



Microarray Clustering in a Multiobjective framework

Dr. Ujjwal Maulik

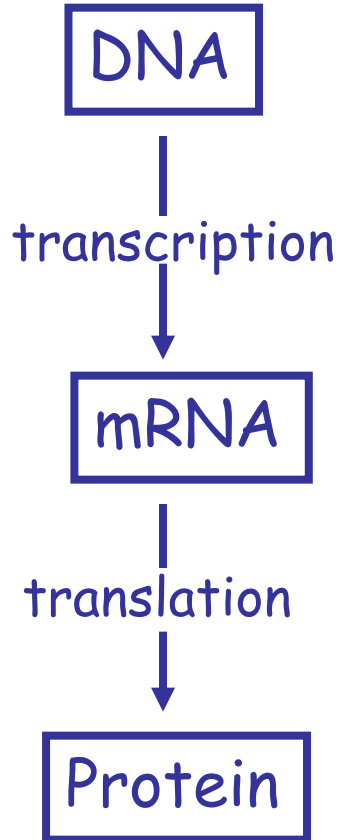
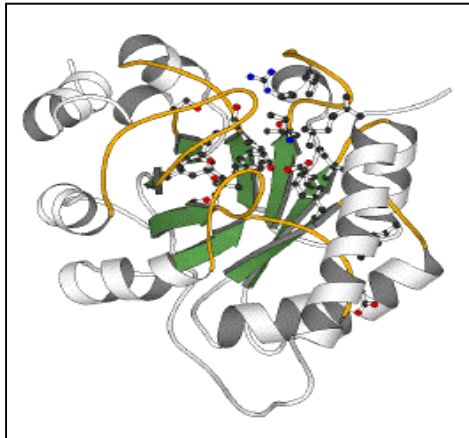
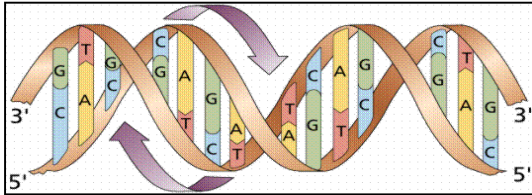
Department of Computer Sc. & Engg.

Jadavpur University

umaulik@cse.jdvu.ac.in

URL: <https://sites.google.com/site/drujjwalmaulik/>

Central Dogma of Molecular Biology



CCTGAGCCAAC TATTGATGAA



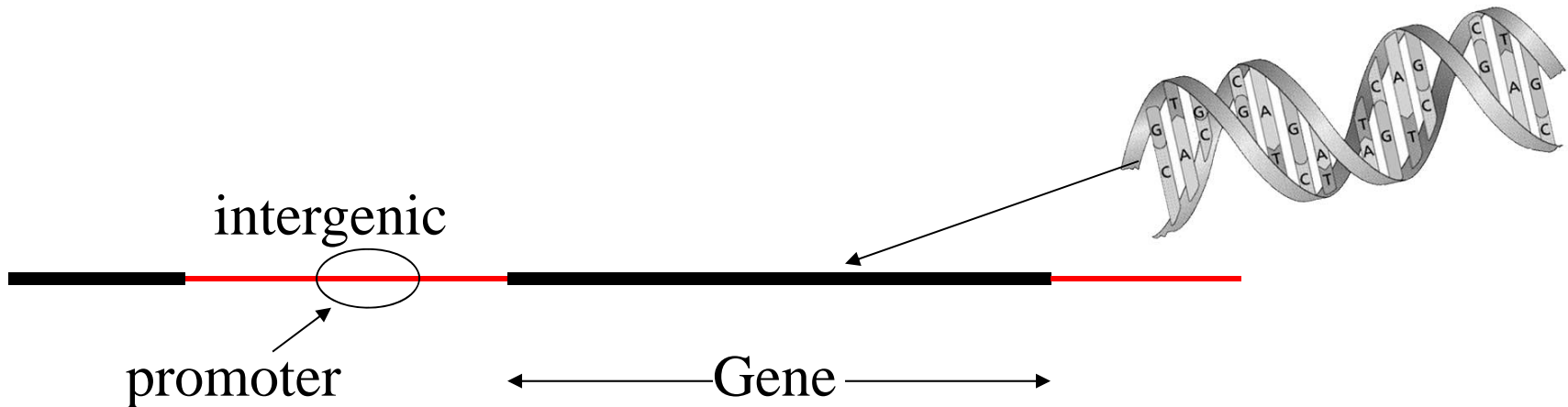
CCUGAGCCAACUAUUGAUGAA



PEPTIDE

Transcription

- Process by which DNA forms RNA



Promoter/transcription factors acts as a switch
turning the gene on or off



Gene Expression

- Genome is in general the same in all the cells
 - Hair, nails, liver, lung, heart
- Then why is the behavior different?
- Not all genes are expressed to the same extent everywhere



- Differential expression of genes
 - not all mRNAs, and hence their protein products, are generated everywhere
 - Expression is tissue specific
 - Level varies from one tissue to the other
 - Expression level of a gene is also dependent on time
 - Amount of mRNA produced varies with time



Gene Expression

- Indicates the amount of mRNA produced from a gene
 - Whether the gene is active or not
 - How active the gene is
- Difference in gene expression causes
 - Functional difference among tissues
 - Multiple abnormalities

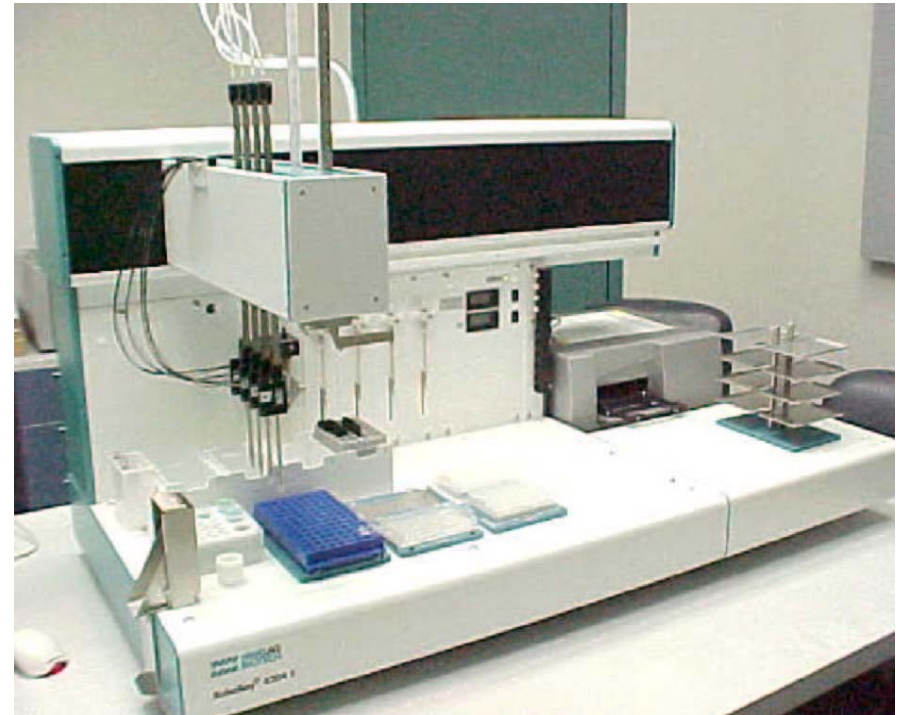


Factors Controlling Gene Expression

- Controlled production of transcription factors
 - Regulatory networks
 - No TFs → no transcription into mRNAs
- Selective transport of mRNAs into the cytoplasm
- Controlled translation
 - mRNA degradation via post transcriptional gene silencing
 - mRNA repression
- Protein activation or degradation

Microarray

- What is it?
 - Technology to simultaneously monitor the expression levels of a large number of genes
- cDNA microarray chip
 - Typically a glass slide, onto which about 10,000 cDNAs (typically 600-2400 nt long) from a library are spotted/attached per sq. cm. using a spotter





cDNA Microarray Chip

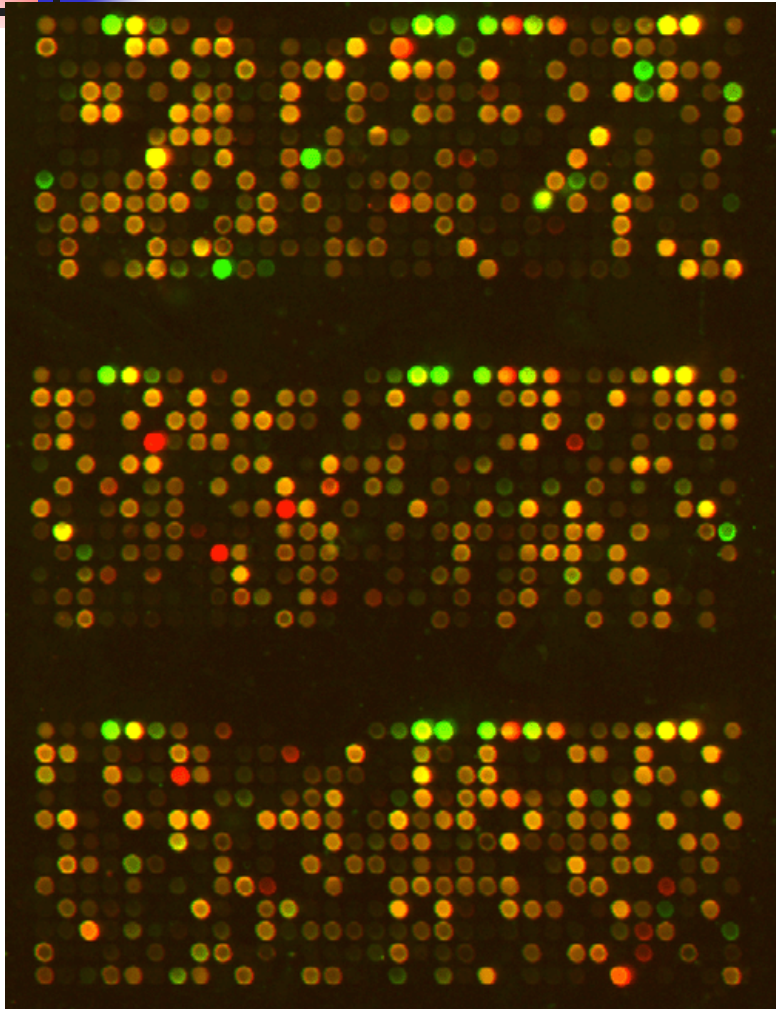
- Preparing the probe: combination of normal (reference) and diseased (test) samples
 - Reference/Control sample
 - mRNA from normal tissues converted to cDNA by reverse transcription and colored with green-fluorescent dye Cy3
 - Experimental RNA samples being investigated
 - mRNA from diseased tissues converted to cDNA by reverse transcription and colored with red-fluorescent dye Cy5
- Both reference and test samples are added on the microarray chip.
 - Hybridization of the probes and the spotted cDNAs takes place
- Chip is washed to remove excess probes (unhybridized ones)
- Two images, in red and green bands, are acquired.
 - That measure the spot intensities using red and green channels
- Gene expression: the Cy5/Cy3 fluorescence ratio



Other Microarray and Issues

- Oligonucleotide microarray (Affymetrix Chips)
 - Simultaneous measurement of a larger number of expression values
 - approx. 250,000 targets per sq. cm.
 - More accurate
 - More expensive
- Several errors might occur in chip generation, hybridization, imaging, etc.
- Hence expression values may differ from one microarray experiment to another.
- Normalization of the data is necessary to account for these variations.

