ABSTRACT

Title of Thesis:	**TIME SERIES TRANSCRIPTIONAL PROFILING ANALYSIS OF THE *Arabidopsis thaliana* USING FULL GENOME DNA MICROARRAY AND METABOLIC INFORMATION.**
	Bhaskar Dutta, Master of Science, 2004
Thesis Directed By:	**Dr. Maria Klapa, Assistant Professor** **Department of Chemical Engineering** **Dr. John Quackenbush, Investigator** **The Institute for Genomic Research**

With the advent of the DNA microarray technology, it became possible to study the expression of entire cellular genomes. Tanscriptional profiling alone can not provide a comprehensive picture of the cellular physiological state and it should be complemented by other cellular fingerprints. Transcriptional profiling combined with metabolic information of a systematically perturbed system can unravel the relationship between gene and metabolic regulation. In this context the transcriptional response of *Arabidopsis thaliana* liquid cultures (grown for 12 days under light and 23^0C) to 1-day treatment with 1% CO_2 was measured by full genome cDNA microarrays. The Time series gene expression profiles were analyzed in the context of the known *Arabidopsis* *thaliana* physiology using multivariate statistics. Data analysis revealed an increase in the rate of CO_2 fixation, biomass production and cell wall growth. The breadth of the information obtained from a single experiment validated the significance of the high throughput transcriptional profiling.

TIME SERIES TRANSCRIPTIONAL PROFILING ANALYSIS OF THE *Arabidopsis thaliana* USING FULL GENOME DNA MICROARRAY AND METABOLIC INFORMATION

By

Bhaskar Dutta

Thesis submitted to the Faculty of the Graduate School of the
University of Maryland, College Park, in partial fulfillment
of the requirements for the degree of
Master of Science
2004

Advisory Committee:
Professor Maria Klapa, Chair
Dr. John Quackenbush
Professor William E. Bentley

Dedication

To my parents

Acknowledgements

I would like to take this opportunity to thank my advisors, Professor Maria Klapa and Dr. John Quackenbush, for providing me with the opportunity to work on such an exciting and cutting-edge research project and their constant encouragement and valuable advice. Working with them has allowed me to become more thorough in my conceptual understanding of the field of systems biology in general and functional genomics in particular. In addition, I would like to extend thanks to my other committee member, Professor William E. Bentley.

I would also like to thank Tara Vantoai, Fenglong Liu and Linda Moy along with my advisors for conducting the experiments.

Table of Contents

List of Tables

List of Figures

Chapter 2: Transcriptional Profiling

Chapter 3: Plant Physiology Under Conditions of Elevated CO_2

Chapter 4: Results

Chapter 5 Discussion

Chapter 6 Future work

1. Introduction

Advent of DNA microarray [Brown P. et al., 1999; Fodor et. al., 1997; Schena et. al., 1996] enables the simultaneous measurement of the expression of thousands of genes. DNA microarray provides vast amount of data which has to be processed and analyzed effectively to obtain underlying biological information. Several multivariate statistical techniques are used for data analysis [Quackenbush, 2002]. Data analysis can identify patterns of gene expression. Grouping of genes according to the expression pattern can provide greater insight about their biological function because, genes that are functionally related are expected to have similar expression pattern [Eisen, 1998].

Transcriptional profiling alone cannot delineate the cellular function [Klapa et al, 2003]. A comprehensive analysis of the biological systems requires finger printing of the cellular responses at different levels of cellular function

To study a system, whose comprehensive first principle model is not known, it is a common practice to systematically perturb the system and see how it behaves. Systematically perturbing a system means, intentionally perturbing one or combination of input variables of the system to a certain level, keeping others constant, so that its effect in the output variables can be observed. This helps to find a relation between the variable that was perturbed and the variables that reflected the change. Similarly studying multiple perturbations involving different variables can be used to generate a comprehensive idea of the system.

1.1 Motivation:

High throughput analysis has driven the research from hypothesis based to data driven. The advantage of data driven analysis is, no prior hypothesis is required. Full genome DNA microarray being a high-throughput technology, can be used for simultaneous measurement of the _expression of all the genes of a cellular genome. The results obtained from a single experiment can not only verify the results obtained from multiple conventional experiments, but can also reveal lot of new information.

Genes that have similar expression pattern have a very high chance to be co-expressed or functionally related. So, comparison of expression patterns can reveal valuable information about transcriptional regulation and gene interaction. To the best of our knowledge, this was the first effort, to conduct a time series experiment in plants which can study dynamic response in terms of expression profiles. Time series analysis also reveals the time scale response of genes and metabolites and difference between them.

