PointCloudXplore: A Visualization Tool for 3D Gene Expression Data

Oliver Rübel[*,1,2], Gunther H. Weber[2,3], Soile V.E. Keränen[3], Charless C. Fowlkes[4], Cris L. Luengo Hendriks[3], Lisa Simirenko[3], Nameeta Y. Shah[2], Michael B. Eisen[3], Mark D. Biggin[3], Hans Hagen[1], Damir Sudar[3], Jitendra Malik[4], David W. Knowles[3], and Bernd Hamann[1,2]

[1] International Research Training Group "Visualization of Large and Unstructured Data Sets," University of Kaiserslautern, Germany

[2] Institute for Data Analysis and Visualization, University of California, Davis, CA, USA

[3] Life Sciences and Genomics Divisions, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

[4] Computer Science Division, University of California, Berkeley, CA, USA

Abstract: The Berkeley Drosophila Transcription Network Project (BDTNP) has developed a suite of methods that support quantitative, computational analysis of three-dimensional (3D) gene expression patterns with cellular resolution in early *Drosophila* embryos, aiming at a more in-depth understanding of gene regulatory networks. We describe a new tool, called PointCloudXplore *(PCX)*, that supports effective 3D gene expression data exploration.

PCX is a visualization tool that uses the established visualization techniques of *multiple views*, *brushing*, and *linking* to support the analysis of high-dimensional datasets that describe many genes' expression. Each of the views in PointCloudXplore shows a different gene expression data property. Brushing is used to select and emphasize data associated with defined subsets of embryo cells within a view. Linking is used to show in additional views the expression data for a group of cells that have first been highlighted as a brush in a single view, allowing further data subset properties to be determined. In PCX, physical views of the data are linked to abstract data displays such as parallel coordinates. Physical views show the spatial relationships between different genes' expression patterns within an embryo. Abstract gene expression data displays on the other hand allow for an analysis of relationships between different genes directly in the gene expression space. We discuss on parallel coordinates as one example abstract data view currently available in PCX. We have developed several extensions to standard parallel coordinates to facilitate brushing and the visualization of 3D gene expression data.

[*] oliverruebel@web.de

1 Introduction

Animal embryos comprise dynamic 3D arrays of cells that express gene products in intricate spatial and temporal patterns that determine the shape of the developing animal. Biologists have typically analyzed gene expression and morphology by visual inspection of photomicrographic images. Yet, to understand animal development, we need good methods to computationally describe gene expression data. To address this challenge, the BDTNP has developed image analysis methods to extract information about gene expression from imaging data, using early *Drosophila melanogaster* embryos as a model. Confocal image stacks of blastoderm stage *Drosophila* embryos are converted into matrices specifying the position of nuclei and the expression levels of select genes around each nucleus, (see Section 3). The resulting new datasets, called PointClouds, promise to be an invaluable resource for studying development. Since available visualization tools are insufficient for comparing and analyzing 3D PointCloud data, we have developed PointCloudXplore as a tool to help biologists explore these datasets.

During embryogenesis complex regulatory networks are built up as transcription factors cross-regulate the expression of other transcription factors as well as enzymes, structural proteins, etc., guiding the development of animals [WKS+03, Law92, SL05]. Since spatial regulation of gene expression directs animal morphogenesis, a major goal of the BDTNP is to decipher how spatial patterns of target gene expression are directed by the expression patterns of the transcription factors that regulate them. Because gene regulation depends on combinatorial inputs from many transcription factors, simultaneous analysis of many expression patterns is required. Therefore, PCX includes multiple visualization methods to allow specific relationships to be seen within highly complex expression data for many genes.

