Lecture 12
Small & Large Scale Expression Analysis

M. Saleet Jafri

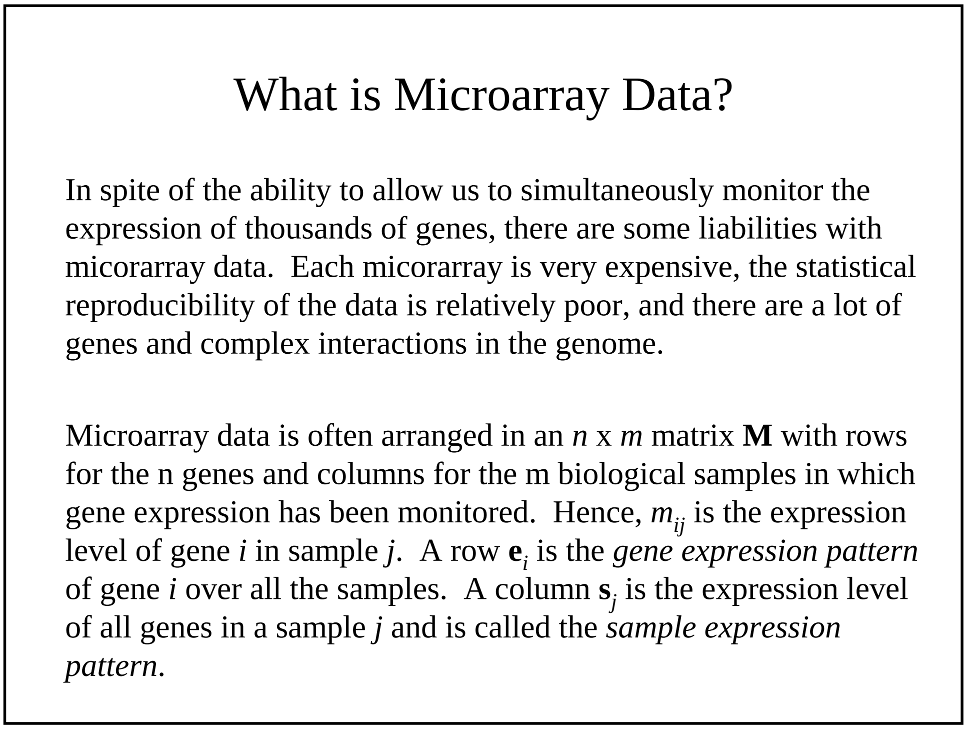
Program in Bioinformatics and Computational Biology

George Mason University

Microarray Data Analysis

Gene chips allow the simultaneous monitoring of the expression level of thousands of genes. Many statistical and computational methods are used to analyze this data. These include:

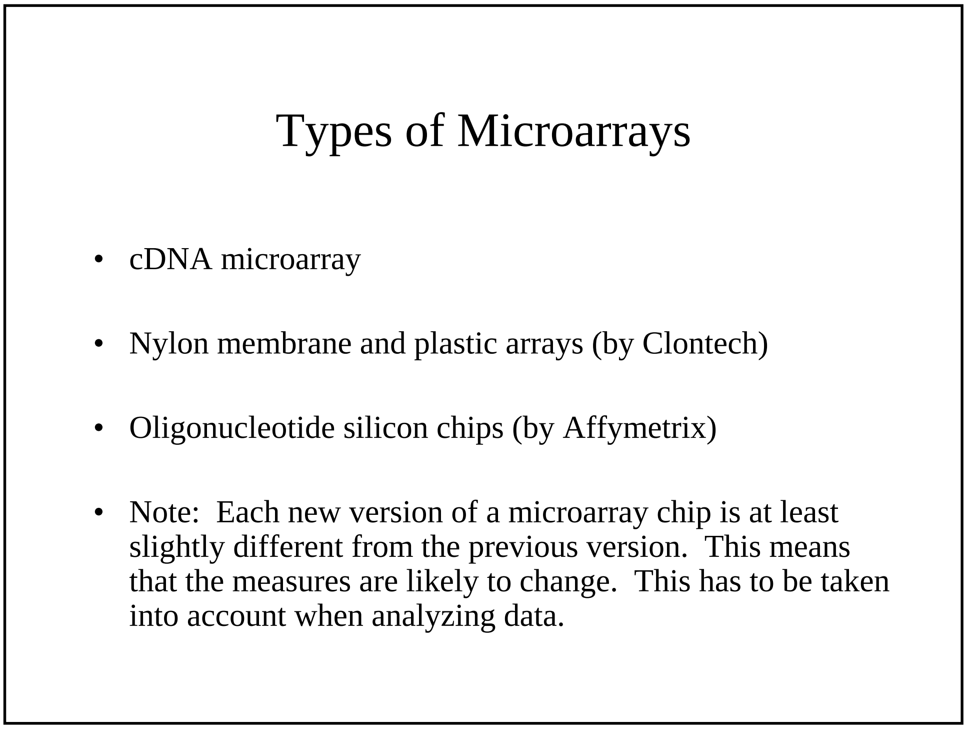
- statistical hypothesis tests for differential expression analysis
- principal component analysis and other methods for visualizing high-dimensional microarray data
- cluster analysis for grouping together genes or samples with similar expression patterns
- hidden Markov models, neural networks and other classifiers for predictively classifying sample expression patters as one of several types (diseased, ie. cancerous, vs. normal)