Existing information about the metabolic pathways was used in conjunction with gene annotation to make use of the metabolic information for better understanding of the transcriptional profiling results. *Arabidopsis thaliana* was chosen as model plant because, it has fully sequenced genome and it is well studied. So the obtained results can be compared with the existing literature. CO_2 perturbation allowed the known central carbon metabolism to be perturbed, so that effect of the perturbation can be studied at the transcriptional level in the context of the known metabolic pathways

1.2 Objective and specific aims:

Main objective of this thesis is to study the transcriptional profiling analysis of short term *Arabidopsis thaliana* response to elevated CO_2 and how the knowledge about the metabolic network structure and regulation might be used in improving transcriptional profiling. To achieve this objective the following specific aims were pursued:

1) To study short term time series response of *A. thaliana* liquid cultures to elevated CO_2

2) Use of full genome DNA microarray to study the response at genomic level.

   a) Use of TIGR TM4 open-source software for DNA microarray data analysis.

3) Use of information obtained from well characterized metabolic pathways to better elucidate the clustering obtained from statistical techniques.

4) Discuss the results in the context of the known *Arabidopsis Thaliana* physiology.

1.3 Description of the thesis:

Chapter 1: Describes the main objective and specific aims of the presented work, in the context of the transcriptional profiling research. A short description of each chapter of the thesis is also provided. .

Chapter 2: It provides a brief introduction to DNA microarray technology and a detailed description of different normalization and clustering techniques used for microarray data processing and analysis. The techniques described were used to analyze the data of the present study.

Chapter 3: It describes the mechanism of the CO_2 fixation in plants how it is affected by elevated CO_2 based on previous studies. The consequences of the elevated CO_2 to other pathways of the plants are also discussed.

Chapter 4: Experimental design was explained. Data analysis steps were discussed in detail in the context of TIGR TM4 software. Results obtained were discussed in reference to established metabolic pathways and metabolic profiling results obtained for the same experiment.

Chapter 5: The results obtained in chapter 4 was discussed in the context of the known *Arabidopsis thaliana* physiology and were compared with these from previous studies (as the latter were discussed in Chapter 2). Possible limitations of the statistical techniques used in data analysis and the experimental setup are also mentioned.

Chapter 6: Ideas for future work based on the conclusions derived in the present one, concerning both the experimental design and the data analysis, are discussed.

2. Transcriptional Profiling:

The advent of oligo-nucleotide arrays and cDNA microarrays has enabled biologists to measure the expression levels of thousands of genes [Brown et al., 1999] in parallel. The wealth of data generated from DNA microarray can be used to develop a more complete understanding of the gene function, regulation and interactions. The most powerful applications of expression data (gene expressions obtained from microarray) is to study of patterns of gene expression across many experiments that study a wide array of cellular responses, phenotypes and conditions [Quackenbush 2001]. Identifying patterns of gene expression and grouping the genes according to the pattern might provide us much greater insight about their biological function and relevance. The genes with similar expression pattern are called co-expressed.

Gene expression analysis is based on two main "assumptions":

1. Genes that are functionally related are expected to be co-expressed. For example Eisen et al. (1998) has shown that genes encoding for parts of a protein complex had similar expression patters.

2. Genes that are co-regulated in the gene regulation network are expected to be co-expressed.

2.1 DNA microarray technology:

Two different technologies are used for microarray slide preparation [Vivian et al., 1999]. Commercially it is manufactured by Affymetix [http://www.affymetrix.com]. It is produced by adding nucleotides sequentially using photolithographic technique to get desired sequence of oligo-nucleotides attached to the plate. The other technology cDNAs are printed onto chemically modified glass slides with the help of an arraying robot [Brown et al., 1999] and called spotted arrays. For this experiment spotted arrays printed in TIGR were used. In the rest of the document microarray refers to spotted array.

The first step in the preparation of microarray slides is proper probe (the sequence that are arranged on the microarray) selection. Then the probes are spotted. The arrayed genes are probes that can be used to query pooled, differentially labeled targets derived from RNA samples from different cellular phenotypes to determine the relative expression levels of each gene.

Two mRNA samples, one for control and another for query, from the tissues of interest are labeled with two different fluorescent dyes Cy3 and

Cy5. Then they are purified and hybridized on the arrays. After hybridization, slides are scanned and independent images for control and query channels are generated. The relative fluorescence intensities give us a measure of relative amount of mRNA in control and query. After image processing data are normalized. Normalization adjusts for differences in labeling and detection efficiencies for fluorescent labels and for difference in the quantity of initial mRNA from the two samples [Quackenbush, 2001].

The normalized value of the expression level for a particular gene in the query sample divided by its normalized value for the control is called “expression ratio” [Quackenbush, 2001]. Logarithm of the expression ratio is used because it is easy to understand. Genes that are up-regulated by a factor of two have a expression ratio of 2, hence $\log_2$(expression ratio) will have a value of 1. Similarly the genes that are down-regulated by the same factor will have a expression ratio 0.5 and $\log_2$(expression ratio) as -1. If the logarithm of expression level ranges between 1 to -1 then the expression level varies within 2 fold. So taking the logarithm of the expression makes the expression profile symmetric for a certain factor of up and down regulation.

There are number of data analysis steps followed in sequence after the microarray slides are hybridized and scanned. TIGR TM4 software was used for microarray data analysis and the steps will be discussed in this context.

