

12.1 Introduction

Microarray is one of the high throughput and large scale technologies available for study of genetics. In the past, hybridization test is used to determine the existence of a DNA/mRNA sequence in a solution. This is known as the "one DNA/mRNA in one experiment" approach. With the emergence of microarray, hybridization tests of thousands of genes can be done in one experiment.

12.2 Hybridization Test

Hybridization test is a technique used in the past to find a subsequence, S , which is unique to the target DNA/mRNA, T . This is enabled by creating the reverse complement S' of S and mixing S' with the solution. S' will hybridize with T if the solution contains T . The sequence S' is known as probe and T is known as the target sequence. This technology makes use of the fact that C basepair with G and A basepair with U. An example is shown in Figure 12.1.

mRNA	5'	GAACCUGAUCAUCCAAGUGG.....	3'
Probe		3'-CUUGGACUAGUAGGUUCACC-	5'

Figure 12.1: The probe can hybridize to the mRNA. Note that the probe is a reverse complementary of a substring of the mRNA.

The problem with this test is that it allows only detection of one DNA/mRNA in one experiment, thus making it difficult to get the whole picture.

12.3 Microarray

The emergence of microarray technology allows simultaneous hybridization tests in a single experiment. A microarray contains thousands ($\approx 60,000$) of probes,

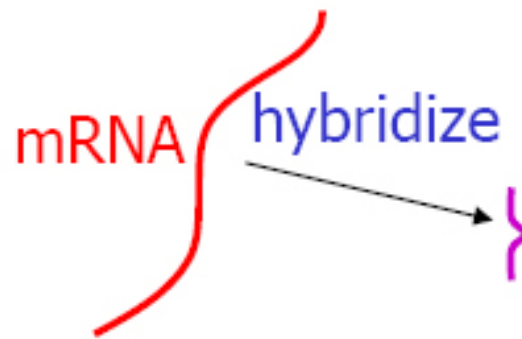


Figure 12.2: The mRNA hybridizes with the probe, which is a reverse complementary substring of the mRNA.

which is about 20-70bp long. Each probe is designed to target its gene of interest and immobilized on a solid surface. The respective target mRNA/DNA hybridizes to their respective probes and emits fluorescence, indicating their presence in the sample. An expression profile (or spectrum) can then be obtained by measuring the level of fluorescence. Expression profiles are presented in an array of numbers, with each element of the array indicating level of fluorescence of each probe.

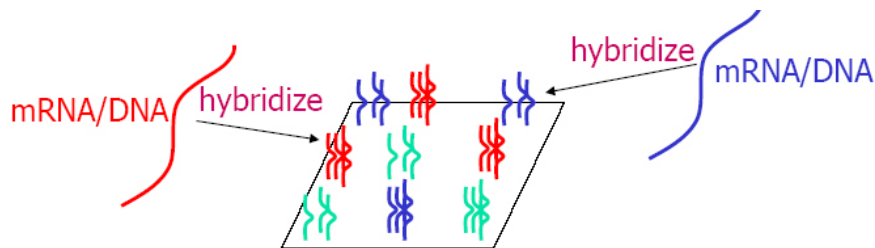


Figure 12.3: The probes immobilized on the array will only hybridize to their target mRNA DNA

12.3.1 Different Types of Microarray

There are different types of microarray. Some examples are DNA hybridization arrays, cDNA microarrays, protein arrays, tissue arrays and combinatorial chemistry arrays.

12.3.2 Applications of Microarray

Microarray has many potential applications, including identification of complex genetic diseases, drug discovery and toxicology studies, mutation Polymorphism

detection (SNPs), pathogen analysis and differing expression of genes over time, between tissues and disease state

All the mentioned applications are made possible by observing and analysing gene expression profile through microarray techniques.

12.3.3 Gene Expression

According to the central dogma, information is passed in the order of DNA to mRNA, then to protein. Base on the assumption that all expressed genes will be translated to protein, gene expression profile can indicate the amount of protein produced. Expression profile of cells is defined as the level of expression of every gene in the cells.

In any cell and at any time, different genes are expressed at different levels. It is the difference in expression profiles that provides us with information on the activities occurring in the cells.

