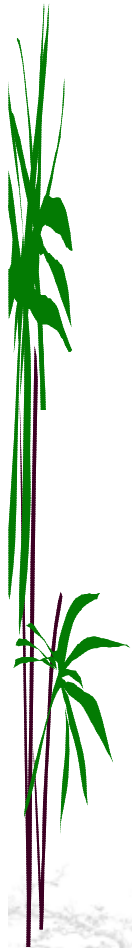




Towards Interactive Exploration of Gene Expression Patterns

Author: Daxin Jiang, Jian Pei, Aidong Zhang

Presented by: Lu-An Tang





Outline

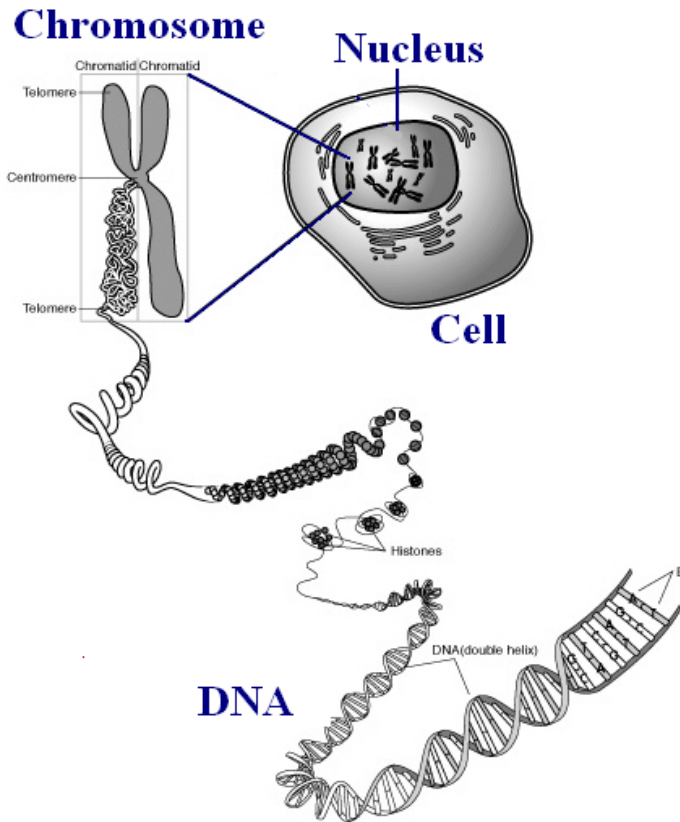


- Introduction and Backgrounds
- Problem of Microarray Data Analysis
- The Attraction Tree Approach
- Index the Attraction Trees
- Experiment Evaluation
- Related Studies and Future works





Basic Concepts



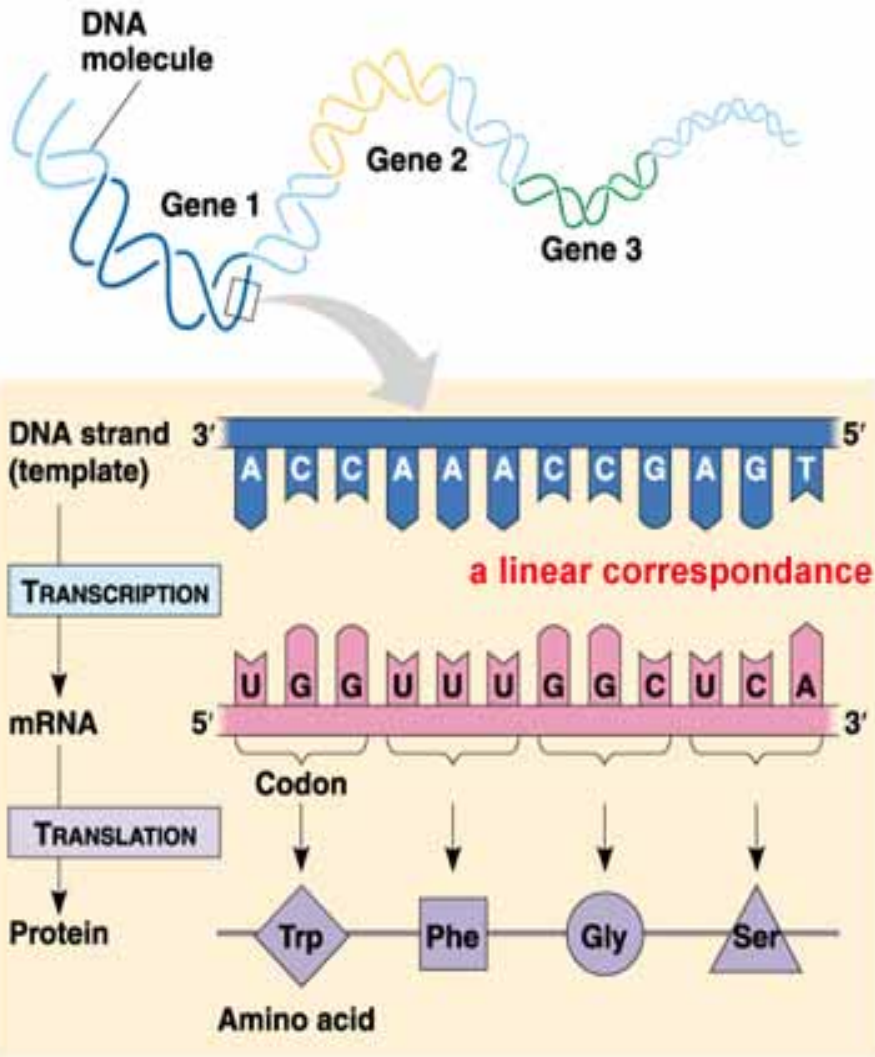
● **Chromosome:** an organized structure of DNA that is found in cells

● **Gene:** a part of DNA with heredity information

● **Protein:** organic compounds forming the basis of life, the essential parts of cells



Basic Concepts cont.d



Gene Expressoin

- messenger RNA(mRNA) gets the heredity information from Gene, and transcribe it to the template for protein

Information Flow:

Gene/DNA → mRNA → Protein

Software Engineering:

Requirements → Model → Codes

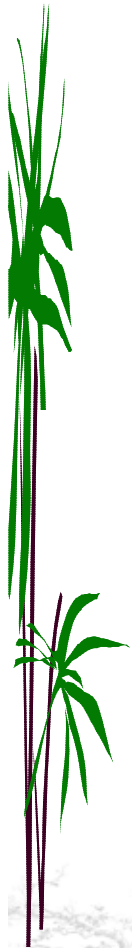
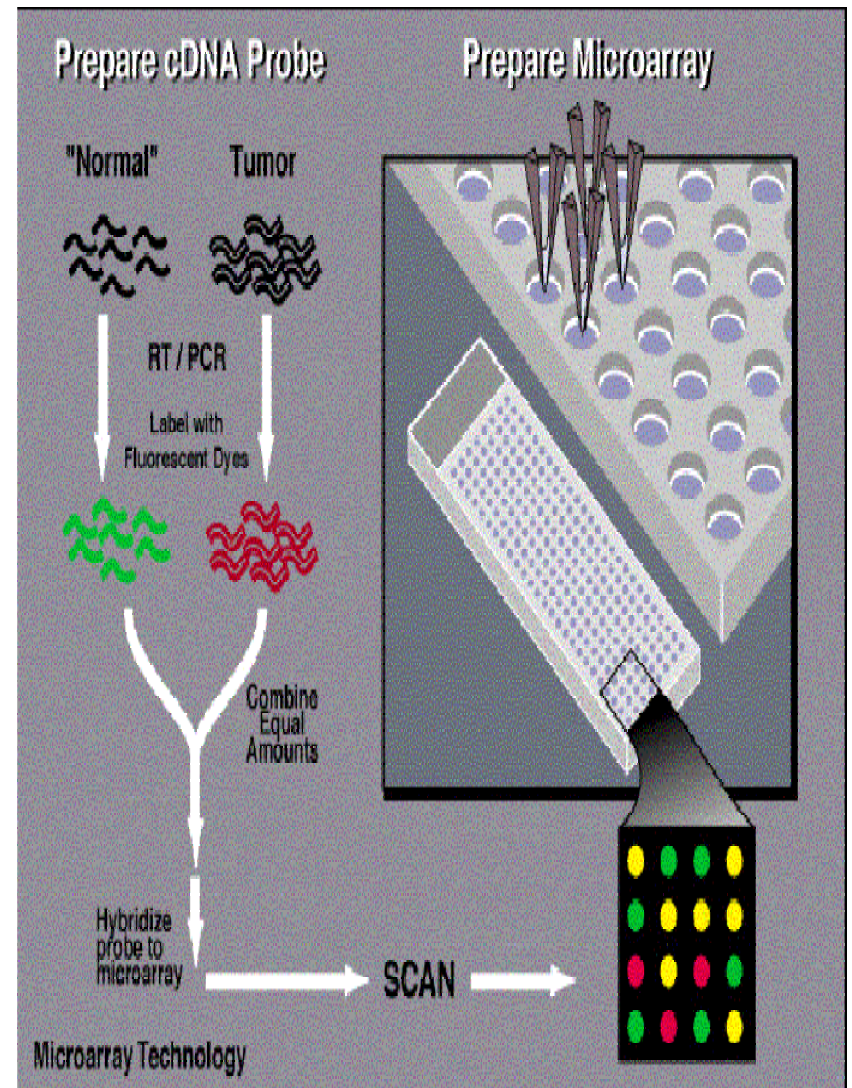
Key steps: Gene → mRNA



Gene Expression and Microarray Experiment



- Is gene expression only influenced by itself?
 - No, the gene expression is related to the environment (e.g., enzymes) and other genes
- Microarray: Find the issues that influence gene expression by large scale experiment
 - Complete thousands of experiment in one time





History of Microarray Technology



- 1990 The Microarray Chip technology is proposed
- 1996 First paper using microarray technique for tumor analysis
- 1997 Y.Chen et al propose the techniques to read the expression value by fluorescence – the experiments are widely carried out and huge amount of data are collected
- 1997—2000, main researchers are biologists
- Most influence paper: Alizadeh et al: Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling **Nature 2000**
- 2001, the data mining researchers start this topic

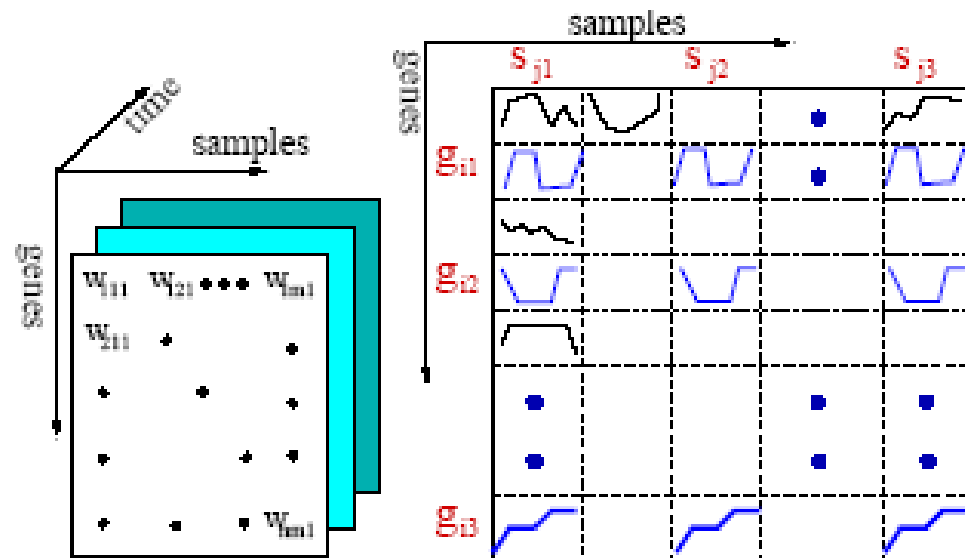
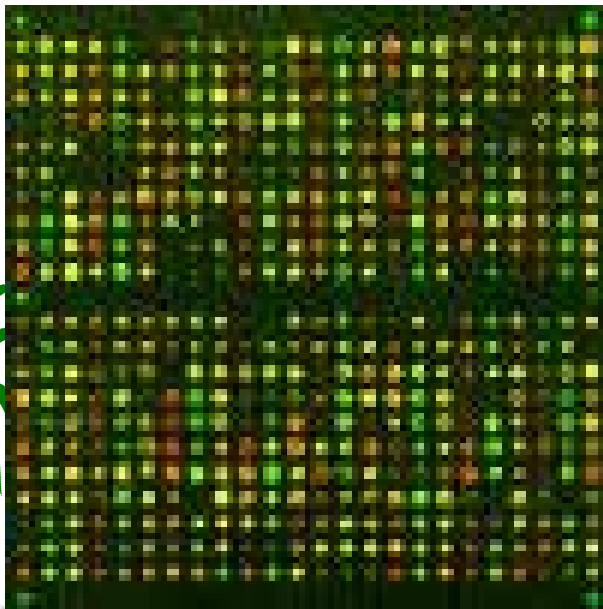




Features of Micro-Array Data



- Each row is a gene, totally 5000—50,000 rows
- Each column is a sample (e.g., environment, enzyme, etc, very expensive, usually 10-20)
- Time: An experiment usually takes about 24-72 hours, record each 0.5-3 hours
- A matrix of time series in high dimension (gene as dimension)





Outline



- Introduction and Backgrounds
- **Problem of Microarray Data Analysis**
- The Attraction Tree Approach
- Index the Attraction Trees
- Experiment Evaluation
- Related Studies and Future works

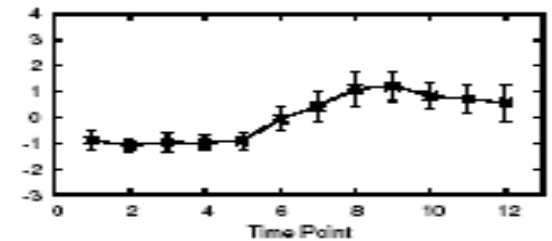
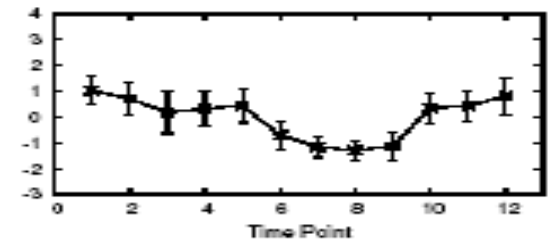
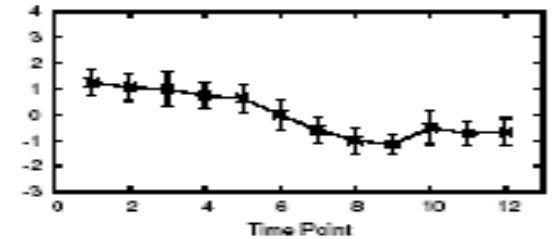