2.2 Image Processing

TIGR TM4 software spotfinder was used for image processing. The TIFF image files generated from the scanning of hybridized files is used for image processing. Image processing software takes the scanned image of both the dyes corresponding to each slide. Spotfinder generates TAV file which contains the information like position of the spot on the slide, intensity of the two dyes for each spot and whether the spot should be rejected or not.

2.3 Data Normalization

In many field comparisons are needed to extract conclusions, for an effective comparison appropriate normalization of the data is needed. In the context of DNA microarray analsis there is need for comparison among

i. Two different dyes

ii. Gene spots on the same slide

iii. Gene spots on different slides

In this process the source of systematic error that introduces difference between comparable data should be taken into consideration, so that data are compared only with respect to experimental perturbation. In the case of cDNA microarray analysis, such sources of systematic error arise in the experimental process of cDNA microarray development and hybridization. Following are sources of systematic error:

- Unequal quantities of starting RNA: in cDNA microarray RNA concentration of sample set is measured with respect to a reference. Equal amount of sample and reference RNA is taken so that they can be compared get relative expression of the sample with respect to reference.
- Difference in labeling efficiencies: fluorescent dye is attached to a mRNA sample through a biochemical reaction. Some dye can have preferential binding to one of the mRNA samples. Hence that mRNA sample will always be shown at higher abundance compared to the other mRNA sample.
- Difference in scanning efficiencies: sample and reference are attached with two different dyes and after hybridization the slide is scanned for two different dye intensities in two different channels. Difference in sensitivity of the scanner for the two dyes can cause one of the dyes to be detected more effectively.
- Variation of the intensity across the slide: cDNA microarray is printed by a pen assembly and different parts (metablocks) of a microarray are

printed by different pens. If there is variation among pens, this will translate into variation in the spots printed by different pens.

To account for the systematic errors various normalization methods have been proposed. In the rest of the text only those used in the present analysis in the context of MIDAS (TIGR TM4 software for normalization) are explained in greater detail.

2.3.1 Total intensity normalization:

Total intensity normalization can eliminate the biases caused by difference in labeling and scanning efficiencies of the two dyes. It can also compensate for the unequal quantities of starting mRNA of the two sets. The total intensity normalization is based on the following hypothesis [Quackenbush 2002]. If the two samples to be compared have equal weight of mRNA, if the average mass of each molecule is approximately the same then each sample will have equal number of mRNA. It is also assumed that arrayed genes on the microarray slide equally interrogate the two mRNA samples. Hence the total number of mRNA molecules attached to the microarray slide is same for the two samples. Intensity of a spot is proportional to the amount of mRNA bound to the spot. As the total amounts of mRNA with two different dies are equal, the total fluorescent intensity for each die will also be equal. This can be checked by calculating the ratio of sum of intensities of two dyes, called normalization factor and is given by

$$N_{total} = \frac{\sum_{i=}^{N_{array}} R_i}{\sum_{j=1}^{N_{array}} G_j} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.. (2.1)$$

where R_i and G_i corresponds to the intensity of the red and green dye (two dyes used for two samples) for ith gene and N_{array} is total number of genes in the slide. In absence of any systematic error N_{total} value should be 1. When the value is not 1, then one of the samples (depending on which one is taken as reference) is scaled up or down depending on the value of N_{total}, so that, after the scaling the sum of the intensities of both the dyes are same. This process is equivalent to subtracting a constant from the logarithm of expression ratio.

$$\log_2(t_i) = \log_2(T_i) - \log_2(N_{total}) \qquad \ldots..\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(2.2)$$

where, t_i is normalized expression ratio and is given by

$$t_i = \frac{R_i}{\mathrm{N}_{\mathrm{total}} G_i} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(2.3)$$

T_i is expression ratio before normalization and is given by

$$T_i = \frac{R_i}{G_i} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots.(2.4)$$

N_{array} can be the number of genes on a section of the slide, a whole slide or number of slides. In the same way as above, in stead of comparing mean intensities, median intensities of the two samples can also be equated.

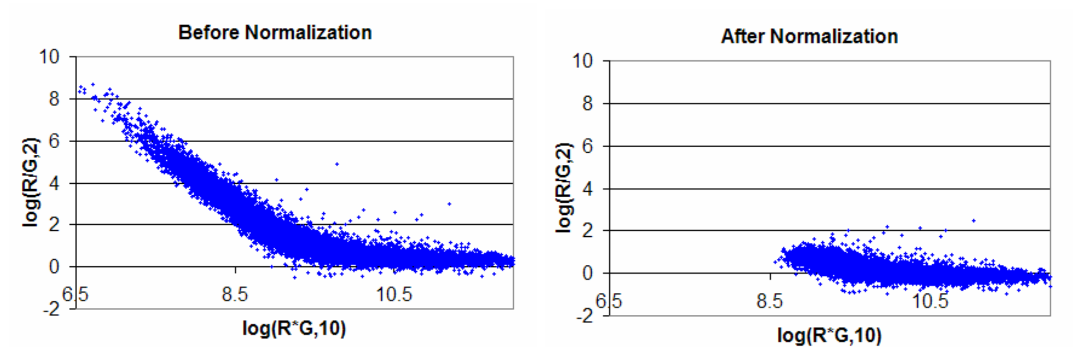


Figure 2.1: RI plot before after total intensity normalization. (R-I plot obtained from the data of one of the time points of the experiment, displaying the ratio of the intensities ($\log_2(R_i/G_i)$) as a function of the product of the intensities ($\log_{10}(R_i*G_i)$) before and after total intensity normalization.)

