

Review

Microarray cluster analysis and applications

Instructor: Prof. Abraham B. Korol

Institute of Evolution, University of Haifa

Date: 22 – Jan – 2003

Submitted by: Erika Shay

Table of Contents

Summary.....	3
Background.....	4
Microarray preparation.....	6
Probe preparation, hybridization and imaging	7
Low level information analysis.....	8
High level information analysis	10
Cluster analysis.....	17
Distance metric.....	17
Different distance measures	17
Clustering algorithms	22
Difficulties and drawbacks of cluster analysis.....	30
Alternative method to overcome cluster analysis pitfalls	31
Microarray applications and uses	36
Conclusions	38
Appendix	39
General background about DNA and genes.....	39
References	41
Glossary	43

Summary

Microarrays are one of the latest breakthroughs in experimental molecular biology, that allow monitoring of gene expression of tens of thousands of genes in parallel. Knowledge about expression levels of all or a big subset of genes from different cells may help us in almost every field of society. Amongst those fields are diagnosing diseases or finding drugs to cure them. Analysis and handling of microarray data is becoming one of the major bottlenecks in the utilization of the technology.

Microarray experiments include many stages. First, samples must be extracted from cells and microarrays should be labeled. Next, the raw microarray data are images, have to be transformed into gene expression matrices. The following stages are low and high level information analysis.

Low level analysis include normalization of the data. One of the major methods used for High level analysis is Cluster analysis. Cluster analysis is traditionally used in phylogenetic research and has been adopted to microarray analysis. The goal of cluster analysis in microarrays technology is to group genes or experiments into clusters with similar profiles.

This survey reviews microarray technology with greater emphasis on cluster analysis methods and their drawbacks. An alternative method is also presented. This survey is not meant to be treated as complete in any form, as the area is currently one of the most active, and the body of research is very large.

Background

Most cells in multi-cellular eukaryotic organisms contain the full complement of genes that make up the entire genome of the organism. Yet, these genes are selectively expressed in each cell depending on the type of cell and tissue and general conditions both within and outside of the cell. Since the development of the recombinant DNA and molecular biology techniques, it has become clear that major events in the life of a cell are regulated by factors that alter the expression of genes. Thus, understanding of how expression of genes is selectively controlled has become a major domain of activity in modern biological research. Two main questions arise when dealing with gene expression: how does gene expression reveal cell functioning and cell pathology. These questions can be further divided into:

- How does gene expression level differ in various cell types and states?
- What are the functional roles of different genes and how their expression varies in response to physiological changes within the cellular environment.
- How is gene expression effected by various diseases? Which genes are responsible for specific hereditary diseases.
- What genes are affected by treatment with pharmacological agents such as drugs.
- What are the profiles of gene expression changes during a time dependent series of cellular events?

Prior to the development of the microarrays, a method called "differential hybridization" was used for analysis of gene expression patterns. This method generally utilized cDNA probes (representing complementary copies mRNA), that were hybridized to replicas of cDNA libraries to identify specific genes that are expressed differentially. By utilizing two

sets of probes, an experimental and a control probe, differences in expression patterns of genes were identified. Although this method was useful, it was limited in scope generally to a small sample of the whole spectrum of genes.

Microarray method that has been developed during the course of the past decade represents a new technique for rapid and efficient analysis of expression patterns of tens of thousands of genes simultaneously. Microarray technology has revolutionized analysis of gene expression patterns by greatly increasing the efficiency of large-scale analysis using procedures that can be automated and applied with robotic tools.

A microarray experiment requires a large array of cDNA or oligonucleotide DNA sequences that are fixed on a glass, nylon, or quartz wafer (adopted from the semiconductor industry and used by Affymetrix, Inc.). This array is then reacted generally with two series of mRNA probes that are labeled with two different colors of fluorescent probes. After the hybridization of the probes, the microarray is scanned using generally a laser beam to generate an image of all the spots. The intensity of the fluorescent signal at each spot is taken as a measure of the levels of the mRNA associated with the specific sequence at that spot. The image of all the spots is analyzed using sophisticated software linked with information about the sequence of the DNA at each spot. This then generates a general profile of gene expression level for the selected experimental and control conditions.

Thus, in brief, a microarray experiment includes the following steps:

1. Microarray preparation.
2. Probe preparation, hybridization.

3. Low level information analysis.
4. High level information analysis.

Microarray preparation

Microarrays are commonly prepared on a glass, nylon or quartz substrate. Critical steps in this process include the selection and nature of the DNA sequences that will be placed on the array, and the technique of fixing the sequences on the substrate. Affymetrix company that is a leading manufacturer of gene chips, uses a method adopted from the semiconductor industry with photolithography and combinatorial chemistry. The density of oligonucleotides in their GeneChips is reported as about half a million sequences per 1.28 cm^2 (Affymetrix web site).

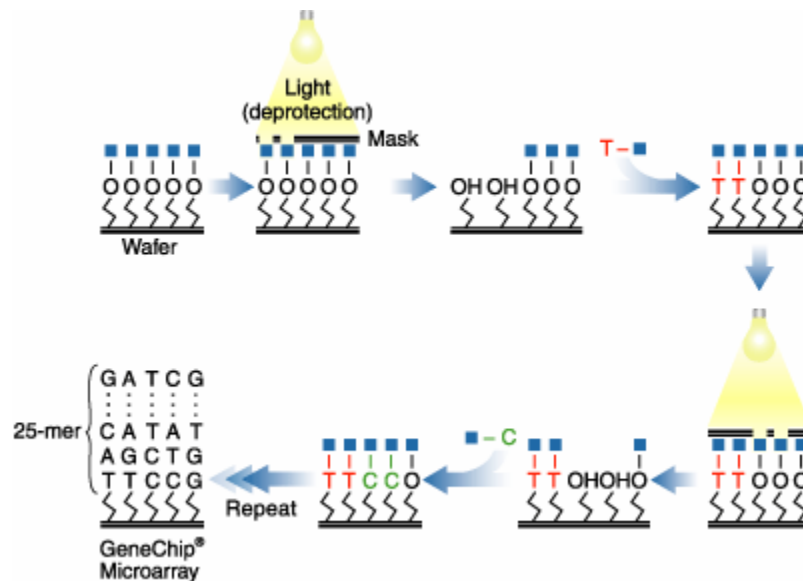


Figure 1: Lithographic process of GeneChip microarray production used by Affymetrix (<http://www.affymetrix.com/technology/manufacturing/index.affx>). The method shown is used to produce chips with oligonucleotides that are 25 base

long. In products prepared by other approaches long sequences in the range of hundreds of nucleotides can be fixed on the substrate.

Probe preparation, hybridization and imaging

To prepare RNA probes for reacting with the microarray, the first step is isolation of the RNA population from the experimental and control samples. cDNA copies of the mRNAs are synthesized using reverse transcriptase and then by in vitro transcription cDNA is converted to cRNA and fluorescently labeled. This probe mixture is then cast onto the microarray. RNAs that are complementary to the molecules on the microarray hybridize with the strands on the microarray. After hybridization and probe washing the microarray substrate is visualized using the appropriate method based on the nature of substrate. With high density chips this generally requires very sensitive microscopic scanning of the chip. Oligonucleotide spots that hybridize with the RNA will show a signal based on the level of the labeled RNA that hybridized to the specific sequence. Whereas the dark spots that show little or no signal, mark sequences that are not represented in the population of expressed mRNAs.