A Typical cDNA Microarray

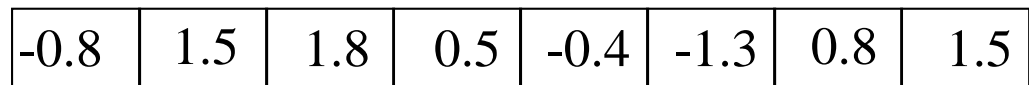


- **Red:** Gene over-expressed in diseased (test) sample than in normal (reference) sample.
- **Green:** Gene under-expressed in diseased (test) sample than in normal (reference) sample.
- **Yellow:** Expression level of test and normal (reference) gene same

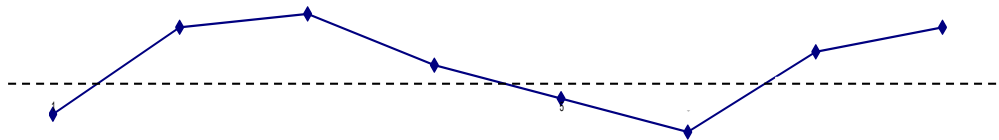
Expression Vectors

Gene Expression Vectors encapsulate the expression of a gene over a set of experimental conditions or sample types.

Numeric Vector



Line Graph



Heatmap



Ack:



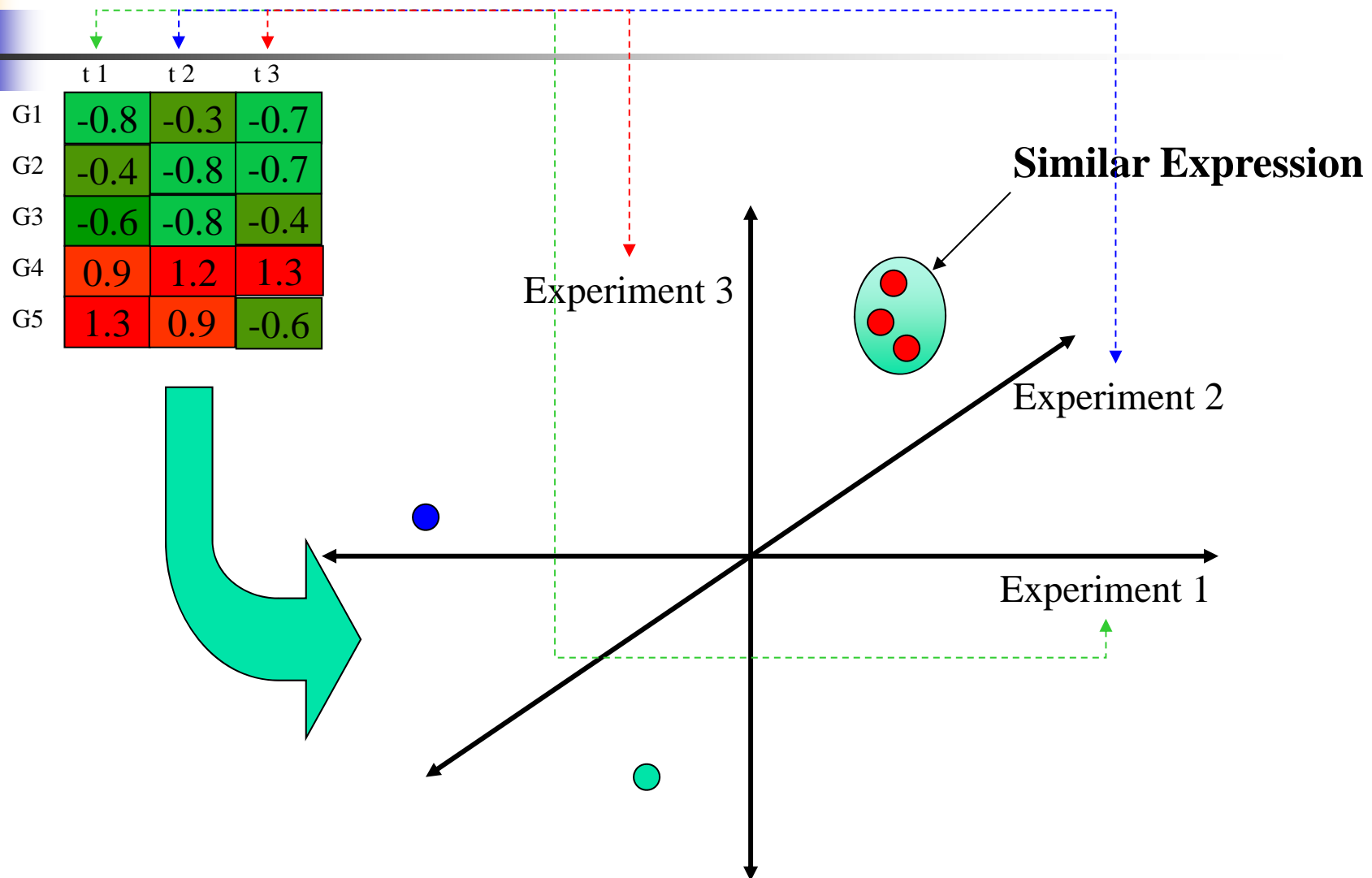
ParaBioSys

Parallel Biological Systems

Mass. General Hospital • Harvard Medical School • Boston University



Expression Vectors As Points in 'Expression Space'

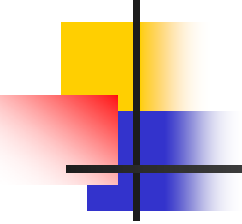




Distance and Similarity

- The ability to calculate a distance (or similarity, it's inverse) between two expression vectors is fundamental to many algorithms
- Selection of a distance metric defines the concept of distance

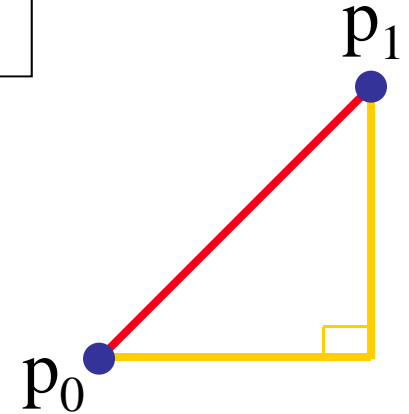
Some Distance Measures



	Exp 1	Exp 2	Exp 3	Exp 4	Exp 5	Exp 6
Gene A	x_{1A}	x_{2A}	x_{3A}	x_{4A}	x_{5A}	x_{6A}
Gene B	x_{1B}	x_{2B}	x_{3B}	x_{4B}	x_{5B}	x_{6B}

Some distances: (MeV provides 11 metrics)

1. Euclidean: $\sqrt{\sum_{i=1}^6 (x_{iA} - x_{iB})^2}$
2. Manhattan: $\sum_{i=1}^6 |x_{iA} - x_{iB}|$
3. Pearson correlation





Potential Microarray Applications

- Drug discovery / toxicology studies
- Mutation/polymorphism detection
- Differing expression of genes over:
 - Time
 - Tissues
 - Disease States
- Sub-typing complex genetic diseases



Microarray Data Analysis

- Data analysis consists of several post-quantization steps:
 - Statistics/Metrics Calculations
 - Scaling/Normalization of the Data
 - Gene Selection
 - Classification
 - Clustering Gene Expression Data
 - Biclustering
- Most software packages perform only a limited number of analysis tasks



Popular Methods of Clustering of Gene Expression Data

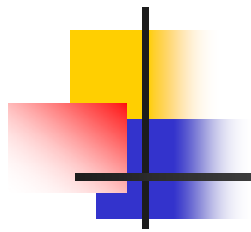
- Hierarchical methods
 - Single link, average link, complete link
 - dendrogram
- Self-Organizing Maps
- k-means Clustering



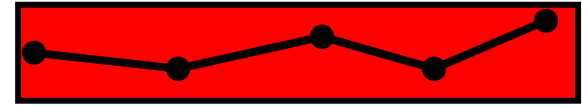
Hierarchical Clustering

- IDEA: Iteratively combines genes into groups based on similar patterns of observed expression
- By combining genes with genes OR genes with groups algorithm produces a dendrogram of the hierarchy of relationships.
- Display the data as a heatmap and dendrogram
- Cluster genes, samples or both