2.3.2 Lowess:

It is observed very often that $\log_2(R_i/G_i)$ values can have a systematic dependence on intensity [Yang Y. et al., 2002 and Yang I. et al., 2002], which most commonly appears as a deviation from zero for low or high intensity spots. This leads to a long tail in R-I plot (plot of ratio of the intensities ($\log_2(R_i/G_i)$) as a function of the product of the intensities, $\log_{10}(R_i*G_i)$). Locally weighted regression (Lowess) [Cleveland et al 1979] can take care of this systematic error in microarray data. It carries out a locally weighted regression between $\log_{10}(Ri*Gi)$and $\log_2(Ri/Gi)$ and gets the best fit curve which predicts $\log_2(Ri/Gi)$ as a function of $\log_{10}(Ri*Gi)$. Best fit curve, which captures the systematic error in the data, is subtracted from each data ($\log_2(Ri/Gi)$) point to remove the systematic error in the data. The weights assigned in this locally weighted regression are function of the distance of the data points from the fitted curve. If a point is far from the curve then it has very low weight, as the point has more chance of being an outlier. Lowess carries out the regression for each block of the microarray slide separately. Lowess can also be applied globally by considering whole data set (all the spots of the microarray slide).

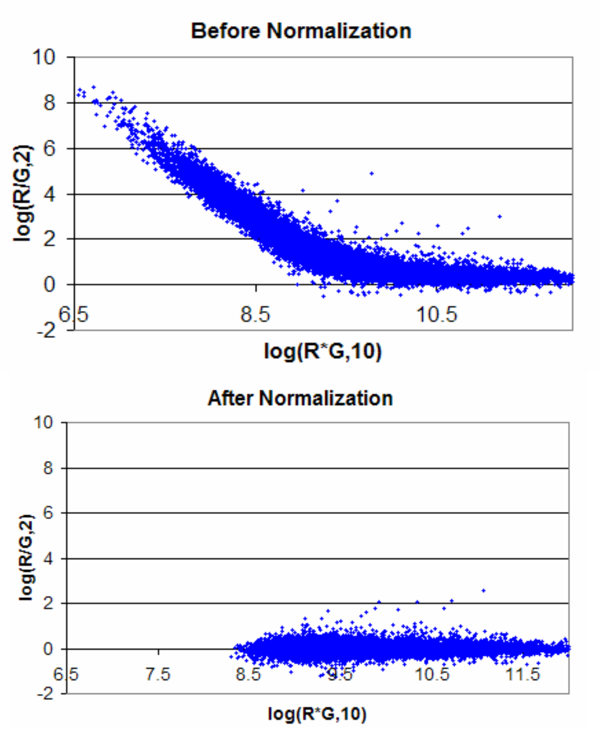


Figure 2.2: RI plot before and after lowess normalization

The data after total intensity normalization in Fig 2.2 shows a systematic bias in RI plot. The plot is showing a small tail at low intensity values due to systematic error. This error is eliminated in the data after lowess normalization (Fig 2.2).

2.3.3 Standard Deviation Regularization:

In the above normalization methods mean intensity of the two sets are equated. How the points are scattered around the mean is also an important criterion to study. In a spotted array different meta-blocks are printed by different pens, so the spots may vary slightly from meta-block to meta-block due to difference in pen. Standard deviation regularization scales the data so that there is same variation for all the meta-blocks. [Yang Y. et al., 2002],

It is assumed that the mean of $\log_2$(ratio) is already zero for each meta block, by applying the normalization methods discussed above. So the variance of the nth meta-block will be given by

$$\sigma^2{}_n = \sum_i^N \left(\log_2(T_i)\right)^2 \qquad \text{.............................(2.5)}$$

where T_i is ratio of the dye intensity for ith gene and is given by

$$T_i = \frac{R_i}{G_i} \qquad \text{.............................(2.6)}$$

N is the number of spots in a meta-block. Appropriate scaling factor for the jth meta-block is given by

$$a_j = \frac{\sigma_j{}^2}{\left[\prod_{k=1}^{N_{metablock}} \sigma_k{}^2\right]^{1/N_{metablock}}} \qquad \text{.............................(2.7)}$$

where $N_{metablock}$ is the number of meta-blocks in a slide. All the elements of the jth meta-block is scaled by dividing them with the scaling factor. Hence

$$\log_2(T_i) = \frac{\log_2(T_i)}{a_j} \qquad \text{.............................(2.8)}$$