2 Previous Work

Generally, data can be displayed in multiple formats, or *views*, that each allow different relationships between the data components to be observed. The *linking* of multiple views is a well-established visualization method [BMMS91]. For example, it has been shown that linking abstract data displays, such as scatter plots, with physical data views, such as a 3D model of a catalytic converter, can improve data analysis significantly and provide insight into complex physical phenomena [KSH04, PKH04, DGH03]. Hauser et al. [HLD02] proposed integrating parallel coordinates with physical views for a better understanding of high-dimensional phenomena. Gresh et al. [GRW+00] used parallel coordinates linked to physical views for visualizing biological data sets describing cardiac measurement and simulation experiments.

Parallel coordinates were proposed contemporaneously by Inselberg [Ins84] and Wegman [Weg90] and are a common information visualization technique for high-dimensional data sets. In a parallel coordinate view, a data set consists of a set of *samples*, which in our case are the cells in a *Drosophila* embryo. Each sample (cell) has a set of associated

quantities, which in our case are the relative expression levels for multiple genes. Expression data for each gene corresponds to a dimension in the data set, with data for each gene being represented by one of a series of parallel vertical axes. Each sample (cell) defines a *data line*, i.e., a zig-zag line connecting adjacent parallel axes. The intersection point of the data line with each vertical axis corresponds to the value of the sample for the corresponding dimension (i.e., the relative expression level for the corresponding gene in that cell).

Many extensions to standard parallel coordinates have been developed to make them more useful for practical applications. Fua et al. [FWR99] proposed hierarchical parallel coordinates, including several techniques for visualization of selected subsets of the data. Distortion operations, such as dimensional zooming, for example, support a more detailed analysis of data subspaces. Color is widely used for improving parallel-coordinate views since dedicated line coloring eases following the course of data lines. Fua et al. [FWR99] and Novotny [Nov04] proposed the use of color bands for visualization of brushes in parallel coordinate views. Wegman and Luor [WL97] proposed the application of transparency and "over-plotting" translucent data points/lines. This method highlights dense areas while sparse areas fade away, thus revealing inherent data characteristics.

3 Gene Expression Data and Visualization Pipeline

A single PointCloud file contains the x,y,z location of each nucleus in one embryo and the relative concentration of gene products (mRNA or protein) associated to each nucleus [FLHK$^+$05]. These files are created in the following manner (see Figure 1). Embryos are fixed, stained, and mounted, then imaged using a confocal microscope (Figure 1, *IA*). The obtained images are processed to detect all nuclei and measure the associated gene expression levels (Figure 1, *IS*). Embryos are typically labeled with one fluorophore to detect the nuclei, and with two others to detect two gene products. It is not practical to obtain the expression of more than a few genes in a single embryo, due to the limited number of different fluorophores that can be distinguished by the microscope, as well as the difficulty in adding these labels to the embryos. Since it is critical to compare the relationships between transcription factors and many of their target genes in a common co-ordinate framework, a set of PointClouds using both morphology and a common reference gene to determine correspondences (Figure 1, *ER*) are registered into Virtual PointClouds [FLHK$^+$05]. The resulting Virtual PointCloud contains averaged expression levels for many genes mapped on the nuclei of one of the embryos in the set. PCX, see Figure 1, can be used for visualization of both single-embryo PointClouds and Virtual PointClouds.

4 Physical Views

We have developed several physical views (models) of the embryo to support analysis of spatial gene expression patterns. In all these Embryo Views, each cell is represented by

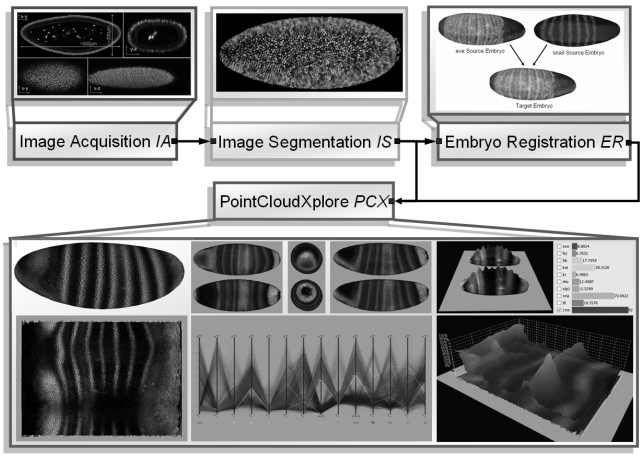


Figure 1: Gene Expression Data and Visualization Pipeline: PCX is used to visualize data from single embryo PointClouds and Virtual PointClouds.