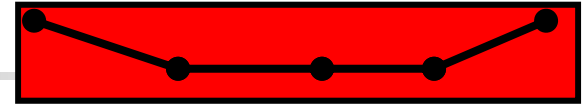
Hierarchical Clustering



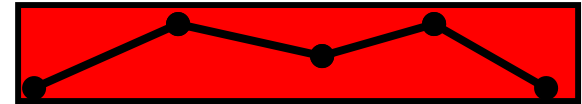
Gene 1



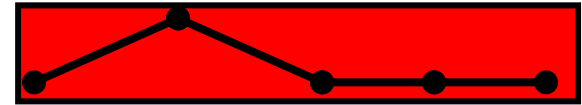
Gene 2



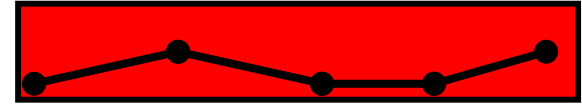
Gene 3



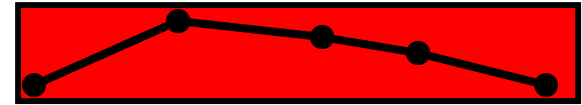
Gene 4



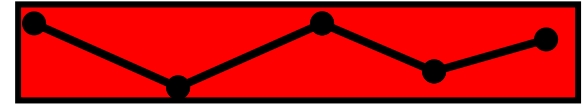
Gene 5



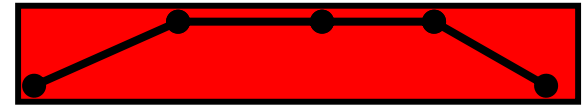
Gene 6



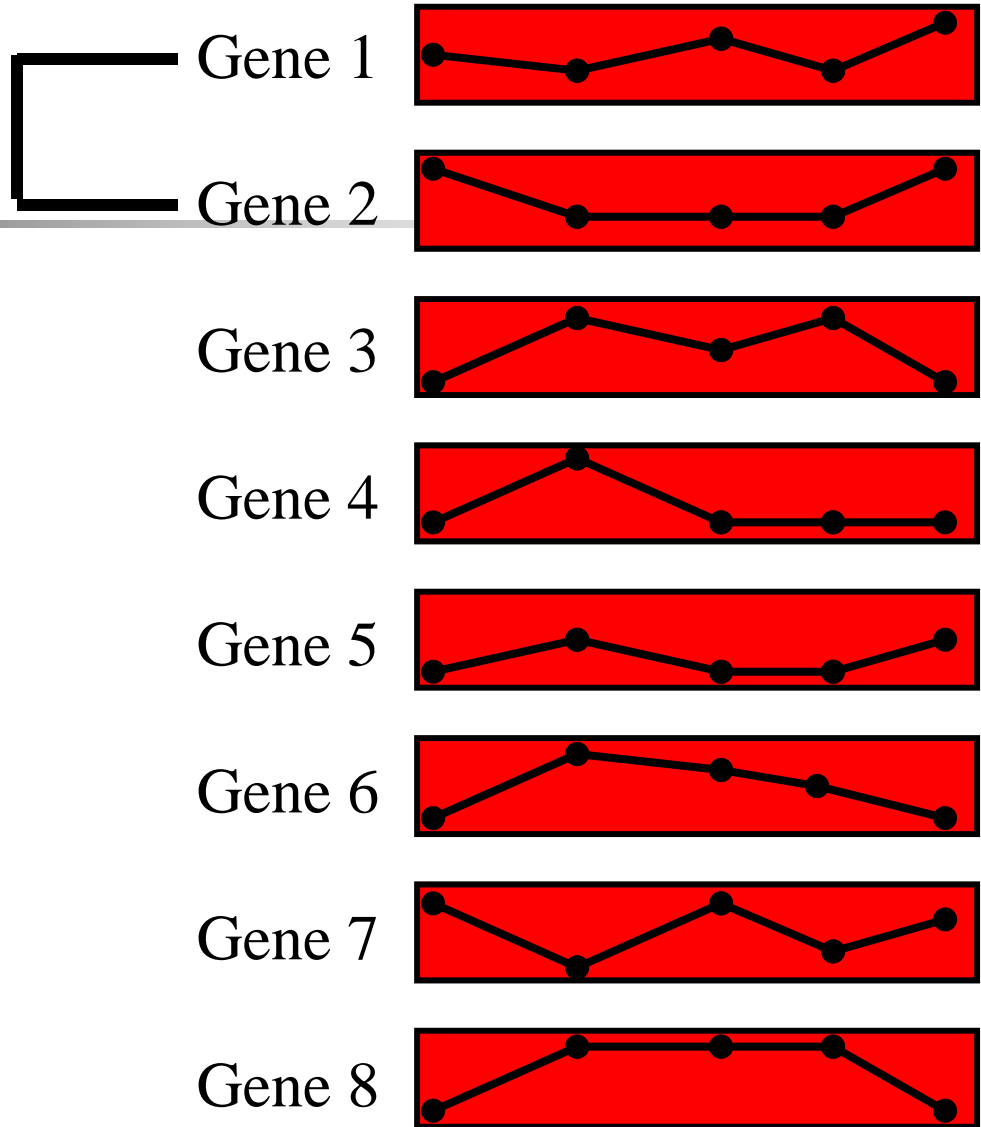
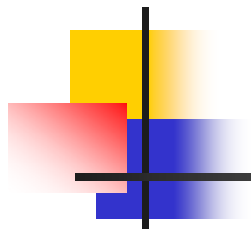
Gene 7



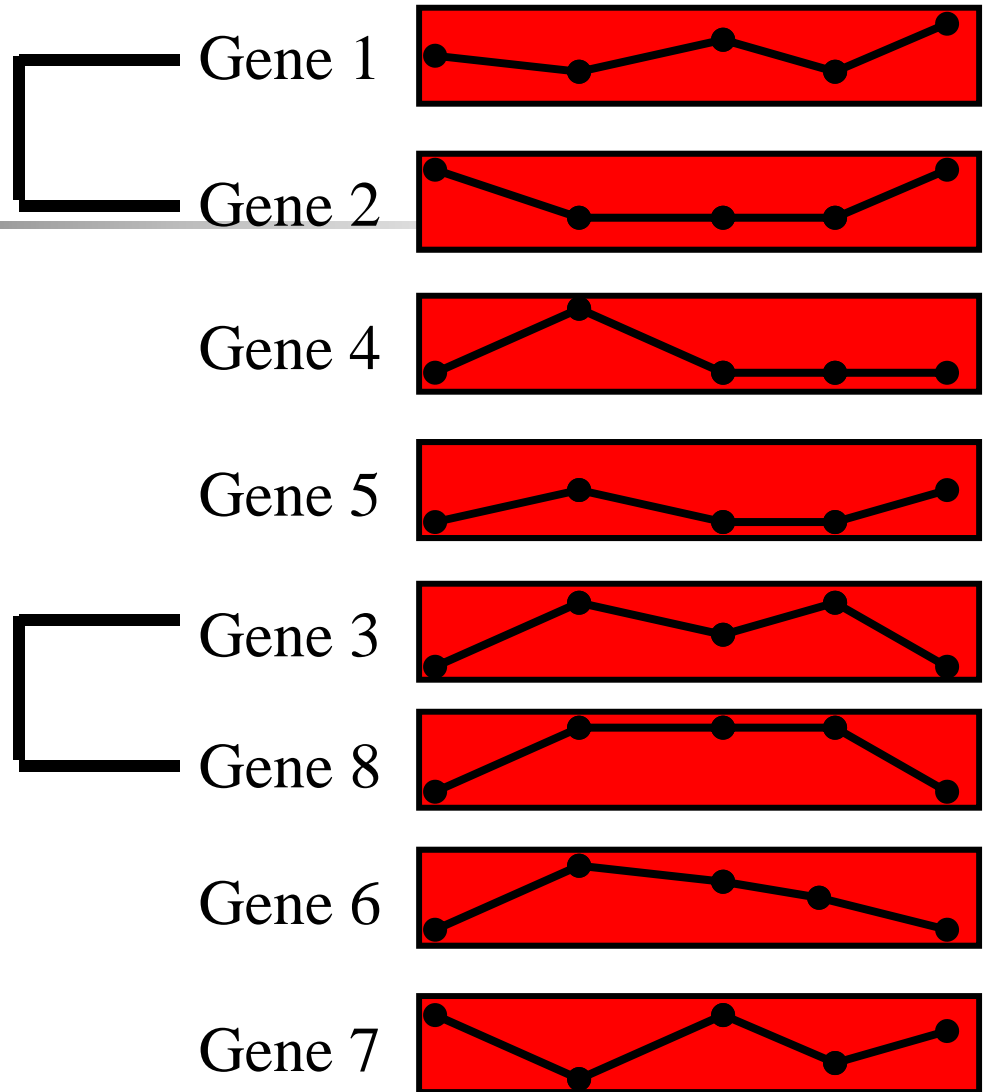
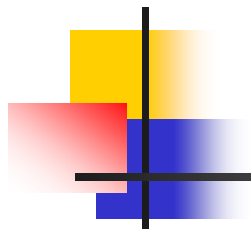
Gene 8



