Multidimensional Visualization and Clustering

Presentation for Visual Analytics
of Professor Klaus Mueller

Xiaotian (Tim) Yin
04-26-2007

Paper List

- HD-Eye: Visual Mining of High-Dimensional Data
- Value and Relation Display for Interactive Exploration of High Dimensional Datasets
- PointCloudXplore: Visual Analysis of 3D Gene Expression Data Using Physical Views and Parallel Coordinates

HD-Eye: Visual Mining of High-Dimensional Data

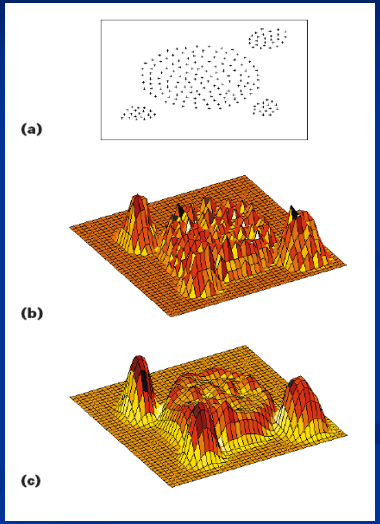
Alexander Hinneburg, Daniel A. Keim, and Markus Wawryniuk
University of Halle, Germany

Introduction

- Motivation:
  - Efficient clustering for high-dimensional data under noise
- Key:
  - Interactive !
  - advanced clustering algorithm + new visualization methods → interactive clustering tool

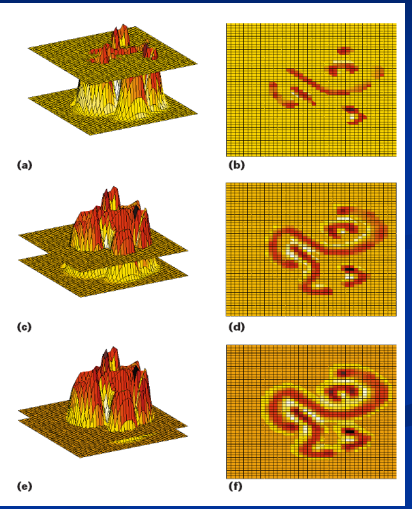
What is a cluster?

- Local maxima of density function



What is a cluster?

- Uni-centered vs multi-centered



How to cluster?

- Idea
  - First, project to lower dimension
  - Second, build separators and multidimensional grid in projected space

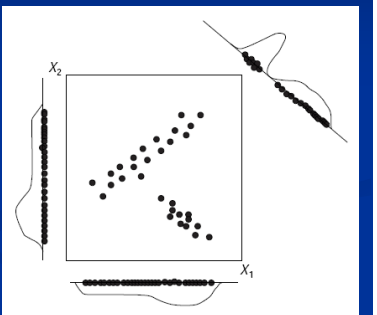
  Lemma:

  separation in contracting projection (w/ small error) →

  separation in original data space (w/o large error)
- Hierarchically
  - Independent choice of projections and separators in different nodes of the hierarchy
- Difficulty:
  - Choosing good projections
  - Choosing good separators
  - ***Need visual assistance!***

