shown in front of the plot are used to further highlight the different clusters.

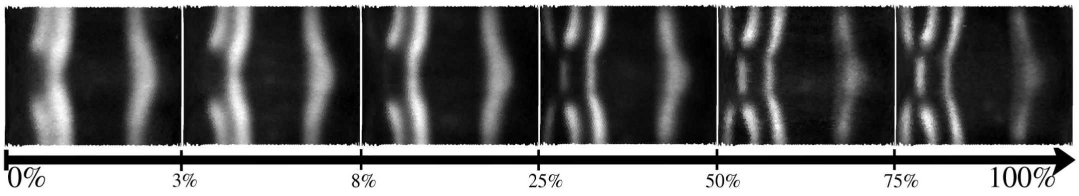


Fig. 15. The expression pattern of *giant (gt)* shown at six different time cohorts of stage 5 of embryo development.

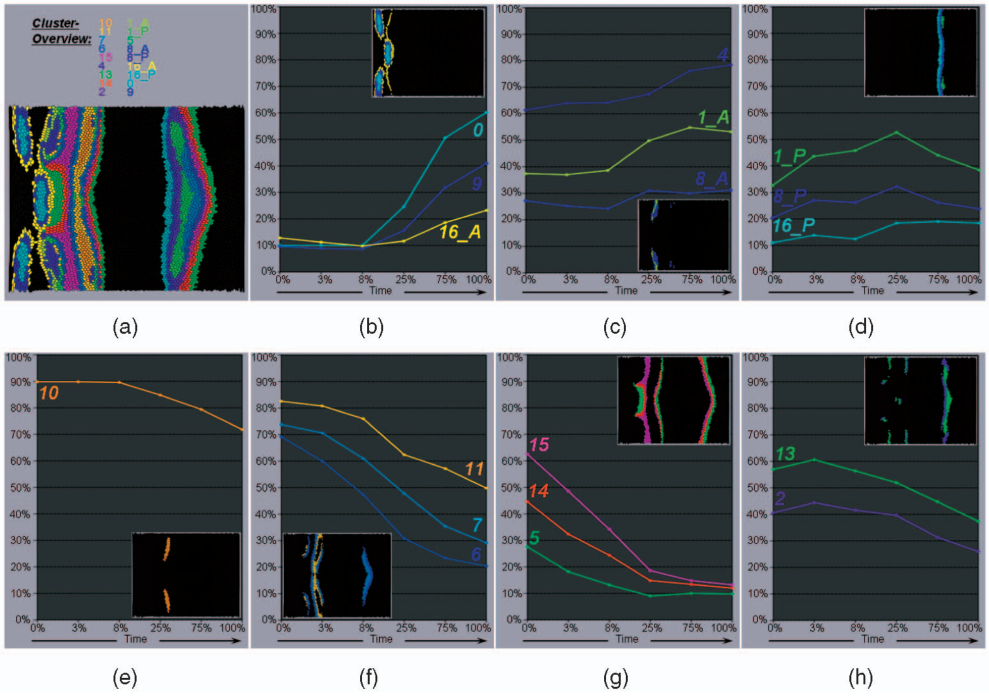


Fig. 16. Based on the patterns of *gt* shown in Fig. 15, cells were classified into 17 clusters, as suggested by $\varepsilon_{\text{scatter}}$ and ε_{exp}. Two clusters selected cells showing only background expression of *gt* at all time steps and are therefore not shown here. Clusters 1, 8, and 16 were each split into their two main spatial components. The remaining clusters were not split, since no significant divergence in the temporal expression profile between their main spatial components could be identified. (a) An unrolled view showing all 18 clusters of interest. (b), (c), (d), (e), (f), (g), (h) The user grouped the 18 temporal clusters into seven main groups based on their average temporal expression profiles in *gt*. The six time steps are shown on the x-axis, and the expression level is shown on the y-axis of each plot. The spatial patterns defined by the different clusters are displayed in the accompanying unrolled view plots.

and visual quantification of temporal change is not accurate. For example, the pattern of *giant (gt)* expression can be seen to change between six time cohorts within 1 hour, but it is not possible to rigorously describe how (see Fig. 15). To show how PCX can assist in the analysis of the spatiotemporal expression pattern of genes, we have used clustering to classify cells into groups of similar temporal behavior.