What is Microarray Data?

In spite of the ability to allow us to simultaneously monitor the expression of thousands of genes, there are some liabilities with micorarray data. Each micorarray is very expensive, the statistical reproducibility of the data is relatively poor, and there are a lot of genes and complex interactions in the genome.

Microarray data is often arranged in an *n* x *m* matrix **M** with rows for the n genes and columns for the m biological samples in which gene expression has been monitored. Hence, m_{ij} is the expression level of gene *i* in sample *j*. A row $\mathbf{e}_i$ is the *gene expression pattern* of gene *i* over all the samples. A column $\mathbf{s}_j$ is the expression level of all genes in a sample *j* and is called the *sample expression pattern*.



Types of Microarrays

- cDNA microarray
- Nylon membrane and plastic arrays (by Clontech)
- Oligonucleotide silicon chips (by Affymetrix)
- Note: Each new version of a microarray chip is at least slightly different from the previous version. This means that the measures are likely to change. This has to be taken into account when analyzing data.

cDNA Microarray

- The expression level e_{ij} of a gene i in sample j is expressed as a log ratio, $\log(r_{ij}/g_i)$, of the log of its actual expression level r_{ij} in this sample over its expression level g_i in a control.
- When this data is visualized e_{ij} is color coded to a mixture of red ($r_{ij} >> g_i$) and green ($r_{ij} << g_i$) and a mixture in between.

Nylon Membrane and Plastic Arrays (by Clontech)

- A raw intensity and a background value are measured for each gene.
- The analyst is free to choose the raw intensity or can adjust it by subtracting the background intensity.

Oligonucleotide Silicon Chips (by Affymetrix)

- These arrays produce a variety of numbers derived from 16-20 pairs of perfect match (PM) and mismatch (MM) probes.
- There are several statistics related to gene expression that can be derived from this data. The most commonly used one is the *average difference* (AVD), which is derived from the differences of PM-MM in the 16-20 probe pairs.
- The next most commonly used method is the *log absolute value* (LAV), which comes from the ratios PM/MM in the probe pairs.
- Note: The Affymetrix gene-chip software has a absent/present call for each gene on a chip. According to Jagota, the method is complex and arbitrary so they usually ignore it.

For What Do We Use Microarray Data?

- Genes with similar expression patterns over all samples – We can compare the expression patterns e_i and $e_{i'}$ of two genes i and i' over all samples.
- If we use cluster analysis, we can separate the genes into groups of genes with similar expression patterns (trees).
- This will allow us to find what unknown genes have altered expression in a particular disease by comparing the pattern to genes know to be affiliated with a disease.
- It can also find genes that fit a certain pattern such as a particular pattern of change with time.
- It can also characterize broad functional classes of new genes from the known classes of genes with similar expression.

For What Do We Use Microarray Data?

- Genes with unusual expression levels in a sample – In contrast to standard statistical methods where we ignore outliers, here outliers might have particular importance. Hence, we look for genes whose expression levels are very different from the others.
- Genes whose expression levels vary across samples – We can compare gene expression levels of a particular gene or set of genes in different samples. This can be used to look compare normal and diseased tissues or diseased tissue before and after treatment.

For What Do We Use Microarray Data?