Problem Description



- Only focus on “Gene - Time”
- Mining Correlations among Genes – most useful
 - Which genes’ expression are positively correlated
 - Which of those are negatively correlated
- Major Solution: Clustering
 - Find out the co-expressed genes (similar values and trends of change)
 - Generalize **Coherent Gene Expression Pattern**

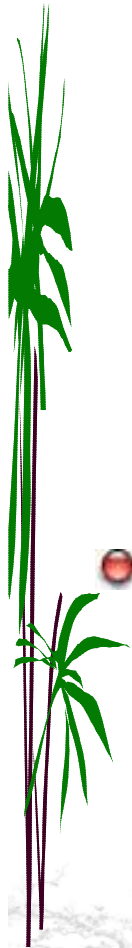




K Means Clustering



- Each gene expression is a vector $g (v_1, v_2, \dots, v_n)$
- K-means clustering proceeds by repeated application of a three-step process where:
 - 1) the mean vector for all items in each cluster is computed
 - 2) items are reassigned to the cluster whose center is closest to the item
 - 3) repeat
- The parameters controlling k-means clustering are the number of clusters (K)

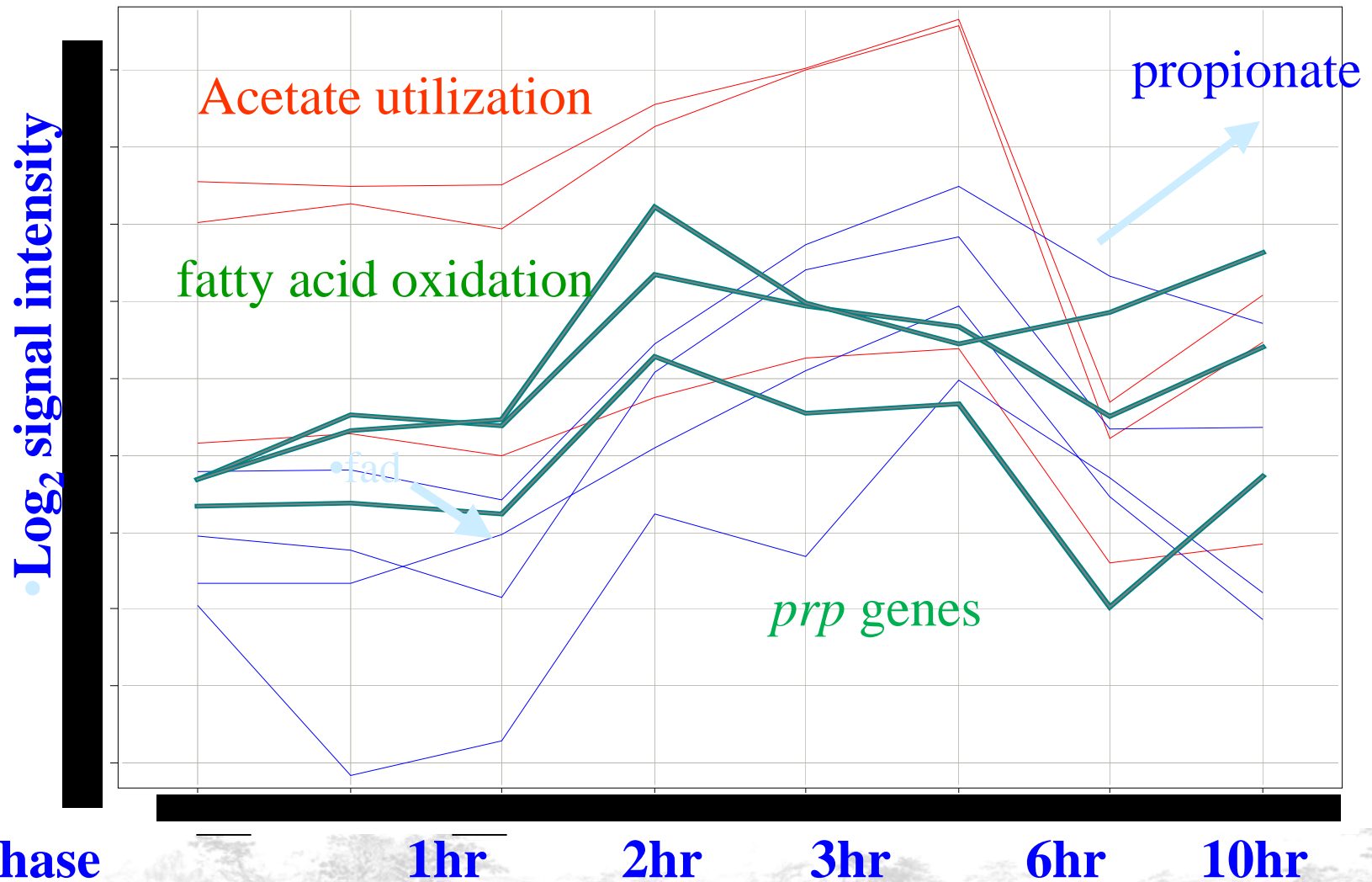




Clustering Gene Expression Data



Profile Chart



log phase

2010-3-19

From Jeremy Glasner slides

DAIS UIUC

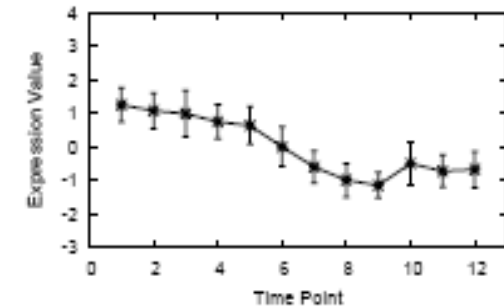
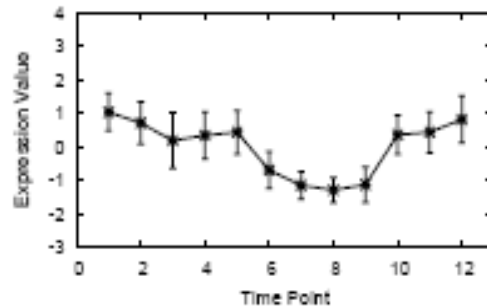
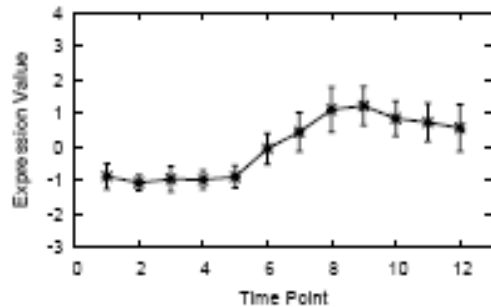
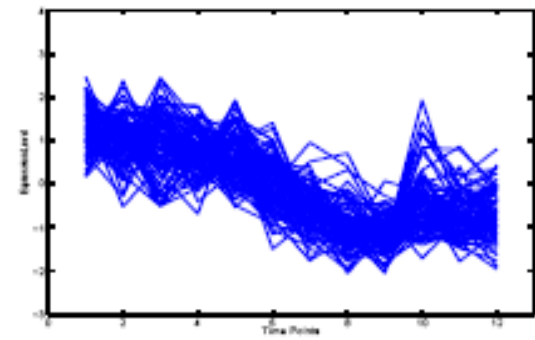
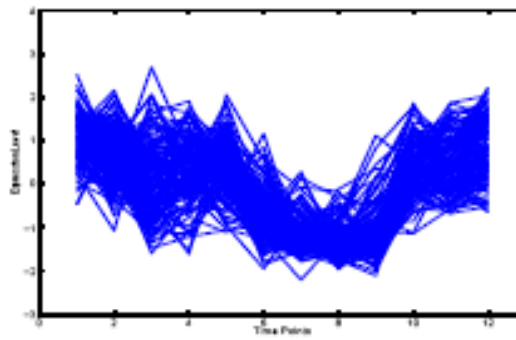
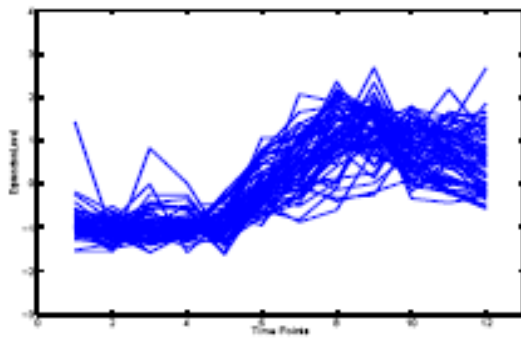
11



Problem of Clustering Algorithm



- Generate results in **single** level
- Do not utilize domain knowledge – some genes are already know to be related





Challenges I



- Main Problem: The biologists and computer scientists have different interests
 - Computer scientists: how efficient the algorithm is? What is the time and space complexity of the algorithm? Can it discover all the targets? What is the precision? Can it run in parallel manner? ... -- Cares more on the **process**
 - Biologists: What are the co-expressed genes? – Cares only one **results**
- **Problem:** the fancy algorithms by computer scientists cannot satisfy the biologists





Challenges II

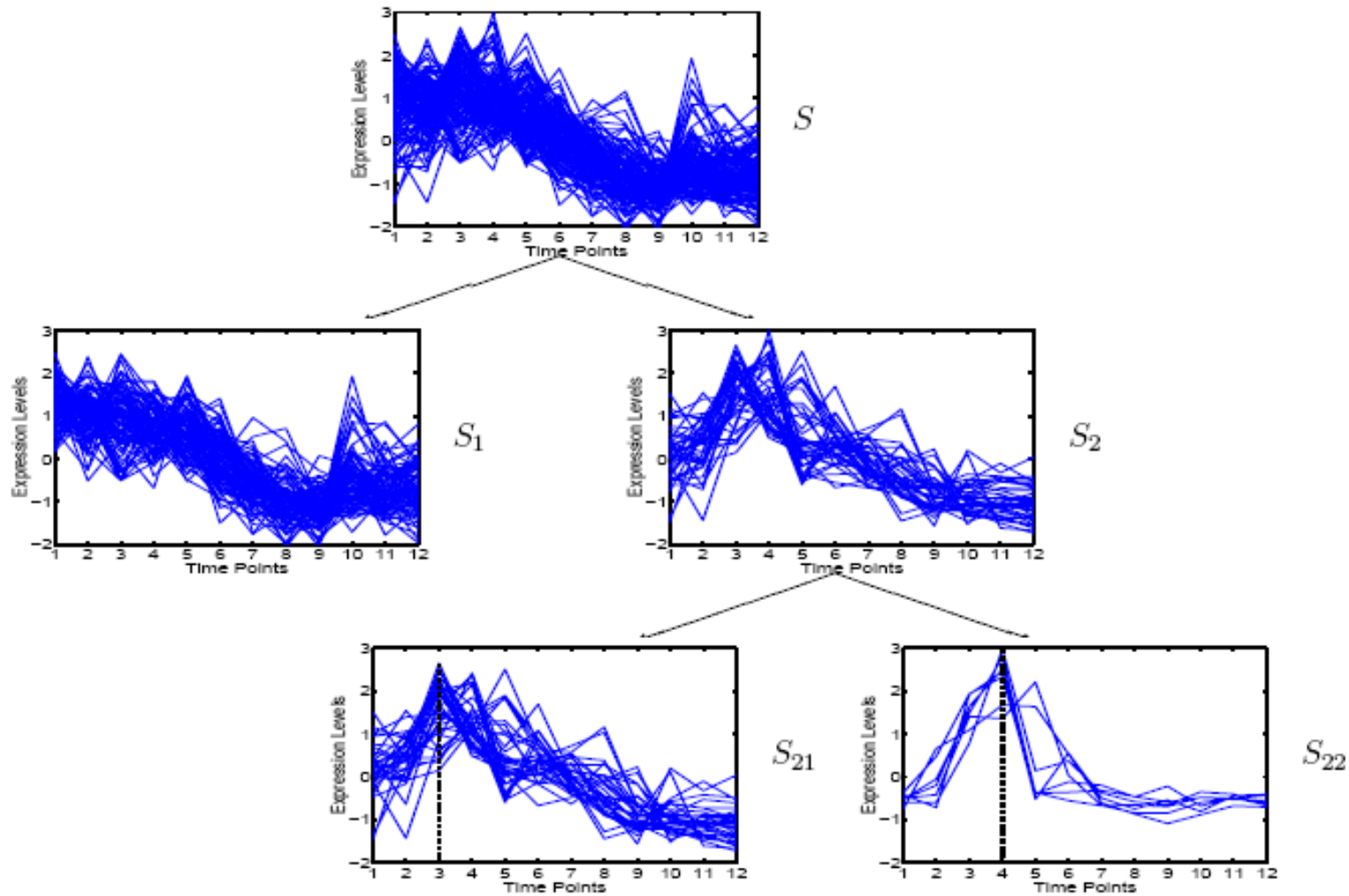


- The **granularity** and **interesting point** are determined by the biologists/ medical researchers, but they could not provide such thing in advance – actually they **do not know** what to mine, but want to **judge** the results
- Unsupervised clustering: Too many and too complex result – usually they are not interested in
- Problem: no interaction with biologists in the process
- **Not Data Mining, but information visualization**