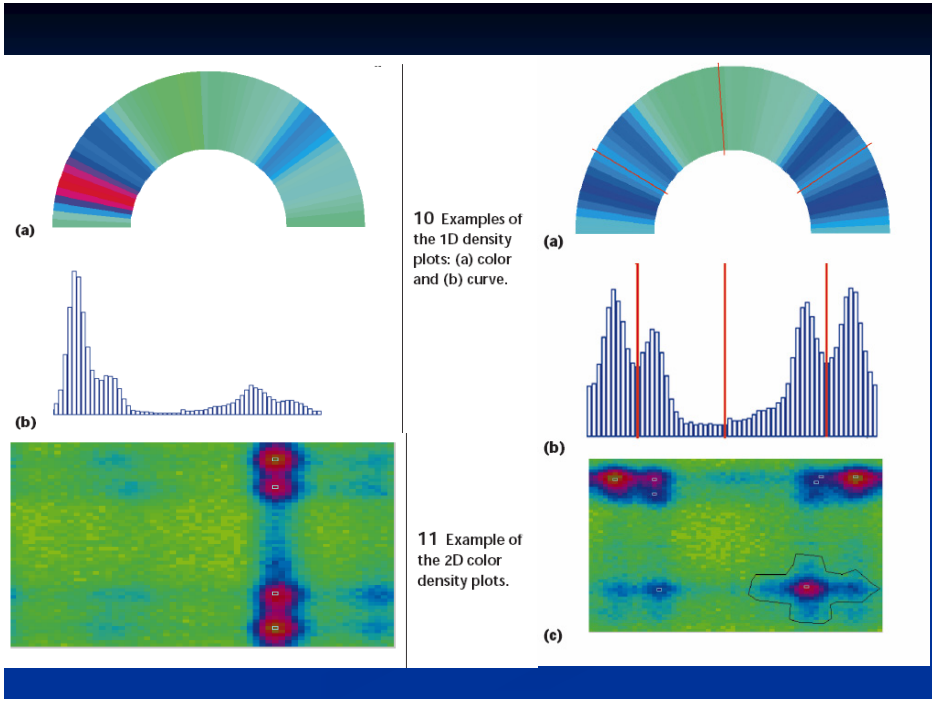
Visual Finding of Projections and Separators

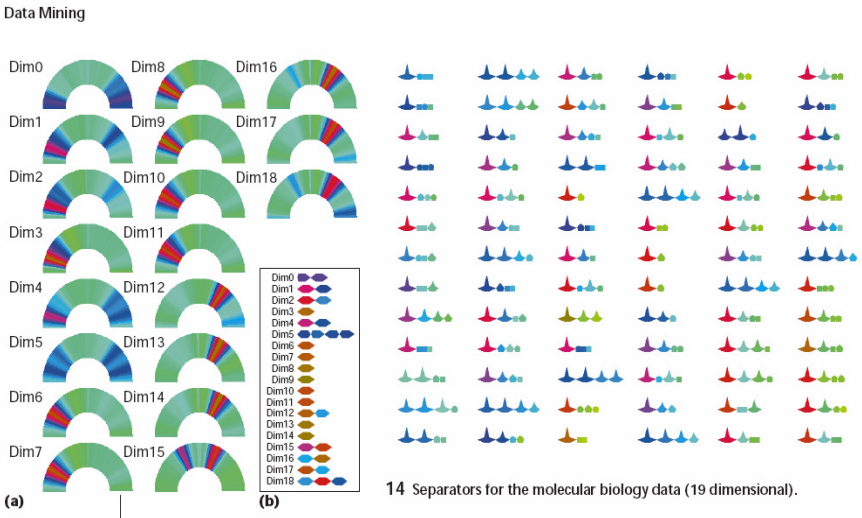
- Axes-parallel projection
  - only good for detecting center-defined clusters with no linear dependencies between the attributes.
- Good projections in general
  - Should contain well-separated clusters.
  - Do not require (like PCA) one projection separates all.



Visual Finding of Projections and Separators

- Finding projection
  - Initially, the HD-Eye system proposes some projections, like axes parallel projections, diagonal projections and etc.
  - Later on, the user can select the interesting projections, or generate other combinations from the selected ones.
- Finding separators
  - Put separators through low-density region
- After projection, how to visualize the clustering effect ?
  - abstract iconic display
  - color-based point density
  - curve-based point density

