

# Bioinformatics for Biologists

## Microarray Data Analysis: Lecture 2.

Fran Lewitter, Ph.D.  
Director  
Bioinformatics and Research Computing  
Whitehead Institute

## Outline

- Introduction
- Working with microarray data
  - Normalization
  - Analysis
    - Distance metrics
    - Clustering methods
    - Visualization
    - Functional Analysis

WIBR Microarray Course, © Whitehead Institute, May 2004

2

## Normalization revisited

- MGED Data Transformation and Normalization Working Group - Cathy Ball (Stanford) and John Quackenbush (TIGR)
- <http://genome-www5.stanford.edu/mged/normalization.html>

WIBR Microarray Course, © Whitehead Institute, May 2004

3

## Microarray Data

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

Matrix ( $n \times p$ ) =

|          |          |          |          |
|----------|----------|----------|----------|
| $x_{11}$ | $x_{12}$ | ...      | $x_{1p}$ |
| $x_{21}$ | $x_{22}$ | ...      | $x_{2p}$ |
| $\vdots$ | $\vdots$ | $\vdots$ | $\vdots$ |
| $x_{n1}$ | $x_{n2}$ | ...      | $x_{np}$ |

WIBR Microarray Course, © Whitehead Institute, May 2004

4

## 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,  $d_{ij} \geq 0$
  - 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.

WIBR Microarray Course, © Whitehead Institute, May 2004

5

## 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  $(x, y, z)$ .

WIBR Microarray Course, © Whitehead Institute, May 2004

6

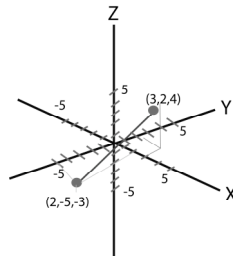
## Distance in 3 dimensions

Let X, Y, Z refer to different experiments (e.g. fetal liver, fetal brain, adult brain)

$$d = \sqrt{\Delta X^2 + \Delta Y^2 + \Delta Z^2}$$

$$d = \sqrt{(3-2)^2 + (2-(-5))^2 + (4-(-3))^2}$$

$$d = \sqrt{1 + 49 + 49} = \sqrt{99}$$



WIBR Microarray Course, © Whitehead Institute, May 2004

7

## 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.

WIBR Microarray Course, © Whitehead Institute, May 2004

8

## Semi-metric distances

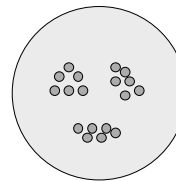
(Adapted from Quackenbush 2001)

- 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

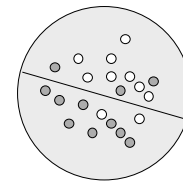
WIBR Microarray Course, © Whitehead Institute, May 2004

9

## Clustering vs Classification



Unsupervised



Supervised

WIBR Microarray Course, © Whitehead Institute, May 2004

10

## Hierarchical methods

(Adapted from Dudoit and Gentleman, 2002)

- Produces a tree or dendrogram
- 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

WIBR Microarray Course, © Whitehead Institute, May 2004

11

## 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

WIBR Microarray Course, © Whitehead Institute, May 2004

12

## 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 dendrogram
- Disadvantages: computational difficulties when considering all possible divisions into two groups

WIBR Microarray Course, © Whitehead Institute, May 2004

13

## Hierarchical Clustering

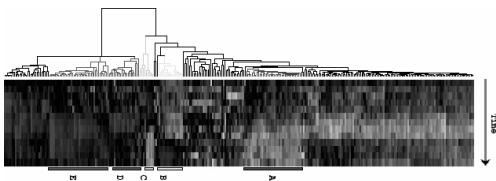
(Adapted from Quackenbush 2001)

- **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

WIBR Microarray Course, © Whitehead Institute, May 2004

14

## Dendrogram



Eisen et al 1998

WIBR Microarray Course, © Whitehead Institute, May 2004

15

## 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  $<$  inter-cluster dissimilarity)

WIBR Microarray Course, © Whitehead Institute, May 2004

16

## 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

WIBR Microarray Course, © Whitehead Institute, May 2004

17

## 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

WIBR Microarray Course, © Whitehead Institute, May 2004

18

[illegible]

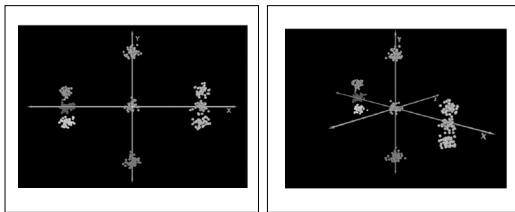
19

(Adapted from Quackenbush 2001)

- WIBR Microarray Course, © Whitehead Institute, May 2004

20

(Quackenbush 2001)



21

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

22

(Adapted from Quackenbush 2001)

- 23

(Adapted from Quackenbush 2001)

- 24

## Citations

---

- Causton H, Quackenbush J, Brazma A. A Beginner's Guide to Microarray Gene Expression Data Analysis, Blackwell publishing, 2003.
- 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.
- Quackenbush J. Microarray data normalization and transformation. *Nature Genetics Supp.* 32:496-501, 2002.
- Dudoit S and Gentleman R. Classification in microarray experiments. Statistics and Genomics Short Course - Lecture 5, January 2002 (<http://www.bioconductor.org/workshop.html>)

WIBR Microarray Course, © Whitehead Institute, May 2004

25

## Lists of Tools

---

- Local WI Page
  - <http://jura.wi.mit.edu/bio/microarrays/biopage5tools.html>
  - WADE
- R Statistics Package Microarray 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/>

WIBR Microarray Course, © Whitehead Institute, May 2004

26

## Exercise 2

### To Do

---

Identify differentially expressed genes

- TTEST, VLOOKUP

Clustering (Eisen's Cluster3)

- Hierarchical and K-means
- Heatmap; JavaTreeView

WIBR Microarray Course, © Whitehead Institute, May 2004

27