# Bioinformatics

# Microarrays: designing chips, clustering methods

Fran Lewitter, Ph.D.
Head, Biocomputing
Whitehead Institute

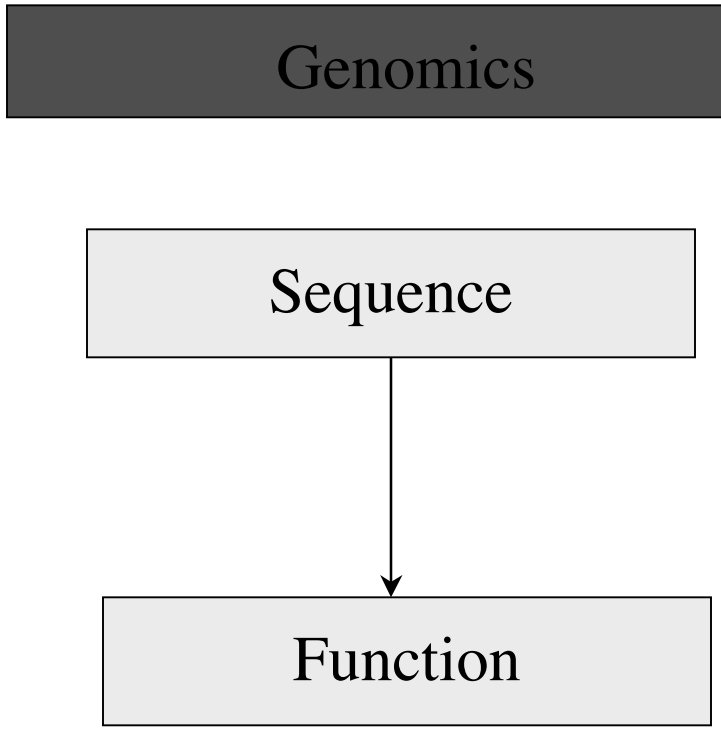
# Course Syllabus

| | |
|---|---|
| Jan 7 | Sequence Analysis I. Pairwise alignments, database searching including BLAST (FL) [1, 2, 3] |
| Jan 14 | Sequence Analysis II. Database searching (continued), Pattern searching(FL)[7] |
| Jan 21 | No Class - Martin Luther King Holiday |
| Jan 28 | Sequence Analysis III. Hidden Markov models, gene finding algorithms (FL)[8] |
| Feb 4 | Computational Methods I. Genomic Resources and Unix (GB) |
| Feb 11 | Computational Methods II. Sequence analysis with Perl. (GB) |
| Feb 18 | No Class - President's Birthday |
| Feb 25 | Computational Methods III. Sequence analysis with Perl and BioPerl (GB) |
| Mar 4 | Proteins I. Multiple sequence alignments, phylogenetic trees (RL) [4, 6] |
| Mar 11 | Proteins II. Profile searches of databases, revealing protein motifs (RL) [9] |
| Mar 18 | Proteins III.Structural Genomics:structural comparisons and predictions (RL) |
| Mar 25 | Microarrays: designing chips, clustering methods (FL) |

# Today's Outline

- Intro
- Designing microarrays
- Working with microarray data
  - Normalization
  - Analysis
    - Distance metrics
    - Clustering methods

# Research Trends




- How are genes regulated?
- How do genes interact?
- What are the functional roles of different genes?
- How does expression level of a gene differ in different tissues?

# Transcriptional Profiling

(Adapted from Quackenbush 2001)

- Study of patterns of gene expression across many experiments that survey a wide array of cellular responses, phenotypes and conditions
- Simple analysis - what's up/down regulated?
- More interesting - identify patterns of expression for insight into function, etc.