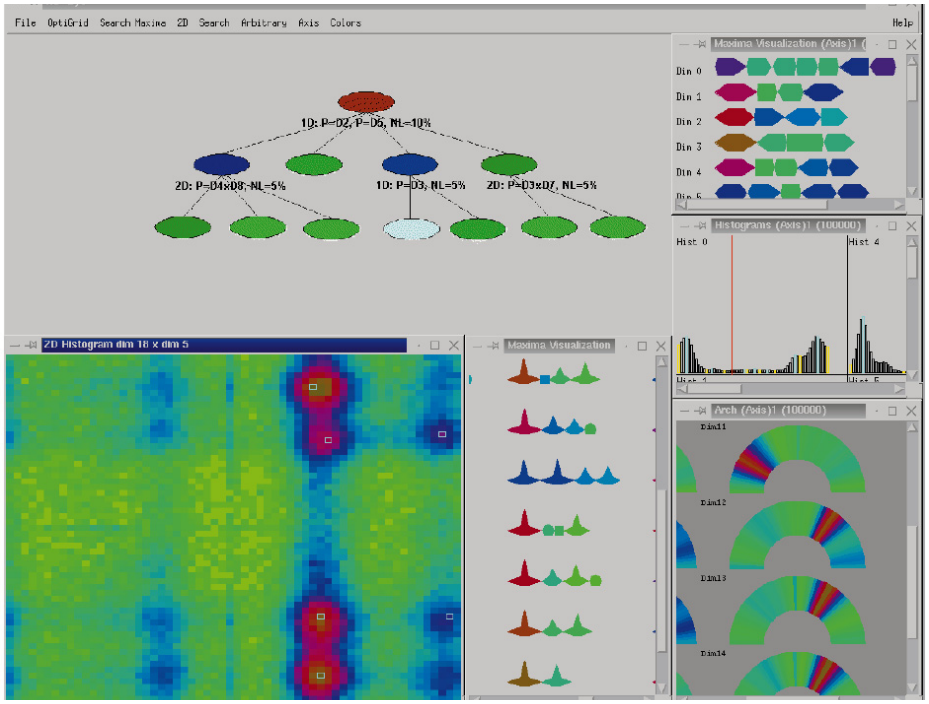




14 Separators for the molecular biology data (19 dimensional).

Overall Pipeline

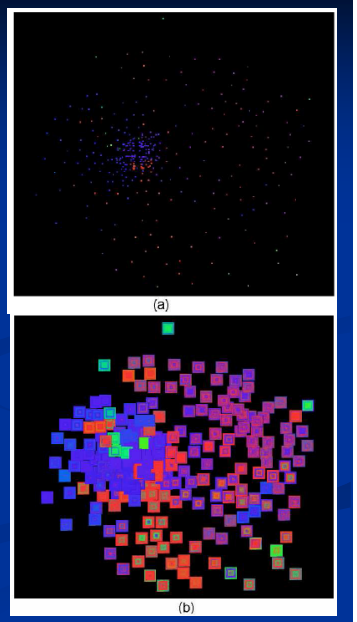
- 1. Initialize the cluster hierarchy with *root* = *D*
  - *Note:*
    - *node in hierarchy* ←→ *a region in the original space*
    - *Going down the hierarchy, regions are subdivided iteratively*
- 2. While a node *v* with the data set *Dv* can be split
  - Visually find projections $P = \{ P1, \ldots, Pk \}$
  - Visually find separators $H = \{ H1, \ldots, Hr \}$
  - Partition the region (associated with node *v*) into a multidimensional grid *G*, and insert data points of *Dv* into *G*
  - pick highly populated grid cells in *G* (i.e. containing clusters), add the cells as child nodes of *v* in the hierarchy



Value and Relation Display for Interactive Exploration of High Dimensional Datasets

Jing Yang, Anilkumar Patro, Shiping Huang, Nishant Mehta,
Matthew O. Ward and Elke A. Rundensteiner
Computer Science Department
Worcester Polytechnic Institute

- VaR
  - A new multi-dimensional visualization technique
- At high level:
  - map dimensions to a 2D space
- At low level:
  - map data value within a single dimension into a "glyph" (subwindow)
- Interactive tools:
  - For navigation:
  - For selection:



Glyph Positioning

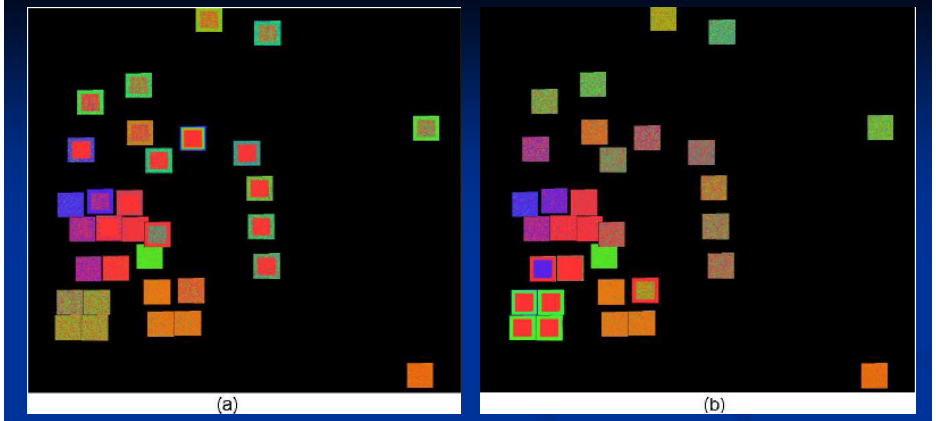
- Where to put glyphs ?
  - Step1: Build a *distance matrix* that captures the correlation between each pair of dimensions in the dataset.
  - Step2: Apply MDS on the distance matrix to get a set of positions in a 2D space, where each position corresponds to a dimension.
- How to build the "distance matrix" ?
  - Determined by certain *correlation measures*
  - a good *correlation measures* should keep most (least) related dimensions close to (far from) each other;
  - That is, the *distance matrix* should have maximum variance among all non-diagonal elements.

Glyph Positioning

- A de facto problem …
  - Two dimensions might be closely related only in part of the data items.
  - Need build histogram of the data value difference s between each pair of dimensions,
  - and partition each histogram into "bins".

Pixel Arrangement in Glyphs

- Idea
  - Use 2D texture to reveal data patterns in each dimension
- VaR uses "spiral arrangement"
  - Map each *data item* to a *pixel*
  - *Order* the data items by their values in a certain *base dimension*
  - place the ordered pixels from the center to the outside of a square *spirally*

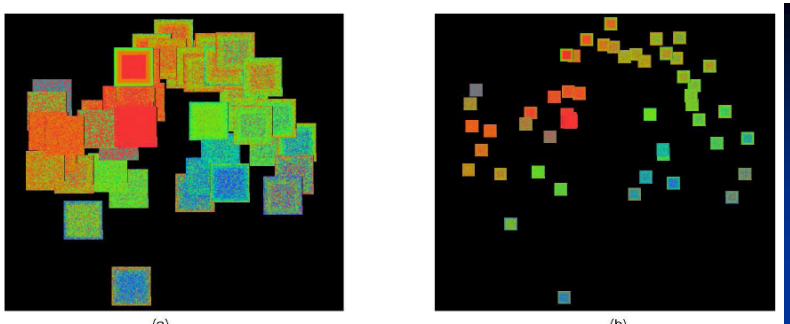


- Different choice of base dimension → different information conveyed by VaR
  - (a): dimensions in the top are closely related to base dimension a.
  - (b): dimensions in the bottom left are closely related to base dimension b.

Interactive Tools

- Navigation Tools
  - Goal
    - To solve the "overlapped glyphs" problem
  - Operations:

    Showing name, layer reordering, manual relocation, extent scaling, dynamic masking, automatic shifting, distortion, zooming and panning, manual pixel reordering, comparing, refining …
  - Example …
    - Oringinal, extent scaling, automatic shifting, distortion