12.3.4 cDNA Microarray

A cDNA microarray contains probes that are short segments of cDNA clone. Each probe represents a gene. To find the expression profile of cells, mRNAs are extracted from the cells and introduced to the array and allowed to hybridize with the probes on the array. Cell expression profile is then obtained by measuring the level of expression of each gene.

12.3.5 Expression Level Measurement Problems

There are two problems encountered on measuring expression levels and using these levels as indicators of expression profiles. First, the mRNA can form secondary structure. Secondary structure of mRNAs may obstruct hybridization, resulting in lower expression level measured. Second, the amount of cDNA segments in different spots are different. Spots with higher amount will have higher measured level, resulting in misleading data.

12.3.6 Comparative Genomic Hybridization

Comparative genomic hybridization has been used to overcome the above mentioned problems. This is done by measuring ratio of expression levels between target cells and reference cells. For example, for analysis of diseases, the target cells are the disease cells and the reference cells are the normal cells. The mRNAs of target cells and normal cells are labeled with different colours, green and red.

The ratio of the expression levels of the two cell types can then be obtained by computing the ratio of red and green colors for every spot.

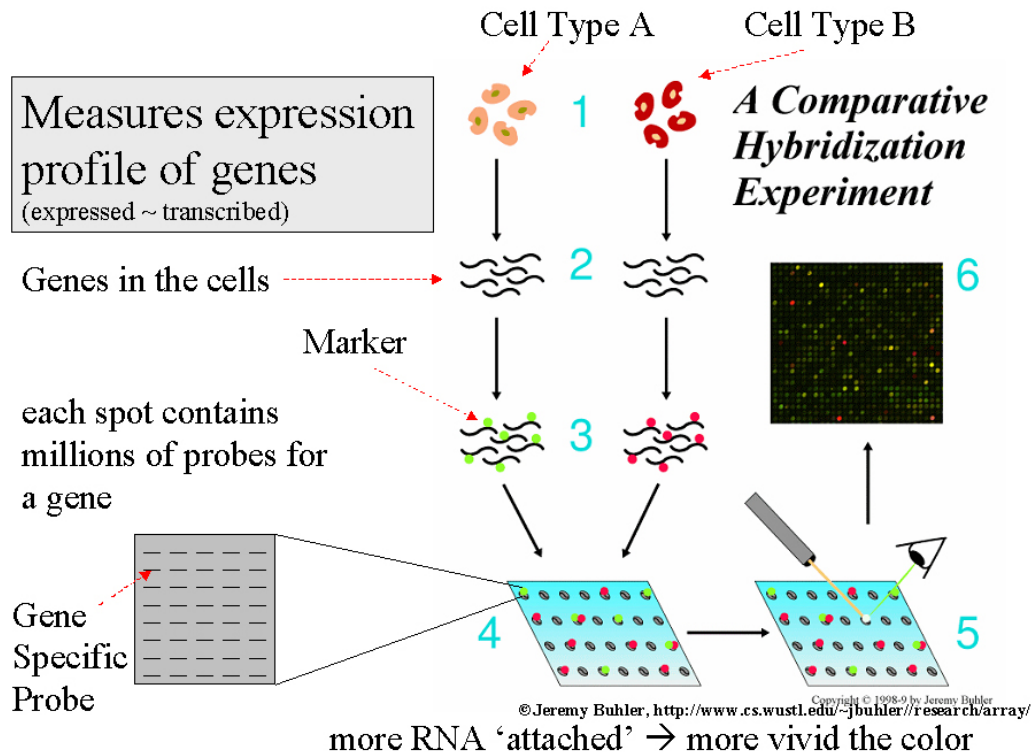


Figure 12.4: **Expression Profile Measurement** mRNAs of Cell type A and cell Type B are extracted and labeled with different labels or markers. Equal amount of both cell types are incubated with microarray to allow hybridization. The ratio between the two markers are then computed.

12.3.7 Applications of cDNA Microarray

There are several applications of cDNA microarray, such as classification, clustering and gene network.

Classification: Sometimes, it is of interest to know the different gene expression pattern of cancer tissue and normal tissue. The difference in gene expression profile of cancer tissue and normal tissue can provide clue in detecting cancer. In some cases, it may be an indication to detect effects of drugs in gene expression by comparing expression pattern of cancer tissue of a patient taking the drugs and that of a patient not taking the drug.