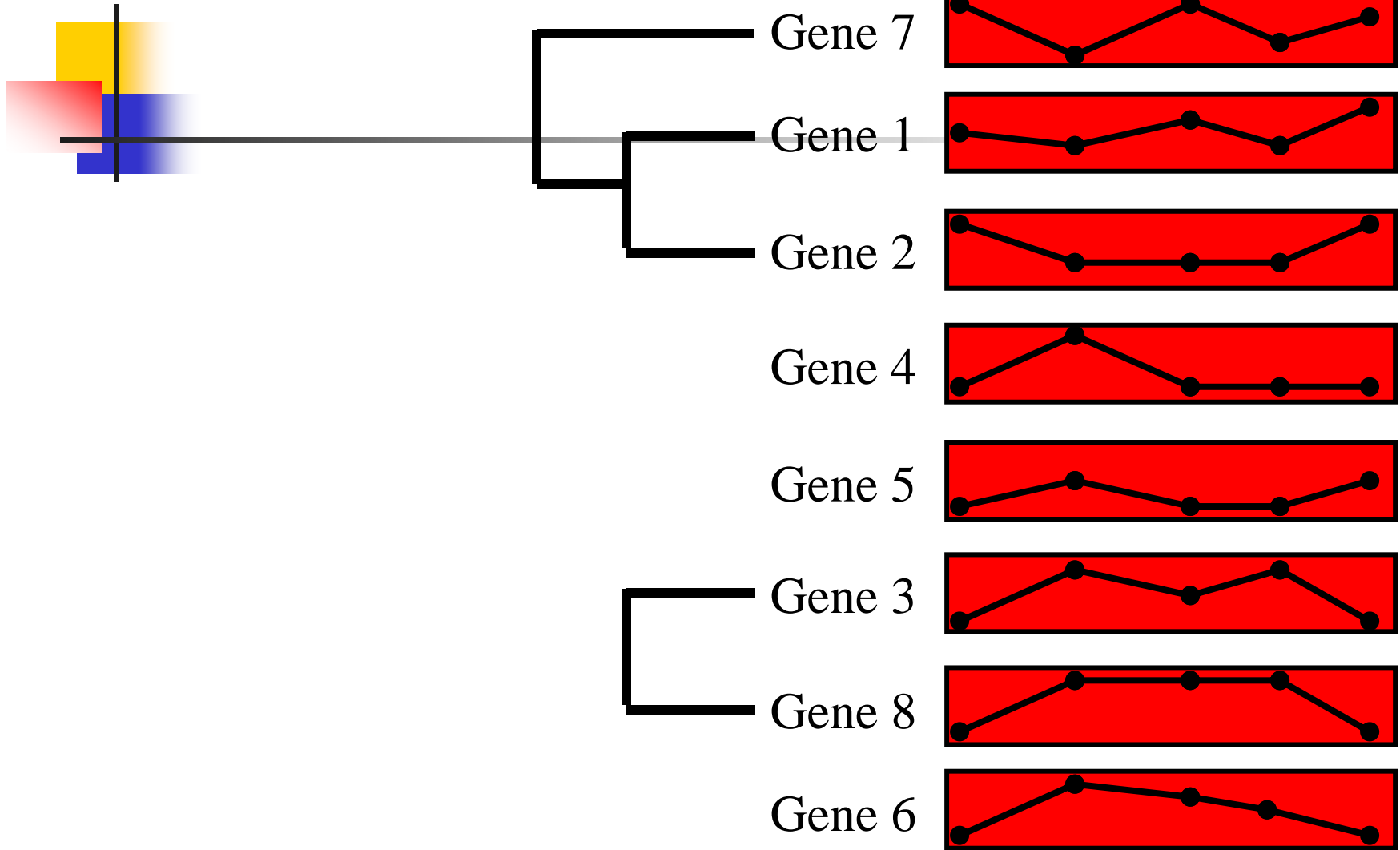
Hierarchical Clustering



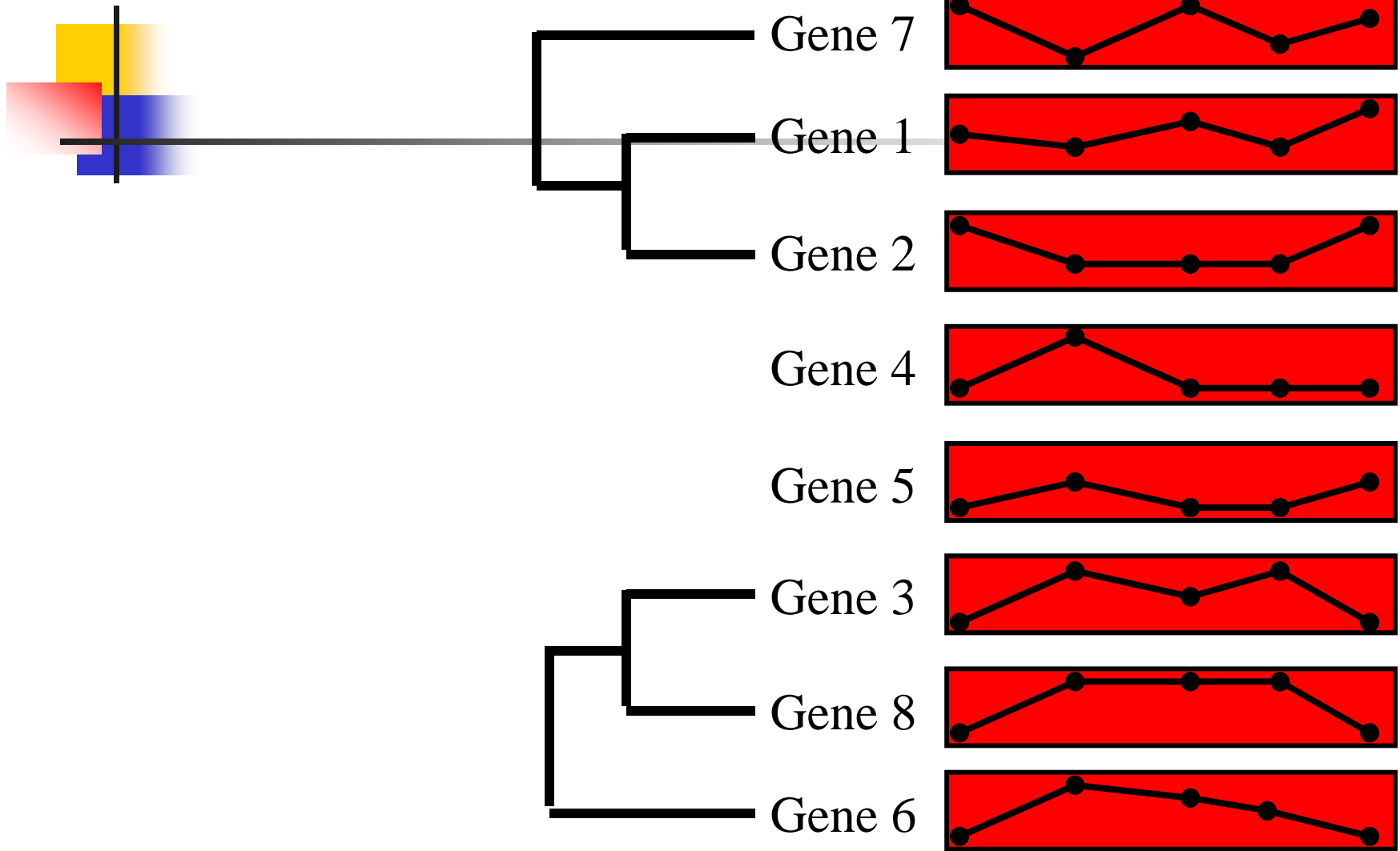
Hierarchical Clustering



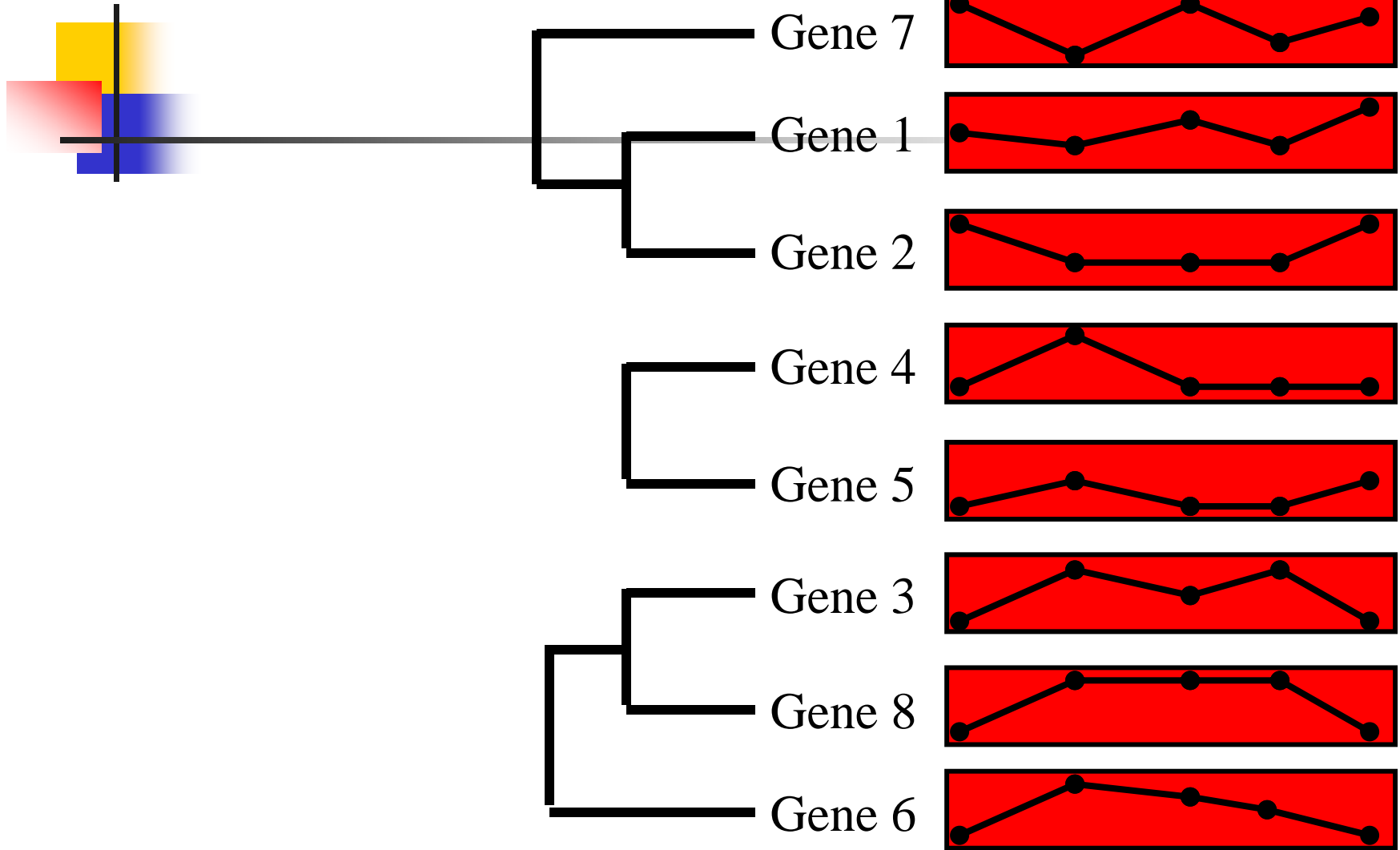
Hierarchical Clustering



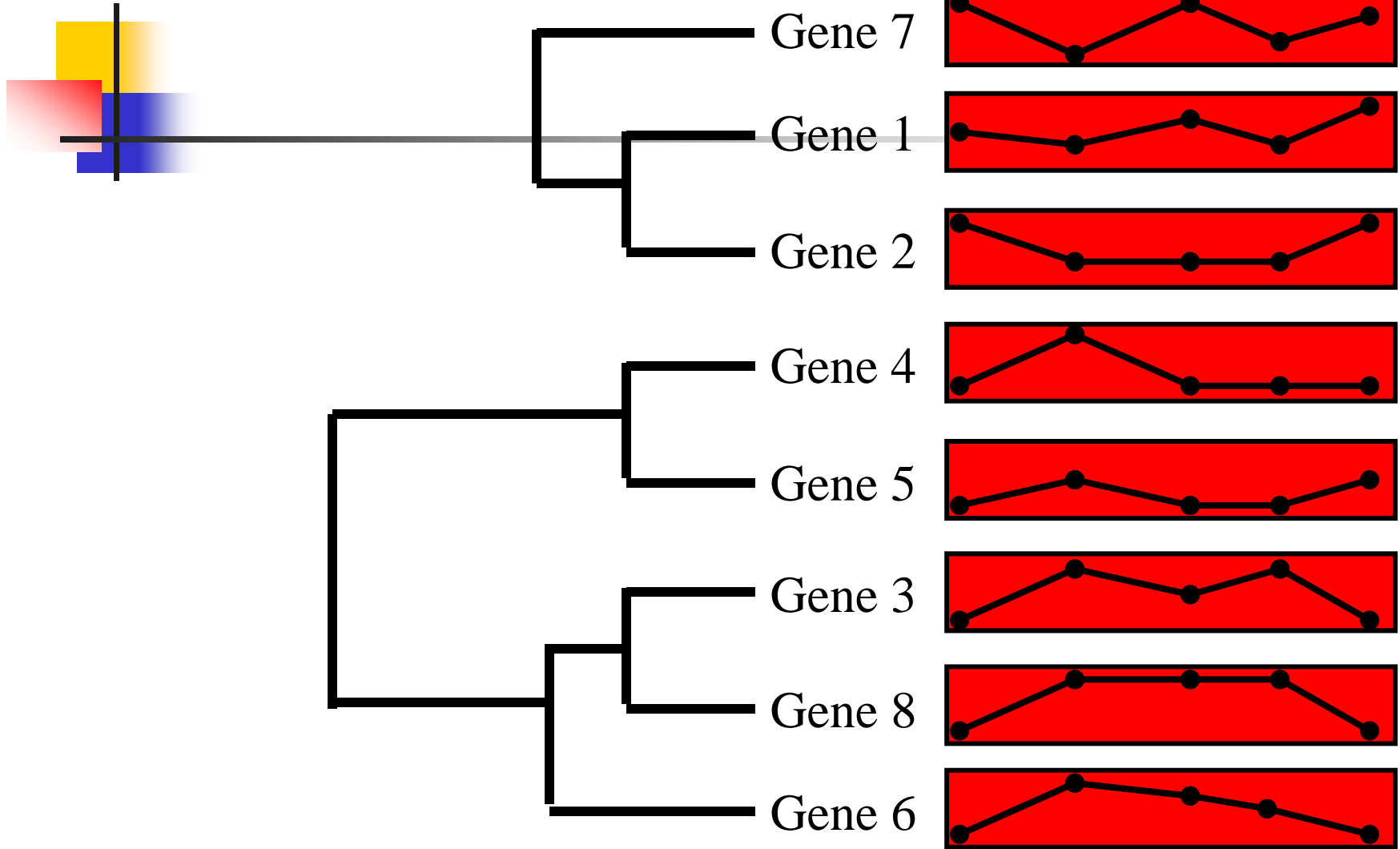
Hierarchical Clustering



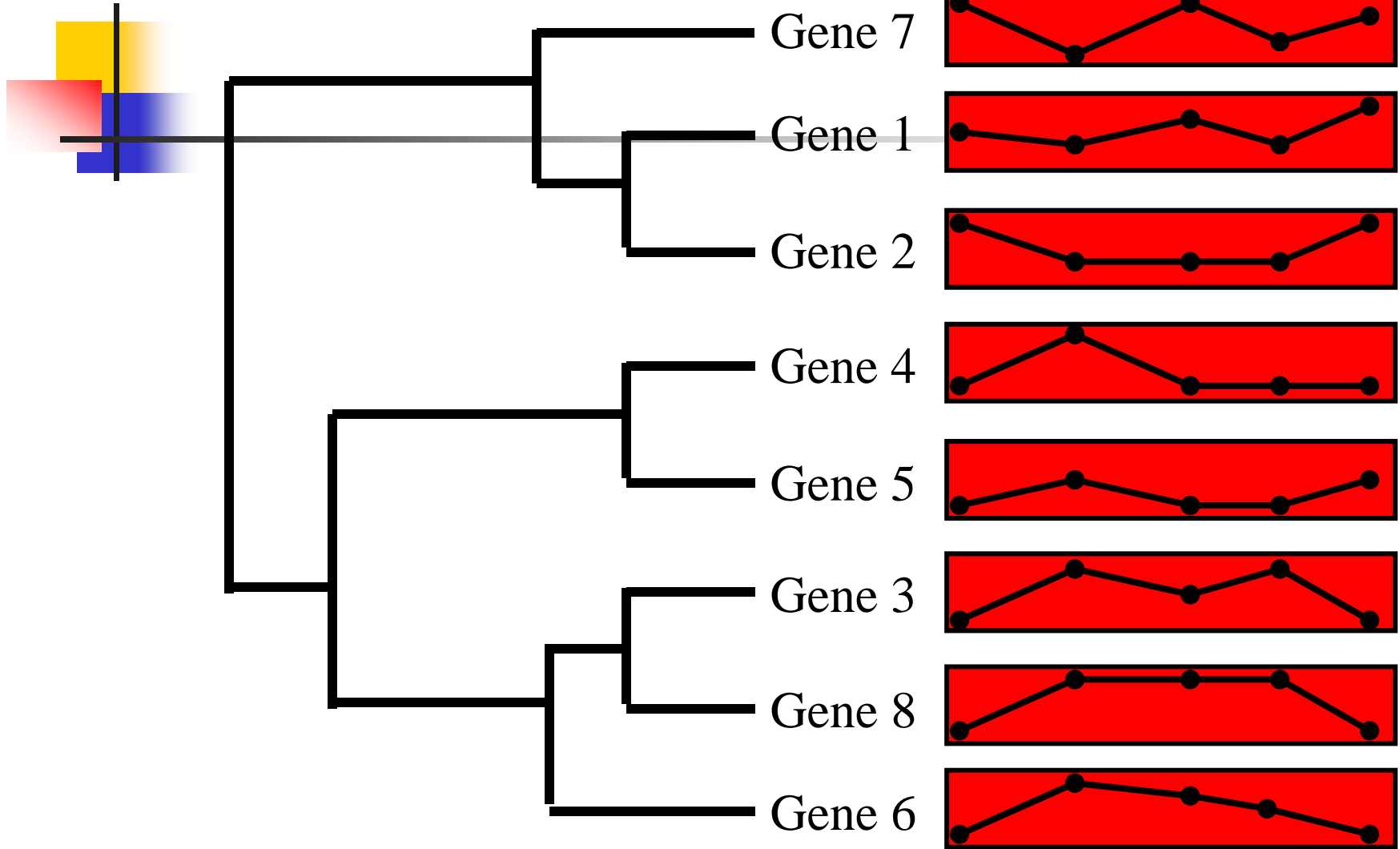
Hierarchical Clustering



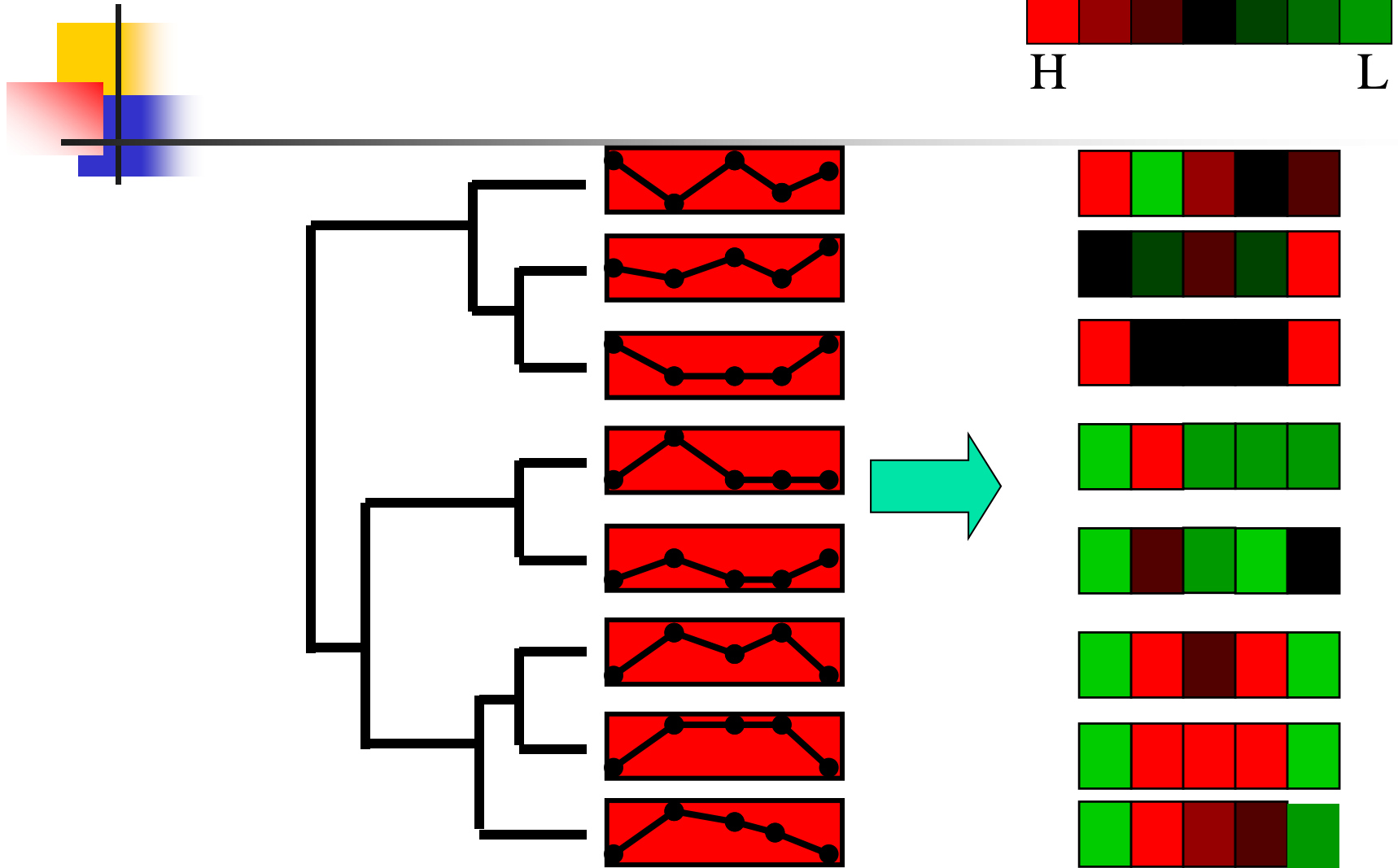
Hierarchical Clustering



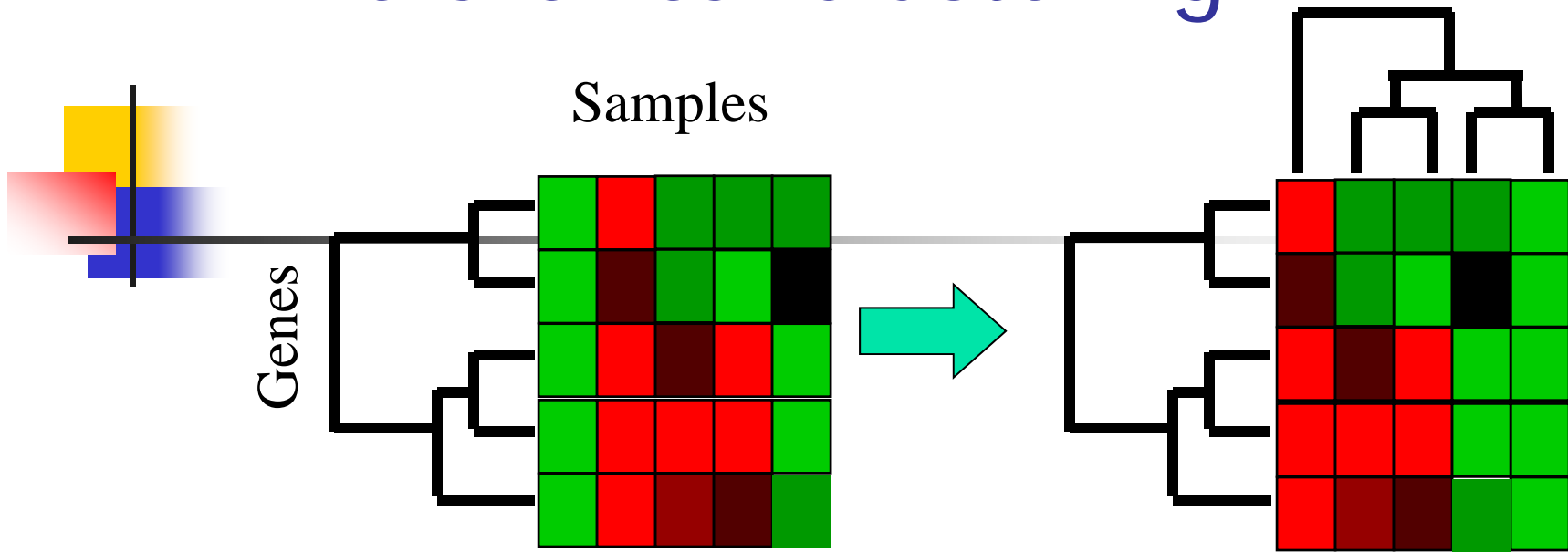
Hierarchical Clustering



Hierarchical Clustering



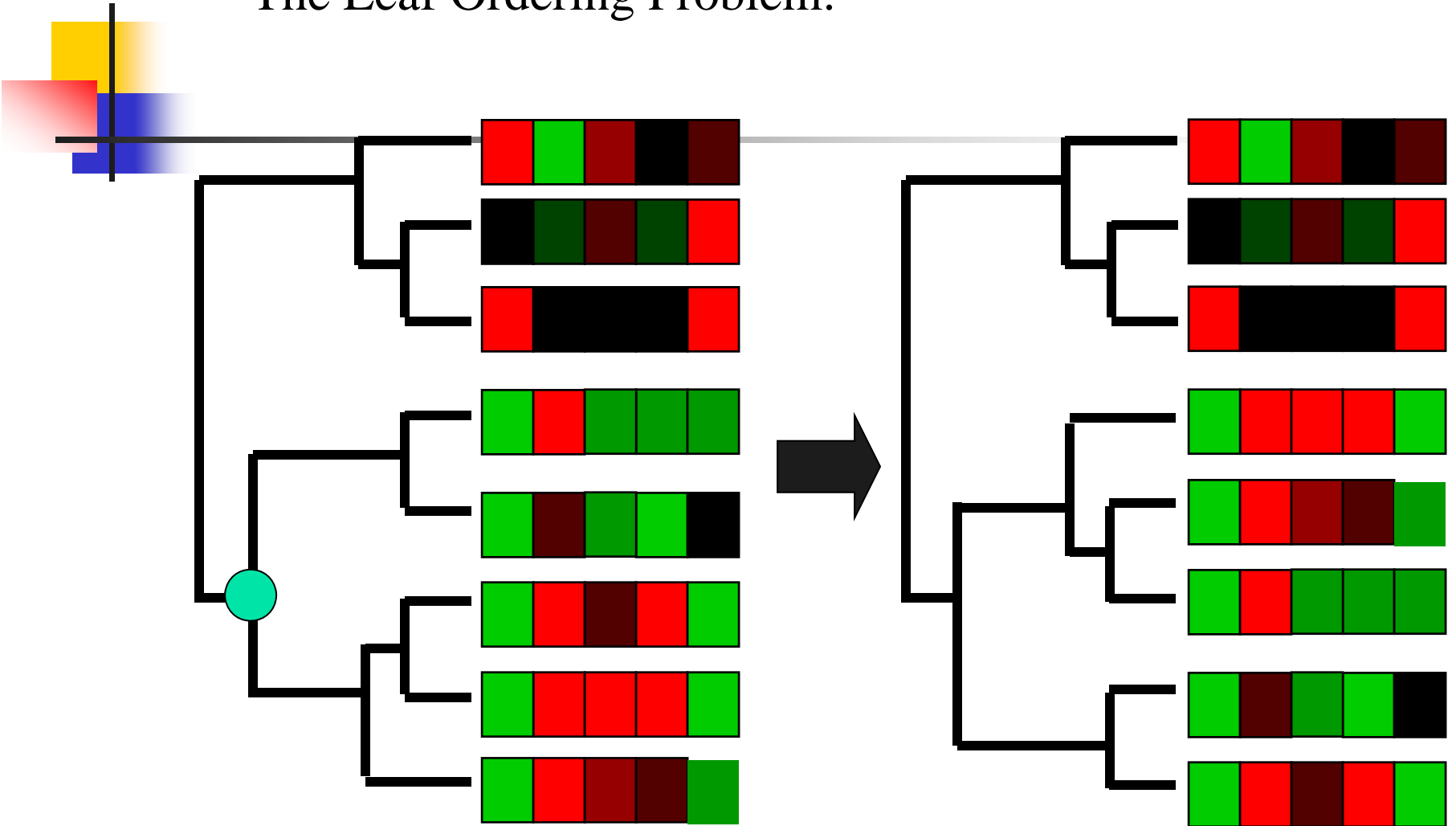
Hierarchical Clustering



The Leaf Ordering Problem:

- Find ‘optimal’ layout of branches for a given dendrogram architecture
- 2^{N-1} possible orderings of the branches
- For a small microarray dataset of 500 genes there are 1.6×10^{150} branch configurations

The Leaf Ordering Problem:

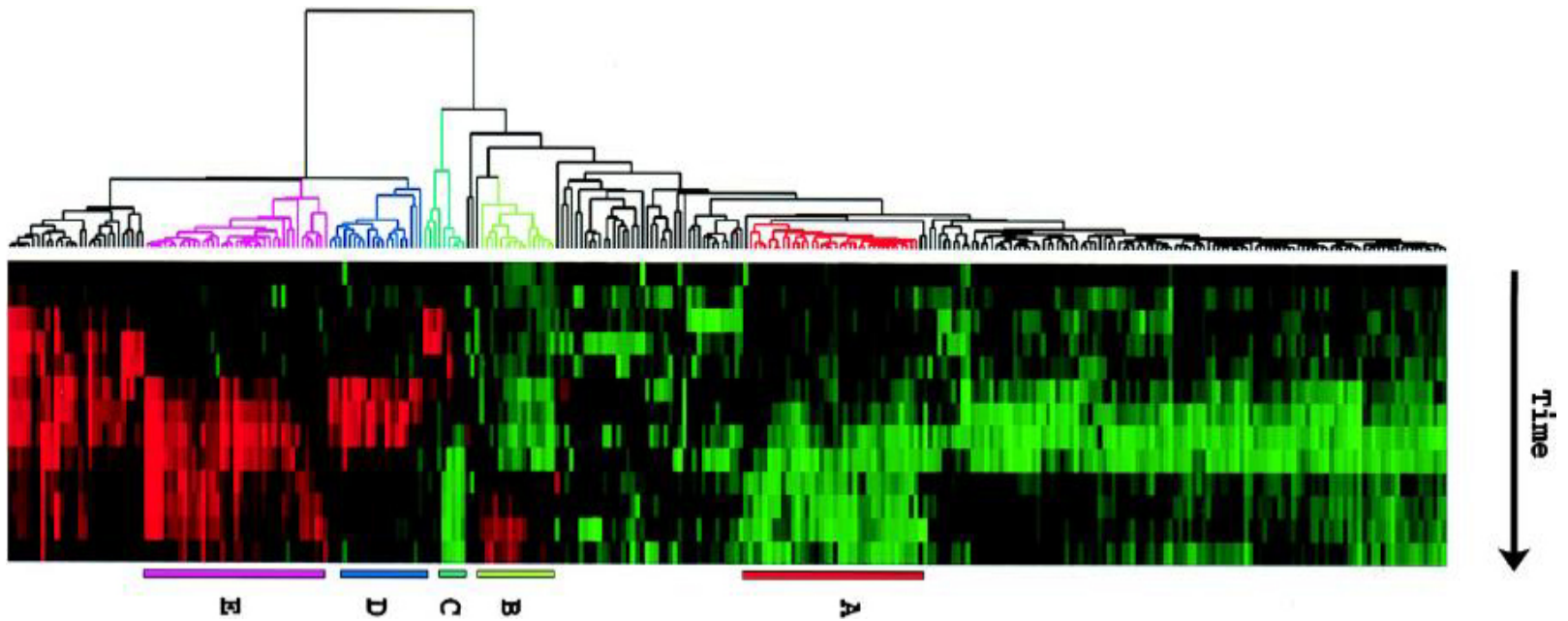




Hierarchical Clustering

- Pros:
 - Commonly used algorithm
 - Simple and quick to calculate
- Cons:
 - Real genes probably do not have a hierarchical organization

Hierarchical Clustering





GA based Fuzzy Clustering

- Automatic evolution of clusters