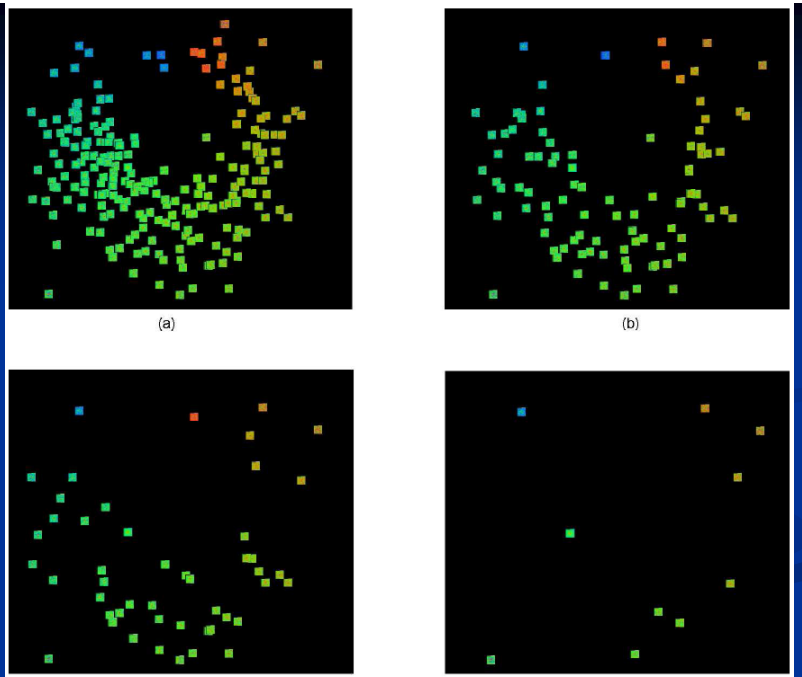
Interactive Tools

- Selection Tools
  - Goal
    - Allow users to select dimensions of interests for further exploration.
    - i.e. dimension reduction
  - Operations:

    Automatic selection of related dimensions

    Automatic selection of separated dimensions

    Manual selection of arbitrary dimensions
  - Example …
    - Automatic selection of separated dimensions with increasing correlation thresholds



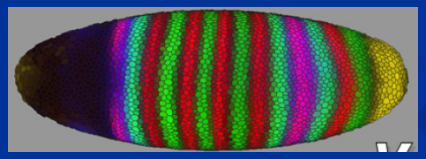
PointCloudXplore: Visual Analysis of 3D Gene Expression Data Using Physical Views and Parallel Coordinates

O. Rübel1, G.H. Weber2, S.V.E. Keränen3, C.C. Fowlkes4, C.L. Luengo Hendriks3, Lisa Simirenko3, N.Y. Shah2,
M.B. Eisen3, M.D. Biggin3, H. Hagen1, D. Sudar3, J. Malik4, D.W. Knowles3 and B. Hamann2

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

Motivation

- Objects under Study
  - An *embryo surface* has multiple *cells*
  - Each cell has n *genes* (each of which has a value of *gene expression level*)



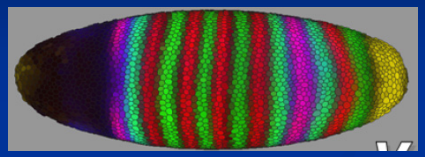
  - *Therefore each cell is assigned with a n-tuple (l1, l2, … ln);*
  - *The set of n-tuples from all cells constitute a gene expressions pattern for the whole embryo surface*

Motivation

- Goal:
  - Visualizing the gene expression pattern for a given embryo surface
  - That is, visualizing the set of (l1, l2, …, ln) values of all the cells on the embryo surface.
- Approaches:
  - Visualize in the *physical space*
  - Visualize in the *gene expression space*
  - Build *link* between these two spaces

Physical Space

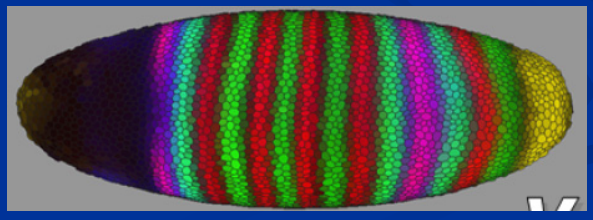
- Physical Space
  - Data space:
    - 3D space, one dim per gene
  - Data point:
    - A 3D position
    - with coordinate (x, y, z)
  - Sample point:
    - a cell (or its nucleus) on the embryo surface
    - whose coordinate (x, y, z) is the physical position of the cell (or of its nucleus)



- Visualization Techniques
  - Color encoding
  - Multi-views (orthographic or unrolled)
  - Expression surfaces