Clustering: Gene expression profiles can also be used to understand gene function. It is believed that co-expressed genes have similar function and base on this belief, gene functions may be deduced by clustering genes which are co-expressed. Another application of cDNA microarray is to understand protein-protein interactions. A paper states that there is correlation between co-expression and protein-protein interaction.

12.4 Probe Design

Given a genome with a set of genes, it is important to design probes such that each probe is a unique signature for its target gene. The length of a probe can be 25, 50 or 70. A good probe should be able to hybridize with the target mRNA at the experiment temperature, which is around 42°C . It should not form secondary structure and it must only hybridize with the target mRNA.

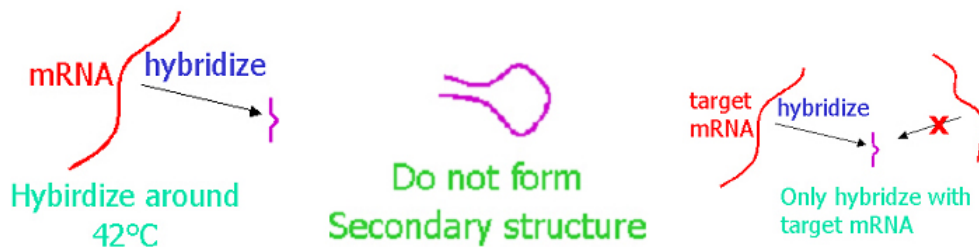


Figure 12.5: A good probe should hybridize to mRNA at 42°C , does not form secondary structure and should only hybridize to target mRNA. These criteria are known as homogeneity, sensitivity and specificity respectively.

12.4.1 Probe Design Problem

The problem with probe design is to identify a probe, p_i for every mRNA X_i . The probe must be a reverse complementary substring of X_i . The probe must not form secondary structure, can hybridize at experimental temperature and does not cross hybridize with other mRNAs. An microarray is created when probes p_i for all mRNAs, X_i are identified.

12.4.2 Cross-hybridization

Consider q is a substring of mRNA X . A probe p is r -cross-hybridized with mRNA X if p and q have less than r mismatches for some q of X (hamming distance). Table 12.1 shows the default r value for different length of probes.

	Length of probe	r
1	20	5
2	50	14
3	70	20

Table 12.1: Default r value for different length of probes

12.4.3 Rules of Lockhart

Lockhart's rules identify probes base on the criteria of homogeneity, sensitivity and specificity. A homogeneous probe will hybridizing at experimental temperature. A sensitive probe will not form secondary structure. A specific probe will hybridize only to its target mRNA.

Example: Given that there exist 3 genes of interest, with sequences CAAGACCGGA, TTACCGATAGGA and GTTATCATC. A suitable probe designed to target the first gene sequence is a sequence complementary to CAAG, since it is homogeneous, sensitive and has no cross-hybridization. A sequence complementary to ACCG cannot be used as a probe since this sequence appears in the first two gene sequences.

12.4.4 Brute-force Algorithm

Brute-force algorithm is used to design probes. In the algorithm, for every mRNA X_i , all possible k-length probes that satisfies homogeneity and sensitivity are checked against r-cross-hybridization with other mRNAs. A probe, p, which satisfies the above stated conditions will be selected as a candidate probe for X_i .

Time analysis: If n is the total length of all mRNAs, there will be $O(n)$ probes. It takes $O(k)$ time to check homogeneity and sensitivity, and $O(n)$ time to check for cross-hybridization. Thus, it takes $O(n^2)$ time in total.

More results: Brute-force approach is too slow. Some better probe designers are proposed and their performances are shown in Table 12.2.