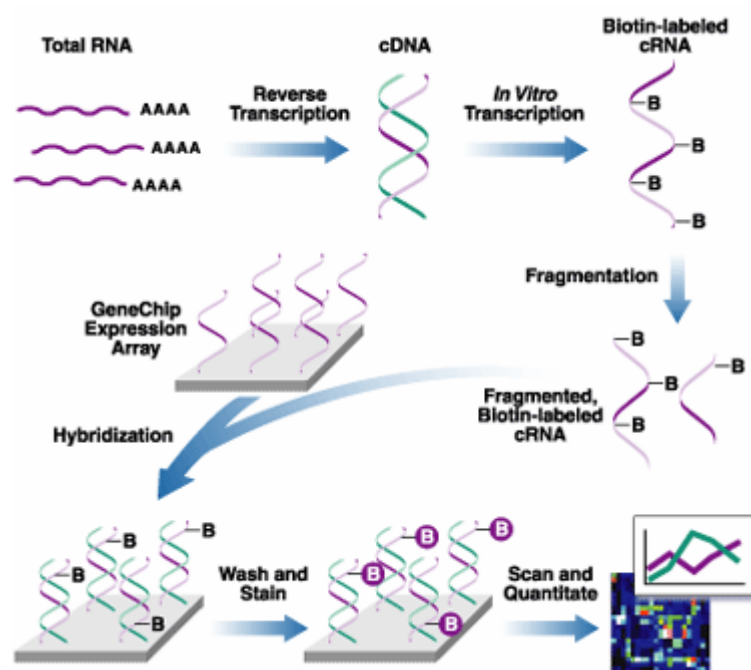


Figure 2: The process of fluorescently labeled RNA probe production (From Affymetrix web site).

Low level information analysis

Microarrays measure the target quantity (i.e. relative or absolute mRNA abundance) indirectly by measuring another physical quantity – the intensity of the fluorescence of the spots on the array for each fluorescent dye (see figure 3). These images should be later transformed into the gene expression matrix. This task is not a trivial one because:

1. The spots corresponding to genes should be identified.
2. The boundaries of the spots should be determined.
3. The fluorescence intensity should be determined depending on the background intensity.

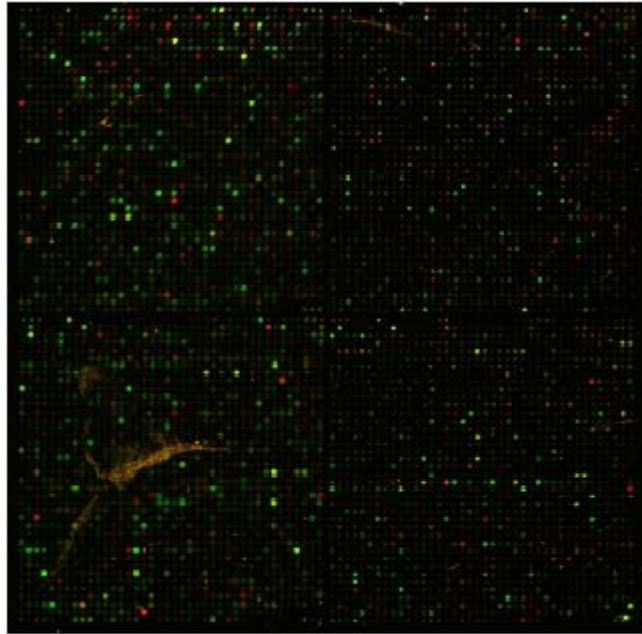


Figure 3: Gene expression data. Each spot represents the expression level of a gene in two different experiments. Yellow or red spots indicate that the gene is expressed in one experiment. Green spots show that the gene is expressed at same levels in both experiments.

We will not discuss the raw data processing in detail in this review. A survey of image analysis software may be found at http://cmpteam4.unil.ch/biocomputing/array/software/MicroArray_Software.html. It is also important to know the reliability for each data point. The reliability depends upon the absolute intensity of the spot, the higher the intensity, the more reliable is the data, the uniformity of the individual pixel intensities and the shape of the spot. Currently, there is no standard way of assessing the spot measurement reliability.

In conclusion, microarray-based gene expression measurements are still far from giving estimates of mRNA counts per cell in the sample. The samples are relative by nature. In addition, appropriate normalization should be applied to enable gene or samples

comparisons. It is important to note that even if we had the most precise tools to measure mRNA abundance in the cell, it still wouldn't provide us a full and exact picture about the cell activity because of post-translational changes.

High level information analysis

There are various methods used for analysis and visualization:

Box plots

A box plot is a plot that represents graphically several descriptive statistics of a given data sample. The method is usually used for finding outliers in the data. The box plot contains a central line and two tails. The central line in the box shows the position of the median. The box will represent an interval that contains 50% of the data. The interval may be changed by the user of the software. Data points that fall beyond the box's boundaries are considered outliers.

Gene pies

Gene pies are visualization tools most useful for cDNA data obtained from two color experiments. Two characteristics are shown in gene pies: absolute intensity and the ratio between the two colors. The maximum intensity is encoded in the diameter of the pie chart while the ratio is represented by the relative proportion of the two colors within any pie chart. When determining the ratio between the two colors, a special care should be given to the absolute intensity. The ratio is most informative if the intensities are well over background for both colored samples, because if one of the genes is below background the ratio might vary greatly with small changes in the absolute intensity values.

Scatter plots

The scatter plot is a two or three dimensional plot in which a vector is plotted as a point having the coordinates equal to the components of the vector. Each axis corresponds to an experiment and each expression level corresponding to an individual gene is represented as a point. In such a plot, genes with similar expression levels will appear somewhere on the first diagonal (the line $y=x$) of the coordinate system. A gene that has an expression level that is very different between the two experiments will appear far from the diagonal. Therefore, it is easy to identify such genes very quickly. Scatter plots are easy to use but may require normalization of the data points in order to acquire accurate results. The most evident limitation of scatter plots is the fact that they can only be applied to data with two or three components since they can only be plotted in two or three dimensions. To overcome this problem the researcher may use the PCA method.

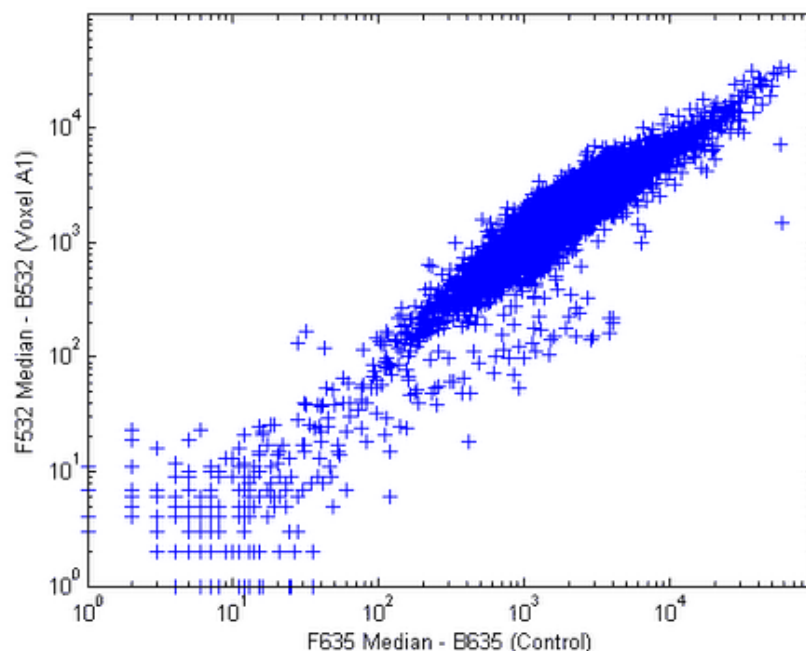


Figure 4(5): A scatter plot describing the expression levels of different genes in two experiments. Zero expression levels should be discarded since they probably are spots that failed to hybridize.