# Microarray Data

Collect data on *n* DNA samples (e.g. rows, genes, promotors, exons, etc.) for *p* mRNA samples of tissues or experimental conditions (eg. columns, time course, pathogen exposure, mating type, etc)

Matrix (*n x p)* =

$$\begin{bmatrix} x_{11} & x_{12} & \ldots. & x_{1p} \\ x_{21} & x_{22} & \ldots. & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ x_{n1} & x_{n2} & \ldots. & x_{np} \end{bmatrix}$$

# Multivariate Analysis

Concerned with datasets with more than one response variable for each observational or experimental unit (e.g. matrix X with *n* rows (genes) and *p* columns (tissue types))

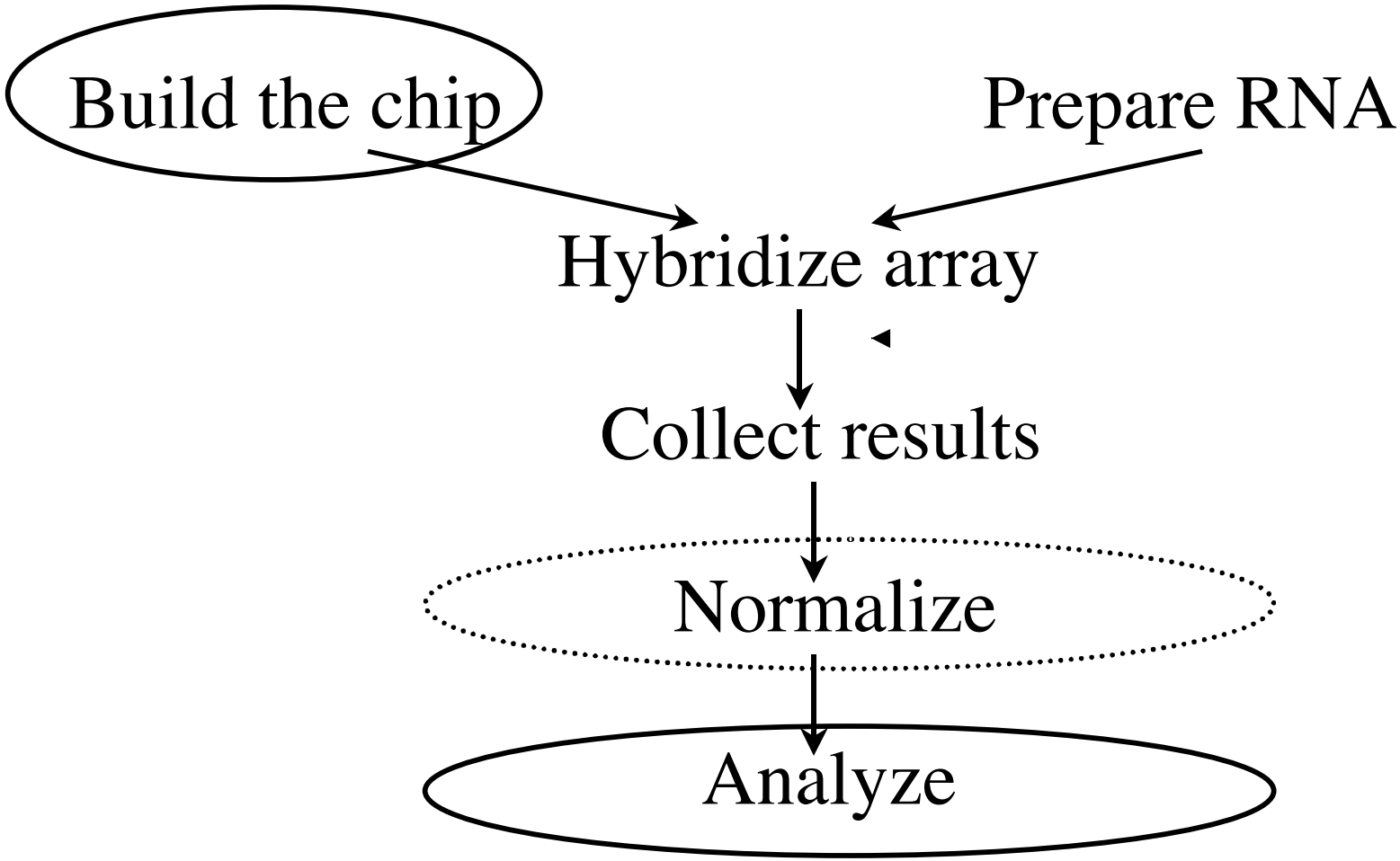
- Hierarchical (phylogenetic trees) vs non-hierarchical (k-means)
- Divisive vs agglomerative
- Supervised vs unsupervised
  - Divide cases into groups vs discover structure of data