one 3D graphical object, positioned in space according to the physical position of the cell it represents. Gene expression values are visualized using color and also by height, in the case of views using gene Expression Surfaces (Section 4.3), . During the developmental stage that the BDTNP is currently investigating — the blastoderm — all cells studied by the BDTNP are located on a surface in shape very similar to an ellipsoid. Orthographic projection is used to display views from fixed angles that display expression and morphology along the three coordinate axes (Section 4.2). In another view, a cylindrical projection is used to map cells onto a plane to gain a global overview of the entire embryo (Section 4.2). Cells of interest can be selected in any of these views to create a so called *brush*, just by drawing on the surface of the embryo. The selected cells that comprise the brush are highlighted using color. The user can interact with all embryo views via interactive zooming, panning, and rotation.

4.1 3D Physical View

In all our Embryo Views, each cell is represented by a polygon on the embryo surface, using the Eigencrust method [KSO04] for constructing an approximation of the Delauny triangulation of the surface of a PointCloud. The dual mesh of the triangulation is a tesselation similar to a Voronoi diagram, defined on the embryo surface [dBvKOS00]. Each cell is represented by a Voronoi polygon with exactly one original data point in its center. The polygon sizes depend on the cells' distribution in the embryo, whereas the shape of the polygons has no direct meaning. This results in a 3D model of the embryo which provides an intuitive way to look at the data (see Figure 2).

Each polygon is colored according to expression values measured in the cell it represents. The color mapping is based on the HSV color model [FDFH95]. The basic color hue, H,

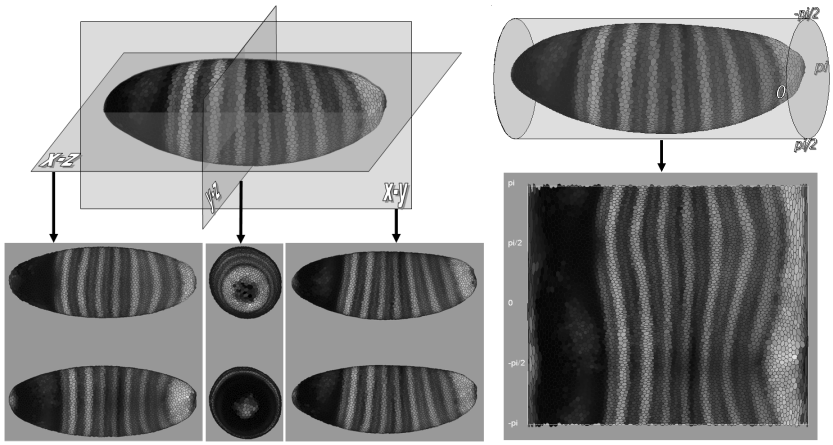


Figure 2: 2D Orthographic View (left); Unrolled View (right).

of each gene is defined by the user. Gene expression values are mapped linearly to color brightness V. Saturation of color is always one, unless specified differently. For each gene, a minimum and maximum value can be defined independently by the user. All expression values below the minimum are mapped to black, and all values above the maximum are mapped to full intensity.

4.2 2D Physical Views

To allow one to obtain a much quick overview of all cells several 2D visualizations of the embryo have been developed. As shown in Figure 2 (left), by centering the main coordinate system within the 3D embryo model, orthographic projection can be used to create three 2D views of the embryo showing the dorsal/ventral, anterior/posterior, and the left/right sides of the embryo. These Orthographic Views provide an overview of all cells while preserving the shape of the embryo as a frame of reference for a biologist. The curvature of the embryo leads to high densities of cells on the projection borders, but provides an impression of depth and shape. A general overview of all cells in an embryo can be obtained by switching between the three different Orthographic Views.