PCA

A major problem in microarray analysis is the large number of dimensions. In gene expression experiments each gene and each experiment may represent one dimension. For example, a set of 10 experiments involving 20,000 genes may be conceptualized as 20,000 data points (genes) in a space with 10 dimensions (experiments) or 10 points (experiments) in a space with 20,000 dimensions (genes). Both situations are beyond the capabilities of current visualization tools and beyond the visualization capabilities of our brains.

A natural solution would be to try to reduce the number of dimensions by eliminating those dimensions that are not “important”. PCA does exactly that by ignoring the dimensions in which data do not vary much. PCA calculates a new system of coordinates. The directions of the coordinate system calculated by PCA are the eigenvectors of the covariance matrix of the patterns. An eigenvector of a matrix A is defined as a vector z such as:

$$Az = \lambda z$$

where λ is a scalar called eigenvalue. For instance, the matrix:

$$A = \begin{bmatrix} -1 & 1 \\ 0 & -2 \end{bmatrix}$$

has the eigenvalues $\lambda_1 = -1$ and $\lambda_2 = -2$ and the eigenvectors $z_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $z_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$.

In intuitive terms, the covariance matrix captures the shape of the set of data points. PCA captures, by the eigenvectors, the main axes of the shape formed by the data diagram in an n-dimensional space. The eigenvalues describe how the data are distributed along the eigenvectors and those with the largest absolute values will indicate that the data have the largest variance along the corresponding eigenvectors. For instance, the figure below shows a data set with data points in a 2-dimensional space. However, most of the variability in the data lies along a one-dimensional space that is described by the first principal component (P_1). In this example the second principle component (P_2) can be discarded because the first principle component captures most of the variance present in the data.

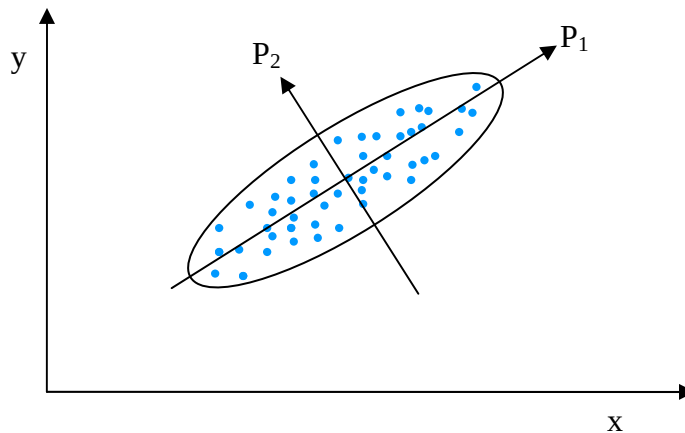


Figure 5: Each data point in this diagram has two coordinates. However, this data set is essentially one dimensional because most of the variance is along the first

eigenvector p_1 . The variance along the second eigenvector p_2 is marginal, thus, p_2 may be discarded.

It is important to notice that in some circumstances, the direction of the highest variance may not be the most useful. For example, in gene expression diagram which describes gene expression levels from two samples, the PCA would capture two axes. One axis would represent the within-experiment variation, while the other would represent the inter-experiment variation. Although the within-experiment axis could show much more variance than the inter-experiment axis, the within-experiment axis is of no use for us. This is because we know *a priori* that genes will be expressed at all levels¹.

The dimensionality reduction is achieved through PCA by selecting a small number of directions (e.g. 2 or 3) and look at the projection of the data in the coordinate system formed with only those directions.

In spite of its usefulness, PCA has also limitations. Those limitations are mainly related to the fact that PCA only takes into consideration the variance of the data which is a first-order statistical characteristic of the data. Another major limitation is that PCA takes into account only the variance of the data and completely discards the class of each data point. In some cases, such handling of the data will not produce the required result as the classes would not be defined by the PCA. Furthermore, PCA may fail to distinguish between classes when the classes' variance is the same. PCA's limitations may be overcome by an alternative approach called ICA.

Independent component analysis (ICA)

ICA is a technique that is able to overcome the limitations of PCA by using higher order statistical dependencies like skew¹ and kurtosis². ICA has been successfully used in blind source separation problem. The problem is to identify the n sources of n different signals.

Cluster analysis

Clustering is the most popular method currently used in the first step of gene expression matrix analysis. Clustering, much like PCA that is discussed above, reduces the dimensionality of the system and by this allows easier management of the data set. The goal of clustering is to group together objects (i.e. genes or experiments) with similar properties.

There are two straightforward ways to study the gene expression matrix:

1. Comparing expression profiles of genes by comparing rows in the expression matrix.
2. Comparing expression profiles of samples by comparing columns in the matrix.

By comparing rows we may find similarities or differences between different genes and thus to conclude about the correlation between the two genes. If we find that two rows are similar, we can hypothesize that the respective genes are co-regulated and possibly functionally related. By comparing samples, we can find which genes are differentially expressed in different situations.

Unsupervised analysis

Clustering is appropriate when there is no a priori knowledge about the data. In such circumstances, the only possible approach is to study the similarity between different

samples or experiments. Such an analysis process is known as unsupervised learning since there is no known desired answer for any particular gene or experiment. Clustering is the process of grouping together similar entities. Clustering can be done on any data: genes, samples, time points in a time series, etc. The algorithm for clustering will treat all inputs as a set of n numbers or an n -dimensional vector.

Supervised analysis

The purposes of supervised analysis are:

1. Prediction of labels. Used in discriminant analysis when trying to classify objects into known classes. For example, when trying to correlate gene expression profile to different cancer classes. This is done by finding a “classifier”. The correlation may be, later, used to predict the cancer class from gene expression profile.
2. Find genes that are most relevant to label classification.

Supervised methods include the following:

1. Gene shaving.
2. Support Vector Machine (SVM).
3. Self Organizing Feature Maps (SOFM).

Cluster analysis

When trying to group together objects that are similar, we should define the meaning of similarity. We need a measure of similarity. Such a measure of similarity is called a distance metric. Clustering is highly dependent upon the distance metric used.

Distance metric

A distance metric d is a function that takes as arguments two points x and y in an n -dimensional space $\mathbb{R}^n$ and has the following properties (1, p. 264-276):

1. **Symmetry.** The distance should be symmetric, i.e.:

$$d(x, y) = d(y, x)$$

2. **Positivity.** The distance between any two points should be a real number greater than or equal to zero:

$$d(x, y) \geq 0$$

3. **Triangle inequality.** The distance between two points x and y should be shorter than or equal to the sum of the distances from x to a third point z and from z to y :

$$d(x, y) \leq d(x, z) + d(z, y)$$

Different distance measures

The distance between two n -dimensional vectors $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and $\mathbf{y} = (y_1, y_2, \dots, y_n)$, according to different methods, is:

Euclidean distance

$$d_E(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2} = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

The Euclidean distance takes into account both the direction and the magnitude of the vectors.