# Multivariate Methods

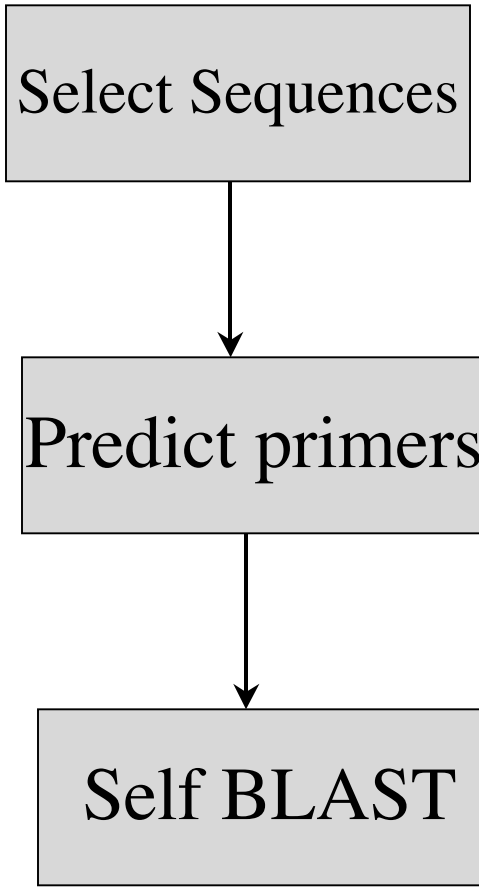
- Cluster analysis - discover groupings among cases of X
  - Hierarchical produces dendograms
  - K-means - choose a prespecified number of clusters
  - Self Organizing Maps

- Principal component analysis (PCA)
  - Linear method, unsupervised, seeks linear combinations of the columns of X with maximal (or minimal) variance (graphical)

# DNA Microarrays



# Designing a Microarray




- Download sequences from Entrez
- Predict genes from genomic sequence
- Locate/predict specific region of genes (e.g. promoter, exons, etc)

WIBR Primer3 (Rozen & Skaletsky)

# Sources of variation

wanted vs unwanted

Across experimental conditions

Chip, slide

Hybridization conditions

Imaging

# Normalization

(Adapted from Quackenbush 2001)

Compensate for experimental variability

- Total intensity normalization
  - assumes the quantity of initial mRNA is same for labelled samples
- Normalization using regression technique
- Normalization using ratio statistics

# After normalization

(Adapted from Quackenbush 2001)

- Data reported as an "expression ratio" or as a logarithm of the expression ratio

- Expression ratio is the normalized value of the expression level for a particular gene in the query sample divided by its normalized value for the control

- Use log of expression ratio for easier comparisons

# Distance Metrics

(Adapted from Quackenbush 2001)

- Metric distances - $d_{ij}$ between two vectors, $i$ and $j$,must obey several rules:
  - Distance must be positive definite, $dij \geq 0$
  - Distance must be symmetric, $d_{ij} = d_{ji}$, so that the distance from $i$ to $j$ is the same as the distance from $j$ to $i$.
  - An object is zero distance from itself, $d_{ii} = 0$.
  - When considering three objects, $i$, $j$ and $k$, $d_{ik} \leq d_{ij} + d_{jk}$. This is sometimes called the 'triangle' rule.

# Distance Metrics

(Adapted from Quackenbush 2001)

- The most common metric distance is Euclidean distance,which is a generalization of the familiar Pythagorean theorem. In a three-dimensional space, the Euclidean distance, $d_{12}$, between two points, $(x_1,x_2,x_3)$ and $(y_1,y_2,y_3)$ is given by:

$$d_{12} = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2},$$

- where $(x_1,x_2,x_3)$ are the usual Cartesian coordinates $(x,y,z)$.

# More on distance

(Adapted from Quackenbush 2001)

The generalization of this to higher-dimensional expression spaces is straightforward.

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2},$$

where $x_i$ and $y_i$ are the measured expression values, respectively, for genes X and Y in experiment $i$, and the summation runs over the $n$ experiments under analysis.