- Cluster centers encoded in chromosome
- Fitness computed by cluster validity index
 - Xie-Beni Index (XB)

$$XB(U, Z; X) = \frac{\sum_{i=1}^K \sum_{k=1}^n u_{ik}^2 D^2(z_i, x_k)}{n \times \min_{i \neq j} \{\|z_i - z_j\|^2\}}$$

- Genetic operations.
 - Conventional Roulette wheel selection followed by single point crossover and mutation



Necessity of having multiple objectives

- In general, clustering is a simple but difficult problem
 - For many data sets no unambiguous partitioning of the dataset exists.
 - Even if there is an unambiguous partitioning of the data set, clustering algorithms may fail
 - because those are based only on one objective function which measures either spatial separation or the compactness of the clusters.



Necessity of having multiple objectives

- Use of MOO provides a means to overcome some of the limitations of current clustering algorithm.
- If there are several objective functions for clustering
 - They indicate different characteristics of a partitioning
 - simultaneous optimization of all these objectives may lead to higher quality solutions and an improved robustness towards different data properties.



Multiobjective optimization: Mathematical definition

- The multiobjective optimization can be formally stated as: Find the vector of decision variables

$$\mathbf{x} = [x_1, x_2, \dots, x_n]$$

- which will satisfy the m inequality constraints:

$$g_i(\mathbf{x}) \geq 0, \quad i = 1, 2, \dots, m,$$

- And the p equality constraints

$$h_i(\mathbf{x}) = 0, \quad i = 1, 2, \dots, p.$$

- And simultaneously optimizes M objective functions

$$f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_M(\mathbf{x}).$$



Domination Relation and Pareto Optimality

- Let us consider two solutions \mathbf{a} and \mathbf{b} . Then \mathbf{a} is said to dominate \mathbf{b} iff

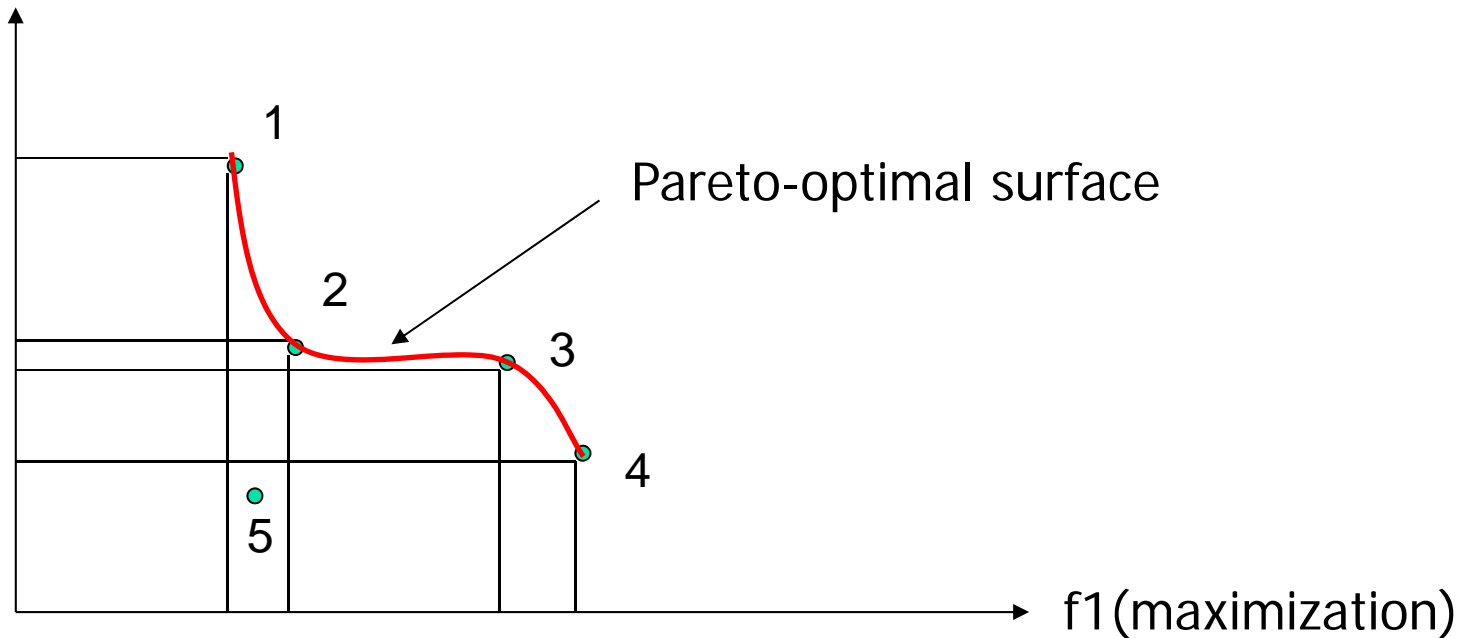
$$\forall i \in 1, 2, \dots, M, f_i(\mathbf{b}) \leq f_i(\mathbf{a}) \quad \text{and} \quad \exists j \in 1, 2, \dots, M, f_j(\mathbf{b}) < f_j(\mathbf{a})$$

i.e., for all functions f_i , \mathbf{a} has a higher or equal value than that of \mathbf{b} and also there exists at least one function f_j for which \mathbf{a} 's value is strictly greater than that of \mathbf{b} .