Manhattan distance

$$d_M(\mathbf{x}, \mathbf{y}) = |x_1 - y_1| + |x_2 - y_2| + \dots + |x_n - y_n| = \sum_{i=1}^n |x_i - y_i|$$

where $|x_i - y_i|$ represents the absolute value of the difference between x_i and y_i . The Manhattan distance represents distance that is measured along directions that are parallel to the x and y axes meaning that there are no diagonal direction (See figure 2).

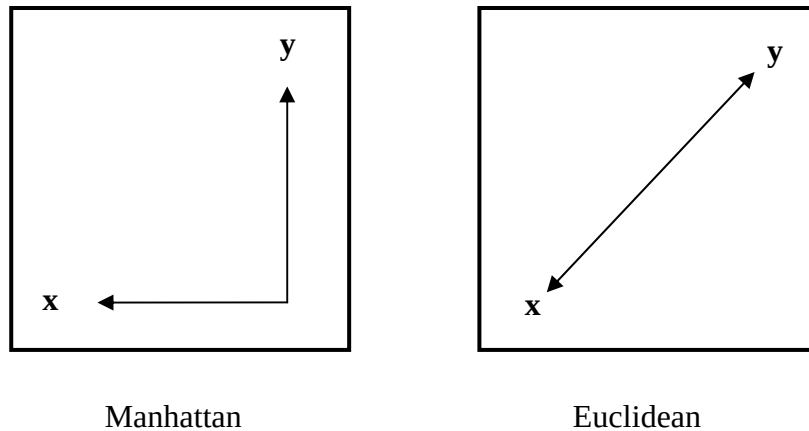


Figure 6(3): The Manhattan vs. Euclidean distance. It is evident that the Manhattan distance is greater than the Euclidean because of the Pythagorean Theorem.

Data which is clustered using this distance metric might appear slightly more sparse and less compact than the Euclidean distance metric. In addition, This metric is less robust regarding miscalculated data than is the Euclidean distance metric.

Chebychev distance

$$d_{\max}(\mathbf{x}, \mathbf{y}) = \max_i |x_i - y_i|$$

The Chebychev distance will simply pick the largest distance between two corresponding genes. This implies that any changes in lower values will be discarded. This kind of metric is very resilient to any amount of noise as long as the values don't exceed the maximum distance.

Angle between vectors

$$d_{\alpha}(\mathbf{x}, \mathbf{y}) = \cos(\theta) = \frac{\sum_{i=1}^n x_i y_i}{\sqrt{\sum_{i=1}^n x_i^2} \cdot \sqrt{\sum_{i=1}^n y_i^2}}$$

This Metric takes into account only the angle and discards the magnitude. Note that if a point is shifted by scaling all its coordinates by the same factors (i.e. noise), the angle distance will not change. This distance not resilient to noise if the noise adds some constant value to all dimensions (assuming different values in different dimensions).

Correlation distance

$$d_R(\mathbf{x}, \mathbf{y}) = 1 - r_{xy}$$

Where r_{xy} is the Pearson correlation coefficient of the vectors $\mathbf{x}$ and $\mathbf{y}$:

$$r_{xy} = \frac{s_{xy}}{\sqrt{s_x} \sqrt{s_y}} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Since the Pearson correlation coefficient r_{xy} takes values between -1 and 1, the distance $1-r_{xy}$ will vary between 0 and 2. The Pearson correlation finds whether two differentially expressed genes vary in the same way. The correlation between two genes will be high if the corresponding expression levels increase or decrease at the same time, otherwise the correlation will be low (see figure 4 for illustration). Note that this distance metric discards the magnitude of the coordinates (or the gene expression absolute values). If the genes are anti-correlated it will not be revealed by the Pearson correlation distance, but rather by the Pearson squared correlation distance(4).

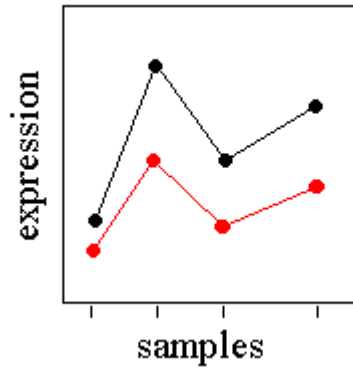


Figure 7(4): The black profile and the red profile have almost perfect Pearson correlation despite the differences in basal expression level and scale.

Squared Euclidean distance

$$d_{E^2}(\mathbf{x}, \mathbf{y}) = (x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2 = \sum_{i=1}^n (x_i - y_i)^2$$

The squared Euclidean distance tends to give more weight to outliers than the Euclidean distance because of the lack of squared root. Data which is clustered using this distance metric might appear more sparse and less compact than the Euclidean distance metric. In addition, This metric is more sensitive to miscalculated data than is the Euclidean distance metric.

Standardized Euclidean distance

This distance metric is measured very similar to the Euclidean distance except that every dimension is divided by its standard deviation:

$$d_{SE}(\mathbf{x}, \mathbf{y}) = \sqrt{\frac{1}{s_1^2}(x_1 - y_1)^2 + \frac{1}{s_2^2}(x_2 - y_2)^2 + \dots + \frac{1}{s_n^2}(x_n - y_n)^2} = \sqrt{\sum_{i=1}^n \frac{1}{s_i^2}(x_i - y_i)^2}$$

This method of measure gives more importance to dimensions with smaller standard deviation (because of the division by the standard deviation). This leads to better clustering than would be achieved with Euclidean distance in situations similar to those illustrated in figure 5.

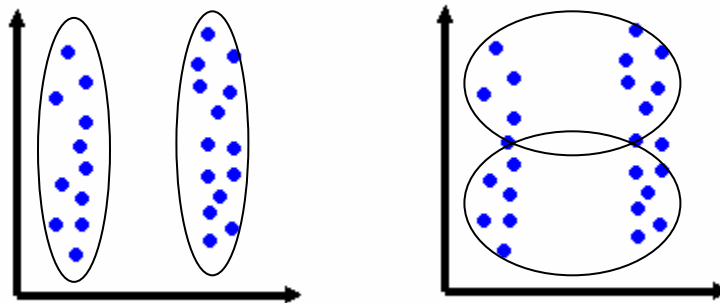


Figure 8: An example of better clustering done when using the Standardized Euclidean distance (left panel) in comparison with the Euclidean distance (right panel). The better results are due to equalization of the variances on each axis.

Mahalanobis distance

$$d_{ml}(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^T S^{-1} (\mathbf{x} - \mathbf{y})}$$

Where S is any $n \times n$ positive definite matrix and $(\mathbf{x} - \mathbf{y})^T$ is the transposition of $(\mathbf{x} - \mathbf{y})$. The role of the matrix S is to distort the space as desired. It is very similar to what is done with the Standardized Euclidean distance except that the variance may be measured not only along the axes but in any suitable direction. If the matrix S is taken to be the identity matrix⁵ then the Mahalanobis distance reduces to the classical Euclidean distance as shown above.

Clustering algorithms

Clustering is a method that is long used in phylogenetic research and has been adopted to microarray analysis. The traditional algorithms for clustering are:

1. Hierarchical clustering.
2. K-means clustering.
3. Self-organizing feature maps (a variant of self organizing maps).
4. Binning (Brazma et al. 1998).

More recently, new algorithms have been developed specifically for gene expression profile clustering (for instance Ben-Dor et al. 1999; Sharan and shamir 2000) based on

finding approximate cliques in graphs. In this section we will focus on the first three traditional clustering algorithms. In addition, we will discuss the main clustering drawbacks and other methods that are used to overcome these drawbacks.