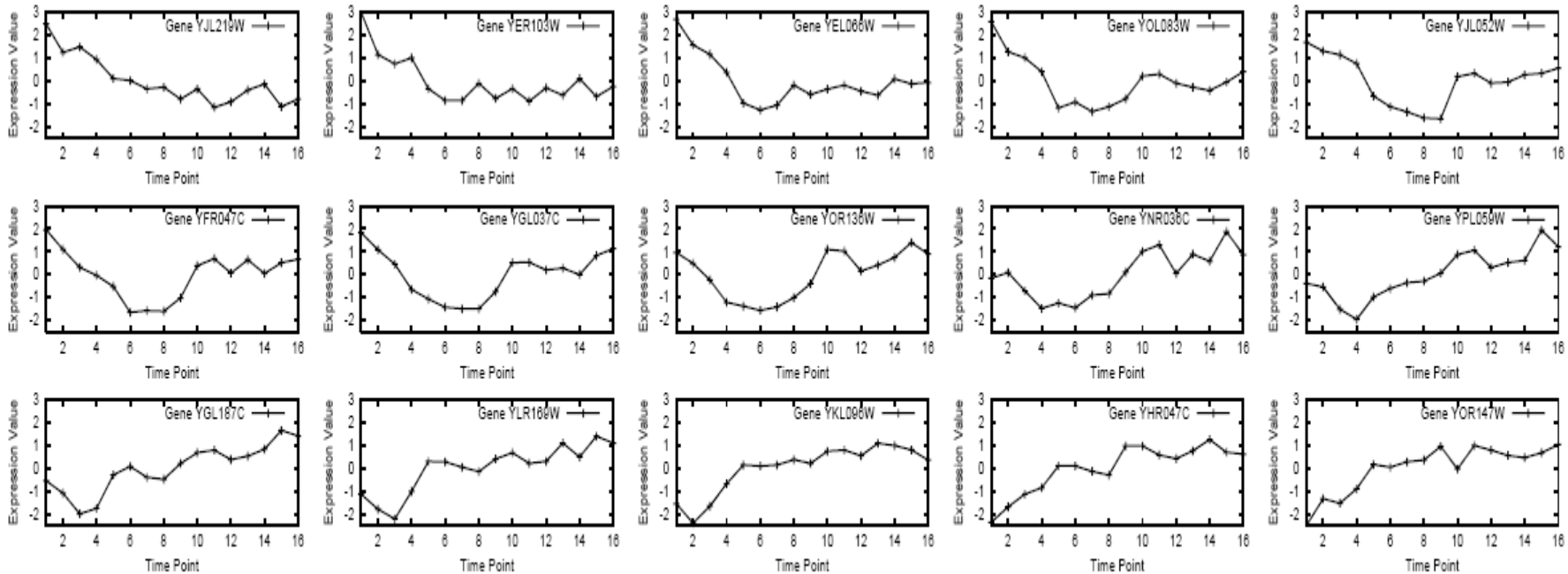
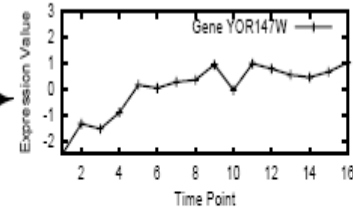
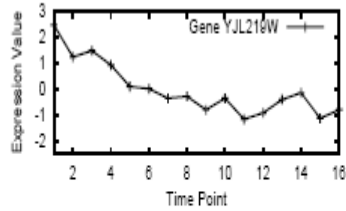


Observation 1: Hierarchy is more helpful





Observation 2: Many expressions are connected

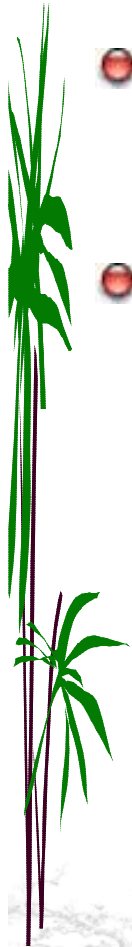




GeneX Framework



- According to the distance of each gene expression vector, clustering the data to an **Attraction Tree**
- Traverse the tree, according to the trends of change, build on “**coherent gene index**”
- Define a parameter to reflect the **correlation** of gene expression, provide operations of **Roll-up** and **Drill-down**





Outline



- Introduction and Backgrounds
- Problem of Microarray Data Analysis
- **The Attraction Tree Approach**
- Index the Attraction Trees
- Experiment Evaluation
- Related Studies and Future works



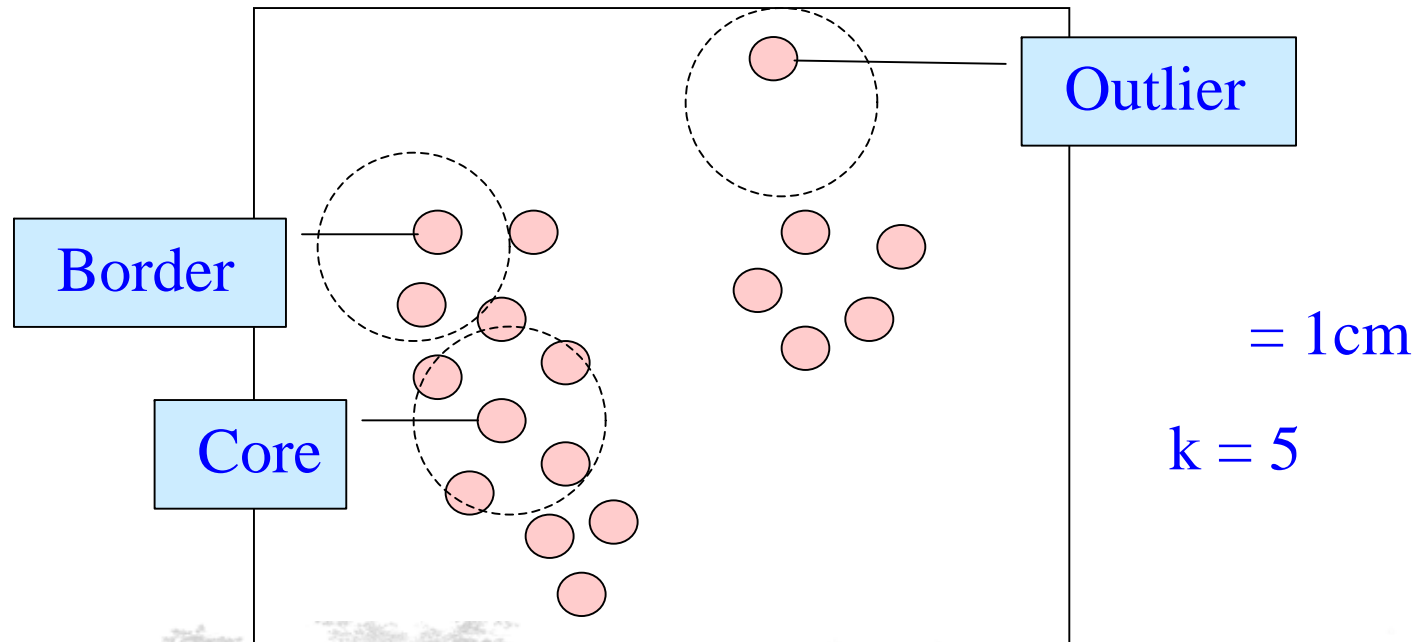


DBSCAN Algorithm



- Two parameter:

- Maximum radius of the neighbourhood
- k – Minimum number of points in an ϵ -neighbourhood of that point

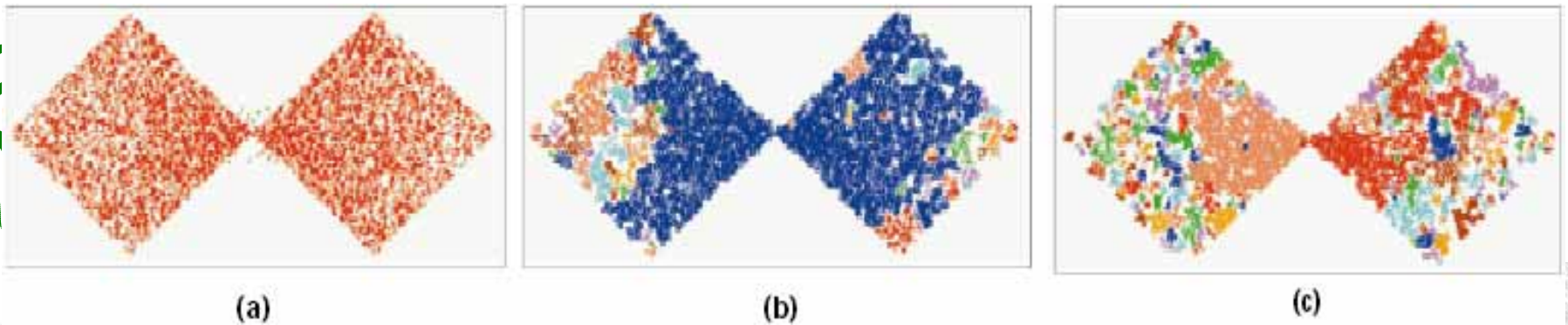
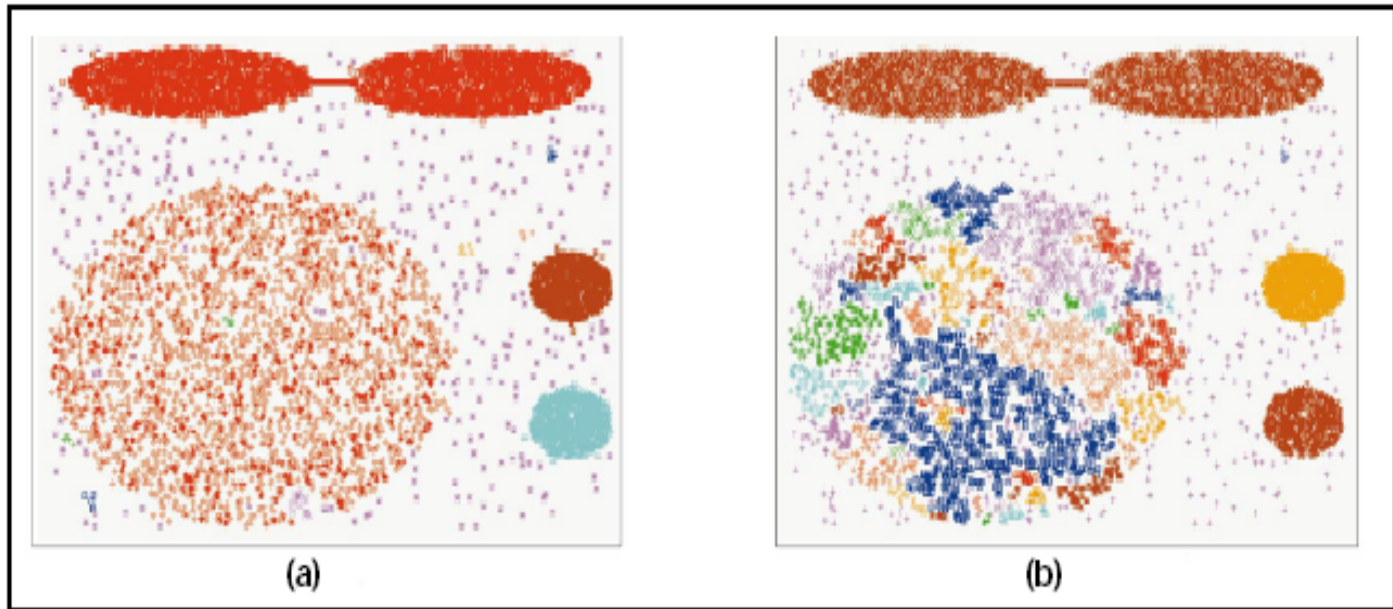




DBSCAN: Sensitive to ϵ and k



Figure 8. DBScan results for DS1 with MinPts at 4 and Eps at (a) 0.5 and (b) 0.4.





Denclue Clustering



- DENsity-based CLUstering(1998): Efficient clustering in noisy and high dimensional data
 - Influence Function: Gaussian Function
 - Similarity Measure: Pearson similarity
 - Distance Measure: Euclidean distance
- For two objects/gene O_i and O_j , the similarity measure

$$f(O_i, O_j) = e^{-\frac{\text{distance}(O_i, O_j)^2}{2\sigma^2}}$$

$$\text{density}(O) = \sum_{O_j \in \mathcal{D}, \text{similarity}(O, O_j) \geq \bar{S}} f(O, O_j)$$



Similarity Measure



- O_i and O_j are neighbors if $f(O_i, O_j) > S$
- The neighbors influence O_i , the influence $I(O_i)$ is determined by

$$I(O_i)^{(d)} = \sum_{O_j \in \mathcal{D}, \text{similarity}(O_i, O_j) \geq \bar{S}} \frac{1}{f(O_i, O_j)} O_j^{(d)} \quad (1 \leq d \leq m),$$

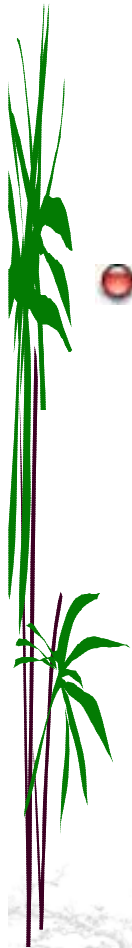
- where m is the number of attributes of object O_j , and O_j^d is the d -th attribute of O_j



Definition of Attraction



- Intuitively, $I(O_i)$ indicates the **dense region** in O_i 's neighborhood: If O_i moves toward the direction of, O_i is likely to reach an area with higher density
- If O_i has a higher density than all of its neighbors, O_i is a **maximum** object. Two cases:
 - O_i has no neighbors at all – it is a noise object
 - O_i has a higher density than other neighbors





Definition of Attraction II



- We call a gene/object O_i is “attracted” by its attractor O_j (denoted by $O_i \rightarrow O_j$) according to the following:

$$\text{Attractor}(O_i) = \begin{cases} O_i & \text{if } O_i = O_{max} \\ \arg \max_{O_1 \in A_1} \text{similarity}(O_1, O_i) & \text{if } O_i \text{ is a noise object} \\ \arg \max_{O_2 \in A_2} \text{similarity}(O_2, I(O_i)) & \text{if } O_i \text{ is not a local maximum} \\ \arg \max_{O_1 \in A_1} \text{similarity}(O_1, I(O_i)) & \text{otherwise.} \end{cases}$$

- A_1 is the set of local maximums O_1
- A_2 is the set of O_i 's neighbors such that O_2 has a higher density than O_i





Attraction Tree

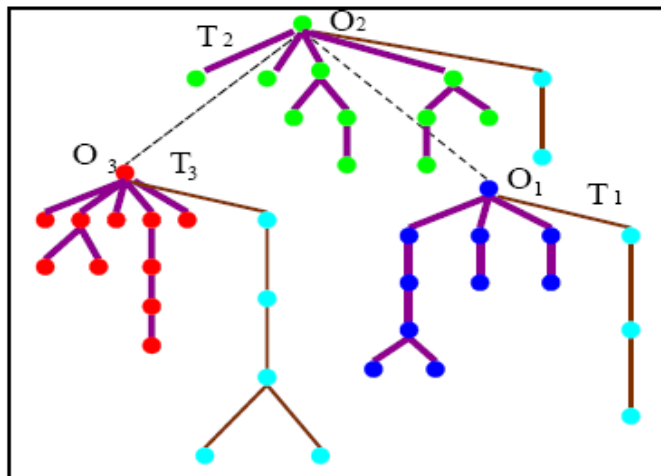
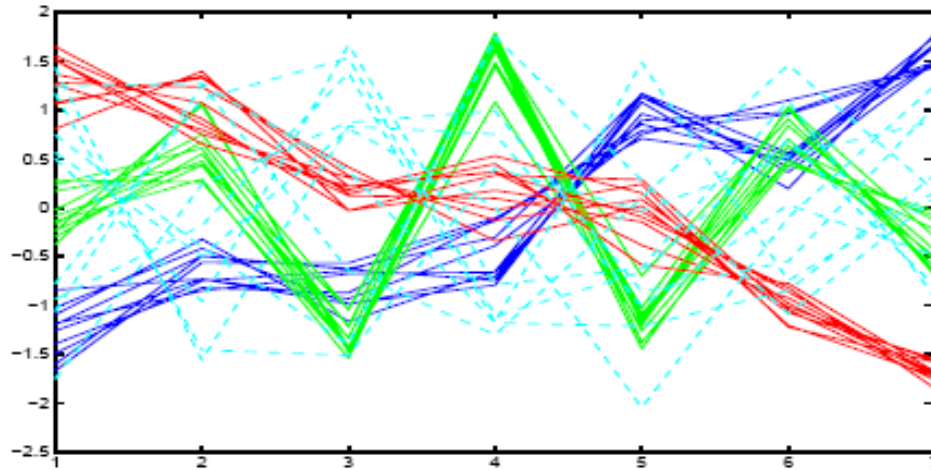


- $O_i \rightarrow O_j$: O_j is the **most dense object** in the kNN of O_i
- Two Extreme case: $k = 1$ (only similarity) and $k = n$ (only density)
- Each cluster generates an attraction tree: the root is the gene with the **highest density**, the first child nodes are those **attracted directly** by the root, etc...

