Micro-array
Microarrays are microscopic array of single-stranded DNA molecules immobilized on a solid surface by biochemical synthesis also called as DNA chips, gene chips, biochips, DNA microarrays or simply the arrays. Basic idea behind microarray technology is to simultaneously measure the relative expression level of thousands of genes within a particular cell population or tissue. Microarray data analysis helps to get a better understanding of gene expression patterns and regulatory mechanisms.

Steps of Microarray Technology
First Step
Samples of DNA clones or polynucleotides (or probes) with known sequence are spotted and immobilized onto a glass slide or other substrate called “microarray”. Each spot in the microarray corresponds to a gene or an EST (expressed sequence tag).

Second Step
Pools of purified mRNA from cell populations under study are reverse-transcribed (mRNA => cDNA). cDNAs are labeled with one of two fluorescent dyes, “red” (Cy5) and “green” (Cy3). Labeled cDNA (or target) from two cell populations (normally reference and test) are combined and applied to the microarray. Each probe on the microarray should be bound to a quantity of labeled target that is proportional to the level of expression of the gene represented by that probe. Unhybridized cDNA (target) is washed off. Slide is dried and scanned with a scanning laser microscope to measure the brightness of each fluorescent spot. Probe spots are identified and fluorescence intensities are quantified from the high-resolution image and intensity of each hybridization signal is translated into a table with numerical measures using software developed for the purpose of the analysis of the data.

![Figure 1. Sketch of cDNA microarray technology](image_url)
Microarray Data Analysis

There are two broad categories of methods of Microarray data analysis i.e. detection of differentially expressed genes and identification of co-regulated genes. For detection of differentially expressed genes, we use various statistical tests. While for identification of co-regulated or co-expressed genes we use cluster analysis. In cluster analysis, we try to cluster the genes on the basis of similar expression patterns. Assuming that genes exhibiting similar expression patterns may be functionally related, we can assign functions to unknown genes. Moreover, we can also find co-regulated genes which control other genes.

Detection of differentially expressed genes-

For detection of differentially expressed genes we go for various test statistics. If we have only two samples as for example normal and disease, then we use following test statistics-

1. Fold method

Fold change is the simplest method for identifying differentially expressed genes. Expression ratio of a gene in test and reference samples is used in this method. Genes that are upregulated two-fold have an expression ratio of two, and the genes that are downregulated two-fold have an expression ratio of 0.5. But it is known to be unreliable because statistical variability is not taken into account.

2. T-test method

This is the classic statistical approach which is used for detecting differences between two groups. Here variance present in the sample is taken into account.

\[
 t_{n_1+n_2-2} = \frac{\bar{X}_1 - \bar{X}_2}{s \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}
\]

Where s is the pooled variance and n₁, n₂ are sample sizes.

For testing the multiple samples, we go for Analysis of Variance (ANOVA) approach.
1. ANOVA method

For ANOVA, we follow following procedure-

First, we look at the effect of TISSU and GENE factors on logsign variations. The corresponding complete ANOVA model is:

\[ Y_{tgi} = \mu + \alpha_t + \beta_g + \gamma_{tg} + \varepsilon_{tgi} \]

- \( Y_{tgi} \): logsign of the \( i^{th} \) observation of the gene \( g \) in tissue \( t \)
- \( \mu \): mean effect
- \( \alpha_t \): effect of tissue \( t \) with \( t \) in \{A, B\}
- \( \beta_g \): effect of gene \( g \) with \( g \) in \{1,..,1920\}
- \( \gamma_{tg} \): interaction between tissue \( t \) and gene \( g \)
- \( \varepsilon_{tgi} \): residual error

Assumption: \( \{\varepsilon_{tgi}\} \) are independent and normally distributed with equal variance and mean zero \( (\varepsilon_{tgi} \sim \mathcal{N}(0,\sigma^2)) \)

Identification of co-regulated genes

For identifying co-regulated or co-expressed genes, we go for cluster analysis which is broadly divided into two parts, i.e. Hierarchical clustering and K-means clustering.

1. Hierarchical Clustering-

Hierarchical clustering algorithms partition the objects into a tree of nodes, where each node represents a cluster. Each node in a tree has zero or more child nodes, which are below it in the tree; by convention, tree grow down, not up as they do in nature. Hierarchical methods include:

a) Agglomerative: Initially, many small clusters are formed, which are merged based on their similarity. Finally one cluster contains all the objects.

b) Divisive: Initially, all objects form one cluster, which is decomposed into smaller clusters.

Steps in Agglomerative Clustering

- Construct the similarity matrix
- Find largest value in similarity matrix
- Join clusters together, which are having largest value
- Recompute matrix and iterate the process till all the clusters join together
Distances between two clusters are determined by following linkage algorithms.