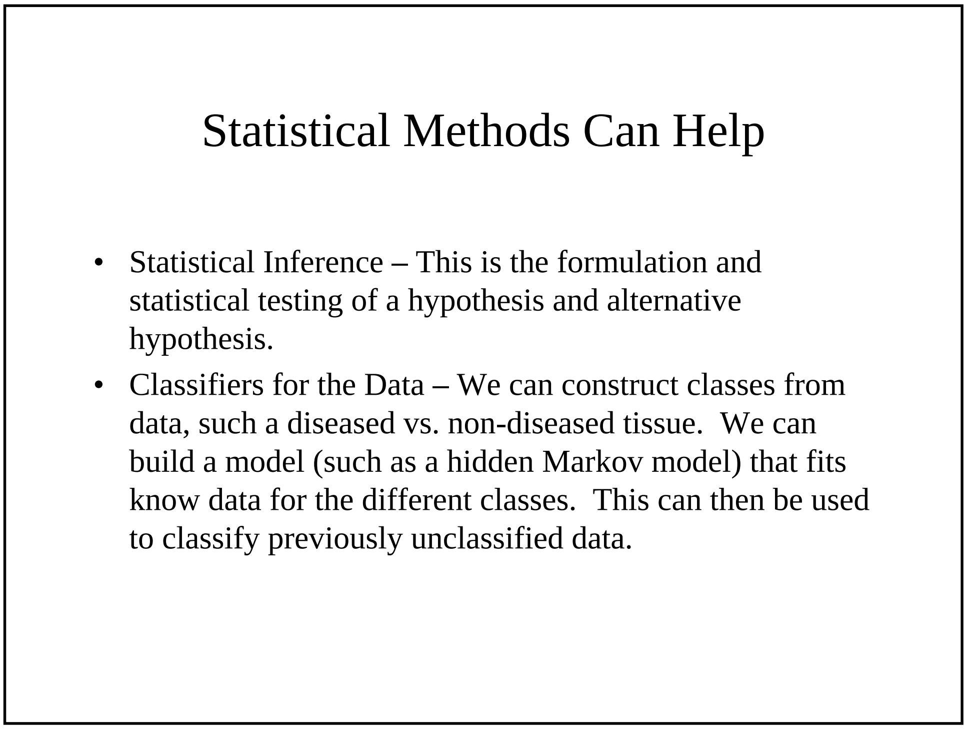
- Samples that have similar expression patterns – We might want to compare the expression patters of all genes between two samples. We might cluster the genes into gene with similar expression patterns to help with the comparison. This can be used to look compare normal and diseased tissues or diseased tissue before and after treatment.
- Tissues that might be cancerous (diseased) – We can take the gene expression pattern of sample and compare it to library expression patterns that indicate diseased or not diseased tissue.

Statistical Methods Can Help

- Experimental Design – Since using microarrays is costly and time consuming, we want to design experiments to use the minimal number of micorarrays that will give a statistically significant result.
- Data Pre-processing – It is sometimes useful to preprocess the data prior to visualization. An example of this is the log ratio mentioned earlier. It is often necessary to rescale data from different microarrays so that they can be compared. This is due to variation in chip to chip intensity. Another type of preprocessing is subtracting the mean and dividing by the variance.