For a scientist who wants an instant overview of the whole blastoderm expression pattern, the Unrolled View uses a cylindrical projection to map all surface cells of the embryo onto a rectangular plane (see Figure 2 (right)). All cells are first projected onto a cylinder, which is unrolled in a plane. In this view, a complete overview of all cells is provided while the relative positions of cells on the anterior/posterior axis and around the embryo are preserved. Due to of the ellipsoid like shape of the embryo, cells in the anterior and posterior of the embryo are distorted by the projection in order to fill the rectangle. Shape and size of cells in the middle part of the embryo are less affected by distortion effects.

All 2D physical views are projections of the original 3D embryo model. To visualize gene expression values and brushes the same color mapping is used as in the 3D physical view (see Section 4.1).

4.3 Expression Surfaces

To support a more quantitative analysis of gene expression data, Expression Surfaces can be defined above either the Orthographic or the Unrolled Views. Each Expression Surface displays data for one gene. The xy positions of Expression Surface points are determined by the positions of cells in the underlying views, whereas the height of an Expression Surface is determined by the expression values measured for the gene it represents. Spatial relationships between several genes' expression patterns can be viewed at once using multiple Expression Surfaces. For example, Figure 3 shows the quantitative relationship between *eve* and *ftz*. The expression levels of these two genes are spatially largely non-overlapping and change relative to one another along each body axis.

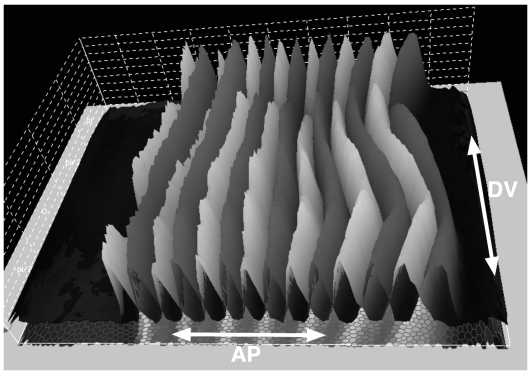


Figure 3: Gene expression surfaces for *eve* (light gray) and *ftz* (dark gray).

5 3D Parallel Coordinates

A limitation of all described Embryo Views is that when more than four or five genes whose expression overlaps are displayed at the same time, it is often not possible to distinguish each gene's expression. Therefore, we have adapted parallel coordinates to create several Parallel Coordinate Views, in which relationships between many genes' expression can be visualized. See Section 2 for an introduction to parallel coordinates. Further, we have linked Parallel Coordinate Views and Embryo Views, ensuring that all brushes defined in the Embryo Views can be displayed in the Parallel Coordinate Views and vice versa. In general, parallel coordinates introduce several other visulaization problems, such as occlusion and cluttering. To reduce these problems, several extensions to standard par-

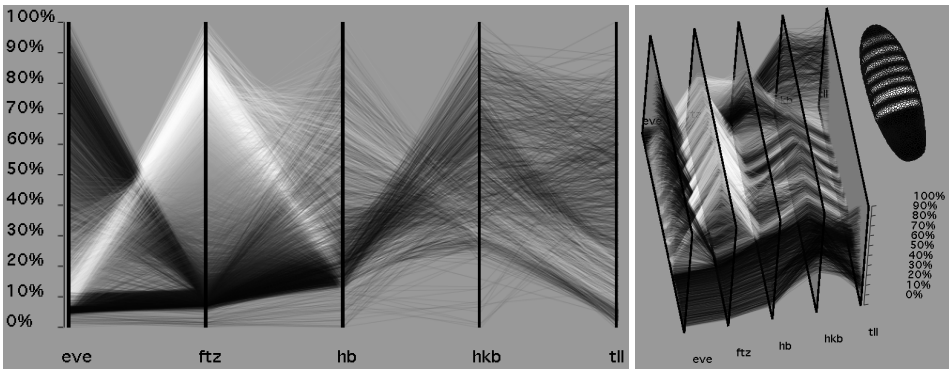
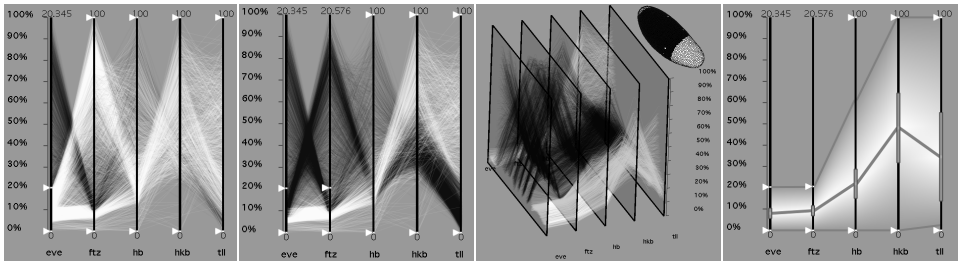