12.5 Clustering

When the expression of two genes are highly similar, it may mean that they have similar function or are co-expressed. Clustering helps to identify genes with similar expression patterns. Data from cDNA array can be expressed as $n \times m$ matrix R where each row represents the expression pattern of a gene and each column represents a condition or cell type. Row i and column j of the matrix

	Li and Storm BIBE 2000	Rouillard, Herber, Zuker Bioinformatics 2000	Kaderali and Schliep Bioinformatics 2002	Rahmann WABI 2002	Sung & Lee CSB 2003
E. Coli 4.6M 4300 genes	23-mers, 1.5days				50-mers, 3.1 mins
S. Cerevisiae 8.9M 6343 genes	24-mers, 4days	50-mers, 1day			50-mers, 49 mins
58 HIV-1 subtypes 600k			20-mers, 9 hours		
Neurospora crassa 38M 10895 genes				25-mers, 4 hours	50-mers, 3.5 hours

Table 12.2: Probe design results

contains the ratio of expression level of gene i under condition j to that of a reference condition. By taking the natural log of these data will yield positive values for upregulated genes and negative values for downregulated genes. Up-regulation of a gene means higher expression level of that gene in that condition compared to reference condition, and vice versa for downregulation. Using the log data to calculate distance or similarity is a better measure due to the fact that downregulated genes will contribute weakly to the calculation if raw ratio is used.

12.5.1 Similarity/Distance Matrix

Preprocessing the input to get a $n \times n$ similarity or distance matrix M before performing clustering. Matrix M contains the similarity/ distance information between two genes, where M_{ij} is the similarity or distance between the expression patterns of Gene i and Gene j . The similarity or distance can be measured by Euclidean distance and Pearson correlation.

Euclidean Distance: One of the methods to calculate distance between two genes is by calculating Euclidean distance. The euclidean distance between gene A and gene B with data (a_1, a_2, \dots, a_k) and (b_1, b_2, \dots, b_k) can be calculated using the equation:

$$\sqrt{\frac{1}{k} \sum_{i=1}^k (a_i - b_i)^2}$$

For example, assume the expression patterns of gene A and gene B are (3, 1, 2, 4) and (2, 2, 2, 3) respectively. The Euclidean distance between them will be:

$$\sqrt{\frac{(3-2)^2+(1-2)^2+(2-2)^2+(4-3)^2}{4}} = 1$$

It is important to note here that using log ratio to calculate Euclidean distance makes more sense here. For example, gene A is upregulated 5 times, having ratio of 5; gene B is downregulated 5 times, having ratio of 0.2; and gene C does not upregulate nor downregulate, having ratio of 1. The distance between gene A and gene C should be the same as the distance between gene B and gene C. However, using this data, the Euclidean distance between gene A and gene C will be 4 and the distance between gene B and gene C will be 0.8. However, if log ratio is used, gene A and gene B distance from gene C will both be $\log 5$. This is illustrated in Table 12.6.

Pearson Correlation: Another measure known as Pearson correlation can also be calculated using the following formula:

$$\frac{1}{k} \sum_{i=1}^k \left(\frac{a_i - \bar{A}}{\sigma_A} \right) \left(\frac{b_i - \bar{B}}{\sigma_B} \right)$$

where \bar{A} is the mean and σ_A is the standard deviation.