Inter-cluster distances

We saw on distance metric function how to calculate the distance between data points. This chapter discusses the main methods used to calculate the distance between clusters.

Single linkage

Single linkage method calculates the distance between clusters as the distance between the closest neighbors. It measures the distance between each member of one cluster to each member of the other cluster and takes the minimum of these.

Complete linkage

Calculates the distance between the furthest neighbors. It takes the maximum of distance measures between each member of one cluster to each member of the other cluster.

Centroid linkage

Defines the distance between two clusters as the squared Euclidean distance between their centroids or means. This method tends to be more robust to outliers than other methods.

Average linkage

Measures the average distance between each member of one cluster to each member of the other cluster.

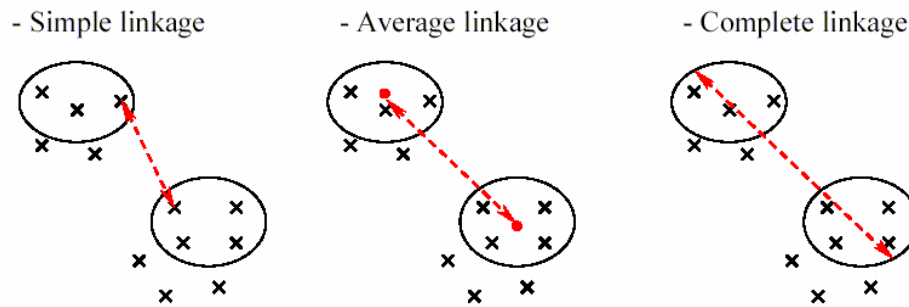


Figure 9(7): Illustrative description of the different linkage methods.

Conclusion

The selection of the linkage method to be used in the clustering greatly affects the complexity and performance of the clustering. Single or complete linkages require the less computations of the linkage methods. However, single linkage tends to produce stringy clusters which is bad. The centroid or average linkage produce better results regarding the accordance between the produced clusters and the structure present in the data. But, these methods require much more computations. Based on previous experience, Average linkage and complete linkage maybe the preferred methods for microarray data analysis⁶.

k-means clustering

A clustering algorithm which is widely used because of its simple implementation. The algorithm takes the number of clusters (k) to be calculated as an input. The number of clusters is usually chosen by the user. The procedure for k-means clustering is as follows:

1. First, the user tries to estimate the number of clusters.
2. Randomly choose N points into K clusters.
3. Calculate the centroid for each cluster.

4. For each point, move it to the closest cluster.
5. Repeat stages 3 and 4 until no further points are moved to different clusters.

The k-means algorithm is one of the simplest and fastest clustering algorithms. However, it has a major drawback. The results of the k-means algorithm may change in successive runs because the initial clusters are chosen randomly. As a result, the researcher has to assess the quality of the obtained clustering.

The researcher may measure the size of the clusters against the distance of the nearest cluster. This may be done to all clusters. If the distances between the clusters are greater than the sizes of the clusters for all clusters then the results may be considered as reliable. Another method is to measure the distances between the members of a cluster and the cluster center. Shorter average distances are better than longer ones because they reflect more uniformity in the results. Last method is for a single gene. If the researcher wants to verify the quality of a certain gene or group of genes, he may do this by repeating the clustering several times. If the clustering of the gene or group of genes repeats in the same pattern, then there is a good probability that the clustering is trustworthy. Although these methods are used widely and successfully, the skeptic researcher may want to obtain more deterministic results which may be done, with some price, by hierarchical clustering.

Hierarchical clustering

Hierarchical clustering typically uses a progressive combination of elements that are most similar. The result is plotted as a dendrogram that represents the clusters and relations between the clusters. Genes or experiments are grouped together to form clusters and clusters are grouped together by an inter-cluster distance to make a higher level cluster.

Thus, in contrast to k-means clustering, the researcher may deduce about the relationships between the different clusters. Clusters that are grouped together at a point more far from the root than other clusters are considered less similar than clusters that are grouped together at a point closer to the root.

The two main methods that are used in hierarchical clustering are bottom-up method and top-down. The bottom-up method works in the following way:

1. Calculate the distance between all data points, genes or experiments, using one of the distance metrics mentioned above.
2. Cluster the data points to the initial clusters.
3. Calculate the distance metrics between all clusters.
4. Repeatedly cluster most similar clusters into a higher level cluster.
5. Repeat steps 3 and 4 for the most high-level clusters.

The approximate computational complexity of this algorithm varies between n^3 , when using single or complete linkage, and n^2 , when using the centroid or average linkage (n is the number of data points).

The top-down algorithm works as follows:

1. All the genes or experiments are considered to be in one super-cluster.
2. Divide each cluster into 2 clusters by using k-means clustering with $k=2$.
3. Repeat step 2 until all clusters contain a single gene or experiment.

This algorithm tends to be faster than the bottom-up approach.

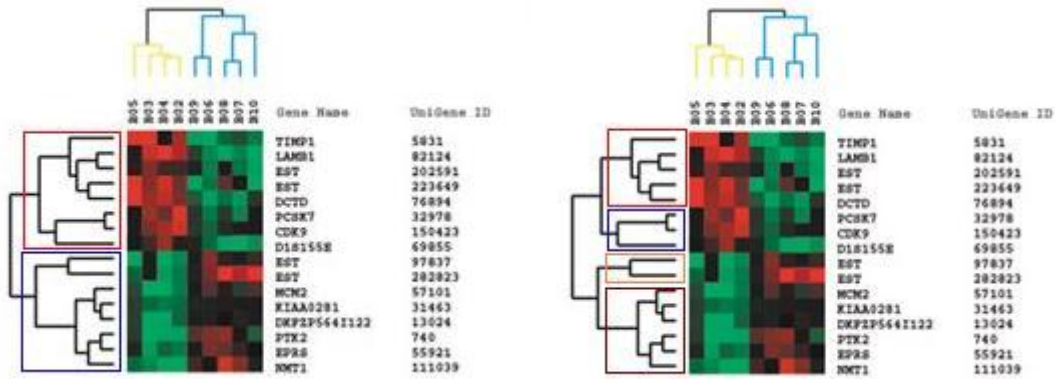


Figure 10: Two identical complete hierarchical trees. The Hierarchical tree structure can be cut off at different levels to obtain different number of clusters. The figure on the left shows 2 clusters while the figure on the right shows 4 clusters indicated by rectangles of different colours.

Self-organizing feature maps

Self-organizing feature maps (SOFM) is a kind of SOM. SOFM as hierarchical and k-means clustering also groups genes or experiments into clusters which represent similar properties. However, the difference between the approaches is that SOFM also displays the relationships or correlation between the genes or experiments in the plotted diagram (see figures 11 and 12). Genes or experiments that are plotted near each other are more strongly related than data points that are far apart. SOFM is usually based on destructive neural network technique (8,9).

Destructive neural network technique is conceptually adopted from the way the brain works. The result of a complex computation is calculated by using a network of simple elements. This is different then conventional algorithms that work by calculating most calculations in one element. An SOFM can use a grid with one, two or three dimensions.

The grid is assembled from simple elements called units. The computational procedure starts with a fully connected grid and reduces (destructs) the number of connections over time in order to better converge to the appropriate classes.