Figure 4: Expression level of four different genes visualized in 2D Parallel Coordinates (left) and 3D Parallel Coordinates (right).

allel coordinates have been developed, which have already been described in more detail in [OGS+06]. By varying color and transparency of data lines it is much easier to gain a fast overview and to detect important features and clusters in the data. Line trace highlighting and animation are additional tools which allow one to follow the course of single data lines through the graph. To allow for detailed analysis of brushes the dimensional zooming technique is used. Brushes define a subspace of the entire gene expression space. By scaling the selected ranges to the entire length of the parallel axis it is possible to analyse details in a user-defined subspace. Statistical properties of brushes such as selected minimum- and maximum values, average expression values, and standard deviations can be analyzed using brush bands (see Fig 5(d)).

Information about the spatial relationships between different genes' expression patterns is essential for the analysis of regulatory networks. Information about relative cell positions along the two main axes of the embryo, anterior/posterior (AP) and dorsal/ventral (DV), can be derived from the Unrolled View described in Section 4.2. To display this information in Parallel Coordinate Views, the coordinate axes have been extruded into the third dimension (Figure 4 (left)). Data lines are ordered from back-to-front according to cell positions along either the AP axis or the DV circumference. Along any given data line, the positional information is constant, such that data lines do not intersect each other in this third dimension. The 3D coordinate axes are drawn highly transparent with active z-buffering to prevent the addition of colors of overlapping parallel axes. This strategy guarantees a complete overview of the entire plot, with no details hidden. In addition, the outer frame of the coordinate axes are drawn with full opacity, which makes it easier to determine the position of the coordinate axis in 3D space.

By using this 3D visualization, spatial data dimensions are clearly separated from gene expression dimensions of the data, and the basic character of the spatial gene expression patterns in one dimension is preserved. For example, if data lines are sorted according to the position of cells along the AP axis, then the stripe patterns of genes like *eve* or *ftz* are visible in the plot (Figure 4 (left)). This 3D view also reveals what is not obvious



(a) (b) (c) (d)

Figure 5: Defining brushes in Parallel Coordinate View. In (a) a brush is defined to exclude all cells expressing *eve* at more than 20%. In (b) this brush is further refined to also exclude all cells expressing *ftz* at a value greater than 20%. (c) shows a 3D Parallel Coordinate View of the brush defined in (b) , where cell locations along the A/P axis of the embryo are shown in the third dimension. (d) Shows a broad color band display of the brush selected in (b), indicating the minimum, maximum, mean, and standard deviations for expression values for each gene.

in the 2D views shown in Figure 4: Cells expressing both *eve* and *ftz* at low levels are mainly at the anterior and posterior of the embryo, and a subset of these cells is found in the principal areas where *hkb* and *tll* are highly expressed. Even if tens of additional gene dimensions were added to the 3D view, these and doubtless other relationships could still be visualized.