In Fig. 17, the curves of the cluster evaluation functions $\varepsilon_{\text{scatter}}$ and $\tilde{\varepsilon}_{\text{exp}}$ are shown. The suggested number of clusters k is 17, which is also a local minimum of $\varepsilon_{\text{scatter}}$ with $\varepsilon_{\text{scatter}}(17) \approx 31.88$ percent. The overall behavior of $\varepsilon_{\text{scatter}}$ indicates that $k = 17$ is the largest k at the particular level of detail for which $\varepsilon_{\text{scatter}}$ is still relatively low. A comparison of $\varepsilon_{\text{scatter}}(17)$ to the next two lower local minima of $\varepsilon_{\text{scatter}}$—with $\varepsilon_{\text{scatter}}(12) \approx 31.21$ percent and $\varepsilon_{\text{scatter}}(10) \approx 29.25$ percent—shows only a moderate increase in $\varepsilon_{\text{scatter}}$. When comparing $\varepsilon_{\text{scatter}}(17)$ to the $\varepsilon_{\text{scatter}}$ values of the next two larger local minima of $\varepsilon_{\text{scatter}}$—with $\varepsilon_{\text{scatter}}(19) \approx 36.34$ percent and $\varepsilon_{\text{scatter}}(22) \approx 40.14$ percent—a significantly higher increase in relative physical cluster scattering is visible. This behavior can be interpreted as an indication that $k = 17$ may also provide a good compromise between a high-level and low-level description of the temporal variation of the *gt* expression pattern. A level of $k = 17$ was also confirmed to be appropriate by users of PCX.

Fig. 16 shows as an example the result for *gt*, in which its expression patterns at six successive time cohorts were classified into 17 clusters using k-means clustering and euclidean distances. Two of the 17 clusters selected cells

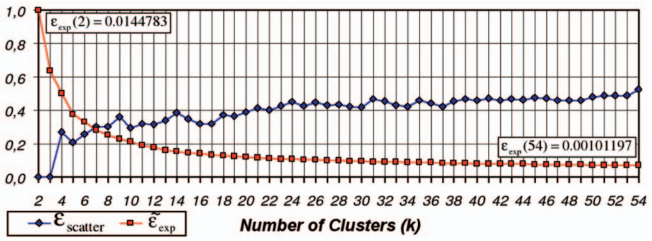


Fig. 17. Cluster evaluation functions $\tilde{\varepsilon}_{\text{exp}}$ (red) and $\varepsilon_{\text{scatter}}$ (blue) for the clustering of the six time steps of *gt* with $w = 10$ and $m = 54$. The suggested k is 17. $\varepsilon_{\text{scatter}}$ further indicates that 17 is the highest k for the particular level of detail with relatively low overall physical cluster scattering.

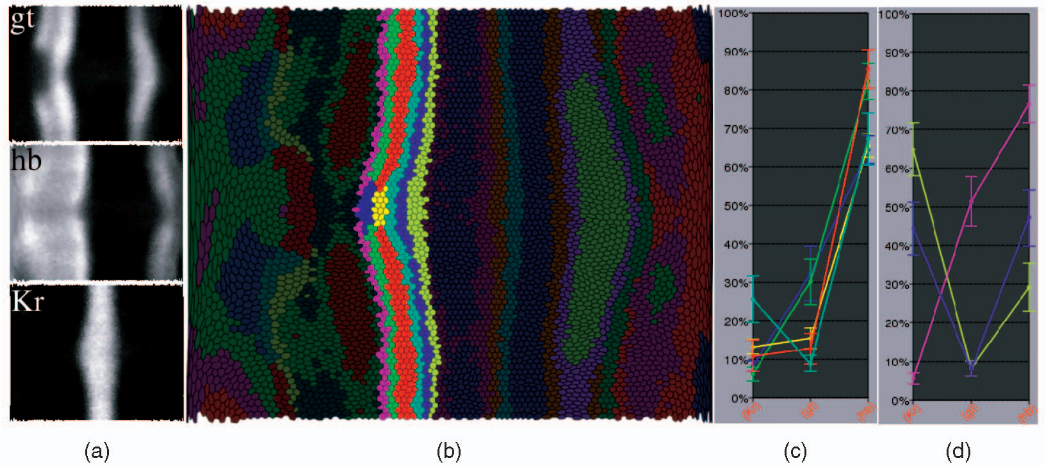


Fig. 18. (a) The transcription factors *gt*, *hb*, and *Kr* at stage 5 : 0 percent-3 percent are used as input to the clustering; their potential target is *eve* stripe 2 at stage 5 : 9 percent-25 percent (see Section 6). (b) Cells were classified into 22 clusters of which eight are of particular interest. Five clusters actually model *eve* stripe 2, and three define the interstripe region between stripes 1 and 2 and stripes 2 and 3. Cluster filtering was applied to three single cells only. Clusters were split in order to separate the stripelike clusters with similar expression profiles from other spatially distant subclusters in the anterior and posterior regions of the embryo. (c) An average curve plot of the five clusters within *eve* stripe 2 showing the characteristic expression profiles of *Kr*, *gt*, and *hb*. (d) Average expression curves for the three interstripe clusters. In both average curve plots, *Kr*, *gt*, and *hb* are shown on the x-axis, and the level of expression is shown along the y-axis.

showing only background expression at all time steps and are not shown. Each of the other 15 clusters show distinct average expression profiles (the differently colored lines plotted in Fig. 16), though some clusters show profiles that are closely related. In the figure, the user has grouped these clusters into seven main subgroups based on their temporal average expression profiles, shown in Figs. 16b, 16c, 16d, 16e, 16f, 16g, and 16h. In addition, clusters 1, 8, and 16 have each been split into two components to separate their anterior and posterior components.