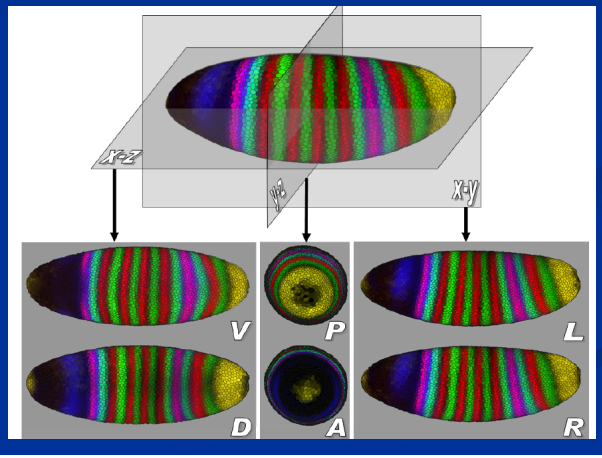
Representing expression with color

- Scheme:
  - One color for each gene,
  - expression values mapped linearly to brightness.
- Example:
  - expression patterns of four genes
  - eve (red), ftz (green), gt (blue), fkh (yellow).



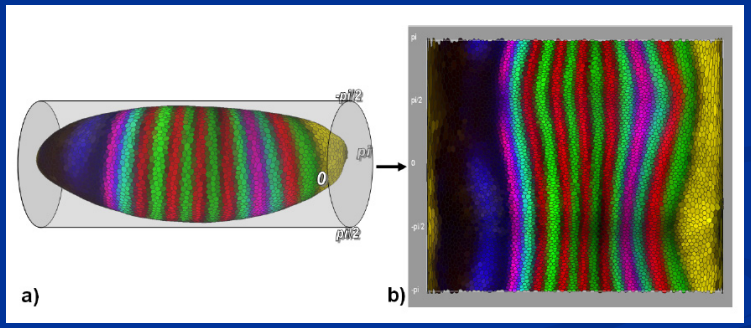
Orthographic views

- 3D → {2D}



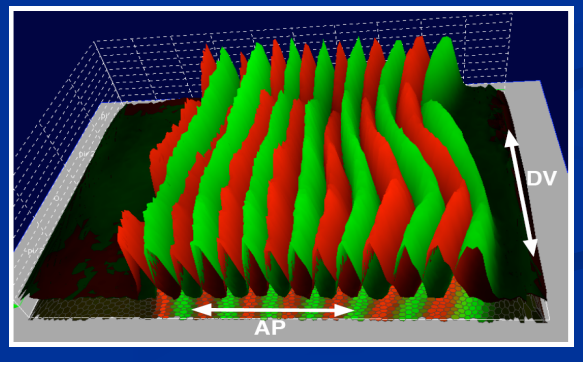
Unrolled Views

- 3D → 2D



Expression Surfaces

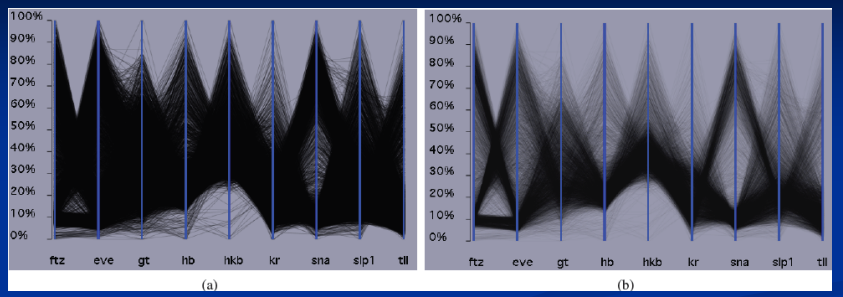
- One surface per gene
  - XY: 2D view (either orthographic or unrolled)
  - Z: value of gene expression level