Where, T_i is the ratio of red to green dye intensity for the i^{th} gene in the j^{th} meta-block. This is same as taking the a_j th root of all the intensities of the j^{th} meta block. So the transformed intensities after the normalization become:

Or $G_i' = [G_i]^{1/a_j}$ and $R_i' = [R_i]^{1/a_j}$ ………………………….(2.9)

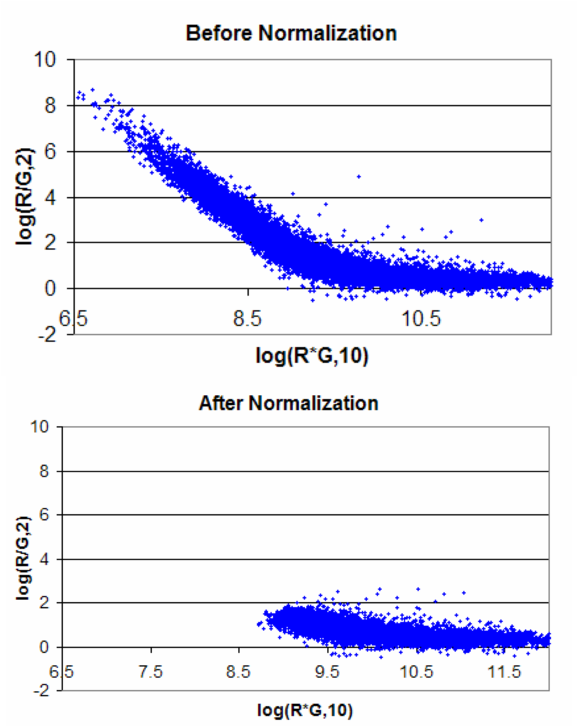


Figure 2.3: RI plot before and after standard deviation normalization.

2.3.4 Flip dye analysis:

By performing a flip dye analysis biases that may occur during labeling and scanning, for example, some die may preferentially bind to mRNAs, can be eliminated [Quackenbush, 2002]. If one of the dyes has higher average intensity over the other, then the sample tagged with that dye will show higher expression, which is misleading. So the same experiment is carried out by swapping the dyes among the samples. If there are two samples A and B, then they can be tagged by two possible combinations, red and green or green and red dye respectively. In the first case when A and B are attached with red and green dye respectively, the ratio will be given by

$$T_{1i} = \frac{R_{1i}}{G_{1i}} = \frac{A_{1i}}{B_{1i}} \qquad \text{…..……………………}(2.10)$$

After the dyes are reversed the ratio will become

$$T_{2i} = \frac{R_{2i}}{G_{2i}} = \frac{B_{2i}}{A_{2i}} \qquad \text{…..……………………}(2.11)$$

As the same experiment is being performed and only the dyes are reversed, $\frac{A_{1i}}{B_{1i}}$ and $\frac{A_{2i}}{B_{2i}}$ are expected to be same. Hence

$$\frac{A_{1i}}{B_{1i}}\frac{B_{2i}}{A_{2i}} = (T_{1i} * T_{2i}) = 1 \qquad \text{…..……………………}(2.12)$$

$$\log_2\left(\frac{A_{1i}}{B_{1i}}\frac{B_{2i}}{A_{2i}}\right) = \log_2(T_{1i} * T_{2i}) = 0 \qquad \text{…..……………………}(2.13)$$

If the measurements are consistent then the value of $\log_2(T_{i1}*T_{i2})$ is expected to be zero, if it is not zero then close to zero. But if the value is far from zero, then the measurements are inconsistent. Either one of the measurements or both could be erroneous. The user can decide how stringent the rejection criteria of the erroneous data would be. Stringent criteria means only a small range of values around zero is acceptable.