Several trends can be readily seen from the different views of the analysis. The unrolled physical views show that clusters with similar average temporal expression profiles are frequently but not always adjacent to one another in the embryo. Expression within a set of clusters in the very anterior of the embryo increases, particularly during the later time cohorts (visible, for example, in Fig. 16b). Expression in the posterior margins of both of the major early *gt* stripes drops rapidly over the time series (Figs. 16f, 16g, and 16h). It is known that the location of the posterior *gt* stripe moves anteriorly during this time series [36], [3], but the data show a much more complex pattern of temporal change than has been observed previously. These results suggest that a complex combination of regulatory interactions drives these patterns.

8 Multiple-Pattern Analysis

To dissect the complex regulatory interactions between genes, the expression patterns of different transcription factors that potentially act together as regulators may be used as input to cluster analysis. Cells are classified into clusters that have similar combinations of expression for the input set of regulators. Each cluster thus describes one potential subpattern that a regulatory network composed of these factors could give rise to. The total number of clusters then gives an approximation of the maximal complexity of the output of the network. The results of such a clustering can also be compared to the expression patterns of suspected target genes to assess possible regulatory relationships.

To provide an example of such multigene clustering, we examined the relationship between the three transcriptional regulators *giant (gt)*, *hunchback (hb)*, and *Krüppel (Kr)* and the second stripe of the *eve* gene. These three factors are well-characterized regulators of this expression stripe; *hb* is an activator, and *Kr* and *gt* are repressors [37]. As discussed in Section 6, the seven stripes of *eve* form characteristic clusters in gene expression space with respect to *gt*, *hb*, and *Kr* expression. By using these three factors' expression patterns as input to a clustering analysis, we can identify the potential expression pattern components that can be defined based on these regulators (see Fig. 18). We used their mRNA expression values from the first temporal cohort (0 percent-3 percent invagination) to simulate their protein expression values at the third temporal cohort (9 percent-25 percent invagination)—the stage of the *eve* comparison target. We have found this lag, on the average, to be optimal for all regulators [7]. In the example, cells are classified into 22 clusters that map to locations throughout the embryo. Eight of these clusters are of interest to the control of *eve* stripe 2, five of which lie within the stripe and three are in the flanking interstripe regions. The five clusters within stripe 2 define the center, the anterior and posterior borders, and a ventral portion of the stripe, suggesting that these characteristic parts of stripe 2 may be different (see Fig. 18b).

To validate the structure formed by the clusters against the target pattern, cluster colors are mapped onto an expression surface of *eve*, in which the height shows the level of expression (Fig. 19). It can be seen that the five clusters fit closely to the expression pattern of the target stripe 2.

Based on the average expression curves, the characteristic expression pattern of the potential regulators in the

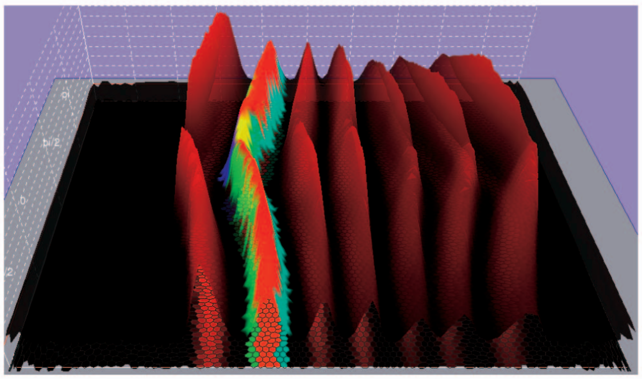


Fig. 19. To validate the structure formed by the five clusters against the target, cluster colors are mapped onto an expression surface of *eve*, where the surface height shows the level of *eve* expression. The visualization shows that the clusters and the target stripe fit closely.

eight clusters that are within and flanking stripe 2 are easily visible (see Figs. 18c and 18d). Here, *hb* is expressed at high levels in all clusters except those that are posterior of stripe 2, consistent with its known role as an activator of stripe 2. *Kr* is expressed at high levels only at the posterior of stripe 2, and *gt* is expressed at high levels only at the anterior of stripe 2, consistent with their known roles as repressors that define the posterior and anterior borders of stripe 2, respectively.