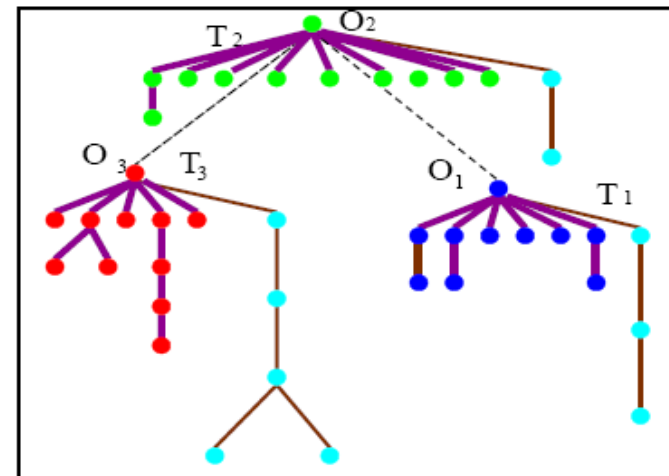
$$Parent(O_i) = \begin{cases} nil & \text{if } O_i = O_{max} \\ Attractor(O_i) & \text{otherwise.} \end{cases}$$



Attraction Tree Example



(a) $K = 1$



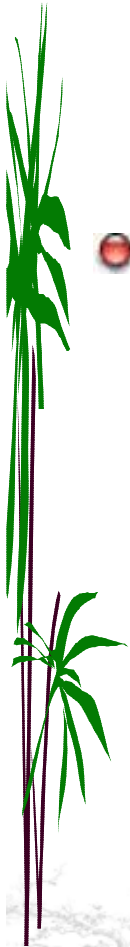
(b) $K = \infty$



Advantages of Attraction Tree



- The attraction tree is **self closed**
 - The gene with the similar expression values share a tree
 - No matter what k is, same groups of objects form the attraction trees
- Robust to the noises: the noises are in the different branch from common genes, could be easily pruned in most of the time





Outline



- Introduction and Backgrounds
- Problem of Microarray Data Analysis
- The Attraction Tree Approach
- **Index the Attraction Trees**
- Experiment Evaluation
- Related Studies and Future works

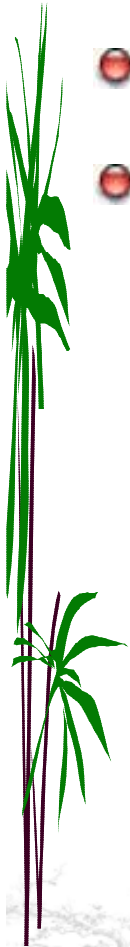




Exploring the gene expression data



- The attraction tree is a data structure that organizes the gene expression data by their similarities
- It provides the foundation to exploring the data
- But that is not enough – Just image in the scenario of libraries, only categorizing the books and put them in shelves is not enough – We have to build something help the users to explore/search/browse them --
indexes

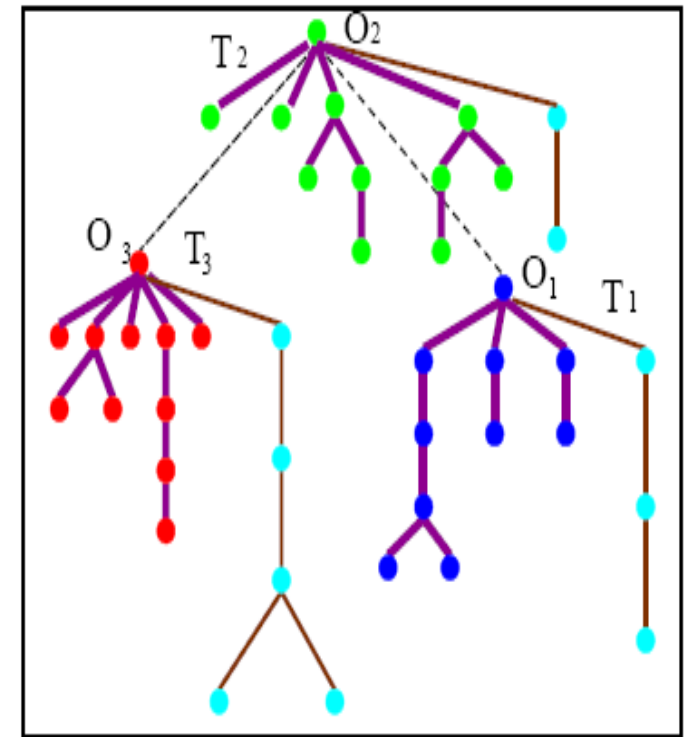




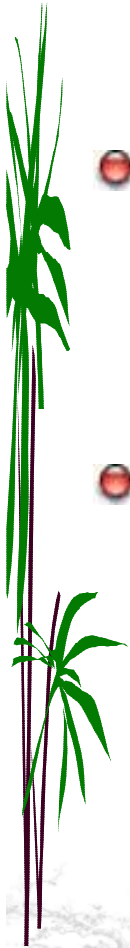
Edges of the Attraction Tree



- Co-expressed genes in a cluster have high similarities (thick lines)
- Noises and intermediate genes have moderate similarities (thin lines)
- Genes in different clusters have the most weak similarities (dashed lines)



(a) $K = 1$

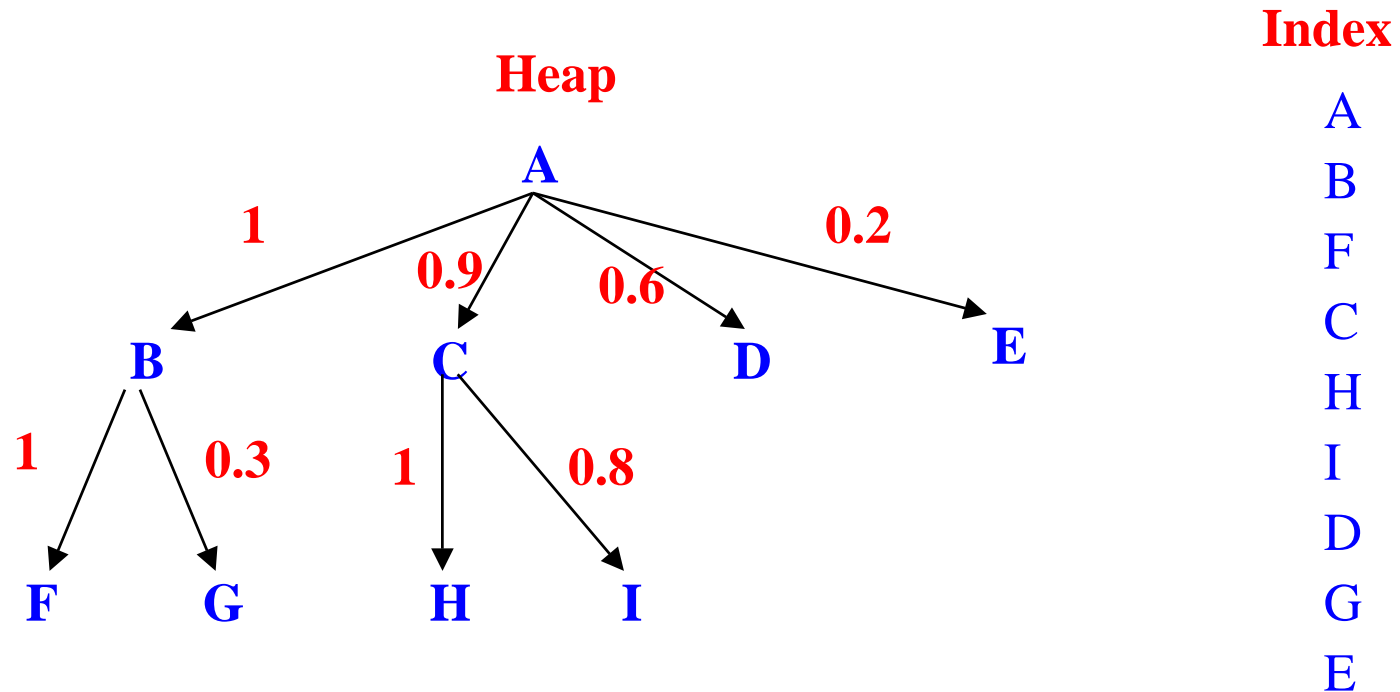




Build an 1-D Index

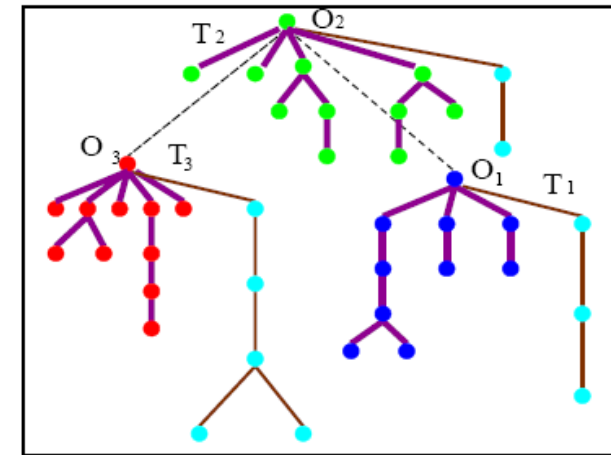
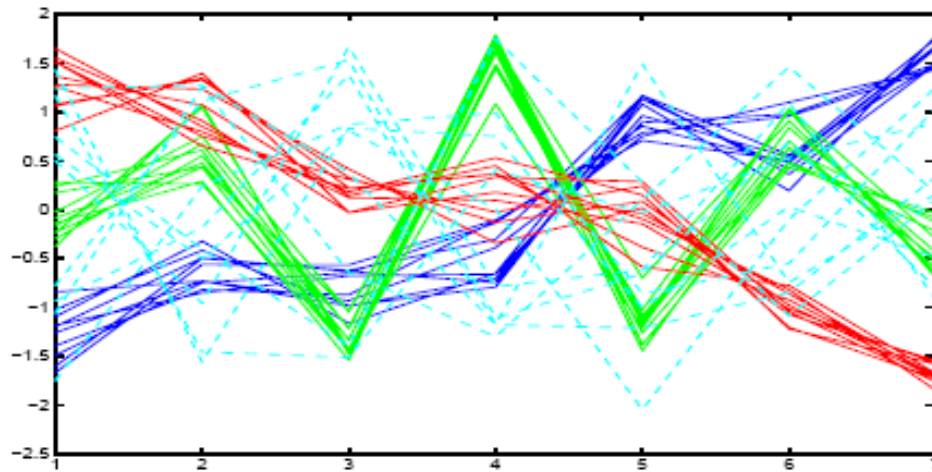


- The edge/link of attraction tree is the similarity measure, representing the change trends of parent gene to children gene
- A 1-D index is constructed by a max heap

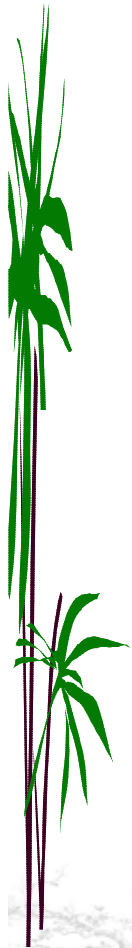
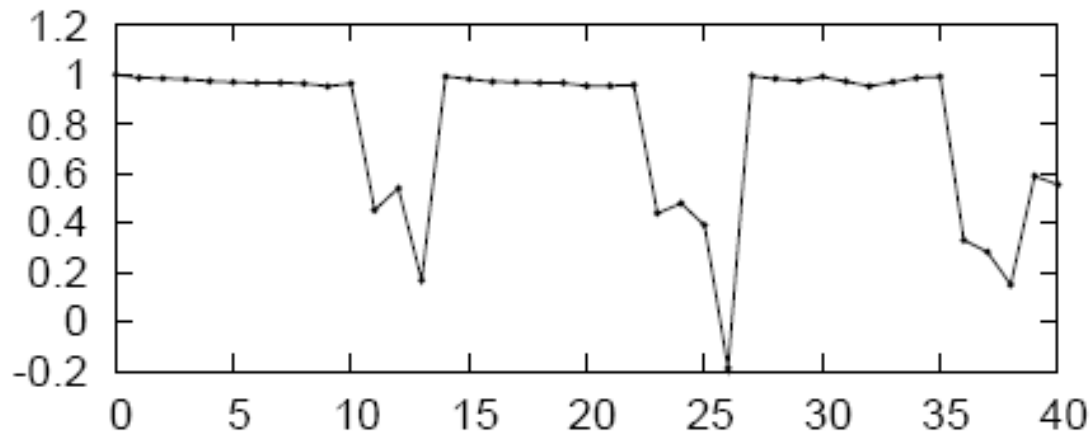




The Co-Express Gene Index



(a) $K = 1$

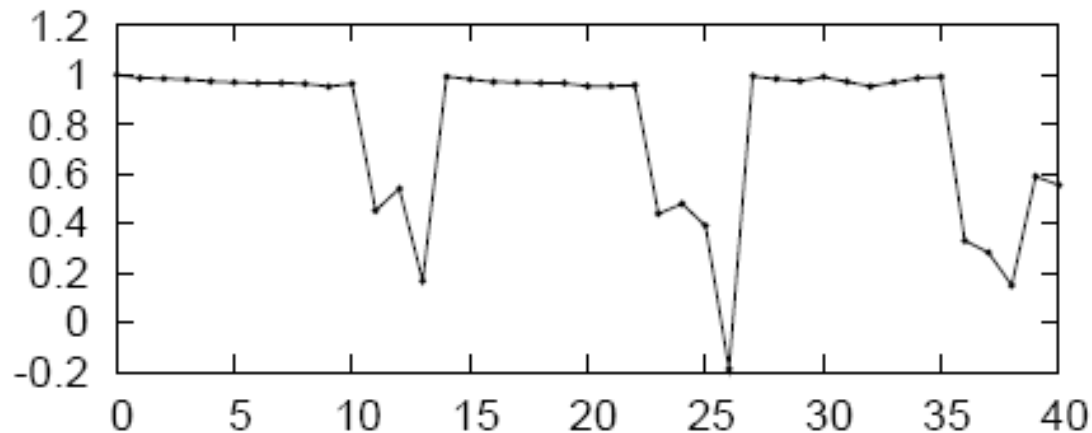




Find the co-expressed gene group



- Key point: Find the **start point and end point**
- Start point of co-expressed genes: before it, the similarity value is low
- End point: after it, the value is low