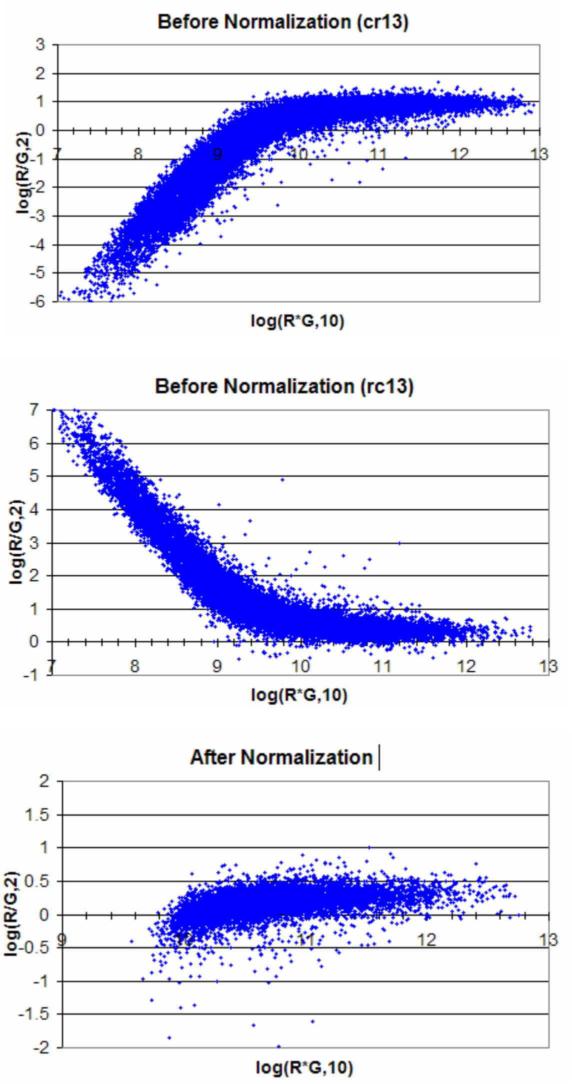


Figure 2.4: RI plot before and after flip dye normalization. Before normalization the RI plots has long tails and look like mirror image with respect to the line y =0 line. After normalization the dye based bias is gone

2.4 Clustering Methods/ Statistical Analysis of DNA Microarray Data:

Several clustering algorithms are used for the identification of the patterns in the gene-expression data. Clustering techniques can be classified as decisive or agglomerative [Quackenbush 2001]. A decisive method begins with all elements in one cluster that is gradually broken down into smaller and smaller clusters. Agglomerative techniques start with single member clusters and gradually fuse them together. There are two types of clustering algorithms supervised or unsupervised [Quackenbush 2001]. Supervised methods use existing biological information about specific genes that are functionally related to 'guide' the clustering algorithm. Most of the algorithms described in this chapter are unsupervised.

2.4.1 Distance Metrics:

Suppose N number of experiments is conducted to study the expression profiles of M genes. Then the expression of a particular gene in N experiments can be represented by a single point in N dimensional space. This is called expression space, as it has the same number of dimension as the number of experiments. Clustering algorithms group the genes together based on their "distance" from each other in the expression space. Distance gives a measure of similarity between the genes. There are various methods for calculating distances.

1. **Euclidean distance** is the most commonly used distance. It is a metric distance. Following are the characteristic of metric distances [Quackenbush 2001]. If d_{ij} is the distance between two vectors i and j,

$$d_{ij} = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})^2} \qquad \text{……………………………(2.14)}$$

where, x_{ik} and x_{jk} are expression level of i^th and j^th genes respectively and n is the number of experiments

- Distance must be positive and definite, d_{ij} >o
- Distance must be symmetric, $d_{ij} = d_{ji}$
- An object is zero distance from itself, $d_{ii} = 0$
- It follows triangular inequality

2. **Manhattan distance** is given by:

$$d_{ij} = \sum_{k=1}^{n} | x_{ik} - x_{jk} | \qquad \text{………………………… (2.15)}$$

where n is the dimension of the expression space [Heyer et al., 1999].

3. **Pearson correlation** is given by [Eisen et al., 1998]

$$S(G_i, G_j) = \frac{1}{n}\sum_{k=1}^{n}\left(\frac{G_{i,k} - G_{i,offset}}{\Phi_i}\right)\left(\frac{G_{j,k} - G_{j,offset}}{\Phi_j}\right) \qquad \text{………………… (2.16)}$$

Where,

$$\Phi_i = \sqrt{\sum_{k=1}^{n}\frac{(G_{i,k} - G_{i,offset})^2}{n}} \qquad \text{……………………… (2.17)}$$

$G_{i,\ offset}$ is the mean and Φ_i is the standard deviation of observation of the ith gene.

4. **Cosine correlation** is given by the following expression [Eisen et al., 1998]

$$C(x_{ik}, x_{jk}) = \frac{\sum_{k=1}^{n} x_{ik} x_{jk}}{\|x_{ik}\| \|x_{jk}\|} \qquad \text{………………………} (2.18)$$

Distance between two clusters can be calculated in different ways:

Average linkage clustering: This is most frequently used. The distance between two clusters i and j is calculated by calculating the average of the distance between each gene of ith cluster with all other genes in the jth cluster. Two clusters with lowest average distance is joined together to form a new cluster.

Complete linkage clustering: Complete linkage clustering is known as the maximum or furthest-neighborhood method. The distance between two clusters is calculated as the greatest distance between the members of relevant clusters. This method often produces clusters that are often similar in size.

Single linkage clustering: The distance between two clusters is calculated as the smallest distance between the members of the relevant clusters. In this method there is a sequential addition of single samples in to an existing cluster. This produces trees with many long, single addition branches representing clusters that have grown by accretion.