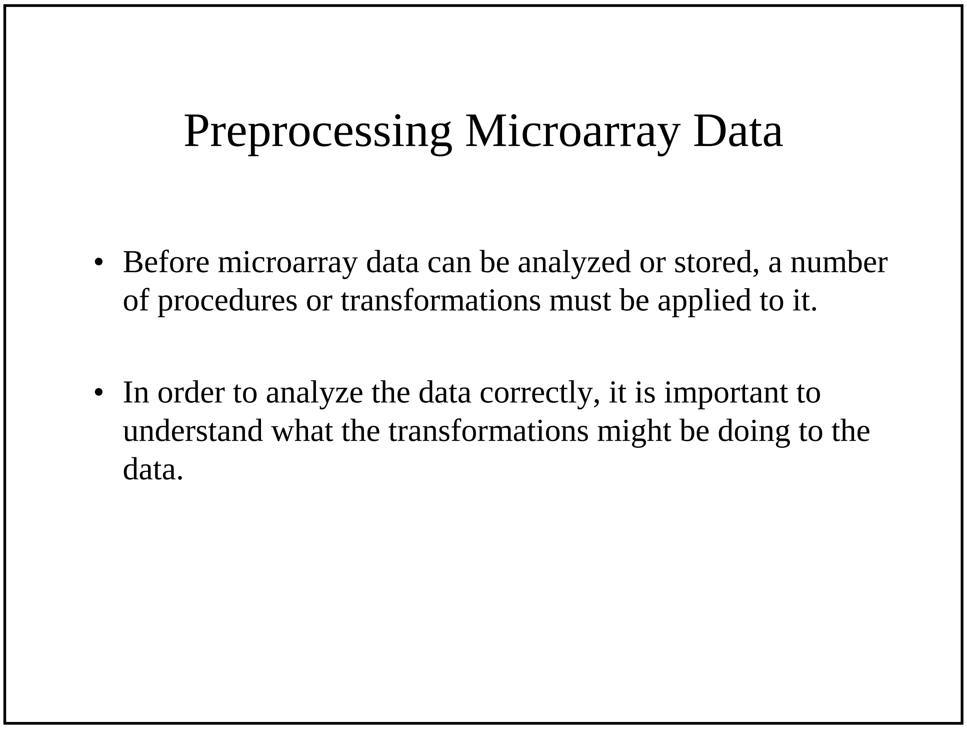
Statistical Methods Can Help

- Data Visualization – Principle component analysis and multidimensional scaling are two useful techniques for reducing multidimensional data to two and three dimensions. This allows us to visualize it.
- Cluster Analysis – By associating genes with similar expression patterns, we might be able to draw conclusions about their functional expression.
- Probability Theory – We can use statistical modeling and inference to analyze our data. Probability theory is the basis for these.



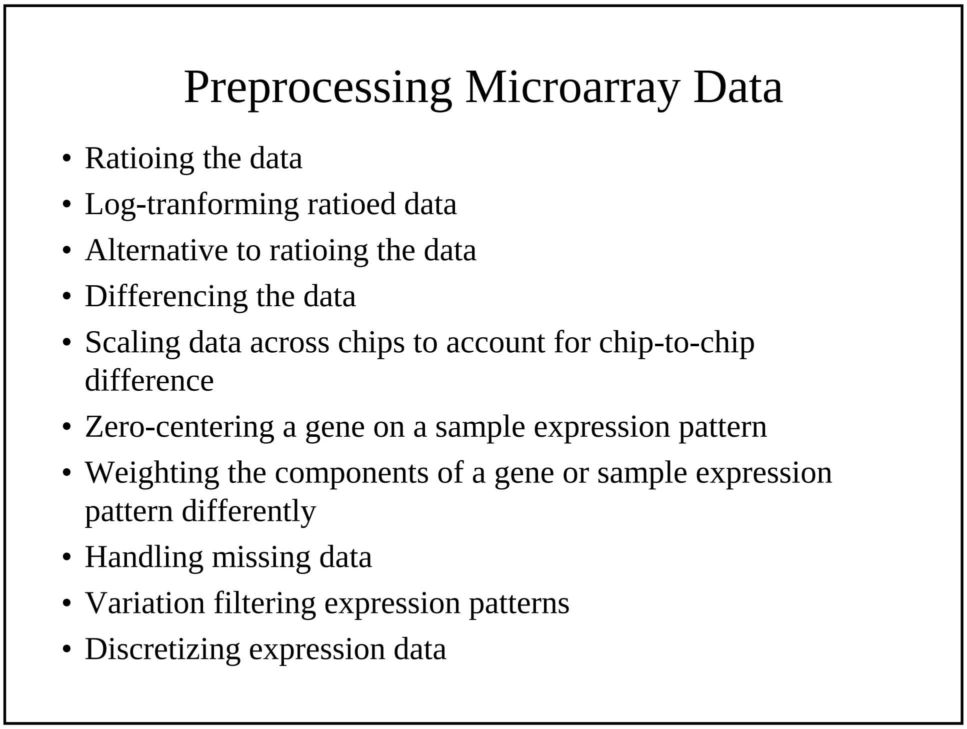
Statistical Methods Can Help

- Statistical Inference – This is the formulation and statistical testing of a hypothesis and alternative hypothesis.
- Classifiers for the Data – We can construct classes from data, such a diseased vs. non-diseased tissue. We can build a model (such as a hidden Markov model) that fits know data for the different classes. This can then be used to classify previously unclassified data.