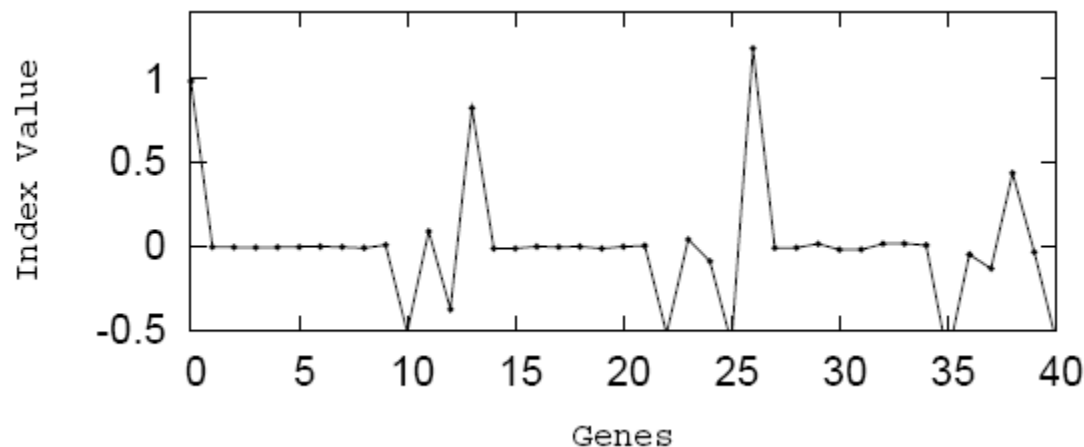
Coherent pattern index



- The coherent pattern index $CPI(g_i)$ is

$$CPI(g_i) = \sum_{j=1}^p Sim(g_{i+j}) - \sum_{j=0}^{p-1} Sim(g_{i-j}).$$

- $COI(g_i)$ = similarity of the p points after g_i – similarity of the p points before g_i
- Start point: peak, end point: valley

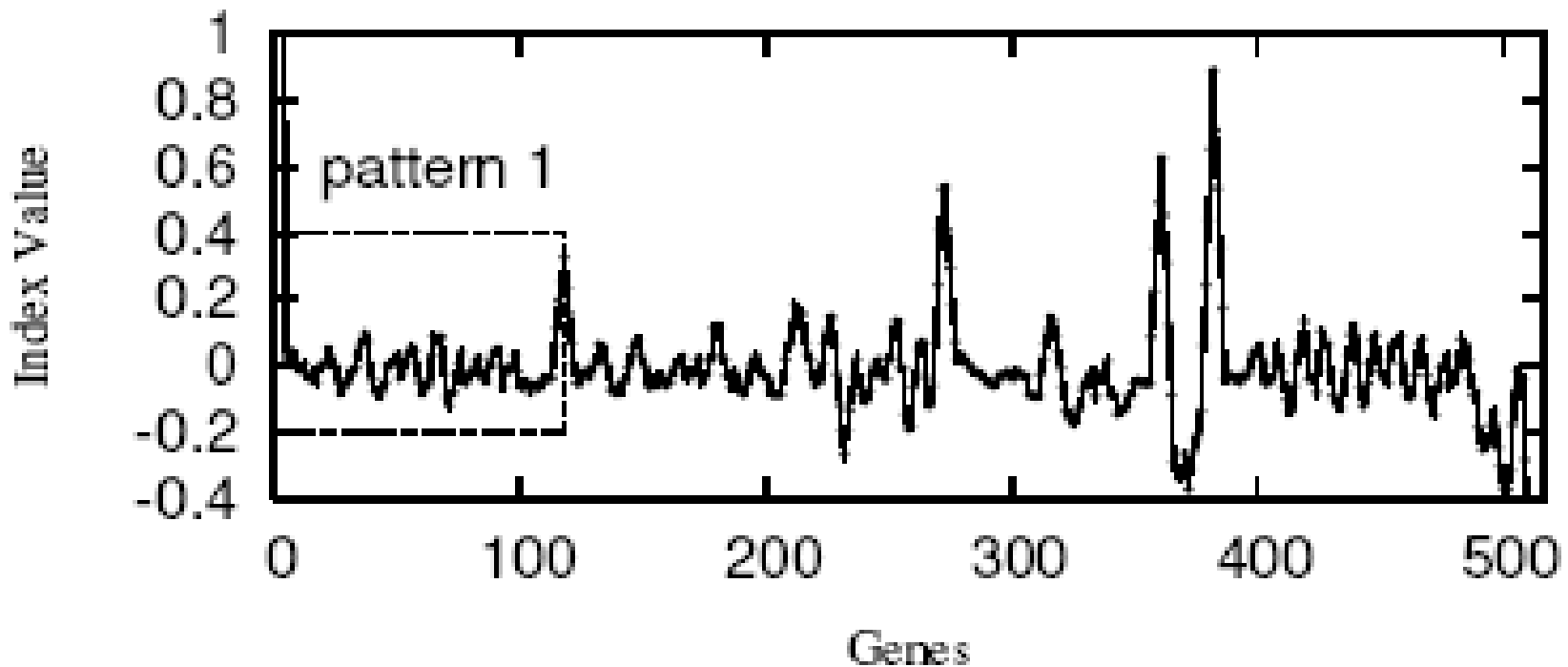




GeneX with $C(p, i)$

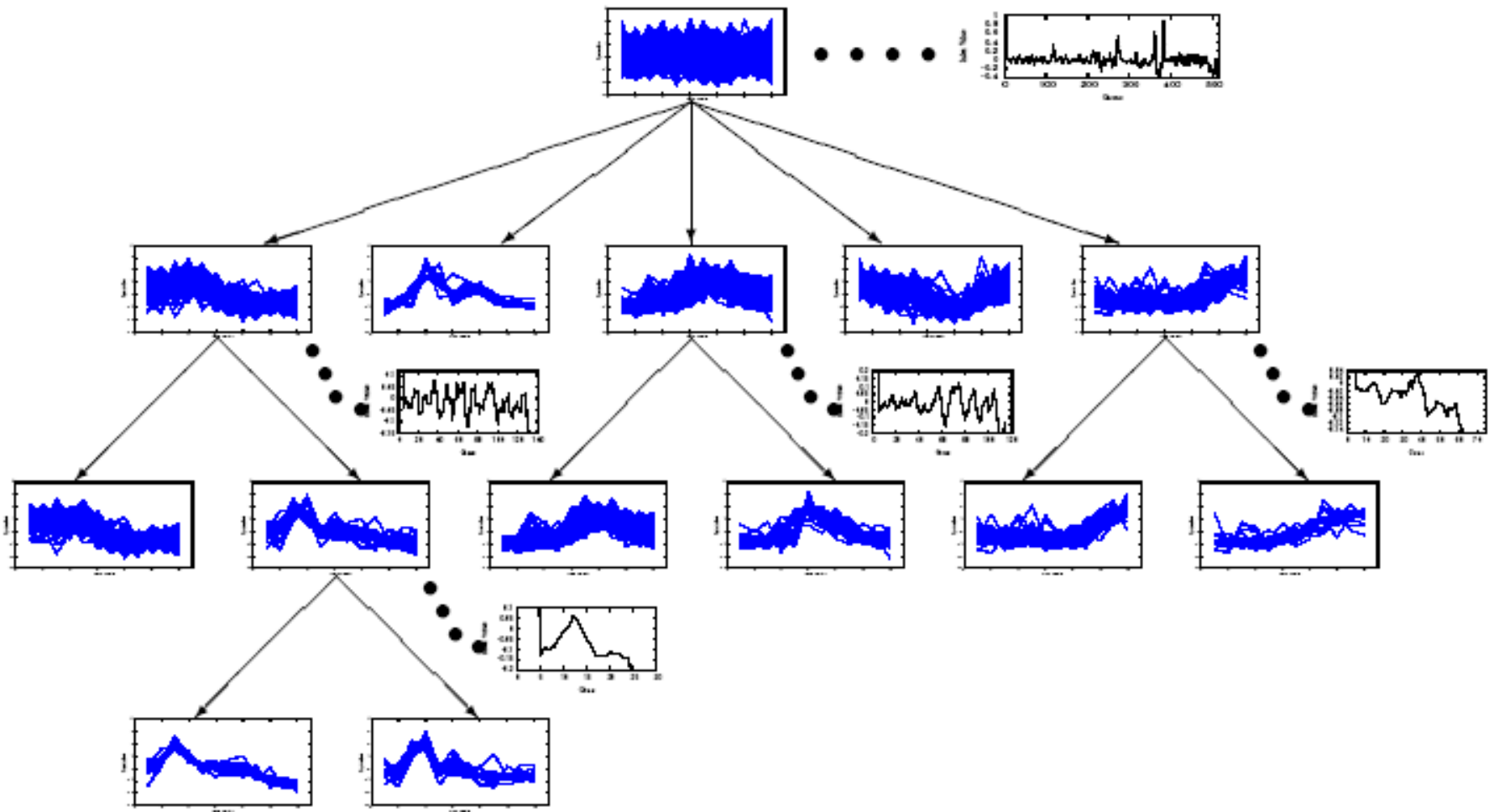


- Each peak is a possible start point for Co-expressed Gene group
- Support drill down on attraction tree





GeneX Effort

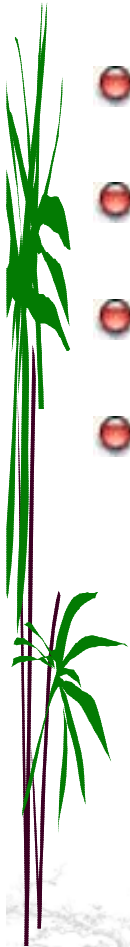




Outline



- Introduction and Backgrounds
- Problem of Microarray Data Analysis
- The Attraction Tree Approach
- Index the Attraction Trees
- **Experiment Evaluation**
- Related Studies and Future works

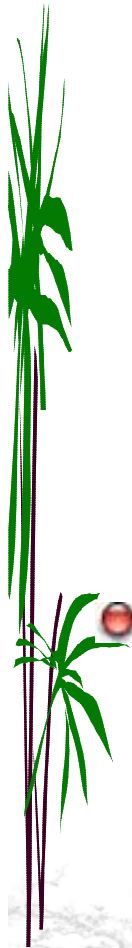




Experiment Setup



- Two real datasets
 - Iyer dataset: 8,600 distinct human genes during a 12-point time series
 - Iyer et al give 10 co-expressed gene group in their papers – ground truth
 - Spellman dataset: 6,220 genes with 5 co-expressed groups as ground truth
- Compared with the clustering algorithm with K-means, SOM, CAST, CLICK, etc

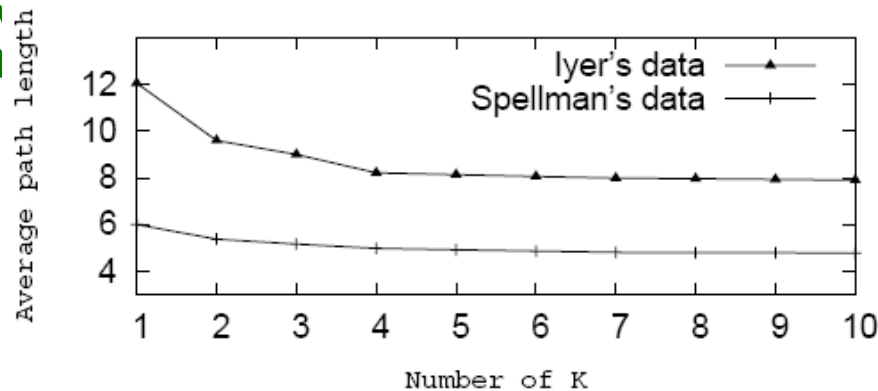




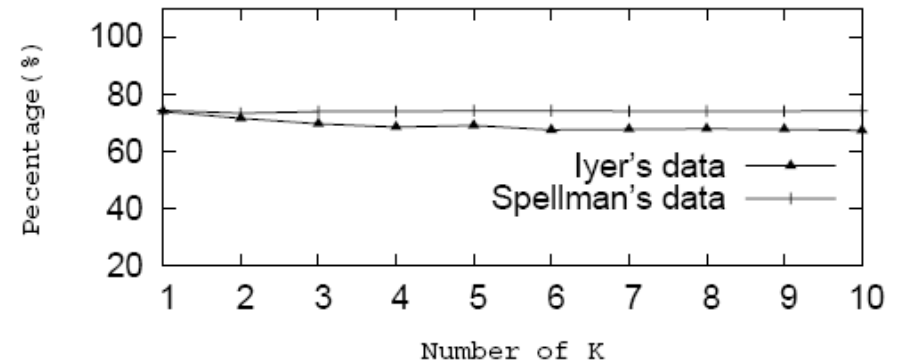
Effectiveness of Parameters



- Average length of the attractor path and the correctness of the attraction tree is insensitive to the parameter k in k NN neighbor



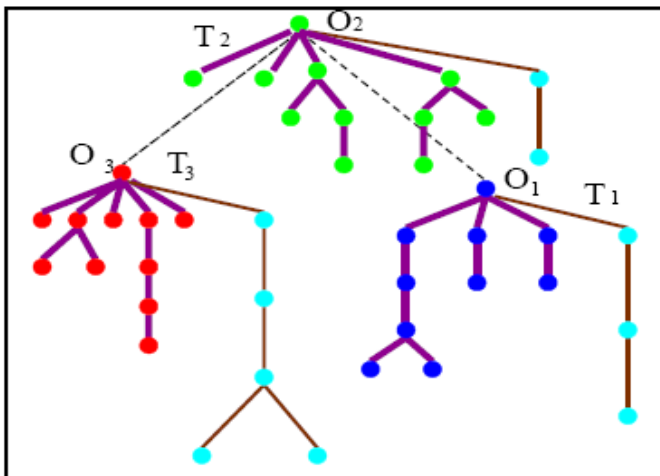
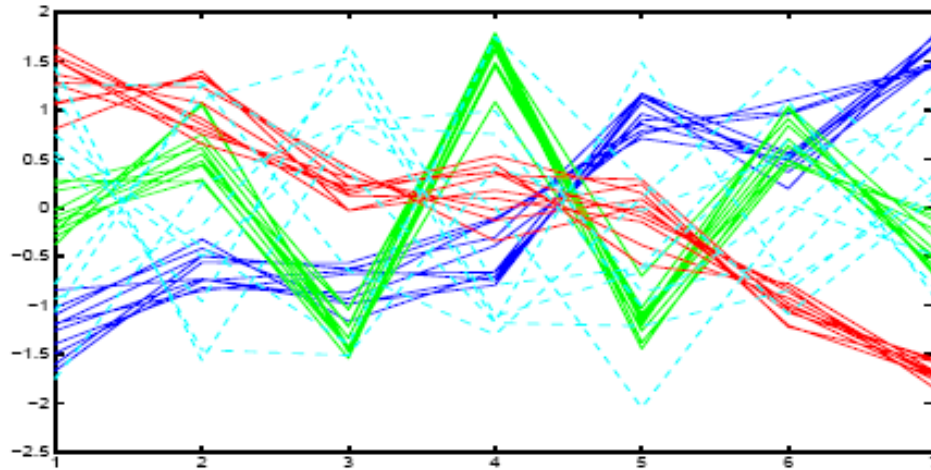
(a) Correctness of attraction trees