Interestingly, the two clusters that form short ventral patches on *eve* stripe 2 (yellow and blue) show a significantly lower expression of *hb* than the two clusters that lie dorsally to them (red and green) (see Figs. 18c and 18d). This correlates with a lower level of *eve* expression in this ventral margin (Fig. 19) and suggests that this reduced expression may be the result of a lower activation by *hb*. *hb* is typically thought of as a regulating gene expression only along the anterior/posterior axis of the embryo. The cluster analysis suggests that it may also be able to mediate differential transcription along the DV axis. However, if we were to add a DV gene such as *snail (sna)* (see Fig. 1) into the analysis, it would be difficult to distinguish if the ventral gap in *eve* stripe 2 resulted from direct inhibition by *sna* if *sna* acted via inhibiting ventral *hb* expression or if all three expression patterns are parallel manifestations of DV patterning systems, each acting separately. Thus, cluster analysis can be used for identifying interesting correlations that might result from novel biological interactions or phenomena, but the analyses should be confirmed by experimental data.

This case study illustrates that clustering the expression patterns of multiple regulators can provide confirmation and additional insights into known regulatory interactions. It is likely that the extension of this strategy to less well-characterized systems will suggest potential regulatory interactions that can then be tested by other means.

9 Conclusions

Our overall objective for this work has been to provide important new capabilities to accelerate scientific knowledge discovery. Our work helps biologists, who aim to discover potentially new experimentally verifiable biological interactions, by providing the ability to define, analyze, and iteratively refine clusters in multiple linked views. For computational biologists, we have presented objective methods for classifying quantitative data points in spatial data sets.

We have shown how data clustering and visualization can be integrated into one framework and how our system can be used effectively to explore and analyze 3D spatial expression data. A system of linked multiple views is used for data exploration and for steering the analysis process, helping bridge spatial patterns of expression with abstract views of quantitative expression information.

Data clustering then provides means for automatically defining cell selections, depicting characteristic data features, and, in this way, improving the visualization. We have shown how dedicated postprocessing of clustering results based on visualization and user knowledge improves the analysis. We have demonstrated how the combination of $\varepsilon_{\text{scatter}}$ as a measure to describe the relative physical scattering of clustering results and ε_{exp} as a measure to suggest a good initial value for k in combination with visual validation of clustering results can be used to determine appropriate values for k.

Analysis of 3D spatial gene expression data is a challenging task, requiring unique strategies not encountered in studies of 1D nonspatial data such as microarray expression data. Using our integrated data visualization and clustering approach, we have shown how the pattern of a gene and its temporal variation can be defined and analyzed. We have shown how suspected relationships between genes can be analyzed to address the question of how the pattern of a gene is created by the action of multiple regulators.

Along with the first release of the BDTNP 3D gene expression database, we have also made a version of PCX freely available to the public [1]. Data clustering and 3D parallel coordinates are currently in active use by BDTNP members and will soon also be included in the public version of PCX.

In PCX, spatial information is incorporated in the analysis process mainly by using cluster postprocessing techniques such as splitting of clusters. Alternatively, the x, y, and z cell positions can be directly added to the cluster analysis. However, this may result in clusters defined by a complex mix of spatial and expression influences, which may not be easy to interpret.

The development of additional analysis techniques that effectively integrate spatial and gene expression information is one focus of future work. Adaptation of spatial clustering methods such as the dual clustering approach proposed by Cheng-Ru et al. [38] is only one promising approach. Alternatively, one could perform clustering based on gene expression information only, then split the resulting clusters into spatially distinct subclusters, and then perform a reclustering based on the centers of the detected subclusters. In PCX, we currently use hierarchical clustering only for the partitioning of the data. By traversing the data hierarchy created in a hierarchical clustering, exploration of the data at multiple levels of detail becomes possible. In addition to clustering of cells, clustering of genes, as well as biclustering, promises to provide further insights into the data. In addition, matrix decomposition techniques such as principal component analysis (PCA) and singular value decomposition (SVD) [39], [40] have successfully been applied to other types of gene expression data. Integration of these and other analysis techniques into PCX should further increase its value for practical use and impact.

Acknowledgments

This work was supported by the National Institutes of Health through Grant GM70444, by the US National Science Foundation through Award ACI 9624034 (CAREER Award) as part of the Large Scientific and Software Data Set Visualization (LSSDSV) program under Contract ACI 9982251, by the Lawrence Berkeley National Laboratory (LBNL) Laboratory Directed Research Development (LDRD) program, and by a large Information Technology Research (ITR) grant. Work at LBNL is conducted under the Department of Energy Contract DE-AC02-05CH11231. The authors thank the members of the Visualization and Computer Graphics Research Group at the Institute for Data Analysis and Visualization (IDAV), University of California, Davis, the members of the BDTNP at LBNL, and the members of the Visualization Group at LBNL.

References

[1] BDTNP, http://bdtnp.lbl.gov/Fly-Net/bioimaging.jsp, 2008.