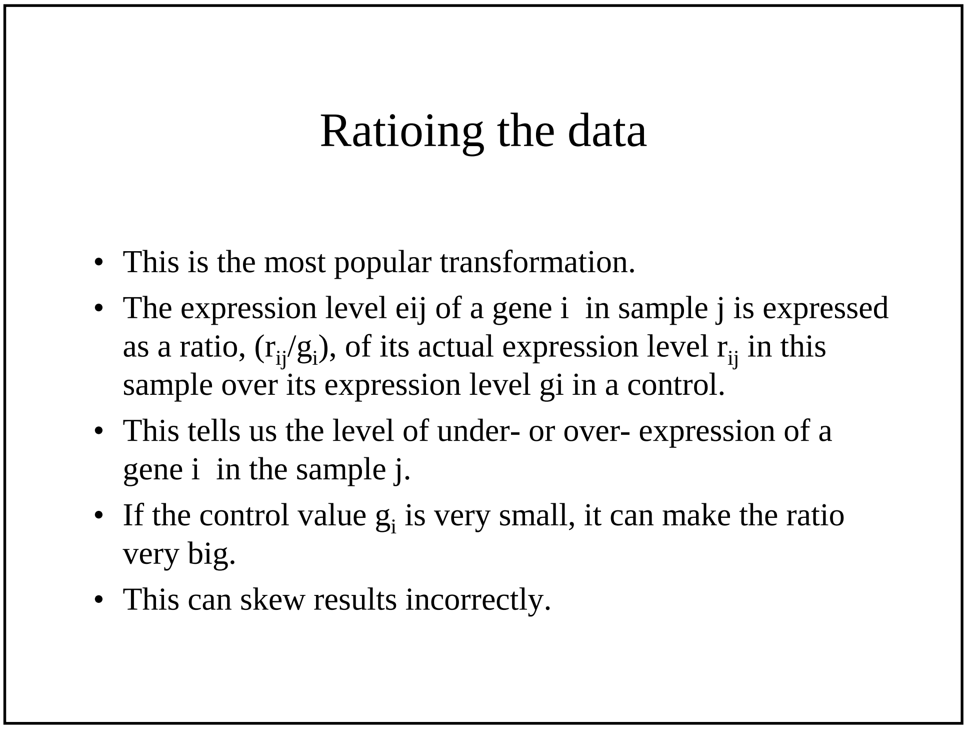
Preprocessing Microarray Data

- Before microarray data can be analyzed or stored, a number of procedures or transformations must be applied to it.
- In order to analyze the data correctly, it is important to understand what the transformations might be doing to the data.



Preprocessing Microarray Data

- Ratioing the data
- Log-tranforming ratioed data
- Alternative to ratioing the data
- Differencing the data
- Scaling data across chips to account for chip-to-chip difference
- Zero-centering a gene on a sample expression pattern
- Weighting the components of a gene or sample expression pattern differently
- Handling missing data
- Variation filtering expression patterns
- Discretizing expression data



Ratioing the data

- This is the most popular transformation.
- The expression level eij of a gene i in sample j is expressed as a ratio, (r_{ij}/g_i), of its actual expression level r_{ij} in this sample over its expression level gi in a control.
- This tells us the level of under- or over- expression of a gene i in the sample j.
- If the control value g_i is very small, it can make the ratio very big.
- This can skew results incorrectly.

Log-tranforming ratioed data

- This is also a popular transformation.
- The expression level e_{ij} of a gene i in sample j is expressed as a log ratio, $\log(r_{ij}/g_i)$, of the log of its actual expression level r_{ij} in this sample over its expression level g_i in a control.
- This will suppress outliers caused when the control value g_i is very small.
- However, it creates a new outlier when r_{ij} is very small.

Alternative to ratioing the data

An alternative that eliminated both of the outlier problems above is

$$\frac{r_{ij}}{r_{ij} + g_i}$$

This gives a value in [0,1] and can be interpreted at the probability of gene i is higher in sample j than in control.

Differencing the data

- Another transformation is to difference the data ie. $r_{ij} – g_i$.
- This is not really appropriate in our previous context.
- However, this is used by Affymetrix in a different context.
- In their data r_{ij} is the strength of the match of the target *i* to a specific probe *j* and g_i is he strength of the match of the target *i* to a control for this probe.

Scaling data across chips to account for chip-to-chip difference

- As mentioned previously, different chips might display different intensities.
- When comparing different chips the data might need to be scaled so that they are on the same scale.
- Alternatively, they can be normalized so that they are between [0,1] and compared.