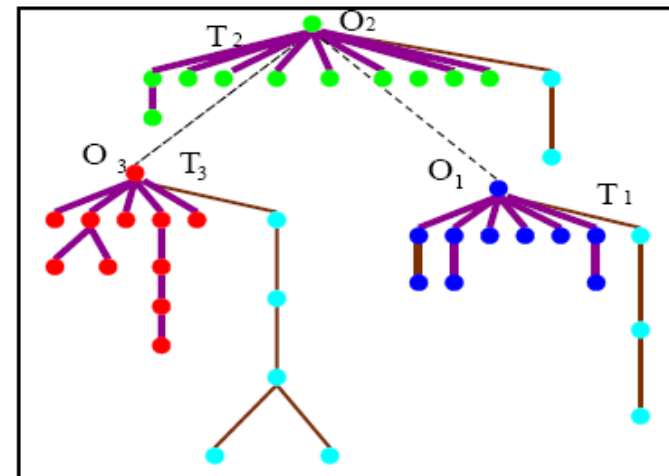
(b) Average length of paths on attraction trees



Attraction Tree Example



(a) $K = 1$



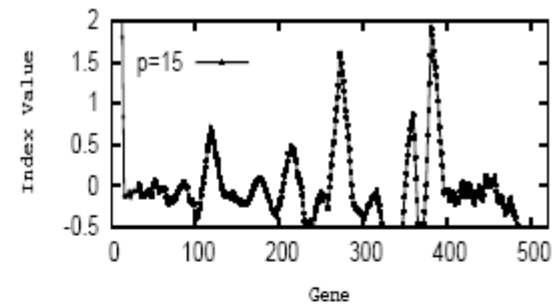
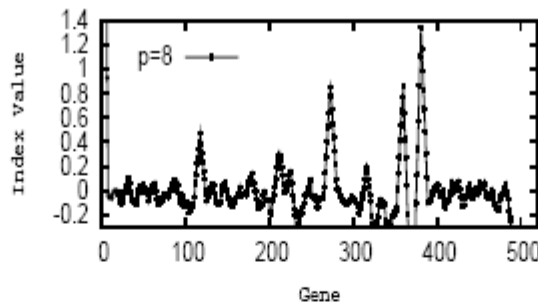
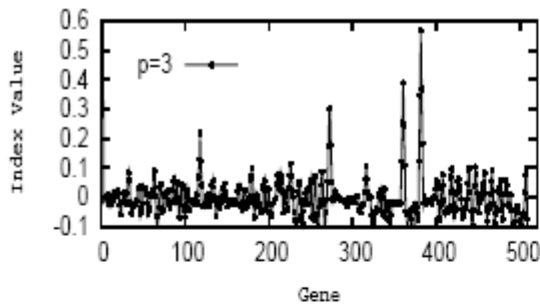
(b) $K = \infty$



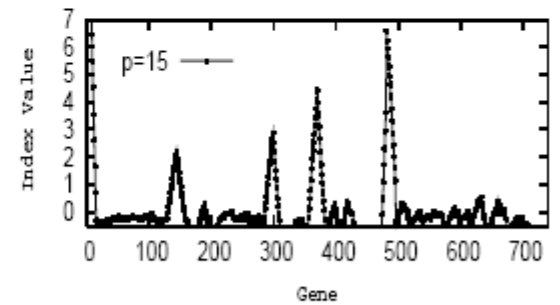
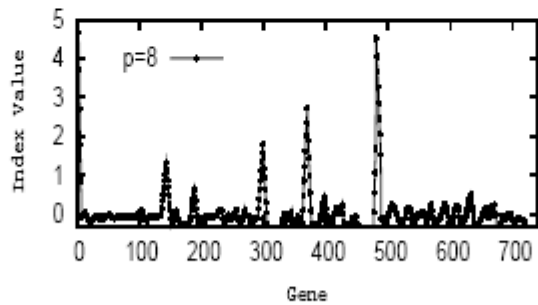
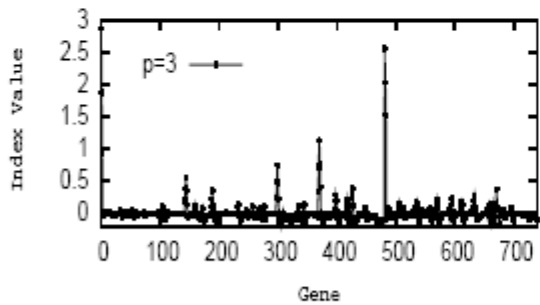
Effectiveness of parameter p



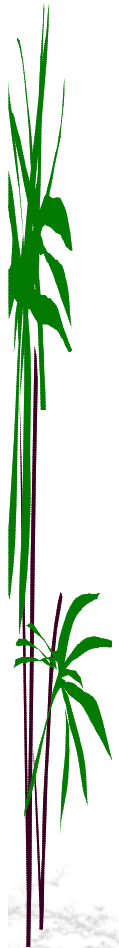
$$CPI(g_i) = \sum_{j=1}^p Sim(g_{i+j}) - \sum_{j=0}^{p-1} Sim(g_{i-j}).$$



(a) Iyer's data

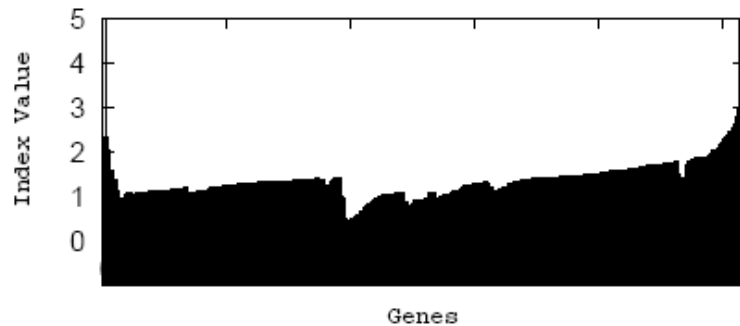


(b) Spellman's data

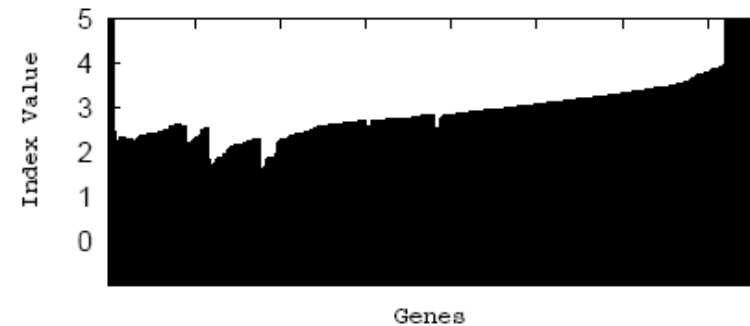




Comparison with OPTICS

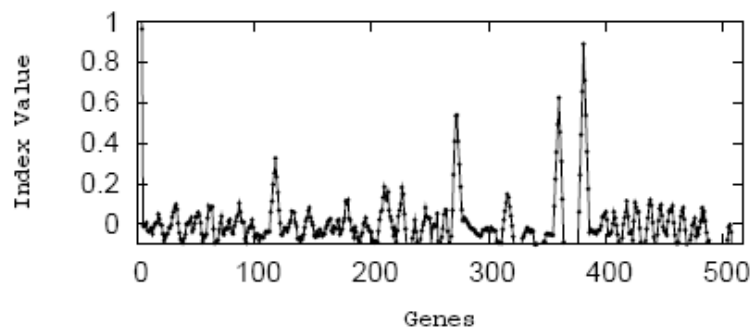


Iyer's data ($\epsilon = 3, MinPts = 5$)

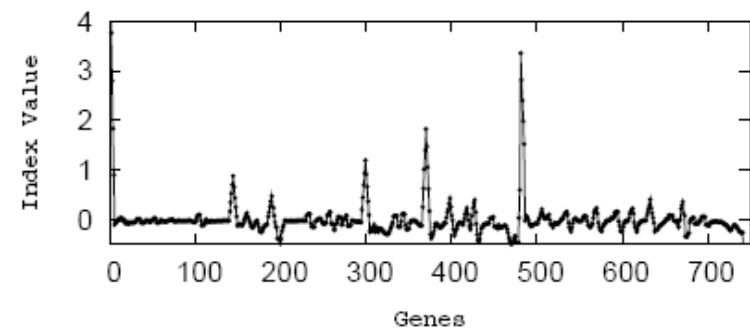


Spellman's data ($\epsilon = 3, MinPts = 5$)

(a) Object ordering generated by *OPTICS*.



Iyer's data ($K = 1, \alpha = 5$)



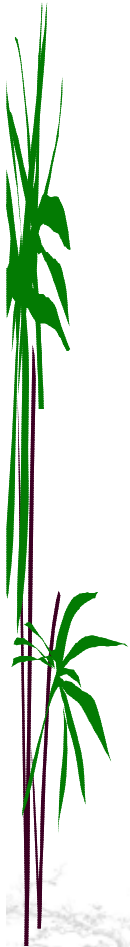
Spellman's data ($K = 10, \alpha = 5$)



OPTICS: A Cluster-Ordering Method (1999)



- OPTICS: Ordering Points To Identify the Clustering Structure
 - Ankerst, Breunig, Kriegel, and Sander (SIGMOD'99)
 - Produces a **special order** of the database wrt its density-based clustering structure
 - This cluster-ordering contains info equiv to the density-based clustering corresponding to a broad range of parameter settings
 - Good for both automatic and interactive cluster analysis, including finding intrinsic clustering structure
 - Can be represented graphically or using visualization techniques





OPTICS: Some Extension from DBSCAN



Index-based:

- $k =$ number of dimensions
- $N = 20$
- $p = 75\%$
- $M = N(1-p) = 5$

• Complexity: $O(N \log N)$

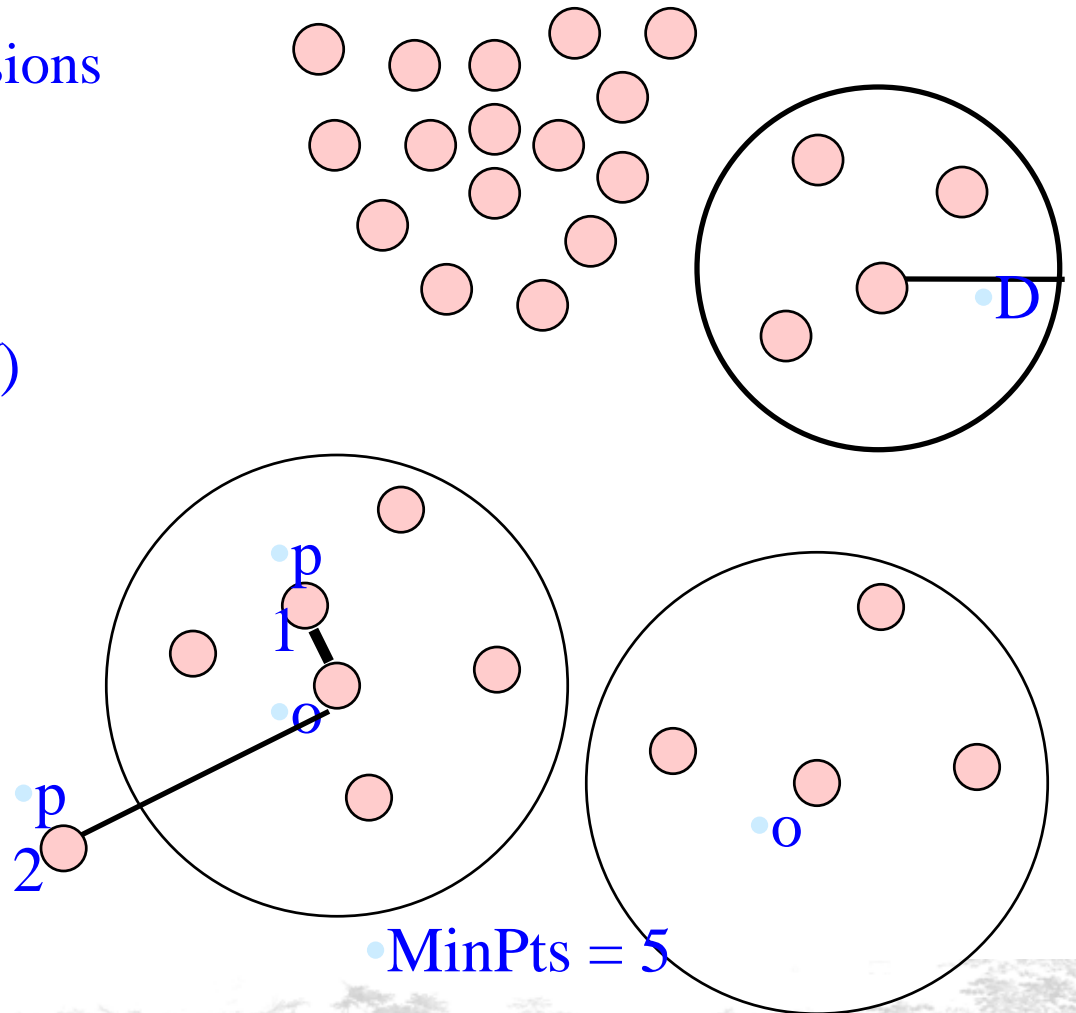
Core Distance:

- min eps s.t. point is core

Reachability Distance

• $\text{Max}(\text{core-distance}(o), d(o, p))$

• $r(p_1, o) = 2.8\text{cm}$. $r(p_2, o) = 4\text{cm}$



• $\text{MinPts} = 5$

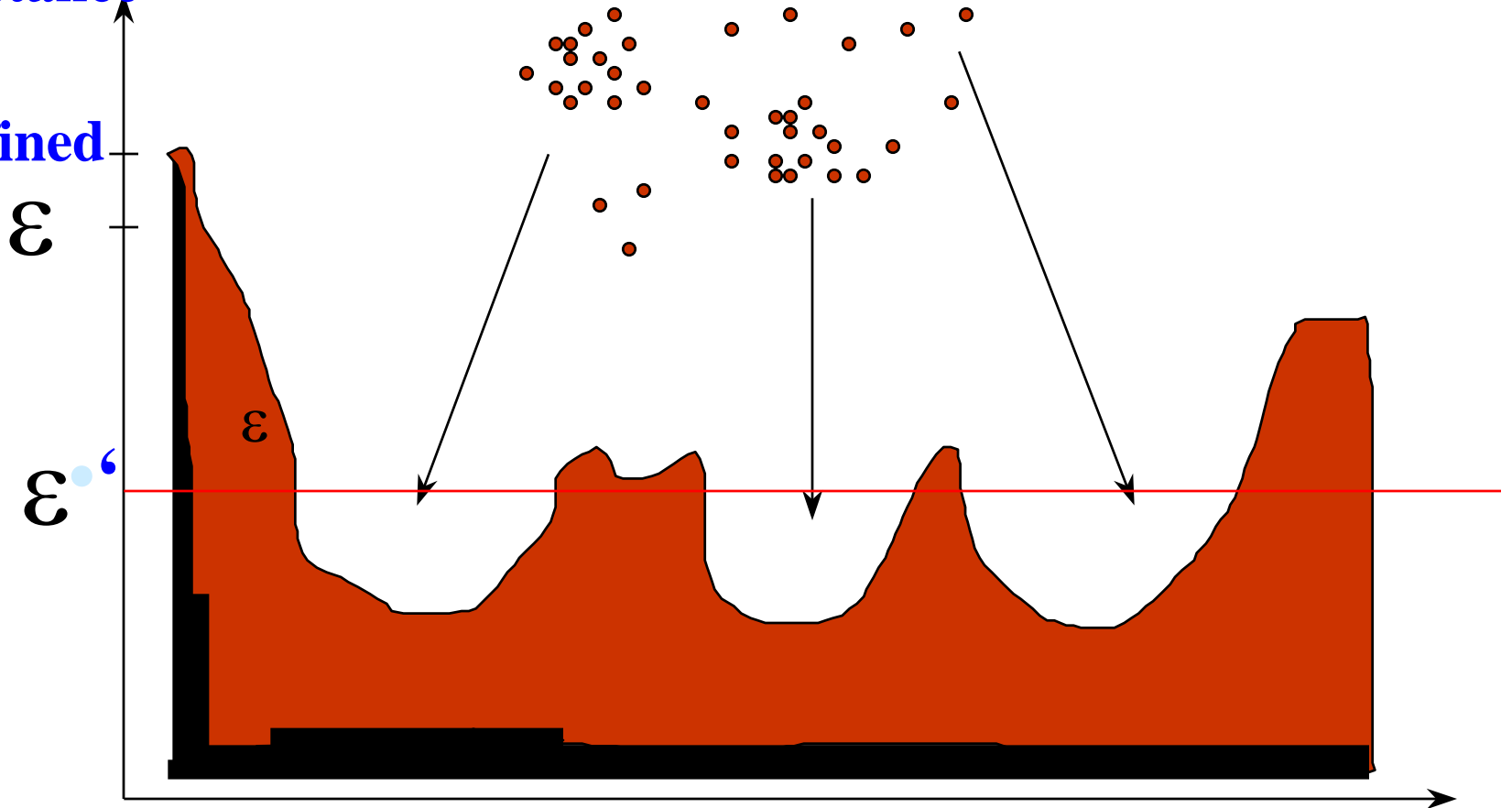


OPTICS & Its Applications



• Reachability distance

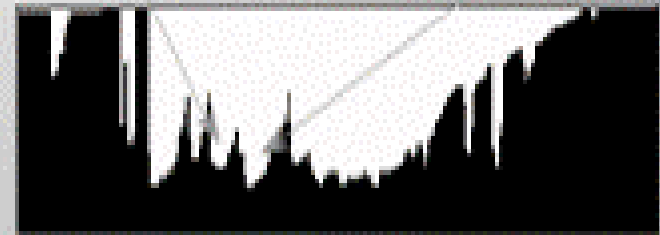
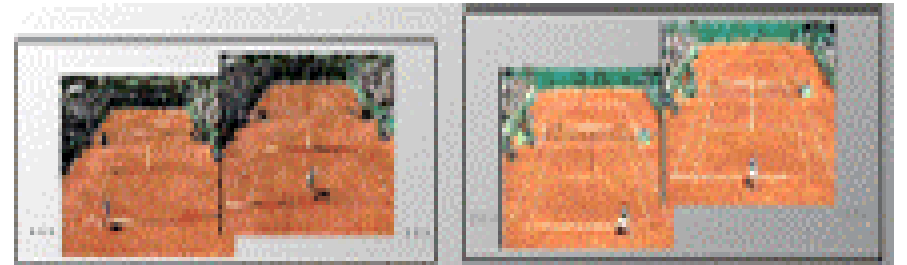
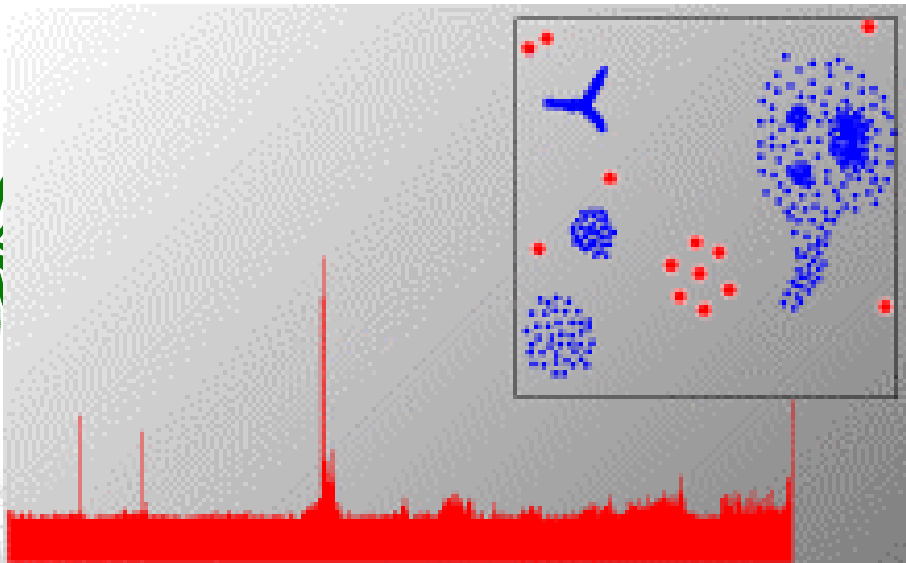
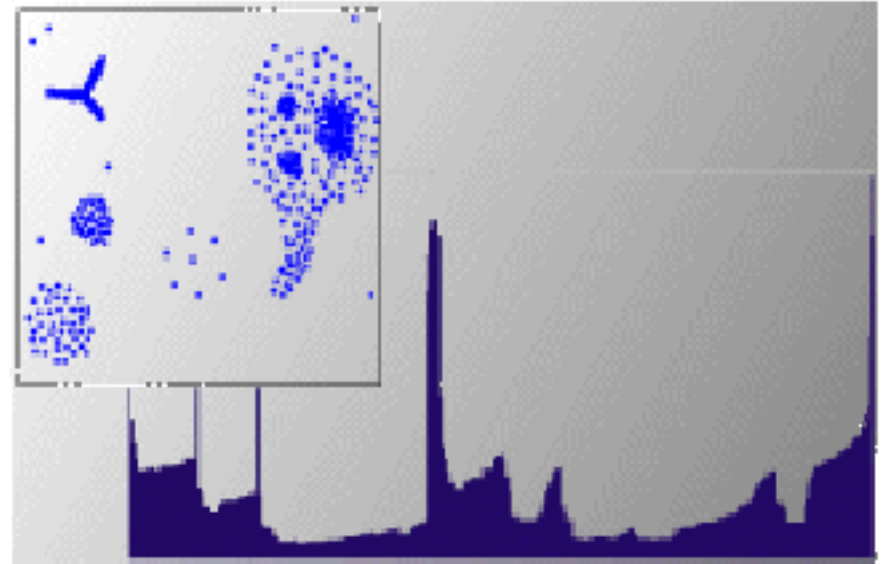
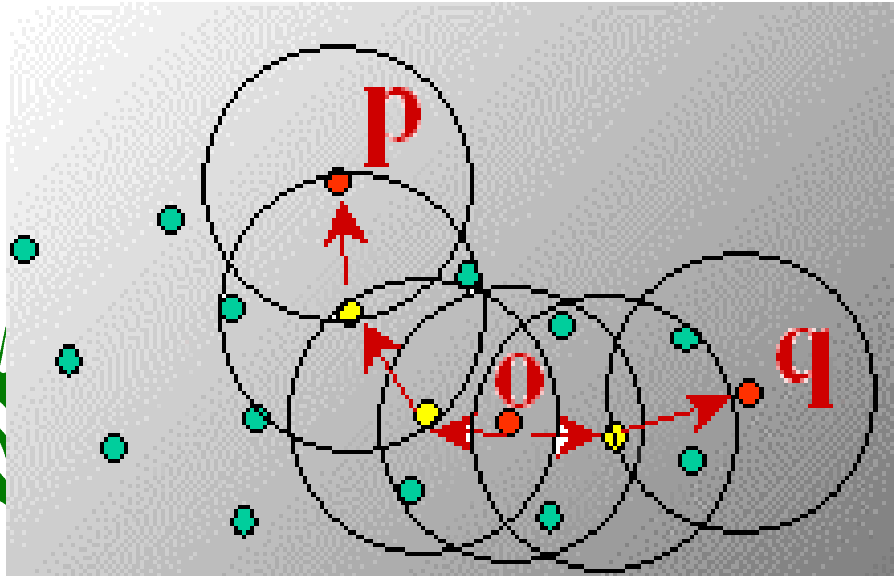
• undefined



• Cluster-order
of the objects

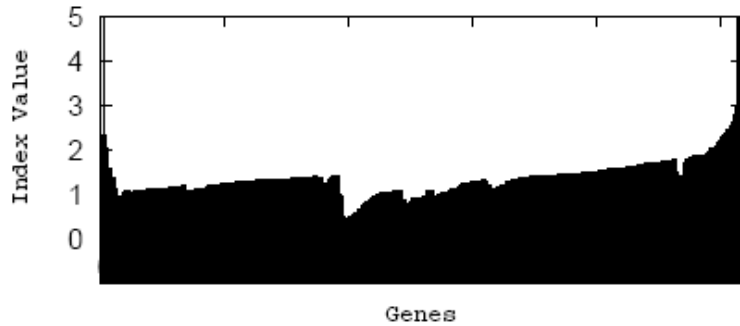


OPTICS & Its Applications

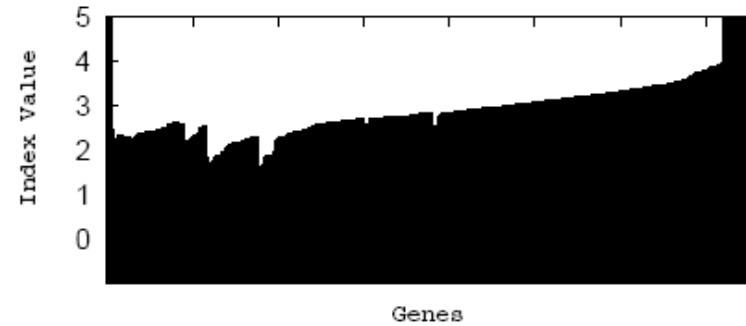




Comparison with OPTICS

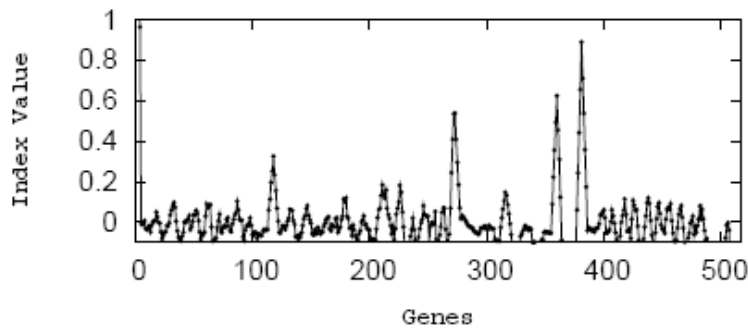


Iyer's data ($\epsilon = 3, MinPts = 5$)

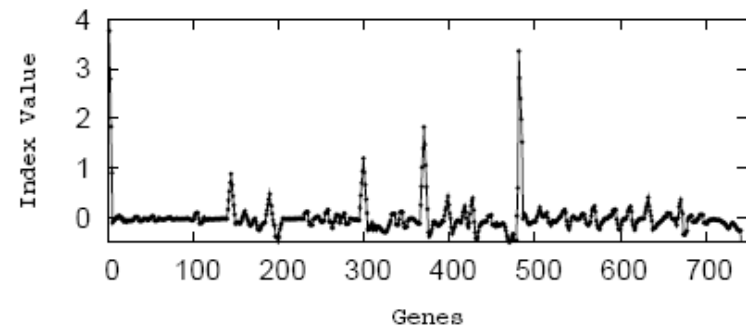


Spellman's data ($\epsilon = 3, MinPts = 5$)

(a) Object ordering generated by *OPTICS*.



Iyer's data ($K = 1, \alpha = 5$)



Spellman's data ($K = 10, \alpha = 5$)



Accuracy test result on Iyer dataset



Pattern	GPX (10)	Kmeans (10)	SOM (10)	ADAPT (11)	CLICK (7)	CAST (9)	SOTA (50)
1	0.998	0.973	0.983	0.956	0.884	0.955	0.962
2	0.996	0.950	0.992	0.911	0.991	0.887	0.936
3	0.993	0.910	0.872	0.993	0.994	0.997	0.947
4	0.995	0.996	0.989	0.984	0.883	0.968	0.955
5	0.964	0.882	0.716	0.868	0.886	0.855	0.962
6	0.940	0.965	0.764	0.989	0.970	0.984	0.972
7	0.972	0.880	0.892	0.976	0.990	0.719	0.988
8	0.995	0.963	0.917	0.997	0.914	0.999	0.958
9	0.907	0.910	0.848	0.824	0.844	0.800	0.940
10	0.987	0.930	0.983	0.981	0.976	0.996	0.960

Figure 19: Coherent patterns discovered in Iyer's data set by different approaches.