- Non-dominated set
 - Among a set of solutions \mathbf{P} , the non-dominated set of solutions \mathbf{P}' are those that are not dominated by any solution in the set \mathbf{P} . A solution \mathbf{a} is called non-dominating with respect to all the solutions if there exists no solution \mathbf{b} that dominates \mathbf{a} .
- Pareto-optimal Set:
 - The non-dominated set of entire search space \mathbf{S} is globally Pareto optimal set.

Example of Dominance and Pareto-Optimality

f2(maximization)



- Here solution 1, 2, 3 and 4 are non-dominating to each other.
- 5 is dominated by 2, 3 and 4, not by 1.



Multiobjective Optimization Using GAs

- Multiobjective GAs are more popular primarily because of their *population based nature*.
- Available Algorithms
 - Non-Pareto approach
 - Vector Evaluated GA (VEGA): non-Pareto
 - Pareto-based approach
 - Non-dominated Sorting GA (NSGA and NSGA-II)
 - Niched Pareto GA (NPGA)
 - Strength Pareto Evolutionary Algorithm (SPEA and SPEA2)



NSGA-II based multiobjective fuzzy clustering algorithm

- Assumption is that total number of clusters present in the data set is known a priori.
- For encoding center-based representation of clusters has been used.
 - Centers of the clusters have been encoded
 - The data points are assigned to that cluster whose center is nearest to the data point among all the centers.
- Two objective functions:
 - XB validity index
 - J_m validity index
 - Both XB and J_m are to be minimized to achieve proper clustering.

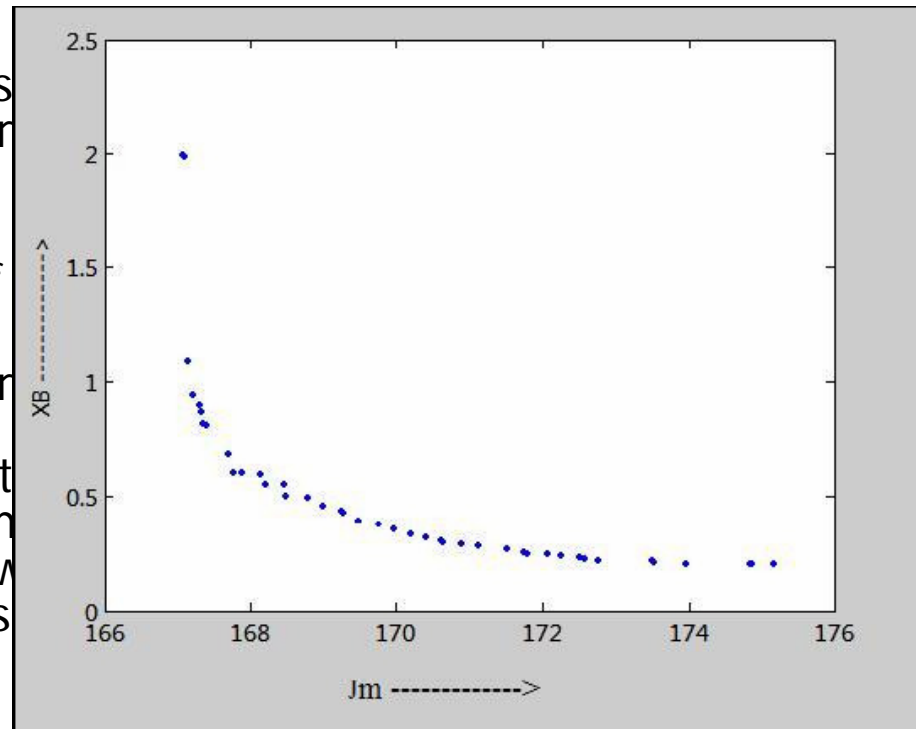


NSGA-II based multiobjective fuzzy clustering algorithm (Cont.)

- The most distinguishing feature of NSGAII is its *elitism* operation, where the non-dominated solutions among the parent and child populations are propagated to the next generation.
- Selection: crowded tournament method
- Conventional crossover and mutation.
- The algorithm is run for fixed number of generations and at each generation, population size is kept constant.

Choice of objectives

- The chosen two objectives XB and J_m are contradictory in nature.
- they represent somehow different characteristics of data.
- J_m computes global cluster variance whereas XB considers both global cluster variance and the minimum separation between any two cluster centers. Hence it is combination of global and worst cases.



Pareto front for Sporulation data



Experimental results

Data Sets	No. of genes	No. of time points	No. of clusters
Yeast Sporulation	6118	7	7
Human Fibroblasts Serum	517	13	10



Experimental results (Cont.)

- The Sporulation data is filtered to ignore the genes whose expression level didn't change significantly across different time points. After filtering, 474 prominently expressed genes are found.
- Both the data set is normalized so that each row has mean 0 and variance 1.



Experimental results (Cont.)

- Performance metric: *Silhouette index*
 - Silhouette width of a point is defined as:

$$s = \frac{b - a}{\max\{a, b\}}$$

- **a**: the average distance of the point from the other points of the cluster to which the point is assigned.
 - **b**: the minimum of the average distances of the point from the points of the other clusters.
- Silhouette index is the average silhouette width of all the data points (genes). It ranges between -1 and 1, and larger value indicates better solution.



Experimental results (Cont.)

- From the final non-dominated set of solutions produced by multiobjective clustering, the solution that gives the best Silhouette index value is chosen.
- Finally a point is assigned to the cluster to which it has highest membership degree.
- Performance has been compared with FCM, single objective clustering that minimizes XB index and Average linkage clustering algorithms.
- Input Parameters:
 - Population size = 50, No. of generations = 100, crossover probability = 0.8 and mutation probability = $1/\text{length of chromosome}$
 - FCM has been run for maximum 100 iterations with $m = 2$.

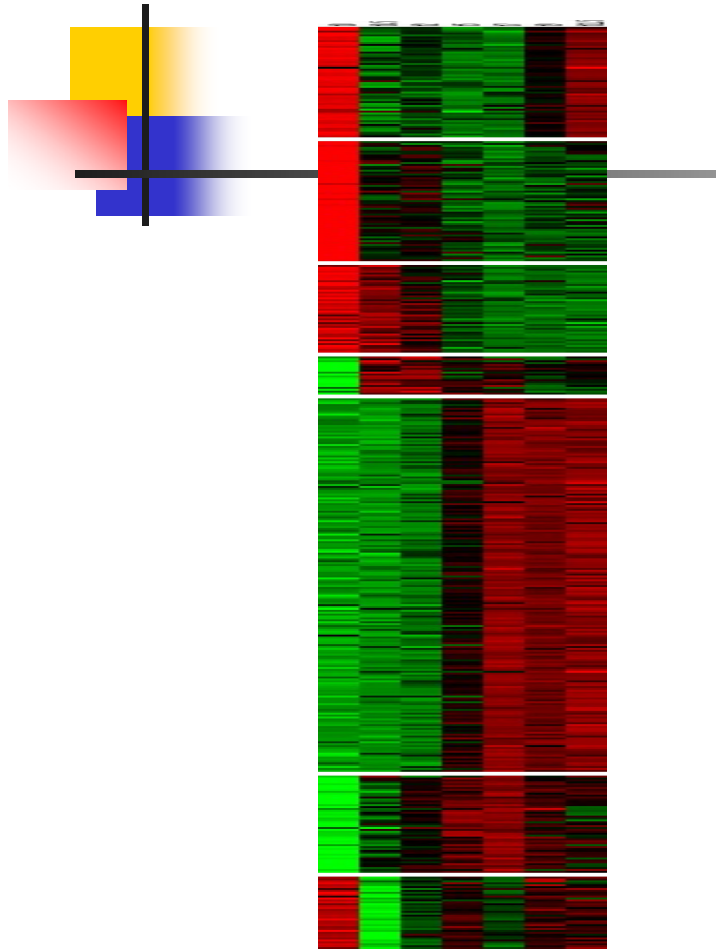


Experimental results (Cont.)

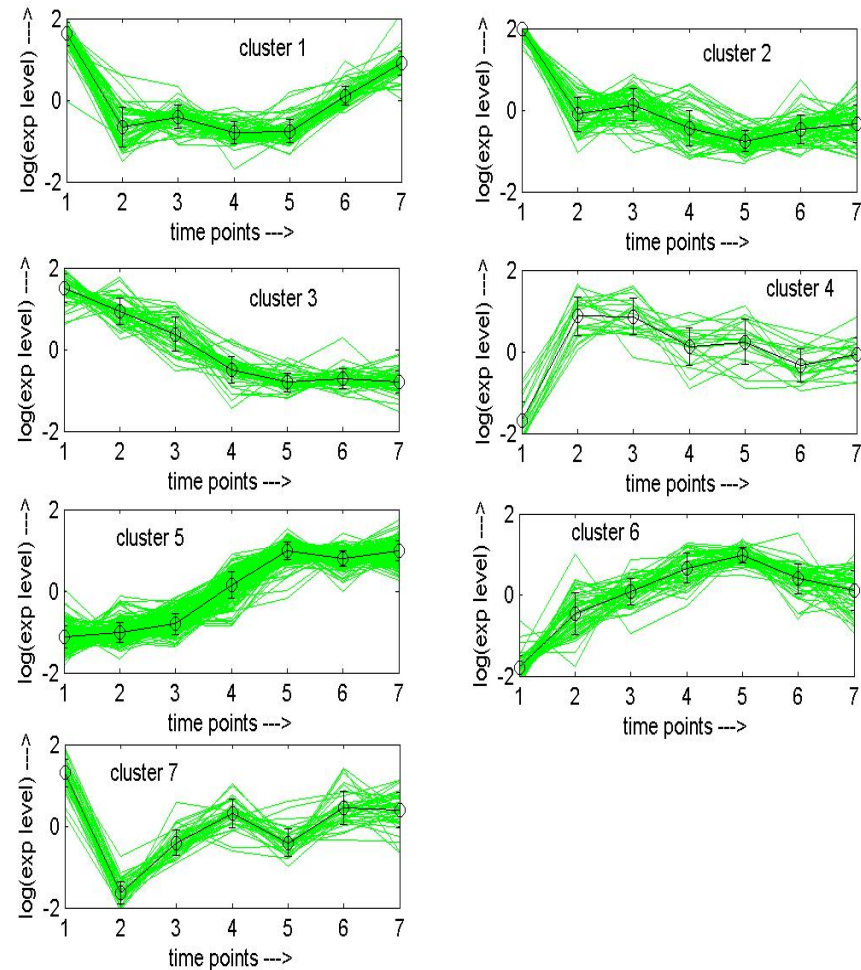
Algorithm	Data set	
	Sporulation	Serum
FCM	0.5879	0.3304
Average Linkage	0.5007	0.2977
Single objective GA minimizing XB index	0.5837	0.3532
NSGAI based multiobjective clustering	0.6465	0.4135