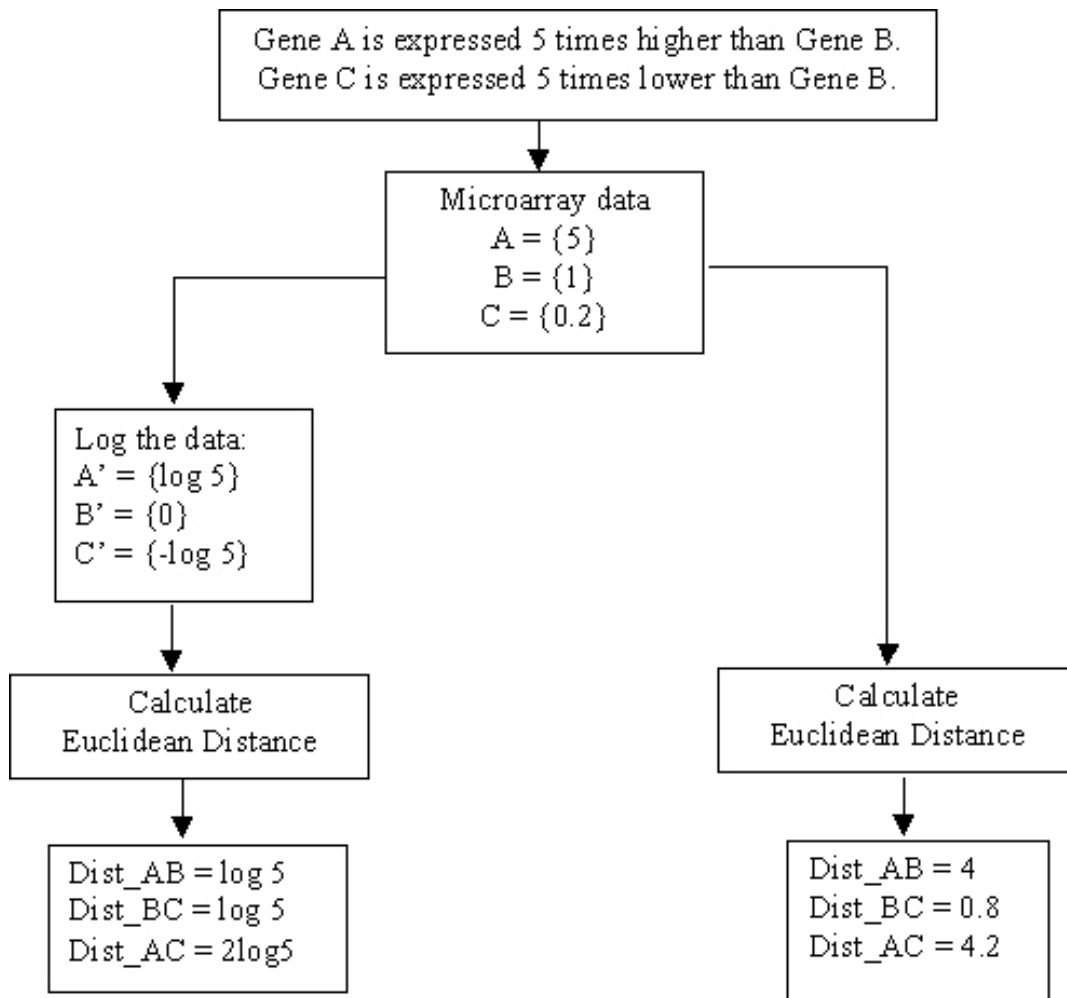


Figure 12.6: This figure illustrates the importance of taking log of the data before calculating score such as Euclidean distance.

12.6 Clustering Methods

After having the similarity matrix for the genes, the next step is to cluster the genes into groups so that genes in the same group have similar expression patterns. Some examples of clustering algorithms are Hierarchical clustering, K-means and self-organizing map (SOM).

Hierarchical Clustering: This approach tries to arrange the genes in the leave of a tree structure so that the closer the genes in the tree, the more similar they are. The tree structure can be constructed using neighbour-joining algorithm. Figure 12.7 shows an example on the clustering result for a microarray experiment.

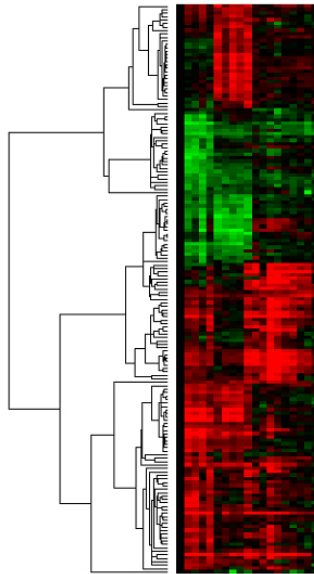


Figure 12.7: An example on hierarchical clustering.

K-means: With input of genes associated with their respective expression pattern, K-means performs clustering of these genes into K clusters/groups.

This algorithm is as follows:

1. Randomly divides the genes into K groups, G_1, G_2, \dots, G_K .
2. The centroid, c_i , for group G_i is computed for $i=1, 2, \dots, K$.
3. A score is calculated base on the formula $\sum_i \sum_{c_i \in G_i} D(g_i, c_i)$, where $D(A, B)$ is the similarity of gene A and gene B.
4. A gene is moved from one group to another and checked if the score can be improved. Perform the best improvement.
5. Goto step 2 until there is no improvement.

References

- [1] WING-KIN SUNG and WAH-HENG LEE, "Fast and Accurate Probe Selection Algorithm for Large Genomes." CSB, 2003.
- [2] TATSUYA AKUTSU, SATORU MIYANO and SATORU KUHARA, "Identification of Genetic Networks from a Small Number of Gene Expression Patterns under the Boolean Network Model." 1999.