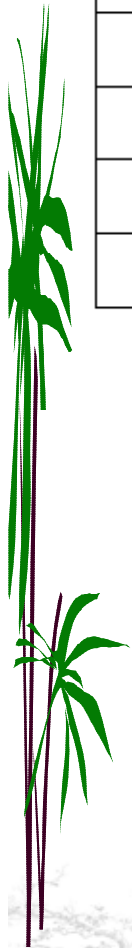


Accuracy test result on Spellman dataset



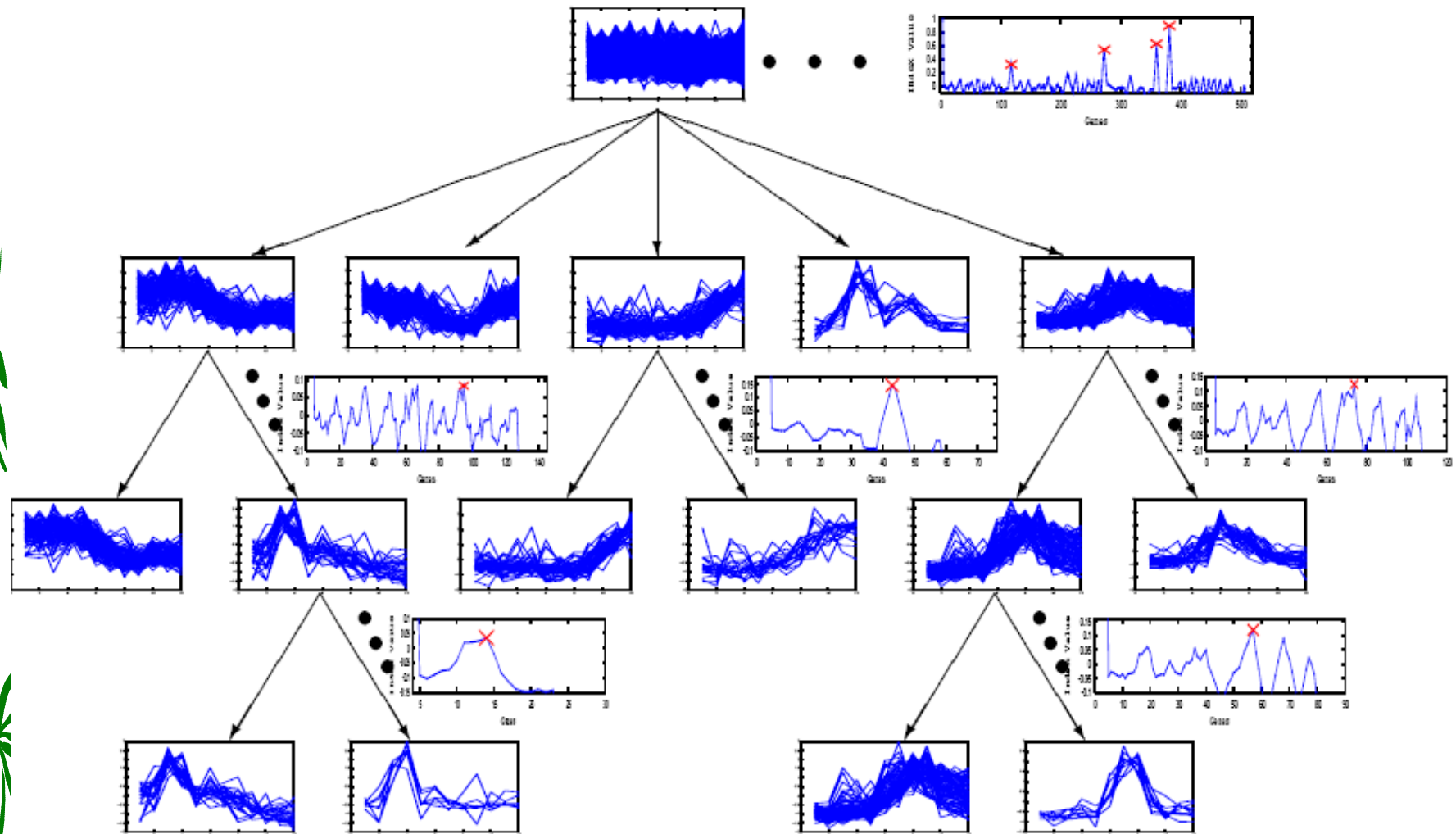
Pattern	GPX (7)	Kmeans (5)	SOM (5)	ADAPT (21)	CLICK (9)	CAST (19)	SOTA (99)
1	0.901	0.928	0.194	0.884	0.855	0.900	0.938
2	0.970	0.976	0.972	0.972	0.978	0.970	0.968
3	0.980	0.950	0.552	0.953	0.970	0.888	0.940
4	0.901	0.773	0.437	0.796	0.984	0.888	0.961
5	0.945	0.965	0.964	0.962	0.978	0.956	0.956

Figure 21: Coherent patterns discovered in Spellman's data set by different approaches.



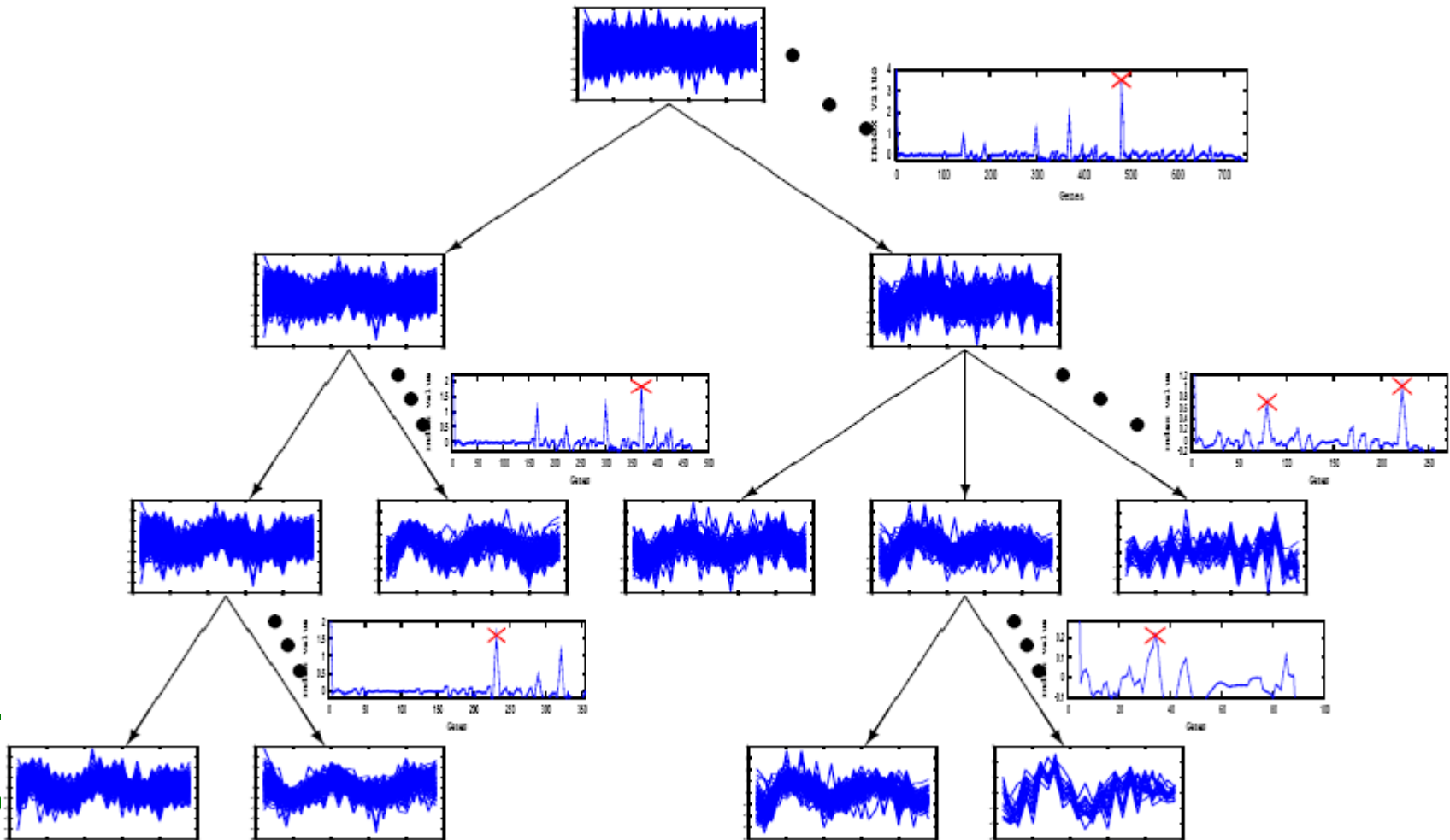


Attraction tree of Iyer dataset





Attraction tree of Spellman dataset

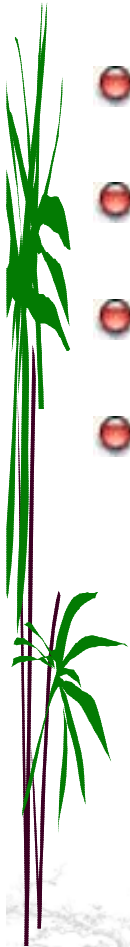




Outline



- Introduction and Backgrounds
- Problem of Microarray Data Analysis
- The Attraction Tree Approach
- Index the Attraction Trees
- Experiment Evaluation
- **Related Studies and Future works**





Discussion and Extension



- Prof. Jian Pei leaved State Univ. of Buffalo and Dr. Daxin Jiang graduated soon after the paper published
- Many future works:
 - Algorithm → Demo → Prototype system
 - Study more dimensions: the sample dimension
 - Integrate with other domain knowledge and tools
- Not a real cube, only one attraction tree, the order of gene is fixed in the index
- develop multiple drill-down paths with different combination of multi-dimensions

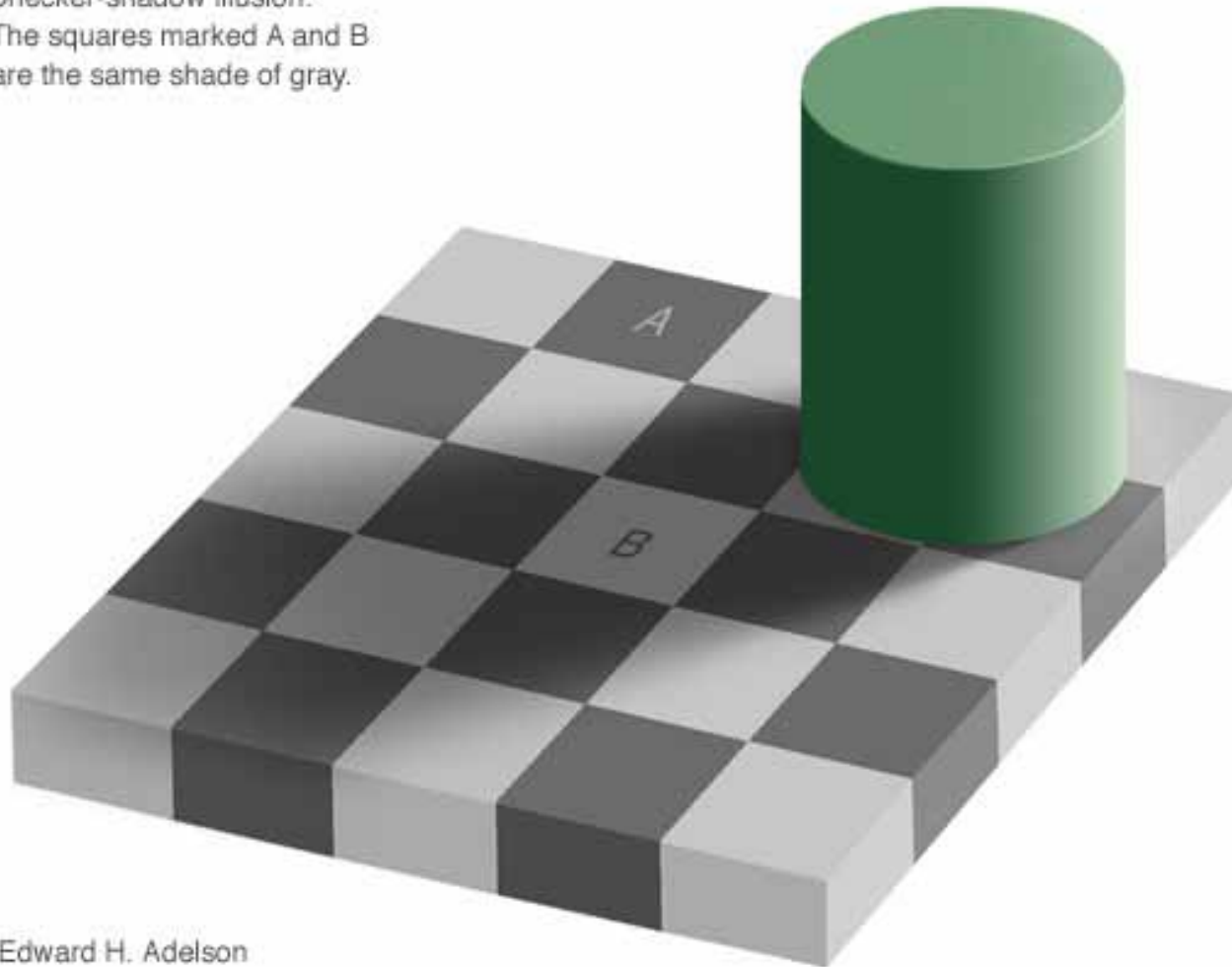




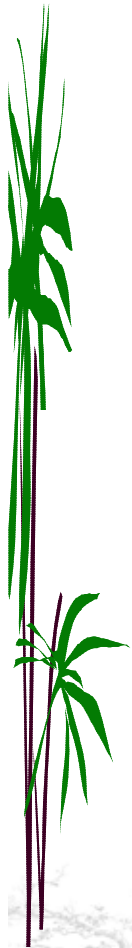
Thanks very much!



Checker-shadow illusion:
The squares marked A and B
are the same shade of gray.



Edward H. Adelson





References



- Jiang, D., Pei, J. and Zhang, A. DHC: A Density-based Hierarchical Clustering Method for Time Series Gene Expression Data. In In Proceedings of the 3rd IEEE Symposium on Bio-informatics and Bio-engineering (BIBE'03), Washington, DC, USA, March 10-12 2003.
- Jiang, D., Pei, J. and Zhang, A. Interactive Exploration of Coherent Patterns in Time-Series Gene Expression Data. In Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'03), Washington, DC, USA, August 24-27 2003.
- D. Jiang , J. Pei and A. Zhang. " Towards Interactive Exploration of Gene Expression Patterns". ACM SIGKDD Explorations (Special Issue on Microarray Data Analysis), Volume 5, Issue 2, page 79 - 90, 2003