[2] C.L. Luengo Hendriks, S.V.E. Keränen, C.C. Fowlkes, L. Simirenko, G.H. Weber, A.H. DePace, C. Henriquez, D.W. Kaszuba, B. Hamann, M.B. Eisen, J. Malik, D. Sudar, M.D. Biggin, and D.W. Knowles, "Three-Dimensional Morphology and Gene Expression in the *Drosophila* blastoderm at Cellular Resolution I: Data Acquisition Pipeline," *Genome Biology,* vol. 7, no. 12, p. R123, http://genomebiology.com/2006/7/12/R123, 2006.

[3] S.V.E. Keränen, C.C. Fowlkes, C.L. Luengo Hendriks, D. Sudar, D.W. Knowles, J. Malik, and M.D. Biggin, "Three-Dimensional Morphology and Gene Expression in the *Drosophila* Blastoderm at Cellular Resolution I: Dynamics," *Genome Biology,* vol. 7, no. 12, p. R124, http://genomebiology.com/2006/7/12/R124, 2006.

[4] O. Rübel, G.H. Weber, S.V.E. Keränen, C.C. Fowlkes, C.L. Luengo Hendriks, L. Simirenko, N.Y. Shah, M.B. Eisen, M.D. Biggin, H. Hagen, J.D. Sudar, J. Malik, D.W. Knowles, and B. Hamann, "Pointcloudxplore: Visual Analysis of 3D Gene Expression Data Using Physical Views and Parallel Coordinates," *Proc. Joint Eurographics-IEEE VGTC Symp. Visualization (EuroVis '06),* B.S. Santos, T. Ertl, and K. Joy, eds., pp. 203-210, 2006.

[5] G.H. Weber, O. Rübel, M.-Y. Huang, A.H. DePace, C.C. Fowlkes, S.V.E. Keränen, C.L. Luengo Hendriks, H. Hagen, D.W. Knowles, J. Malik, M.D. Biggin, and B. Hamann, "Visual Exploration of Three-Dimensional Gene Expression Using Physical Views and Linked Abstract Views," accepted for publication in *IEEE/ACM Trans. Computational Biology and Bioinformatics,* 2008.

[6] C.C. Fowlkes, C.L. Luengo Hendriks, S.V.E. Keränen, M.D. Biggin, D.W. Knowles, D. Sudar, and J. Malik, "Registering *Drosophila* Embryos at Cellular Resolution to Build a Quantitative3D Map of Gene Expression Patterns and Morphology," *Proc. CSB Workshop BioImage Data Mining and Informatics,* Aug. 2005.

[7] C.C. Fowlkes, C.L. Luengo Hendriks, S.V.E. Keränen, G.H. Weber, O. Rübel, M.-Y. Huang, S. Chatoor, L. Simirenko, M.B. Eisen, B. Hamann, D.W. Knowles, M.D. Biggin, and J. Malik, *Constructing a Quantitative Spatio-Temporal Atlas of Gene Expression in the Drosophila blastoderm,* submitted, 2008.

[8] A.K. Jain, M.N. Murty, and P.J. Flynn, "Data Clustering: A Review," *ACM Computing Surveys,* vol. 31, no. 3, Sept. 1999.

[9] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein, "Cluster Analysis and Display of Genome-Wide Expression Patterns," *Proc. Nat'l Academy of Sciences USA,* pp. 14863-14868, 1995.

[10] A.A. Alizadeh, M.B. Eisen, R.E. Davis, C. Ma, I.S. Lossos, A. Rosenwald, J.C. Boldrick, H. Sabet, T. Tran, X. Yu, J.I. Powell, L. Yang, G.E. Marti, T. Moore, J. Hudson, L. Lu, D.B. Lewis, R. Tibshirani, G. Sherlock, W.C. Chan, T.C. Greiner, D.D. Weisenburger, J.O. Armitage, R. Warnke, R. Levy, W. Wilson, M.R. Grever, J.C. Byrd, D. Botstein, P.O. Brown, and L.M. Staudt, "Distinct Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression," *Nature,* vol. 403, pp. 503-511, Feb. 2000.

[11] J. Ihmels, G. Friedlander, S. Bergmann, O. Sarig, Y. Ziv, and N. Barkai, "Revealing Modular Organization in the Yeast Transcriptional Network," *Nature Genetics,* vol. 31, part 4, pp. 370-378, 2002.

[12] M.B. Eisen Cluster 2.20 and Treeview 1.60, http://rana.lbl.gov/EisenSoftware.htm, 2002.

[13] Spotfire, Decision Site, http://www.spotfire.com, 2008.

[14] M. Reich, K. Ohm, P. Tamayo, M. Angelo, and J.P. Mesirov, "Genecluster 2.0: An Advanced Toolset for Bioarray Analysis," *Bioinformatics,* 2004.