**Single linkage:** the distance between two clusters is their minimum distance.

\[
\min_{x \in A, y \in B} d(x, y)
\]

**Complete linkage:** the distance between two clusters is their maximum distance.

\[
\max_{x \in A, y \in B} d(x, y)
\]

**Average linkage:** takes the mean distance between all pairs of objects of two clusters.

\[
\frac{1}{\text{card}(A)\text{card}(B)} \sum_{x \in A, y \in B} d(x, y)
\]

**Partition Clustering**

In this approach, objects are partitioned and may change clusters based on dissimilarity. Partitioning methods are useful for bioinformatics applications where a fixed number of clusters is desired, such as small gene expression datasets.

1. **K-Means** - In this, the user specifies the number of clusters k. Clusters have a representative, which is the mean vector, for finding the closest cluster to an object, which minimizes mean-object dissimilarity with a metric such as Euclidean distance. K-Means iteratively assigns objects to the closest of k clusters, and the means get updated. The iteration continues until the number of objects changing clusters is below
Microarray Data Analysis

a user-specified threshold. It deals with numerical attribute values but it is also applicable to binary datasets. It is unsuitable for discovering arbitrary shaped clusters and handling noisy datasets.

**Steps in K-Means Clustering**
- Randomly generate k clusters and determine the cluster centers or directly generate k seed points as cluster centers
- Assign each point to the nearest cluster center.
- Recompute the new cluster centers.
- Repeat until some convergence criterion is met (usually that the assignment hasn't changed).

![Diagram of K-Means Clustering](image)

1: Initial cluster centers (X) are randomly chosen
2: Each gene (O) is assigned to its closest cluster center
3: The centers of the new clusters are calculated.
4: Each gene (O) is reassigned to its closest cluster center, forming new clusters, etc.

**References**
Alter. 2000, “Singular value decomposition for genome-wide expression data processing and modeling” PNAS, 97, 10101
Su-in lee,(2002) group talk: “Microarray data analysis using ICA”.
A technology for specifically targeting genes has been the ‘Holy Grail’ in genetic medicine. The modular structure of zinc finger domain and modular recognition by zinc finger proteins make them the most versatile of DNA recognition motifs for designing artificial DNA binding proteins. Each zinc finger consists of about 30 amino acids and folds into $\beta\beta\alpha$-structures, which is stabilized by the chelation of a zinc ion by the conserved Cys$_2$His$_2$ residues. Each finger typically recognizes a 3 bp DNA sequence by inserting the $\alpha$-helix into the major groove of DNA. Binding of longer DNA sequences is achieved by linking several of these zinc finger motifs in tandem. Individual zinc finger (ZF) domains that recognize DNA triplets with high specificity and affinity can be used to create designer transcription factors and nucleases that are specific for nearly any site in the genome. The engineered ZFPs thus provide us with a powerful platform technology since other functionalities like non-specific FokI cleavage domain (N), transcription activator domains (A), transcription repressor domains (R) and methylases (M) can be fused to a ZFPs to form respectively zinc finger nucleases (ZFN), zinc finger transcription activators (ZFA), zinc finger transcription repressors (ZFR) and zinc finger methylases (ZFM). Computational informatics could play an important role in predicting ZFP specificity, identifying potential target sites and for designing the amino acid sequence of ZFP predicted to bind to a certain target site.

We have developed an online resource of natural and engineered zinc finger proteins and have included several other associated resources for targeted genomic manipulation (http://web.iitd.ac.in/~sundar/zifbase/). Entries can be retrieved either based on the framework desired or on the number of fingers in the engineered proteins or using both the constraints. We have also included a built-in option to search for the availability of engineered zinc finger proteins recognizing user defined binding sites. This option will give the user the advantage to search for engineered ZFPs for the various possible binding sites recognized by this Tool. Using the tool, the amino acid sequence for a ZFP expected to bind to any chosen target site can be generated (http://web.iitd.ac.in/~sundar/zifpredict/). This tool developed by us can be used for designing zinc finger proteins for the chosen gene target within any genome.

ZifBASE provides an extensive collection of various natural and engineered ZFP. It uses standard names and a genetic and structural classification scheme to present data retrieved from UniProtKB, GenBank, Protein Data Bank, ModBase, Protein Model Portal and the
literature. It also incorporates specialized features of ZFP including finger sequences and positions, number of fingers, physiochemical properties, classes, framework, PubMed citations with links to experimental structures (PDB, if available) and modeled structures of natural zinc finger proteins. ZifBASE provides information on zinc finger proteins (both natural and engineered ones), the number of finger units in each of the zinc finger proteins (with multiple fingers), the synergy between the adjacent fingers and their positions. Additionally, it gives the individual finger sequence and their target DNA site to which it binds for better and clear understanding on the interactions of adjacent fingers. The current version of ZifBASE contains 139 entries of which 89 are engineered ZFPs, containing 3-7F totaling to 296 fingers. There are 50 natural zinc finger protein entries ranging from 2-13F, totaling to 307 fingers. It has sequences and structures from literature, Protein Data Bank, ModBase and Protein Model Portal. The interface is cross linked to other public databases like UniprotKB, PDB, ModBase and Protein Model Portal and PubMed for making it more informative.