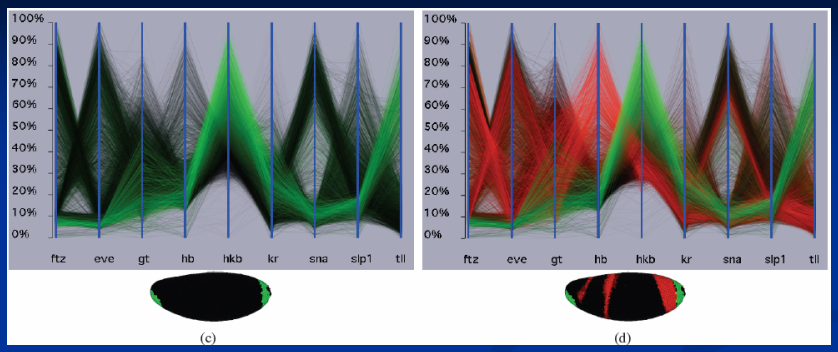
- Physical Space
  - Due to overlap between information from different genes
  - Only allow a few number of genes whose expression can be displayed at the same time
- Gene Expression Space
  - can show relationships between many genes' expression.

Gene Expression Space

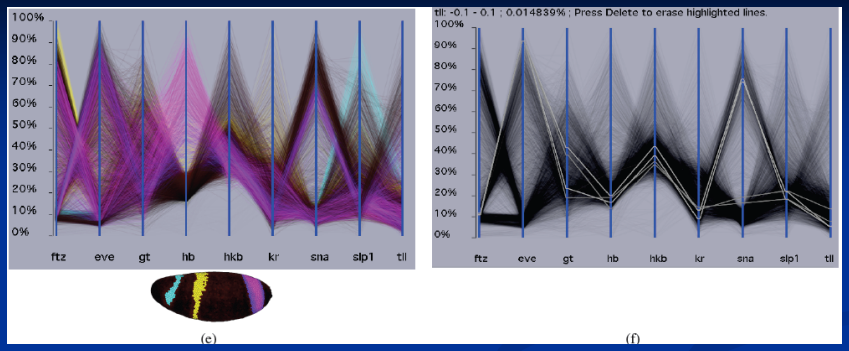
- Gene Expression Space
  - Data space:
    - n-dimesional space, one dim per gene
  - Data point:
    - a combination of expression levels for the set of genes,
    - with coordinate (l1, l2, … ln)
  - Sample point:
    - a cell from the embryo surface,
    - whose coordinate (l1, l2, … ln) is the gene expression levels for this cell
- Visualization Techniques
  - Parallel coordinates



- 2D Parallel Coordinate View for nine genes.
- (a) A view in which data lines have maximum opacity.
- (b) A view in which data line transparancy has been increased.

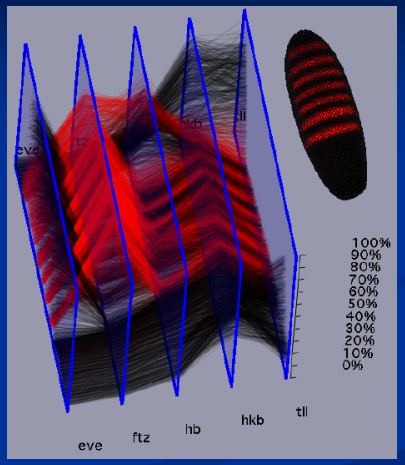


- (c) A view in which color from an Embryo View showing hkb expression (green) is shown.
- (d) A view in which colors from an Embryo view showing both hkb (green) and hb (orange)expression are shown.



- (e) A view showing two brushes drawn on an Embryo View to highlight distinguish between the three hb stripes of expression (light blue, yellow and pink).
- (f) A view showing line traces that highlight data associated with several cells.

- Coupling two spaces !!
- 3D parallel coordinates
  - Paths in red: cells with high expression level of ftz gene.
  - Stretching the 2D parallel coordinates along "anterior" → "posterior" direction



Conclusion

- *HD-Eye: Visual Mining of High-Dimensional Data*
  - Hierarchical clustering techniques with visual assistance
  - Project to low dim space, (visually choosing projections)
  - And partition there (visually choosing separators)
- *Value and Relation Display for Interactive Exploration of High Dimensional Datasets*
  - A new multi-dimensional visualization technique
  - At high level: capture inter-dimension correlation in a 2D space
  - At low level: capture data pattern in each dimension by a "glyph"
  - Interactive tools for navigation and selection
- *PointCloudXplore: Visual Analysis of 3D Gene Expression Data*
  - Application driven
  - Employing physical views and parallel coordinates