[15] A. Saeed, V. Sharov, J. White, J. Li, W. Liang, N. Bhagabati, J. Braisted, M. Klapa, T. Currier, M. Thiagarajan, A. Sturn, M. Snuffin, A. Rezantsev, D. Popov, A. Ryltsov, E. Kostukovich, I. Borisovsky, Y. Liu, A. Vinsavich, V. Trush, and J. Quackenbush, "TM4: A Free, Open-Source System for Microarray Data Management and Analysis," *Biotechniques,* vol. 34, no. 2, pp. 374-378, 2003.

[16] Rosetta Biosoftware, http://www.rosettabio.com, 2008.

[17] J. Seo and B. Shneiderman, "A Knowledge Integration Framework for Information Visualization," *From Integrated Publication and Information Systems to Information and Knowledge Environments,* LNCS 3379, pp. 207-220, 2005.

[18] J. Handl, J. Knowles, and D.B. Kell, "Computational Cluster Validation in Post-Genomic Data Analysis," *Bioinformatics,* vol. 21, no. 15, pp. 3201-3212, 2005.

[19] K.Y. Yeung, D.R. Haynor, and W.L. Ruzzo, "Validating Clustering for Gene Expression Data," *Bioinformatics,* vol. 17, no. 4, pp. 309-318, 2001.

[20] S. Datta and S. Datta, "Comparison and Validation of Statistical Clustering Techniques for Microarray Gene Expression Data," *Bioinformatics,* vol. 19, no. 4, pp. 459-466, 2003.

[21] A. Ben-Dor, N. Friedmann, and Z. Yakhini, "Overabundance Analysis and Class Discovery in Gene Expression Data," technical report, Agilent Laboratories, Palo Alto, 2002.

[22] R. Tibshirani, G. Walther, and T. Hastie, "Estimating the Number of Clusters in a Dataset via the Gap Statistic," Technical Report 208, Dept. Statistics, Stanford Univ., 2000.

[23] G.W. Miligan and M.C. Cooper, "An Examination of Procedures for Determining the Number of Clusters in a Data Set," *Psychometrika,* vol. 50, no. 2, pp. 159-179, June 1985.

[24] M.Q.W. Baldonado, A. Woodruff, and A. Kuchinsky, "Guidelines for Using Multiple Views in Information Visualization," *Proc. Working Conf. Advanced Visual Interfaces (AVI '00),* pp. 110-119, 2000.

[25] C. Henze, "Feature Detection in Linked Derived Spaces," *Proc. IEEE Conf. Visualization (VIS '98),* D. Ebert, H. Rushmeier, and H. Hagen, eds., pp. 87-94, 1998.

[26] D.L. Gresh, B.E. Rogowitz, R.L. Winslow, D.F. Scollan and C.K. Yung, "WEAVE: A System for Visually Linking 3D and Statistical Visualizations, Applied to Cardiac Simulation and Measurement Data," *Proc. IEEE Conf. Visualization (VIS '00),* T. Ertl, B. Hamann, and A. Varshney, eds., pp. 489-492, 2000.

[27] H. Doleisch, M. Gasser, and H. Hauser, "Interactive Feature Specification for Focus+Context Visualization of Complex Simulation Data," *Proc. Eurographics/IEEE TCVG Symp. Visualization (VisSym '03),* G.-P. Bonneau, S. Hahmann, and C.D. Hansen, eds., 2003.

[28] R. Kosara, G.N. Sahling, and H. Hauser, "Linking Scientific and Information Visualization with Interactive 3D Scatterplots," *Short Comm. Papers Proc. 12th Int'l Conf. in Central Europe on Computer Graphics, Visualization, and Computer Vision (WSCG '04),* pp. 133-140, 2004.

[29] H. Piringer, R. Kosara, and H. Hauser, "Interactive Focus+Context Visualization with Linked 2D/3D Scatterplots," *Proc. Second Int'l Conf. Coordinated and Multiple Views in Exploratory Visualization (CMV '04),* pp. 49-60, 2004.

[30] Y.-H. Fua, M.O. Ward, and E.A. Rundensteiner, "Hierarchical Parallel Coordinates for Exploration of Large Datasets," *Proc. IEEE Conf. Visualization (VIS '99)*, pp. 43-50, 1999.

[31] M.J.L. de Hoon, S. Imoto, J. Nolan, and S. Miyano, "Open Source Clustering Software," *Bioinformatics*, vol. 20, no. 9, pp. 1453-1454, 2004.

[32] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrosky, E.S. Lander, and T.R. Golub, "Interpreting Patterns of Gene Expression with Self-Organizing Maps: Methods and Application to Hematopoietic Differentiation," *Proc. Nat'l Academy of Sciences USA*, vol. 96, pp. 2907-2912, Mar. 1999.