If the expression level of a gene at each time point is viewed as a coordinate, then the standardized expression level of each gene at all n time points describes a point in *n* dimensional space, and the Euclidean distance between any two points in this space can be computed. It can be shown that the two points for which the distance is minimized are precisely the points that have the highest correlation. In other words, genes pairs with highly correlated expression pairs are close in expression space. It should be noted that simply using Euclidean distance without standardizing the data is ineffective, because gene pairs whose expression patterns have the same shape but different magnitudes will not score well.

To gauge the measure of a performance, one might consider taking gene pairs those are known to be co-regulated or functionally related, and computing the score (distance or correlation) of each pair. These scores could then be compared with the scores of unrelated gene pairs. The measure that gives high scores only to related genes would be chosen. Unfortunately neither Euclidean distance nor Pearson Correlation consistently gives high scores only to related gene pairs. In fact, not all related genes are coexpressed, and some unrelated genes have similar expression patterns. Because there is a connection between coexpression and functional relation, coexpressed genes provide excellent candidates for further study. However, the connection is complex, and it cannot be derived so easily [Heyer et al.,1999].

Two genes may be close according to one distance definition but may be far apart according to other. So the way we define distance between two expression vectors has a profound effect on the cluster they produce.

To study gene expression patterns statistical and clustering techniques have been proposed. In the rest of the text only the techniques that were used for the resent analysis will be discussed in detail.

2.4.2 Hierarchical Clustering:

Hierarchical clustering is one of the first and widely used clustering techniques for expression data. The reason being, it is simple and the results can be visualized easily. Hierarchical clustering is an agglomerative approach in which expression profiles are joined in groups, which are further joined and this continues till completion, so that finally it forms a single tree. The algorithm of Hierarchical clustering is as follows. Initially each cluster contains a single gene. Then the pair-wise distance is calculated for all of the genes to be clustered. If they are formulated in a matrix form it forms a square matrix which is symmetric. This matrix is called distance matrix or similarity matrix. This matrix is scanned to figure out smallest value (if Euclidean distance is used, because it selects the genes that are closest in the expression space) or highest value (if Pearson correlation distance is used, because it finds the genes that have most similar expression profile). These two genes are most similar or closest, hence they are clustered together. If several pairs have the same separation distance, a predetermined rule is used to decide between alternatives [Quackenbush, 2001]. A node is created joining these two genes, and gene expression profile is computed for the node by averaging observations for the joined elements [Eisen et al., 1998]. The similarity matrix is updated with this new node replacing the two joined element and the process for any set of *n* genes the process repeated n-1 times until only a single cluster remains.

There are several variations in Hierarchical clustering that differs in the rule governing how distances should be calculated among the clusters as they are constructed. There are three ways of calculating distances between two clusters, they are average linkage, complete linkage and single linkage. They are explained in detail in section 2.5.1.

There are several limitations of hierarchical clustering. Decisions to join two elements are based only on the distance between the two elements, and once the elements are joined they can not be separated [Tamayo et al., 1999]. This is a local decision making scheme that doesn't consider the data as a whole, and it may lead to mistakes in the overall clustering.

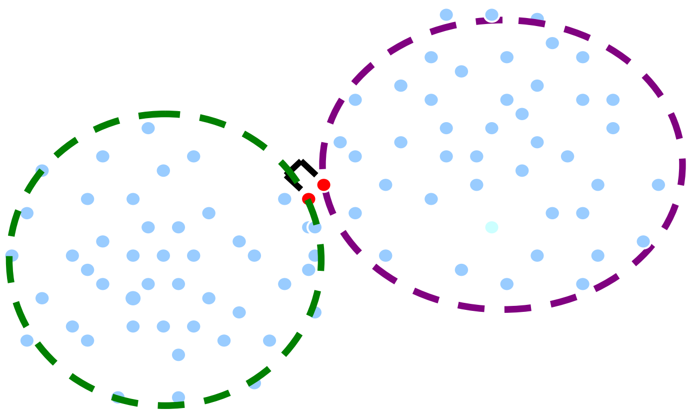


Figure 2.5: Limitation of hierarchical clustering. Hierarchical cluster start growing from the genes closest to each other, but they may belong to different cluster if overall picture is considered.

The Fig 2.5 shows there are two distinct clusters and the red points belong to different clusters but close to each other in expression space. Hierarchical clustering will join the points which are closest to each other in expression space. So the red points will be clustered together. But these points belong to two different clusters. So two points might have minimum distance but that doesn't necessarily mean that they have to belong to the same cluster. Hierarchical clustering has a shortcoming of suffering from lack of robustness and non-uniqueness problems [Tamayo et al., 1999]. An alternative approach to avoid some of the shortcomings are to use decisive clustering approach, such as k-means or self organizing maps, to partition data into groups which has similar expression pattern.