A good description of the basic SOM algorithm is found in Quackenbush's review: "First, random vectors are constructed and assigned to each partition. Second, a gene is picked at random and, using a selected distance metric, the reference vector that is closest to the gene is identified. Third, the reference vector is then adjusted so that it is more similar to the vector of the assigned gene. The reference vectors that are nearly on the two-dimensional grid are also adjusted so that they are more similar to the vector of the assigned gene. Fourth, steps 2 and 3 are iterated several thousand times, decreasing the amount by which the reference vectors are adjusted and increasing the stringency used to define closeness in each step. As the process continues, the reference vectors converge to fixed values. Last, the genes are mapped to the relevant partitions depending on the reference vector to which they are most similar"(11).

SOFMs have some advantages over k-means and hierarchical clustering. SOFM may use a priori knowledge to construct the clusters of genes. This is done by assigning genes with known characteristics to certain units and then inputting the genes with unknown characteristics to the algorithm. The result may supply information about the unknown genes to better understand their functioning or regulation. Other advantages of SOFM method are its low computation complexity and easy implementation.

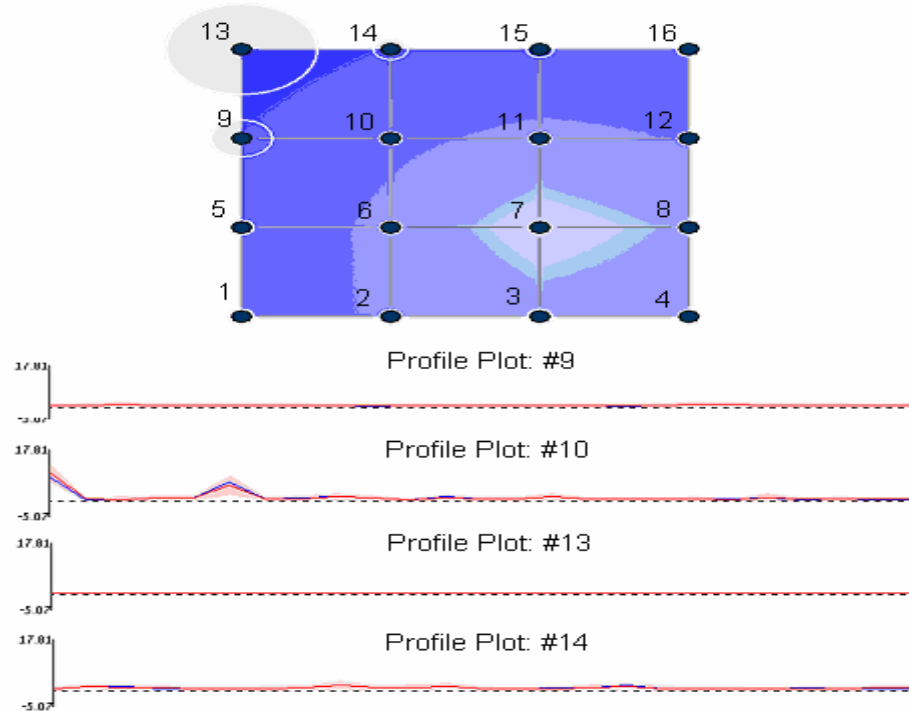


Figure 11: A SOM generated by GeneLinker Platinum™. The clustered data is an example data set. The generated SOM includes 16 clusters numbered 1 to 16. In contrast to the image resulted from k-means or hierarchical clustering, neighbour clusters have similar properties. This can be seen in the profile plots of the neighbour clusters 9, 10, 13 and 14.

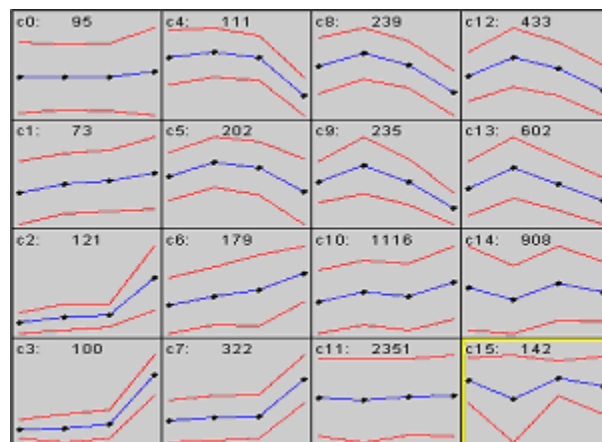


Figure 12: A SOM generated by GeneCluster™. The SOM includes 14 clusters.

It should be noted that neighbouring clusters show similar expression profiles along the experiments. The numbers inside the rectangles represent the number of genes that are clustered in this cluster.

Difficulties and drawbacks of cluster analysis

The clustering methods are easy to implement. However, They have some drawbacks which are inherent in their functioning. K-means have the problem that the k number is not known in advance. In this case the researcher may try different k numbers and then pick up the k number that fits best the data. In addition, k-means clustering may change between successive runs because of different initial clusters. K-means and hierarchical clustering share another problem, which is more difficult to overcome, that the produced clustering is hard to interpret. The order of the genes within a given cluster and the order in which the clusters are plotted do not convey useful biological information. This implies that clusters that are plotted near each other may be less similar than clusters that are plotted far apart.

The essence of the k-means and hierarchical clustering algorithms is to find the best arrangement of genes into clusters to achieve the greatest distance between clusters and smallest distance inside the clusters. However, this problem which is much similar to the TSP⁶ problem is unsolvable in reasonable time even for relatively small data sets. This is the reason that most k-means and hierarchical clustering methods use greedy approach to solve the problem. Greedy algorithms are much faster but, alas, suffer from the problem that small mistakes in the early stages of clustering cause large mistakes in the final

output. This can be partially overcome by heuristic methods that go back in the clustering procedure from time to time to check the validity of the results. Note that this cannot be done optimally because the algorithm would run indefinitely.

Final and very important disadvantage of clustering algorithms is that the algorithm doesn't consider time variation in its calculations. Valafar describes this problem well: "For instance, a gene express pattern for which a high value is found at an intermediate time point will be clustered with another gene for which a high value is found at a later point in time".¹⁰ This problem implies that conventional clustering algorithms cannot reveal causality between genes. One may conclude about causality between genes' expression levels only by considering the time points of genes' expression. A gene expressed at early time point may affect the expression levels of a later expressed gene. The opposite is, of course, impossible. A different approach is needed in order to reveal and illustrate the causality between genes. This may be achieved by a method that is described next.

Alternative method to overcome cluster analysis pitfalls

Reverse engineering of regulatory networks

The methods presented up until now are correlative methods. These methods cluster genes together according to the measure of correlation between them. Genes that are clustered together may imply that they participate in the same biological process. However, one cannot infer, by these methods, the relationships between the genes. The basic questions in functional genomics are: (a) "How does this gene depend on expression of other genes?" and (b) "Which other genes does this gene regulate?" (D'haeseller et al., 2000).

Regulatory networks are also known as genetic networks. These networks' objective is to describe the causal structure of a gene network. Two different ways are used for this purpose: time-series approach and steady-state approach.

Time-series approach

The time-series approach uses the basic assumption that “the expression level of a certain gene at a certain time point can be modeled as some function of the expression levels of all other genes at all previous time points”.¹³ In order to analyze g genes completely we need g^2 linearly independent equations. A linear modeling approach was developed to decrease the dimensionality of the problem. Even so, the number of time points must be at least as large as the number of interactions between the genes studied.