[33] P.L. Rosin, "Unimodal Thresholding," *Pattern Recognition*, vol. 34, no. 11, pp. 2083-2096, 2001.

[34] R. Albert and H. Othmer, "The Topology of the Regulatory Interactions Predicts the Expression Pattern of the Segment Polarity Genes in *Drosophila melanogaster*," *J. Theoretical Biology*, vol. 223, no. 1, pp. 1-18, July 2003.

[35] L. Sanchez and D. Thieffry, "A Logical Analysis of the *Drosophila* Gap-Gene System," *J. Theoretical Biology*, vol. 211, no. 2, pp. 115-141, 2001.

[36] J. Jaeger, S. Surkova, M. Blagov, H. Janssens, D. Kosman, K.N. Kozlov, Manu, E. Myasnikova, C.E. Vanario-Alonso, M. Samsonova, D.H. Sharp, and J. Reinitz, "Dynamic Control of Positional Information in the Early *Drosophila* Embryo," *Nature*, vol. 430, pp. 368-371, http://dx.doi.org/10.1371%2Fjournal.pcbi.0020051, 2004.

[37] P.A. Lawrence, *The Making of a Fly: The Genetics of Animal Design*. Blackwell, 1992.

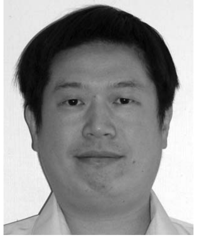
[38] C.-R. Lin, K.-H. Liu, and M.-S. Chen, "Dual Clustering: Integrating Data Clustering over Optimization and Constraint Domains," *IEEE Trans. Knowledge and Data Eng.*, vol. 17, no. 5, pp. 628-637, May 2005.

[39] O. Alter, P.O. Brown, and D. Botstein, "Singular Value Decomposition for Genome-Wide Expression Data Processing and Modeling," *Proc. Nat'l Academy of Sciences USA*, pp. 10101-10106, 2000.

[40] M.E. Wall, A. Rechtsteiner, and L.M. Rocha, "Singular Value Decomposition and Principal Component Analysis," *A Practical Approach to Microarray Data Analysis*, pp. 91-109, Kluwer Academic Publishers, 2003.

Oliver Rübel received the MS degree in computer science from the University of Kaiserslautern, Kaiserslautern, Germany, in 2006. He is currently a PhD student at the University of Kaiserslautern. He is also a student assistant in the Visualization Group, Lawrence Berkeley National Laboratory (LBNL), Berkeley, California, a collegiate of IRTG 1131, University of Kaiserslautern, and a visiting scholar in the Institute for Data Analysis and Visualization (IDAV), University of California, Davis. His current research focus is visualization and analysis of high-dimensional data.

Gunther H. Weber received the Diplom in computer science in 1999 and the PhD degree in computer science in 2003, both from the University of Kaiserslautern, Germany. He is a computer research scientist/engineer at the Lawrence Berkeley National Laboratory (LBNL) and the National Energy Research Scientific Computing Center (NERSC). Furthermore, he holds an appointment as adjunct assistant professor in the Computer Science Department at the University of California, Davis (UC Davis). Prior to his tenure at LBNL, he was first a postdoctoral scholar and later a project scientist at the Institute for Data Analysis and Visualization (IDAV) at UC Davis. He is a member of the IEEE Computer Society.



Min-Yu Huang received the BS degrees in physics and computer science (double major) and the MS degree in computer science from the National Tsing-Hua University, Hsinchu, Taiwan, in 1995 and 1997, respectively. He is currently a PhD candidate in the Department of Computer Science, University of California, Davis, where he is also a member of the Institute for Data Analysis and Visualization (IDAV). He is a student member of the IEEE.

E. Wes Bethel received the MS degree in computer science from the University of Tulsa in 1986. He is a staff scientist at the Lawrence Berkeley National Laboratory, Berkeley, California. He is also a codirector of DoE's SciDAC Visualization and Analytics Center for Enabling Technologies and the founding technical director of R3vis Corp. His research interests include high-performance remote and distributed visualization algorithms and architectures. He is a member of the ACM and the IEEE.

Mark D. Biggin received the BSc degree in biochemistry from Lancaster University in 1981 and the PhD degree in molecular biology from the MRC Laboratory of Molecular Biology, Cambridge University, in 1985. He was a postdoctoral fellow in Robert Tjian's laboratory at the University of California, Berkeley, from 1985 to 1989, before joining the faculty at Yale University as an assistant then associate professor from 1989 to 2000. He moved to the Lawrence Berkeley National Laboratory, Berkeley, California, in 2000 to establish interdisciplinary research projects that seek systems level understandings of animal developmental transcriptional networks and bacterial stress response pathways.