Brushing can be executed in parallel coordinates using two sliders attached to each axis to define ranges in gene expression. In 3D parallel coordinates, two additional sliders are available to allow one to select cells also according to their relative AP- or DV position within the embryo. The physical views and the parallel coordinates are synchronized, i.e., if a brush is edited in one view, then the other view is also updated. If, for example, a brush is changed in parallel coordinates then the user can view in parallel how the spatial pattern the brush defines alters in any physical view. In Figure 5, an example for brushing in parallel coordinates is shown. One can see how additional relationships are revealed in 3D parallel coordinates where the brush splits of into two characteristic regions in the anterior and posterior of the embryo (see Figure 5(c)). Visualizing the brush just as broad color band reveals basic statistical properties of the brush (see Figure 5(d)).

6 Conclusions and Future Work

We have introduced PointCloudXplore and described a subset of its views and functionality. Dedicated physical views of the *Drosophila melanogaster* blastoderm (termed Embryo Views) make the comparison and analysis of spatial gene expression patterns possible. Expression Surfaces provide an effective and intuitive way for quantitative analysis of gene expression data. To support analysis of the relationships between genes directly in gene expression space, we have integrated parallel coordinates into the system (Parallel Coor-

dinate Views). Parallel Coordinate Views have been extended to a 3D rendering, making it possible to present spatial and gene expression data dimensions in one plot, while both dimension types are visually separated and basic spatial properties of gene expression patterns are preserved. All views are linked via brushing. PointCloudXplore makes interactive analysis of 3D expression data possible for the first time.

We plan to integrate automatic data analysis tools into PCX. Unsupervised clustering has been used previously to analyze microarray data and can also be applied to 3D gene expression data. Singular value decomposition (SVD) and other techniques have also been used to analyze gene expression and similar data. For analysis of 3D gene expression data, these techniques need to be modified, and new ones developed. Integration of such tools should further improve the utility of PCX.

7 Acknowledgments

This work was supported by the National Institutes of Health (grant GM70444), by the National Science Foundation (CAREER Award ACI 9624034), through the NSF Large Scientific and Software Data Set Visualization (LSSDSV) program under contract ACI 9982251, and an NSF large Information Technology Research (ITR) grant. Further support was provided by the LBNL Laboratory Directed Research Development (LDRD) program; We thank the members of the Visualization and Computer Graphics Research Group at the Institute for Data Analysis and Visualization (IDAV) at the University of California, Davis, the members of the BDTNP at the Lawrence Berkeley National Laboratory (LBNL), and the members of the Visualization Group at LBNL.

References

[BMMS91] Andreas Buja, John Alan McDonald, John Michalak, and Werner Stuetzle. Interactive data visualization using focusing and linking. In *IEEE Visualization 1991*, pages 156–163. IEEE Computer Society Press, 1991.

[dBvKOS00] Mark de Berg, Marc van Kreveld, Mark Overmars, and Otfried Schwarzkopf. *Computational Geometry*. Springer, second edition edition, February 2000.

[DGH03] Helmut Doleisch, Martin Gasser, and Helwig Hauser. Interactive feature specification for focus+context visualization of complex simulation data. In *Data Visualization 2003 (Proceedings of the EUROGRAPHICS - IEEE TCVG Symposium on Visualization 2003)*, pages 239–248. Eurographics Association, 2003.

[FDFH95] James D. Foley, Andries Van Dam, Steven K. Feiner, and John F. Hughes. *Computer Graphics: Principle and Practice*. Addison-Wesley, second edition in c edition, July 1995.

[FLHK+05] Charless Fowlkes, Cristian L. Luengo Hendriks, Soile Vanamo Elisabeth Keränen, Mark D. Biggin, David W. Knowles, Damira Sudar, and Jitendra Malik. Registering Drosophila Embryos at Cellular Resolution to Build a Quantitative 3D Map of Gene

Expression Patterns and Morphology. In *CSB 2005 Workshop on BioImage Data Minning and Informatics*, August 2005.