The computation of regulatory network in time-series approach is fairly simple, given that enough time-points are given. The procedure is as follows¹³:

1. Compute the system governing the regulation of each gene in each time point with the equation:

$$x_j(t) = \sum_{i=1}^N r_{i,j} x_i(t-1)$$

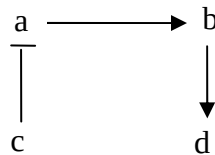
where $r_{i,j}$ is a weight factor representing how gene i affects gene j , positively or negatively.

2. Solve the equation system that is produced in stage 1. Given enough time points this can be done unambiguously. The results may be shown in the following example matrix.

Gene				
Gene	a	b	c	d
a		+		
b				+
c	-			
d				

The pluses in the matrix represent a positive regulation of the horizontal gene upon the vertical gene. The opposite accounts for the minuses.

3. Display the resulted matrix as a regulatory network.



The arrows in the figure represent positive regulation while bars mean negative regulation.

Steady-state approach

The steady-state model measures the effect of deleting a gene on the expression of other genes. If deleting gene *a* causes an increase in expression level of gene *b* than it can be inferred that gene *a* repressed, either directly or indirectly, the expression of gene *b*. Likewise, if deleting gene *a* decreases the expression level of gene *b* than it can be inferred that gene *a* enhanced, either directly or indirectly, the expression level of gene *b*.

The whole regulatory network is constructed by information on the deletion of genes. The resulted regulatory network is a redundant one because many interactions are represented

in many paths. A parsimonious regulatory network may be extracted by deleting arrows which are part of all the paths but the longest one.

Limitations of network modeling

There are many regulatory interactions between proteins. These interactions are not considered at all in the genetic network model. Instead it is assumed that mRNA levels indicate directly the levels of protein products. This suggests that future work should include also posttranslational interactions. Another possible enhancement of the method would be to combine prior biological knowledge, time-series experiments knowledge and steady state experiments results. Last, the results obtained by regulatory networks are practically impossible to validate, because of the immense number of interactions between the genes.

Microarray applications and uses

Microarrays may be used in a wide variety of fields, including biotechnology, agriculture, food, cosmetics and computers. Using the large-scale mRNA measurements we may infer the biological processes in given cells. The cells may be examined a variety of stimuli, at different developmental stages or in healthy against diseased cells. Shedding light on the biological processes within the cells may help us to develop better biological solutions to known problems. We may also use this knowledge to better fit already existing treatments to patients. An example for that is presented next.

There are two distinct types of Lymphoma that conventional clinical methods are unable to distinguish between. Only at very late stages of the disease are the two types distinguishable. With the use of microarrays and building clusters researchers were able to construct groups of gene classifiers to distinguish between the two types of lymphoma even at early stages of the disease. According to different experiments these predictions reach a high confidence of about 90%. The distinction between the two types of lymphoma is very important because the proper treatment can be applied at a stage when the disease can still be healed. The genes in the different clusters may also indicate future research and treatments.

“There are three major tasks with which the pharmaceutical industry deals on a regular basis: (1) to discover a drug for an already defined target, (2) to assess drug toxicity, and (3) to monitor drug safety and effectiveness.”¹⁴ Microarrays may help in all those tasks. By finding genetic regulatory networks, as mentioned above, one can find targets for therapeutic intervention. Drug safety, effectiveness and toxicity also may be examined through the use of microarrays. Thus, the use of microarrays may affect the drug industry

in two ways: shorten the procedure of finding a drug and increase the effectiveness of the drug by fine tuning of its operation.

Microarrays may also help in individual treatments. Drugs that are effective to one patient may not affect another and, even worse, cause unwanted results. With microarray technology, drugs may be costumed to different gene expression profiles. The decrease in the price microarray preparation and analysis can lead to a situation where patient is treated according to his/her gene expression profile. By that side affects may be eliminated and drug effectiveness may be increased.

Conclusions

Microarray is a revolutionary technology. As shown above it includes many stages until a microarray is prepared and further stages until it can be analyzed. All these stages need further research. Currently, microarrays measure the abundance of mRNA in given cells. But, mRNAs go through many stages before they can affect the biological processes in the cell. To mention few, translation, and post-translational changes. A more accurate measurement would be to consider also the abundance of the product of the mRNAs, the proteins and new technologies are under development to take measure of that. Combining these two methods will give more accurate results. The measurement of the mRNAs levels should also be further developed in order to give more credible results.

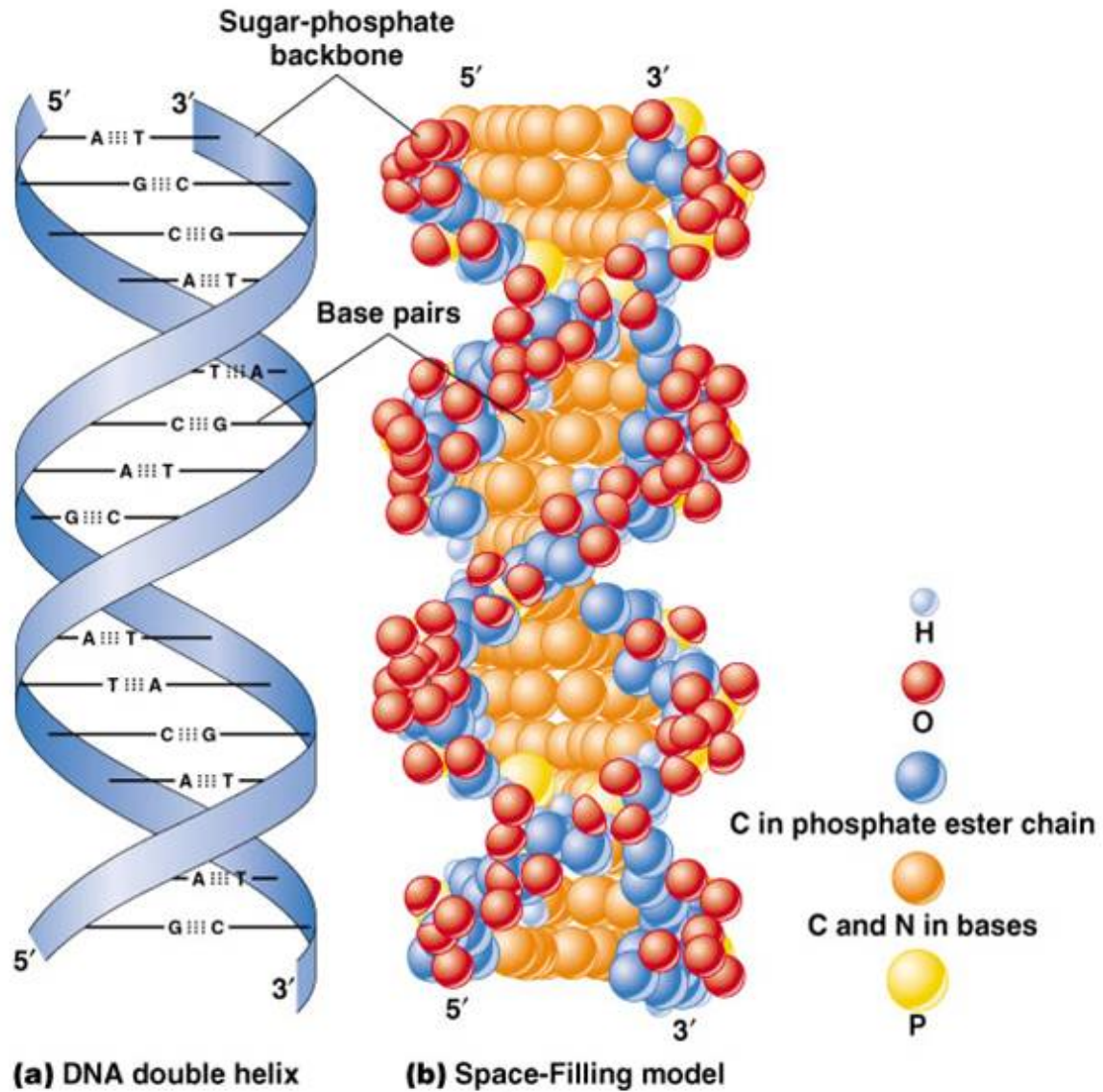
Reaching the interpretation stage also puts many challenges in our way. Clustering methods are fairly easy to implement and, in general, have reasonable computational complexity. However, these methods often fail to represent the real clustering of the data. Clustering methods are, in general, classified as unsupervised methods. Alternative Supervised methods show more accurate results as they include a priori knowledge in the analysis. The undeterministic essence of many clustering methods should also be mentioned as a drawback of the usual clustering method. The researcher may not depend on clustering alone in order to infer anything on the results. “It is a long from finding gene clusters to finding the functional roles of the respective genes, and moreover, to understanding the underlying biological process”.¹² Additional analysis methods should be checked and only then, may conclusions be drawn.