Reference (The full paper given in the end)
1. Introduction
The challenge and responsibility of agricultural scientists is to provide food, fiber and fuel for the ever-increasing populace. Deciphering genome sequences of model plants and crop plants will help agricultural biotechnologists to understand the plant genome architecture and subsequently to identify and employ economically important genes.

The complete genome sequence of model plant Arabidopsis thaliana ignited plant genomics research. Today, hundreds of plant genomes are being sequenced (fully or partially) resulting in an ocean of sequence data. It is a serious challenge for biologists, statisticians and computer scientists alike to interpret the biological meaning of a combination of only four letters. Further, ESTs (Expressed Sequence Tags) provide a snapshot of whole genome with a small percentage of cost of whole genome sequencing. The next generation sequencing output is also a great resource for plant genomics. These technologies will generate enormous amount of sequencing data, which will be beneficial to understand the basic physiology of plants and functions of thousands of genes. Research involving signal transduction (abiotic stress response in plants); non coding RNAs (miRNAs and siRNAs); molecular markers, and tools of transcriptomics, proteomics and metabolomics; and systems biology approach (to study protein-protein interactome in plants) make use of microarray technology apart from sequencing. Hence, the types and the amount of data will always be on the exponential path.

It is here that bioinformatics and computational biology will be a great help in predicting functions of plant genes. It is also realized how in silico methods along with experimental approaches like two hybrid systems, affinity purification and mass spectrometry, biomolecular fluorescence complementation etc. could be used to understand the proteome of plants. As a result, computational tools, methods of statistical analysis and the software packages become very important and felicity in their usage turns out to be extremely significant.

One of such myriad softwares is JMP Genomics. JMP Genomics is a statistical discovery software tool from SAS and JMP. Major objective in using JMP Genomics is to uncover meaningful patterns in high-throughput genetics, expression, copy number and proteomics data. For a biologist, dynamically interactive graphics make it easy to explore data relationships using a comprehensive set of traditional and advanced statistical algorithms. For a statistician, connecting the data, script and results in a single loop for reiterative operations and extrapolations.
JMP Genomics software brings high-powered, sophisticated genomic data exploration and analysis to the desktop. However, the operations are done on SAS for heavy-duty processing of genomics data sets. By dynamically linking advanced statistics with graphics to provide a complete and comprehensive picture of research results, JMP Genomics helps biologists, biostatisticians and statistical geneticists understand data generated from genetics, expression, exon, copy number and proteomics studies. Users enjoy a menu-driven system that simplifies the workflow throughout the research process and interactive data visualization capabilities that let them see and explore their data from every angle, then easily share findings with colleagues and publish in reputed journals.

At present, we use JMP Genomics 4.1, based on functionality from the SAS 9.2 and JMP 8 platforms. In the course of this training program, we shall not be able to master the JMP Genomics. The objective of the session is to:

1. Know about JMP Genomics
2. Know available menus
3. Understand the available options
4. Run selected options
5. Find help

### Definitions of terminologies often mentioned during the discussion

<table>
<thead>
<tr>
<th>Genetics</th>
<th>The branch of biology that studies heredity and variation in organisms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genetic Analysis</td>
<td>The methods as well as results of systematic study of correlation between traits and genes, the implications of such linkage and their applications</td>
</tr>
<tr>
<td>Gene</td>
<td>DNA sequence that conveys the information required to make a molecule, usually a protein or RNA. Gene possesses the complete information to define the composition of the product as well as the signals that control when and where the RNA/protein should be available</td>
</tr>
<tr>
<td>Population Genetic Analysis</td>
<td>The process of making inferences about the evolutionary and demographic history of a gene (or organism) on the basis of data on genetic variation in a species</td>
</tr>
<tr>
<td>Genome</td>
<td>The entirety of an organism's hereditary information in the form of total nuclear DNA sequence typically expressed in number of base-pairs; contained in the haploid set of chromosomes</td>
</tr>
<tr>
<td>Genotype</td>
<td>Genetic constitution of an individual/lineage/organism particularly expressed as the specific allele makeup with reference to a specific trait under consideration; also a group of organisms sharing a specific genetic constitution</td>
</tr>
<tr>
<td>Haplotype</td>
<td>A contraction of the term haploid genotype; haplotype is a combination of alleles at multiple loci/genes/markers/polymorphisms that are transmitted together</td>
</tr>
</tbody>
</table>