Charless C. Fowlkes received the BS degree (with honors) from the California Institute of Technology (Caltech) in 2000 and the PhD degree in computer science from the University of California, Berkeley, in 2005, where his research was supported by a US National Science Foundation Graduate Research Fellowship. He is currently an assistant professor in the Donald Bren School of Information and Computer Sciences, University of California, Irvine. His research interests include the analysis and modeling of spatial gene expression, computer vision, and the ecological statistics of natural images.

Cris L. Luengo Hendriks received the MSc and the PhD degrees from the Department of Applied Physics, Delft University of Technology, The Netherlands, in 1998 and 2004, respectively. While this research was done, he worked as a postdoctoral fellow at Lawrence Berkeley National Laboratory (LBNL), where he developed the software that obtains the Single PointCloud files used in this paper from three-dimensional fluorescence images. He is currently an associate professor at Uppsala University, Sweden. His research interests include image processing, image analysis, and gene expression pattern analysis. He is a member of the IEEE and the IEEE Signal Processing Society.

Soile V.E. Keränen received the PhD degree in genetics from the University of Helsinki. She is a scientist at the Lawrence Berkeley National Laboratory, Berkeley, California, where she is part of the Berkeley Drosophila Transcription Network Project. Her research interests include developing methods for analysis of and discovering rules of spatial pattern formation and evolution of developmental regulatory processes using *Drosophila* embryos and virtual organisms as model systems. She has background as an evolutionary developmental biologist.

Michael B. Eisen received the AB degree in mathematics from Harvard College in 1989 and the PhD degree in biophysics from Harvard University in 1996. His thesis focused on the structure and evolution of influenza A virus proteins. Following a brief stint as a minor-league baseball announcer, from 1996 to 2000, he was a postdoctoral fellow in the School of Medicine, Stanford University, with Patrick O. Brown and David Botstein. In the Brown and Botstein laboratories, he introduced and demonstrated the value of statistical clustering methods in the analysis of genome-wide expression data. His 1998 paper on the subject has more than 6,000 citations. In 2000, he started his own laboratory at the University of California, Berkeley, where he is currently an associate professor of genetics, genomics, and development in the Department of Molecular and Cell Biology and a scientist in the Genomics Division, Lawrence Berkeley National Laboratory. Research in the Eisen Lab focuses on understanding how gene expression patterns are encoded in DNA and dissecting the roles that changes in regulatory sequences have played in organismal diversification.

David W. Knowles received the PhD degree in physics from the University of British Columbia, Canada, in 1992. He came to the Lawrence Berkeley National Laboratory, Berkeley, California, in 1994 as a postdoctoral fellow to unravel the macromolecular interactions of the red blood cell membrane. In 1999, as a scientist, he established the BioImaging Group at the Lawrence Berkeley National Laboratory. His research focuses on developing imaging, image analysis, visual and statistical techniques to quantify and map cellular and subcellular events in biological systems.

Jitendra Malik received the BTech degree in electrical engineering from the Indian Institute of Technology (IIT), Kanpur, in 1980 and the PhD degree in computer science from Stanford University in 1985. In 1986, he joined the University of California, Berkeley, where he is currently the Arthur J. Chick Endowed Professor of the Department of Electrical Engineering and Computer Sciences (EECS). His research interests include computer vision, computational modeling of human vision, and analysis of biological images. He received the gold medal for the best graduating student in electrical engineering from IIT Kanpur in 1980, a Presidential Young Investigator Award in 1989, and the Rosenbaum fellowship for the Computer Vision Programme at the Newton Institute of Mathematical Sciences, University of Cambridge, in 1993. He received the Diane S. McEntyre Award for Excellence in Teaching in 2000. He received a Miller Research Professorship in 2001. At CVPR 2007, he was awarded the Longuet-Higgins Prize for a contribution that has stood the test of time. He is a fellow of the IEEE.

Hans Hagen received the BS degree in computer science and BS and MS degrees in mathematics from the University of Freiburg and the PhD degree in mathematics (geometry) from the University of Dortmund in 1982. He is a professor of computer science at the University of Kaiserslautern, where he teaches and conducts research in the areas of scientific visualization and geometric modeling. He is the P.I. of the DFG International Graduate School for Visualization of KLarge and Unstructured Data Sets. He has written more than 250 scientific articles and edited several books. He closely cooperates with several institutions and universities worldwide. He is a member of the IEEE and received the IEEE Visualization Career Award in 2009.

Bernd Hamann received the PhD degree from Arizona State University in 1991. He is the associate vice chancellor for research and is a professor in the Department of Computer Science at the University of California, Davis. His research and teaching interests are visualization, geometric modeling, and computer graphics. He received a 1992 Research Initiation Award and a 1996 CAREER Award from the National Science Foundation and obtained the 2006 University of California Presidential Chair in Undergraduate Education. He is a member of the IEEE.