Appendix

General background about DNA and genes

DNA is the central data repository of the cell. It is compound of two parallel strands. Each strand consists of four different types of molecules, which are called nucleotides. The four types of nucleotides are marked as: A (Adenine), C (Cytosine), G (Guanine) and T (Thymine). Thus, each strand is a text composed from 4 letters. Nucleotides tend to bond in pairs. T nucleotide bonds with A nucleotide while C nucleotide bonds with G. The double-helix of the DNA is constructed of two complementary strands. In front of every A nucleotide in one strand there exists a C nucleotide in the complementary strand. The same goes to G and C nucleotides.

The double helix of the DNA (see figure #), which is present in every living cell, is a text. This text includes a series of instructions for protein preparation. Each such prescription is called a gene. When a certain protein is required in the cell, an enzyme called RNA polymerase transcribes the appropriate prescription into RNA. The RNA also consists of four different types of molecules called ribonucleotides. These molecules are very similar to the DNA nucleotides. The RNA, in turn, is translated by the ribosome to protein.



Copyright © 2003 Pearson Education, Inc., publishing as Benjamin Cummings.

Figure 14: Structure of double helical DNA

References

1. Draghici S. *Data Analysis Tools For DNA Microarrays*. Chapman and Hall/CRC, London, 2003.
2. Stanford Microarray Database Analysis Help. OncoLink: Analysis Methods. Retrieved Jan 15, 2003, from <http://genome-www5.stanford.edu/help/analysis.shtml>.
3. Manhattan Distance Metric. Retrieved Jan 15, 2003, from http://www.predictivepatterns.com/docs/WebSiteDocs/Clustering/Clustering_Parameters/Manhattan_Distance_Metric.htm Manhattan Distance Metric.
4. Pearson Correlation and Pearson Squared. Retrieved Jan 15, 2003, from http://www.predictivepatterns.com/docs/WebSiteDocs/Clustering/Clustering_Parameters/Pearson_Correlation_and_Pearson_Squared_Distance_Metric.htm.
5. Bioinformatics toolbox. OncoLink: Scatter Plots of Microarray Data Retrieved Jan 15, 2003, from <http://www.mathworks.com/access/helpdesk/help/toolbox/bioinfo/a1060807757b1.shtml>.
6. BarleyBase Homepage. OncoLink: Analysis Retrieved Jan 20, 2003, from <http://barleypop.vrac.iastate.edu/BarleyBase/>.
7. Ludwig institute for cancer research Retrieved Jan 20, 2003, from <http://ludwig-sun2.unil.ch/~apigni/CLUSTER/CLUSTER.html>.
8. M.T. Hagan, H.B. Demuth, and M.H. Beale. *Neural Network Design*. Brooks Cole, Boston, 1995.
9. J.Hertz, A. Krogh, and R.G. Palmer. *Introduction to the theory of Neural Computation*. Perseus Books, 1991.

10. Faramarz Valafar, 2002. Pattern recognition techniques in microarray data analysis: a survey. *Techniques in Bioinformatics and Medical Informatics* (980) 41-64, December 2002.
11. Quackenbush, J. Computational Analysis of Microarray Data. 2001. *Nature Genetics* 2, 418-427.
12. A. Brazma, A. Robinson and J. Vilo. Gene expression data mining and analysis. *DNA Microarrays: Gene Expression Applications*, Chapter 6. Springer, Berlin, 2002.
13. S. Knudsen. A biologist's guide to analysis of DNA microarray data. Wiley liss, New-York, 2002.
14. A. Fadiel and F. Naftolin, 2003. Microarray application and challenges: a vast array of possibilities.
15. Genomic Signal Processing Lab. Retrieved Jan 22, 2003, from <http://gsp.tamu.edu/Research/Highlights.htm>.

Glossary

1. **Skew** - A distribution is skewed if one of its tails is longer than the other.

Distributions with positive skew are sometimes called "skewed to the right" whereas distributions with negative skew are called "skewed to the left". Skew can be calculated as:

$$\text{Skew} = \frac{\sum (X - \mu)^3}{N\sigma^3}$$

Taken from: HyperStat Online Textbook (last updated Dec 18, 2003). OncoLink: Skew. Retrieved Jan 16, 2003, from <http://davidmlane.com/hyperstat/A69786.html>.

2. **Kurtosis** - Kurtosis is based on the size of a distribution's tails. Distributions with relatively large tails are called "leptokurtic"; those with small tails are called "platykurtic". A distribution with the same kurtosis as the normal distribution is called "mesokurtic". The following formula can be used to calculate kurtosis:

$$\text{Kurtosis} = \frac{\sum (X - \mu)^4}{N\sigma^4} - 3$$

Taken from: HyperStat Online Textbook (last updated Dec 18, 2003). OncoLink: Kurtosis. Retrieved Jan 16, 2003, from <http://davidmlane.com/hyperstat/A53638.html>.

3. In linear algebra, the identity matrix⁴ is a matrix which is the identity element under matrix multiplication. That is, multiplication of any matrix by the identity matrix (where defined) has no effect. The *i*th column of an identity matrix is the unit vector *e_i*.

4. Identity matrix – In linear algebra, the identity matrix is a squared matrix which is the identity element under matrix multiplication. That is, multiplication of any matrix by the identity matrix (where defined) has no effect. The diagonal along an identity matrix contains 1's and all other values equal to zero.
5. TSP - The traveling salesperson has the task of visiting a number of clients, located in different cities. The problem to solve is: in what order should the cities be visited in order to minimize the total distance traveled (including returning home)? This is a classical example of an order-based problems (taken from: The Hitch-Hiker's Guide to Evolutionary Computation (last updated Mar 29, 2000). Retrieved Jan 16, 2003, from http://www.cs.bham.ac.uk/Mirrors/ftp.de.uu.net/EC/clife/www/Q99_T.htm#TRAVELLING%20SALESMAN%20PROBLEM). The computational complexity of such a problem is $N!$, where N is the number of cities (genes) to be visited by the salesperson.