Zero-centering a gene on a sample expression pattern

This in effect the same as subtracting the mean expression pattern.

Suppose that **x** is an expression pattern for a particular gene g_i whose components are log-ratios. Let where is the ‘average expression pattern’ or control. Then **x** indicates whether the gene g_i is induced or repressed relative to control. (Remember the **x**’s are vectors).

Subtracting the mean expression pattern and dividing by the standard deviation can accomplish this.

$$x' = \frac{x - \bar{x}}{\sigma}$$

(Remember the **x**’s are vectors).

Weighting the components of a gene or sample expression pattern differently

If we have a matrix of weights **W**=diag(w_1,…,w_n), we can weight the expression patterns by

$$\mathbf{x}_w = \mathbf{Wx}$$

In this way, we can weight the contributions from different genes differently.

Handling missing data

- Sometimes components of an expression pattern **x** are missing.
- To fix this, the missing values can be replaced by the mean over the non-missing values in **x**.

Variation filtering expression patterns

- When we are performing cluster analysis on gene or sample expression patterns, patterns with low variance will all seem sufficiently similar to each other and might form a cluster.
- This cluster will probably not reflect any interesting result.

Discretizing expression data

Sometimes we might want to convert gene or sample expression pattern into discrete values. For example, if we have log-ratio, we may want to simply look at whether something is up- or down-regulated. To do this, we can do the following:

$$x_b = (x_{b,i}) \quad where \quad x_{b,i} = \begin{cases} +1 & when\ x_i > 0 \\ -1 & when\ x_i < 0 \\ 0 & when\ x_i = 0 \end{cases}$$

In this case +1 would indicate up-regulation, 0 would indicate no change and –1 would indicate down-regulation.

Measuring Dissimilarity of Expression Data

- We might want to compare two or more gene or sample expression patterns.
- This might be used to differentiate between diseased and normal cells or finding out the genetic similarity of tissues.
- To do this we need a *distance metric* or a *dissimilarity measure*.

Example Distance Metric

Euclidean Distance – This is the most common distance measure.

$$d(x,y) = \sqrt{\sum_i (x_i - y_i)^2}$$

This should not be used if either

1) Not all components of the vectors being compared have equal weight.

2) There is missing data.

Preprocessing the data can often alleviate these problems.

We can also use the normalized Euclidean distance

$$d(x,y) = \frac{\sqrt{\sum_i (x_i - y_i)^2}}{\sqrt{n}}$$

Example Dissimilarity Measures

- Maximum Coordinate Difference – The following computes the maximum absolute distance along a coordinate

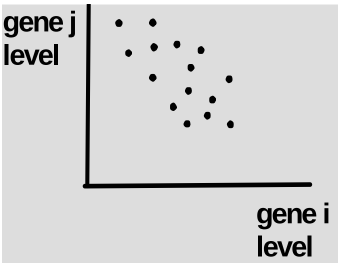
$$d(x,y) = \max_i |\, x_i - y_i \,|$$

- Dot Product –This is a dissimilarity version of the dot product.

$$d(x,y) = -x \circ y$$

Visualizing Micorarray Data

It is usually easiest to understand data if it can be represented in 2 or 3 dimensions. For example, a 2-D scatter plot of the expression levels of genes *i* and *j* over a number of samples can show the relationship between these two genes.

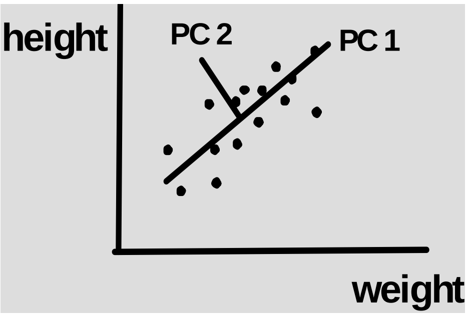


Principal Components Analysis

- In principal components analysis n-dimensional data is converted to d-dimensional data (d<<n) such that the components in the new space are uncorrelated and axis or dimensions of the new space are ordered with respect to the amount of variance they explain.
- The first component explains the most about the data.
- The second component is orthogonal to the first and explains more about the data and so on.

Principal Components Analysis

Example – Consider height and weight data for a group of individuals. This is 2-D data, but there is a correlation between height and weight. We can use this property to reduce the data to 1D. PC1 explains most of the data and PC2 explains the rest.