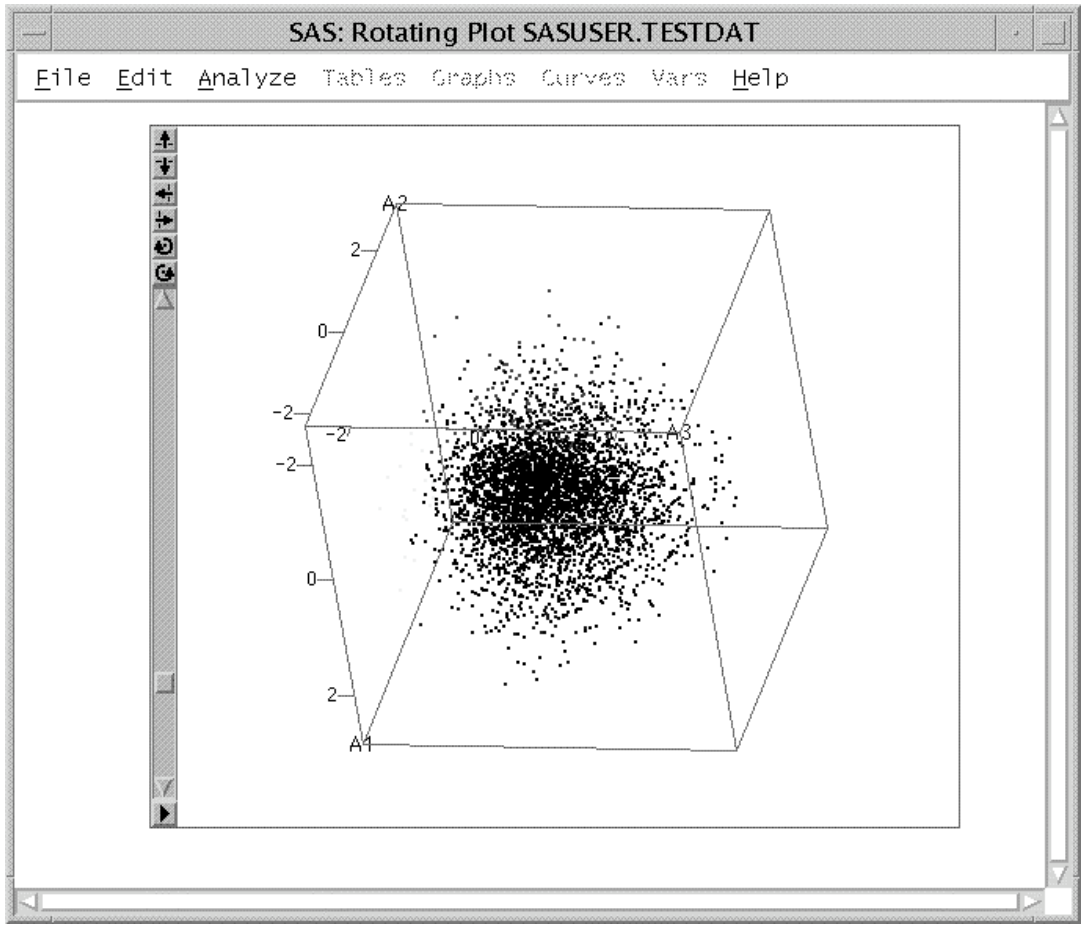
# Semi-metric distances

(Adapted from Quackenbush 2001)

- Distance measures that obey the first three consistency rules, but fail to maintain the triangle rule are referred to as semi-metric.

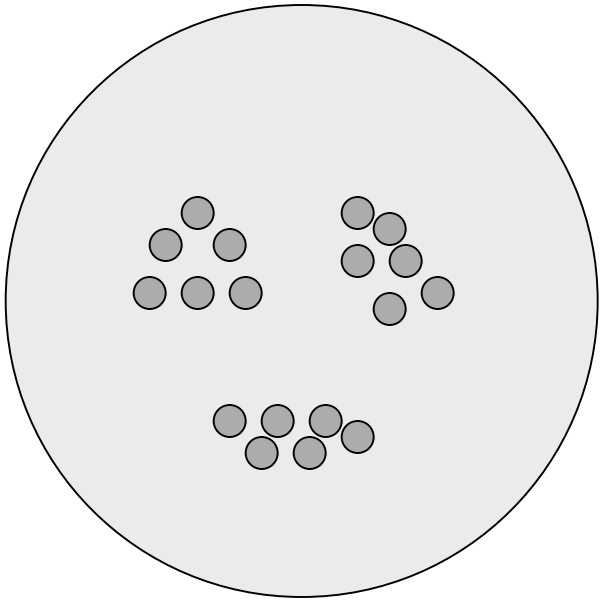
- Pearson correlation coefficient is most commonly used semi-metric distance measure