[FWR99] Ying-Huey Fua, Matthew O. Ward, and Elke A. Rundensteiner. Hierarchical parallel coordinates for exploration of large datasets. In *IEEE Visualization 1999*, pages 43–50. IEEE Computer Society Press, 1999.

[GRW+00] D. L. Gresh, B. E. Rogowitz, R. L. Winslow, D. F. Scollan, and C. K. Yung. WEAVE: A system for visually linking 3-D and statistical visualizations, applied to cardiac simulation and measurement data. In *IEEE Visualization 2000*, pages 489–492. IEEE Computer Society Press, 2000.

[HLD02] Helwig Hauser, Florian Ledermann, and Helmut Doleisch. Angular Brushing of Extended Parallel Coordinates. In *IEEE Symposium on Information Visualization (InfoVis'02)*, pages 127–130. IEEE Computer Society Press, 2002.

[Ins84] Alfred Inselberg. Parallel Coordinates for Multidimensional Displays. In *Spatial Information Technologies for Remote Sensing Today and Tomorrow, The Ninth William T. Pecora Memorial Remote Sensing Symposium*, pages 312–324. IEEE Computer Society Press, 1984.

[KSH04] Robert Kosara, Gerald N. Sahling, and Helwig Hauser. Linking Scientific and Information Visualization with Interactive 3D Scatterplots. In *Short Communication Papers Proceedings of the 12th International Conference in Central Europe on Computer Graphics, Visualization, and Computer Vision (WSCG)*, pages 133–140, 2004.

[KSO04] Ravikrishna Kolluri, Jonathan R. Shewchuk, and James F. O'Brien. Spectral Surface Reconstruction from Noisy Point Clouds. In *Symposium on Geometry Processing*, pages 11–21. ACM Press, July 2004.

[Law92] Peter A. Lawrence. *The Making of a Fly: The Genetics of Animal Design*. Blackwell Science, 1992.

[Nov04] Mateo Novotny. Visually Effective Information Visualization of Large Data. In *Proceedings of Central European Seminar on Computer Graphics (CESCG)*, 2004. Online at `http://www.cg.tuwien.ac.at/studentwork/CESCG/CESCG-2004/`.

[OGS+06] Ruebel O., Weber G.H., Keraenen S.V.E., Fowlkes C.C., Luengo Hendriks C.L., Simirenko L., Shah N.Y., Eisen M.B., Biggin M.D., Hagen H., Sudar J.D., Malik J., Knowles D.W., and Hamann B. PointCloudXplore Visual analysis of 3D gene expression data using physical views and parallel coordinate. In *Sousa Santos, B., Ertl, T. and Joy, K.I., eds., Data Visualization 2006 (Proceedings of EuroVis 2006), Eurographics Association*, pages 203–210, 2006.

[PKH04] Harald Piringer, Robet Kosara, and Helwig Hauser. Interactive Focus+Context Visualization with Linked 2D/3D Scatterplots. In *Proceedings of the 2nd International Conference on Coordinated and Multiple Views in Exploratory Visualization (CMV 2004)*, July 2004.

[SL05] Angelike Stathopoulos and Michael Levine. Genomic Regulatory Networks and Animal Development. *Developmental Cell*, 9(4):449–462, 2005.

[Weg90] Edward J. Wegman. Hyperdimensional Data Analysis Using Parallel Coordinates. *Journal of the American Statistical Association*, 85(411):664–675, September 1990.

[WKS+03] Katrin Weigmann, Robert Klapper, Thomas Strasser, Christof Rickert, Gerd M. Technau, Herbert Jäckle, Wilfried Janning, and Christian Klämbt. FlyMove - a new way to look at development of Drosophila. *Trends in Genetics 19*, pages 310–311, September 2003. Online at `http://flymove.uni-muenster.de/`.

[WL97] Edward J. Wegman and Qiang Luo. High dimensional clustering using parallel coordinates and the grand tour. *Computing Science and Statistics*, 28:361–368, 1997.