Silhouette index values for different algorithms on Sporulation and Serum data sets

Visualizing clustering results



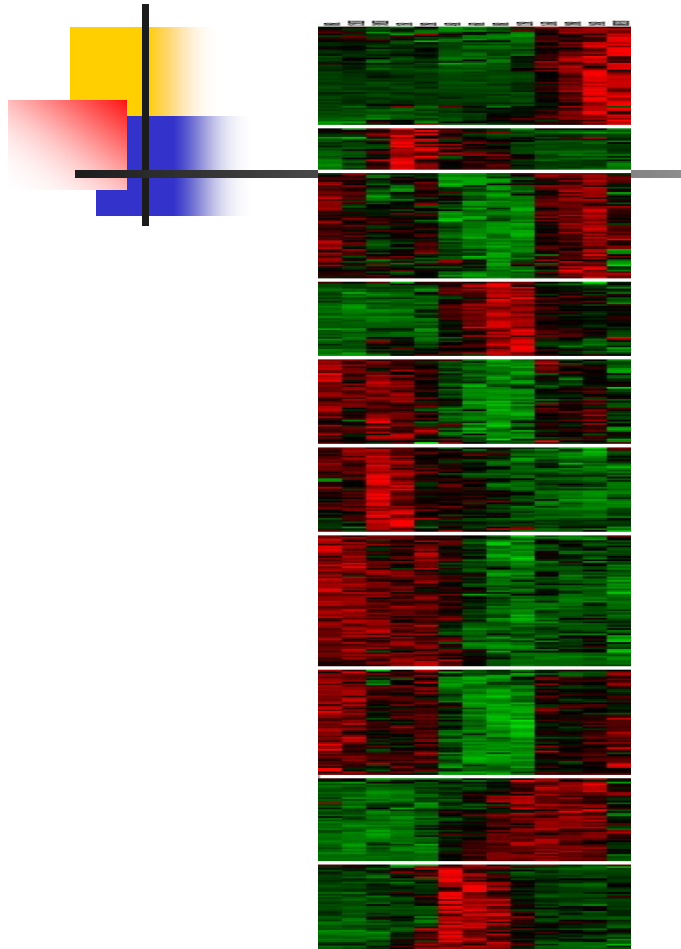
(a)



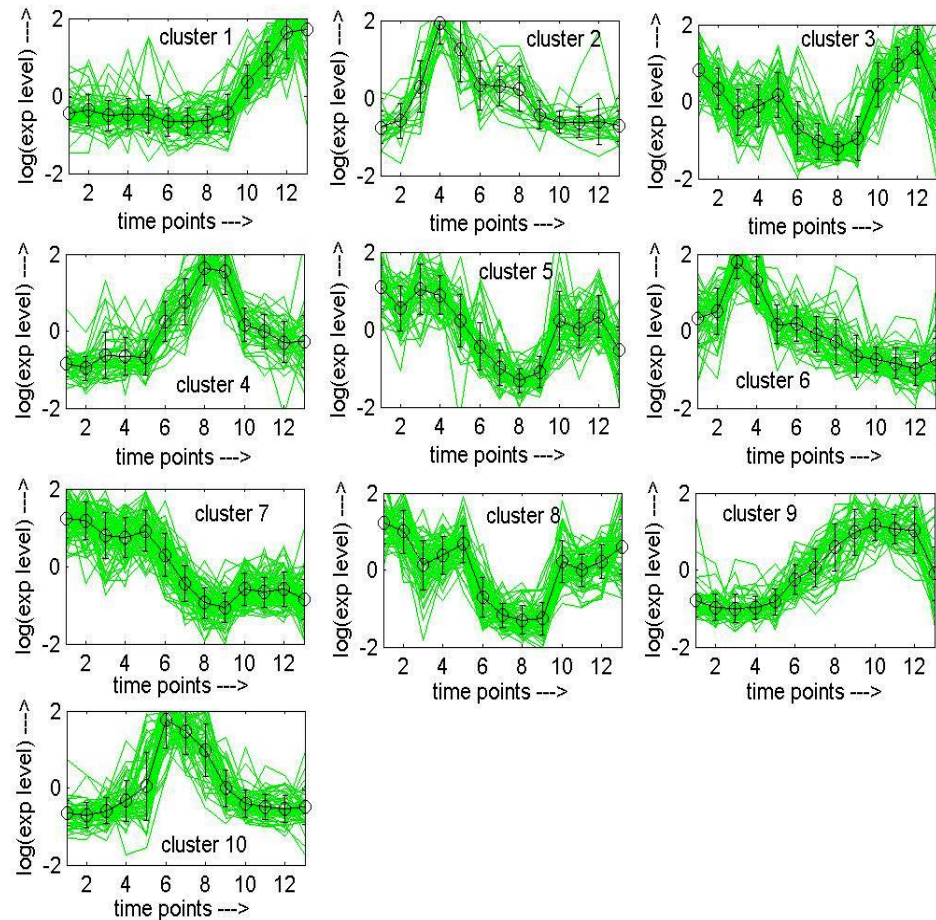
(b)

Sporulation data clustered using multiobjective clustering (7 clusters):
(a) Eisen plot, (b) Cluster profile plots.

Visualizing clustering results (Cont.)



(a)



(b)

Serum data clustered using multiobjective clustering (10 clusters):
(a) Eisen plot, (b) Cluster profile plots.



Conclusion

- NSGA-II based multiobjective fuzzy clustering technique for Microarray data is described.
- Use of other objective functions, may be more than two, needs to be studied.
- Comparative study with other multiobjective optimization strategies is to be made.



References

- M. Schena, "Microarray Analysis", Wiley-Liss, 2002.
- G. J. McLachlan, Analyzing microarray gene expression data, Wiley series in probability and statistics, 2004.
- S. Draghici, Data Analysis Tools for DNA Microarrays, Chapman & Hall /CRC Press, 2003.
- A. A. Alizadeh et al, "Distinct types of diffuse large B-cells lymphomas identified by gene expression profiles", Nature, vol. 403, pp. 503-511, 2000
- *Nature Genetics Supplements*: Special issue on Microarrays, vol. 21, 1999.
- U. Maulik, S. Bandyopadhyay and A. Mukhopadhyay, Multiobjective Genetic Algorithms for Clustering: Applications in Data Mining and Bioinformatics, Springer, Heidelberg, Germany, 2011.
- U. Maulik, S. Bandyopadhyay and J. T. L. Wang, Computational Intelligence and Pattern Analysis in Biological Informatics, Wiley, Singapore, 2010.
- S. Bandyopadhyay, U. Maulik and J. T. L. Wang, Analysis of Biological Data: A Soft Computing Approach, World Scientific, Singapore, 2007.



References

- U. Maulik, A. Mukhopadhyay, D. Chakraborty, "Gene-expression Based Cancer Subtypes Prediction through Feature Selection and Transductive SVM", IEEE Transaction on Biomedical Engineering (accepted).
- A. Mukhopadhyay and U. Maulik, "An SVM-wrapped Multiobjective Evolutionary Feature Selection Approach for Identifying Cancer-MicroRNA Markers", IEEE Transactions on NanoBioScience, 2013 (accepted).
- U. Maulik, A. Mukhopadhyay, M. Bhattacharyya, L. Kaderali, B. Brors, S. Bandyopadhyay, and R. Eils, "Mining Quasi-Bicliques from HIV-1-Human Protein Interaction Network: A Multiobjective Biclustering Approach", IEEE/ACM Transaction on Computational Biology and Bioinformatics (accepted).
- A. Mukhopadhyay, U. Maulik and S. Bandyopadhyay, "An Interactive Approach to Multiobjective Clustering of Gene Expression Patterns", IEEE Transaction on Biomedical Engineering, vol. 60, no. 1, pp. 35-41, 2013.
- U. Maulik, A. Mukhopadhyay and S. Bandyopadhyay, "Finding Multiple Coherent Biclusters in Microarray Data using Variable String Length Multiobjective Genetic Algorithm", IEEE Transactions on Information Technology in BioMedicine, vol. 13, no. 6, pp. 969-975, 2009.



References

- S. Bandyopadhyay, U. Maulik and D. Roy, "Gene Identification: Classical and Computational Intelligence Approaches", IEEE Transactions on Systems, Man and Cybernetics, Part C, pp. 55-68, 2008.
- A. Bhar, M. Haubrock, A. Mukhopadhyay, U. Maulik, S. Bandyopadhyay, and E. Wingender, "Coexpression and Coregulation Analysis of Time-Series Gene Expression Data in Estrogen-Induced Breast Cancer Cell", Algorithms for Molecular Biology, Vol. 8, No. 9, 2013.[Highly Accessed]
- U. Maulik, M. Bhattacharyay, A. Muhopadhyay and S. Bandyopdhyay, "Identifying the Immunodeficiency Gata way Proteins in Human and their involvement in microRNA regulation", Molecular Biosystem (Royal Society of Chemistry), Molecular Biosyst., vol. 7, no. 6, pp. 1842-1851, 2011.
- I. Saha, U. Maulik, S. Bandyopadhyay and D. Plewczynsk, "Fuzzy Clustering of Physicochemical and Biochemical Properties of Amino Acids", Amino Acids, Vol. 43, No. 2, pp. 583-594, 2012.
- A. Mukhopadhyay, S. Bandyopadhyay and U. Maulik, "Multiclass Clustering of Cancer Subtypes through SVM based Ensemble of Pareto-optimal Solutions for Gene Marker Identification", PLoS One, vol. 5, no. 11, 2010.



References

- A. Mukhopadhyay, U. Maulik and S. Bandyopadhyay, "On Biclustering of Gene Expression Data", Current Bioinformatics, vol. 5, no. 3, pp. 204-216, 2010.
- U. Maulik, A. Mukhopadhyay and S. Bandyopadhyay, "Combining Pareto-Optimal Clusters using Supervised Learning for Identifying Co-expressed Genes", BMC Bioinformatics, 10:27, 2009.
- A. Mukhopadhyay, U. Maulik and S. Bandyopadhyay, "A Novel Coherence Measure for Discovering Scaling Biclusters from Gene Expression Data", Journal of Bioinformatics and Computational Biology, vol. 7, no. 5, pp. 853-868, 2009.
- S. Bandyopadhyay, A. Mukhopadhyay and U. Maulik, "An Improved Algorithm for Clustering Gene Expression Data", Bioinformatics, Oxford University Press, vol. 23, no. 21, pp. 2859-2865, 2007.
- S. Bandyopadhyay, A. Bagchi and U. Maulik, "Active Site Driven Ligand Design: An Evolutionary Approach", Journal of Bioinformatics and Computational Biology, vol. 3, no. 5, pp. 1053-1070, 2005.



Questions ???

Thank You