# Example of 3-D Plot

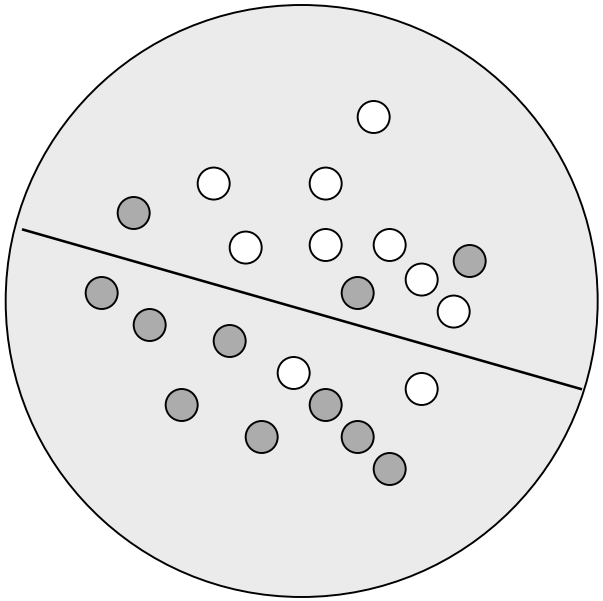


# Clustering Algorithms

(Adapted from Brazma, 2000)



Unsupervised



Supervised

# Hierarchical methods

(Adapted from Dudoit and Gentleman, 2002)

- Produces a tree or dendogram
- Don't need to specify how many clusters
- The tree can be built in two distinct ways
  - bottom-up: agglomerative clustering
  - top-down: divisive clustering

# Agglomerative methods

(Adapted from Dudoit and Gentleman, 2002)

- Start with *n* mRNA sample clusters
- At each step, merge two closest clusters using a measure of between-cluster dissimilarity reflecting shape of the clusters
- Between-cluster dissimilarity measures
  - Unweighted Pair Group Method with Arithmetic mean (UPGMA): average of pairwise dissimilarities
  - Single-link: minimum of pairwise dissimilarities
  - Complete-link: maximum of pairwise dissimilarities

# Divisive methods

(Adapted from Dudoit and Gentleman, 2002)

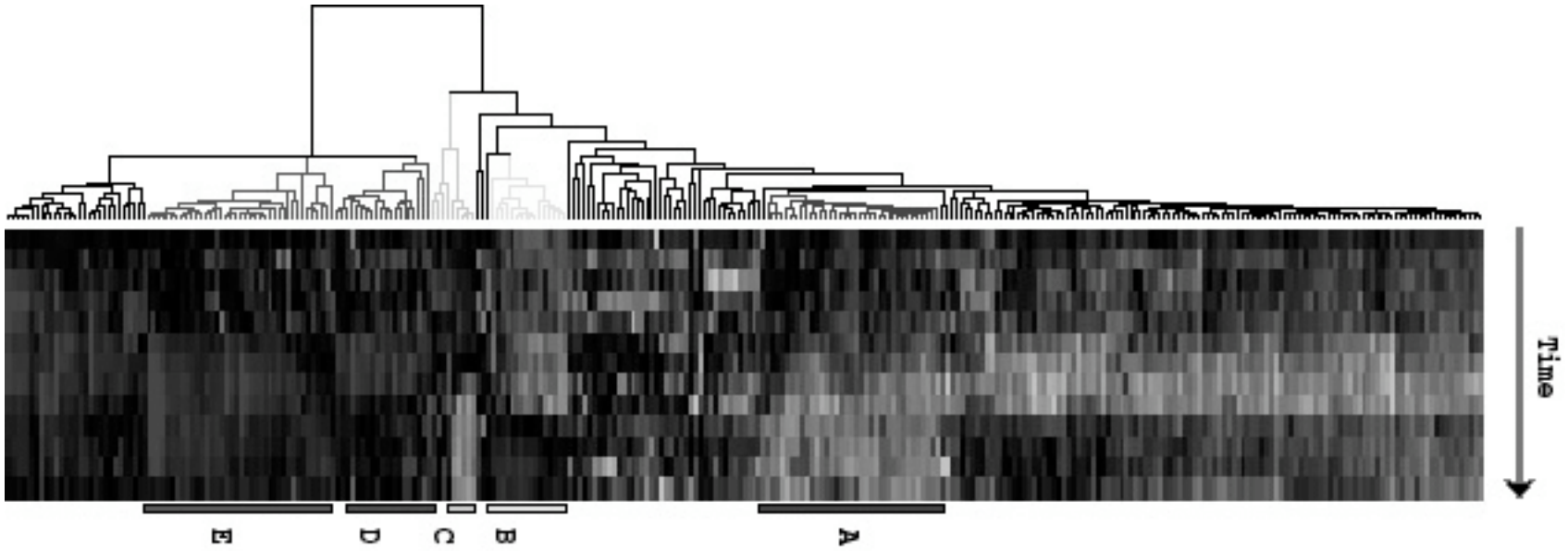
- Start with only one cluster
- At each step, split clusters into parts
- Advantages: obtain main structure of the data, i.e., focus on upper levels of dendogram
- Disadvantages: computational difficulties when considering all possible divisions into two groups