PCA and Microarrays

Sample Application 1 – If we want to compare the sample expression patterns from two groups (diseased vs normal, experimental vs control). If we have n genes, the each pattern is a point in n-dimensional space. Suppose we want to see if the sample expression patterns for these two groups cluster by group. We might want to perform PCA analysis and perform cluster analysis at the top three components.

PCA and Microarrays

Sample Application 2 – On gene chips (such as the one made by Affymetrix), the same gene occupies multiple cells. In theory, the expression level of all cells with the same gene should be perfectly correlated. However, in practice, this is often not the case due to imperfections in the technology or hybridization of the sequence fragment to other genes in the target.

PCA and Microarrays

PCA allows us to see how good the correlation among these cells is. To use PCA, we would hybridize k different samples on the same chip. For each sample, the expression levels of a gene x in the n cells is an n-dimensional vector. Hence, there are k points in n-dimensional space. Using PCA, if most of the variance is explained by the first principal component, the effective dimensionality of the data is 1 and there cells are highly correlated.

PCA Limitations

- Clustering by PCA effectively yields clusters as if the Euclidean distance metric had been used. Hence, it is possible that it might miss clusters.
- The reduction of dimensionality uses all coordinates. If only a few genes out of a thousand differ between two samples (Application 1), clustering by PCA might not yield any meaningful results.

Cluster Analysis of Microarray Data

Recall that microarray data can be thought of as gene expression patterns or sample expression patterns. These can be each considered to be vectors. The first thing we have to do before applying cluster analysis is to find a distance between the various expression pattern vectors. This is done using similarity/dissimilarity measures such as Euclidean distance, Mahalonobis distance, or linear correlation coefficients. Once a distance matrix is computed, the following clustering algorithms can be used. The clusters formed can differ significantly depending upon the distance measure used.

Cluster Analysis of Microarray Data

Hierarchical Clustering – Assume each data point is in a singleton cluster.

Find the two clusters that are closest together. Combine these to form a new cluster.

Compute the distance from all clusters to the new cluster using some form of averaging.

Find the two closest clusters and repeat.

Cluster Analysis of Microarray Data

k-Means Clustering – An alternate method of clustering called k-means clustering, partitions the data into k clusters and finds cluster means μ_i for each cluster. In our case, the means will be vectors also.

Usually, the number of clusters k is fixed in advance. To choose k something must be know about the data. There might be a range of possible k values.

To decide which is best, optimization of a quantity that maximizes cluster tightness ie. minimizes distances between points in a cluster.

Cluster Analysis of Microarray Data

Self-organizing Maps – This is basically an application of neural networks to microarray data. Assume that there is a 2-dimensional grid of cells and a map from a given set of expression data vectors in R^n, ie, there are n nodes in the input layer and a connection neuron from each of these to each cell. Each cell (i, j) gets it own weight from n input neurons. The weight vector μ_{ij} is the mean of the cluster associated with cell (i, j). Each data vector **d** gets mapped to the cell (i, j) that is closest to **d** using Euclidean distance.In order to train the network, the mean vectors μ_{ij} for the cells (i, j) must be learned.

Hidden Markov Models and Microarray Data

We can use Hidden Markov models for pattern recognition in the study of micorarray data. Suppose that we want to consider gene expression data from a tissue sample and want to know if it is control or different from the control (diseased, experimentally altered, responding to drug, etc.). Consider the gene expression data vector as a set of emissions, one for each vector coordinate. Each emission has a value that is defined by some probability distribution function. This can be continuous, or can even discrete. To make it discrete, the data should be preprocessed to indicate, up-regulation, down-regulation, or no significant change.

Finding Genes Expressed Unusually Different in a Population

The following section addresses the question: Is gene g expressed unusually in the sample?

The first thing to do is to come up with a formal mathematical definition for what unusual is. Assume that the microarray data is log transformed ratio data. If a histogram is constructed of the data, it should yield roughly a normal distribution. Anything that is out near either tail can be considered to be unusually expressed. Note that this can be either a high or low expression level.

Finding Genes Expressed Unusually Different in a Population

Calculate the Z-score for the data point considered

$$Z = \frac{e_g - \mu}{\sigma}$$

where e_g is the expression level, μ is the mean and σ is the standard deviation. The Z value will give an indication of the how far the data is toward the tail (α - level).

Use statistical inference (hypothesis testing).