# Hierarchical Clustering

(Adapted from Quackenbush 2001)

- Agglomerative - single expression profiles are joined to form groups….forming a single tree
  - Pairwise distance matrix is calculated for all genes to be clustered
  - Distance matrix is searched for the 2 most similar genes or clusters
  - Two selected clusters are merged to produce new cluster
  - Distances calculated between this new cluster and all other clusters

# Dendogram




Eisen et al 1998

# K-means Clustering

(Adapted from Quackenbush 2001)

- Divisive - good if you know the number (*k*) of clusters to be represented in the data
  - Initial objects randomly assigned to one of *k* clusters
  - Average expression vector calculated for each cluster & compute distance between clusters
  - Objects moved between clusters and intra- and inter-cluster distances are measured with each move
  - Expression vectors for each cluster are recalculated
  - Shuffling proceeds until moving any more objects would make clusters more variable (> intra-cluster distances and decreasing inter-cluster dissimilarity

# Self Organizing Maps (SOM)

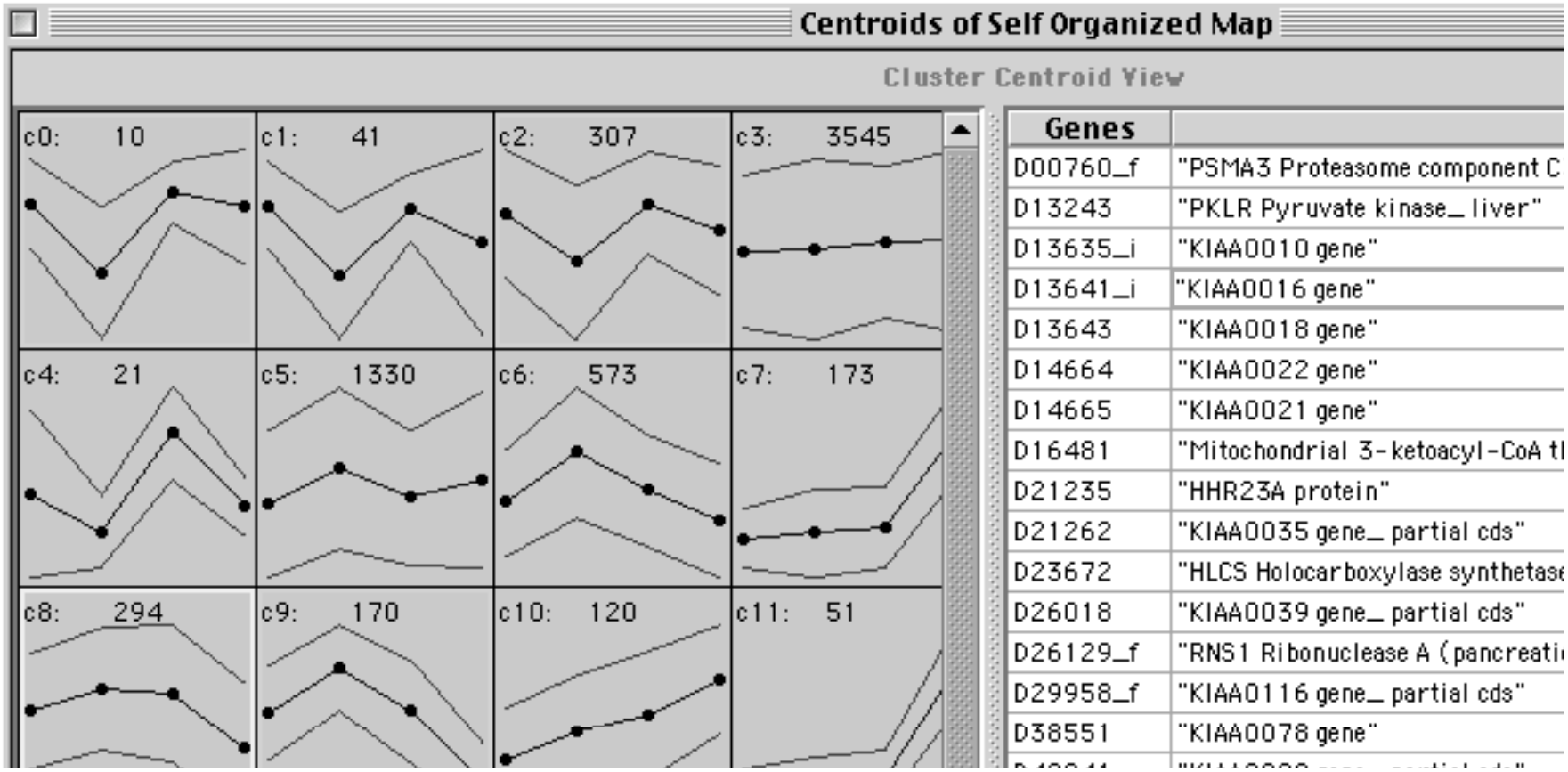
(Adapted from Quackenbush 2001)

- Neural-network based divisive clustering approach
  - Assigns genes to a series of partitions
  - User defines a geometric configuration for the partitions
  - Random vectors are generated for each partition
  - Vectors are first 'trained' using an iterative process until data most effectively separated

# SOMs Continued

- Random vectors are constructed and assigned to each partition
- A gene is picked at random and, using a selected distance metric, the reference vector that is closest to the gene is identified
- The reference vector is then adjusted so that it is more similar to the vector of the assigned gene
- Genes are mapped to relevant partitions depending on the reference vector to which they are most similar

# SOMs from GeneCluster

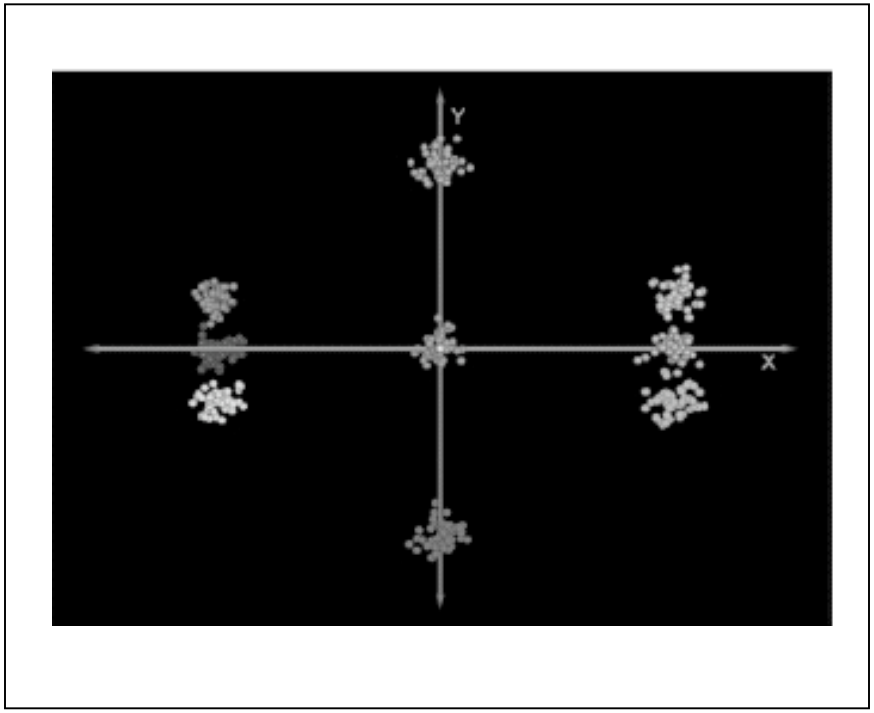


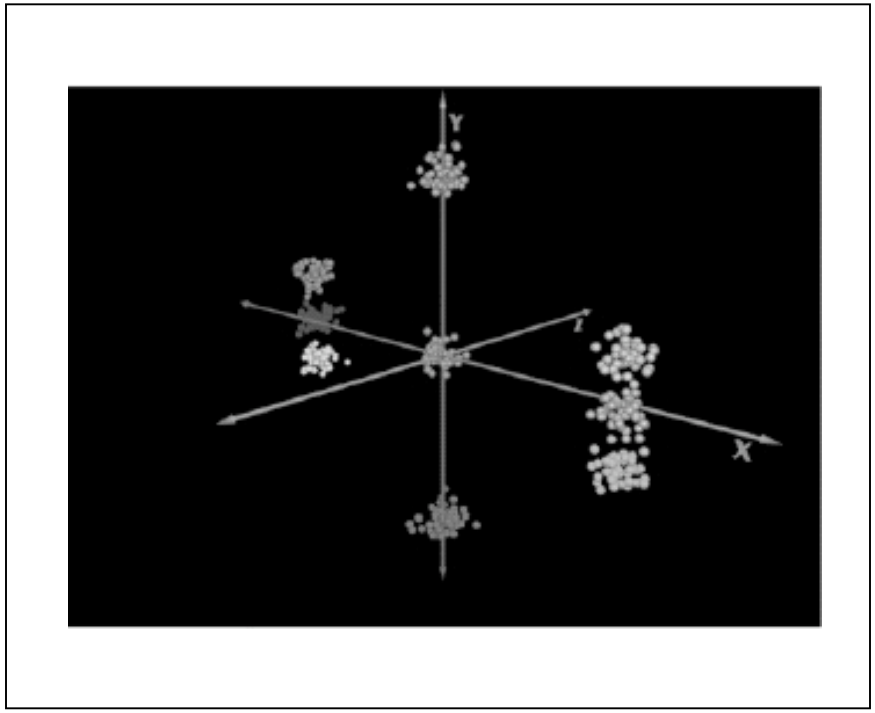
# Principal Component Analysis

(Adapted from Quackenbush 2001)

- Data reduction method
- AKA singular value decomposition
- Used to pick out patterns in data
- Provide projection of complex data sets onto reduced, easily visualized space
- Difficult to define precise clusters but can give you an idea of # of clusters for SOMs or k-means

# Principal Component Analysis




Quackenbush 2001

# Quackenbush 2001

"One must remember that the results of any analysis have to be evaluated in the context of other biological knowledge."

# Supervised Learning

(Adapted from Quackenbush 2001)

- Useful if you have some previous information about which genes are expected to cluster together
- Support Vector Machine (SVM)
- Start with training set (eg. positive and negative examples)
- SVM learns to distinguish between members and non-members of a class

# Warnings

(Adapted from Quackenbush 2001)

- Classification is dependent on
  - clustering method used
  - normalization of data
  - measure of similarity

# Citations


- Brazma A and Vilo J. Minireview: Gene expression data analysis. *FEBS Letters* 480:17-24, 2000.

- Quackenbush J. Computational Analysis of Microarray Data. *Nature Review* | *Genetics* 2:418-427, 2001.

- Dudoit S and Gentleman R. Classification in microarray experiments. Statistics and Genomics Short Course - Lecture 5, January 2002 (http://www.bioconductor.org/workshop.html)

# Local Tools

- GeneCluster (WIBR)
- Cluster & TreeView (Eisen)
- GeneSpring (Silicon Genetics)
- Spotfire (Spotfire)
- R Statistics Package
- Matlab
  - http://www.jax.org/research/churchill/software/anova/

# Lists of Tools

- Rockefeller University
  - http://linkage.rockefeller.edu/wli/microarray/soft.html
- R Statistics Package Microarry Tools
  - http://www.stat.uni-muenchen.de/~strimmer/rexpress.html
- Bioconductor Project
  - http://www.bioconductor.org/
- EBI
  - http://ep.ebi.ac.uk/Links.html, http://ep.ebi.ac.uk/EP/