2.4.3 K- means clustering:

This is a statistical algorithm [Velculescu et al., 1995] by which objects are partitioned into a fixed number (k) of clusters, such that the clusters are internally similar but externally dissimilar. If the advance knowledge of the number of clusters is known then k-means can separate the objects effectively. K-means clustering uses a supervised clustering algorithm that is conceptually simple but computationally intensive [Quackenbush 2001]. First all initial objects are randomly assigned to one of the k clusters. Then an average expression vector is calculated for each cluster which is eventually used to compute the distance between the clusters. Using an iterative method, objects are moved between clusters and intra and inter cluster distances are measured with each move. Objects are allowed to remain in the new cluster only if they are closer to it than to their previous cluster. After each move, the expression vectors for each cluster are recalculated. The shuffling proceeds until moving any more objects will increase the intra-cluster distances and decrease inter-cluster dissimilarity.

Tavazoie (1999) used data gathered by Cho (1998) and applied k-means clustering algorithm and found the members of each cluster to be significantly enriched for genes with similar functions. They used k means algorithm to cluster 3000 genes into different regulation classes. Algorithm was repeated for 200-400 iterations and partitioned the data into 10, 30 and 60 clusters. It was observed that by 200 iterations the algorithm was converged. They finally chose 30-cluster partitioning because it provided the best compromise between number of clusters and separation between them.

2.4.4 Principal Components Analysis (PCA):

Principal Components Analysis (PCA) is a statistical technique that allows the key variables (or combination of variables) in a multidimensional data set to be identified. PCA determines those key variables in the data set that best explains the difference in the observations [Raychaudhuri et al., 2000].

PCA is very effective when some of the data might contain redundant information. For example if a group of experiments are more closely related than we had expected, we could ignore some of the redundant experiments or can take some average vale of the data without losing any information[Qucakenbush 2001]. PCA projects a high dimensional data into a lower dimensional space so that we can find the view, that gives the best separation of the data.

Given a matrix of expression data, A, where each row corresponds to a different gene and each column corresponds to one of several different experimental conditions. The a_{it} entry of the matrix corresponds to i^{th} gene's relative expression ratio with respect to a control population under condition *t.* Using PCA each of the n components can be calculated for a given gene. To compute the principal components, the n (smallest of the number of experiments or number of genes) eigenvalues and their corresponding eigenvectors are calculated from the *n x n* covariance matrix of experimental conditions or time points. Each eigenvector defines a principal component.

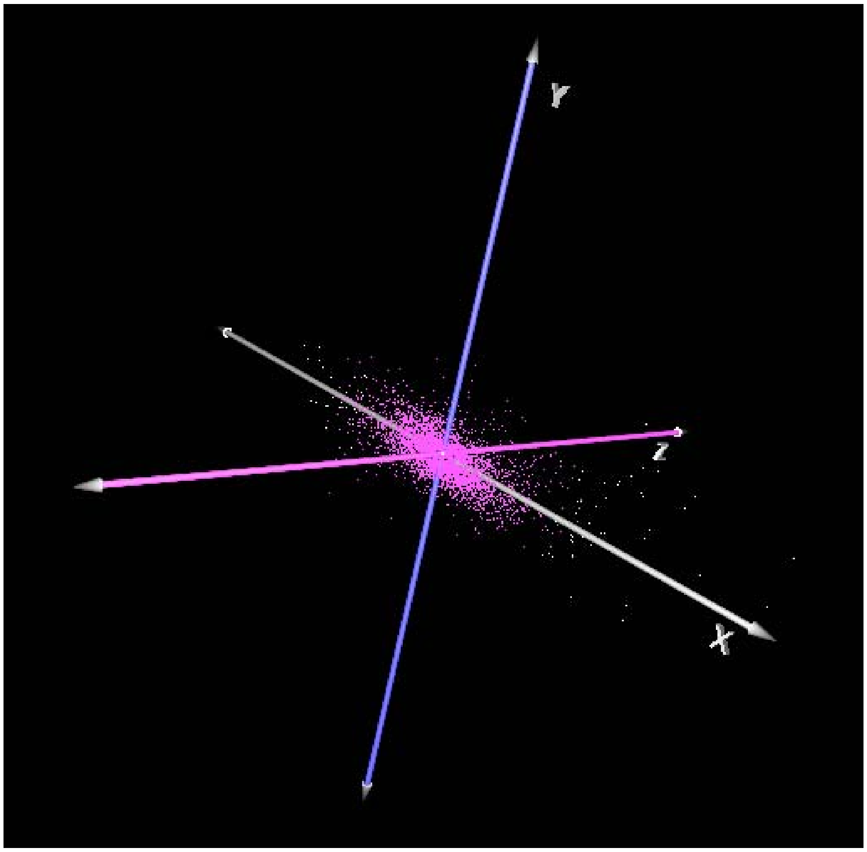


Figure 2.6: PCA of genes using TIGR TM4 software.

A component can be viewed as a weighted sum of the conditions (or time points) where the coefficients of the eigenvectors are the weights. Consequently, the eigenvectors with large eigenvalues are the once that contain most of the information; eigenvectors with small eigenvalues are uninformative [Raychaudhuri et al., 2000]. Data can be converted in terms of principal components from the following relation

$$a_{ij}' = \sum_{t=1}^{n} a_{it} v_{tj} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(2.19)}$$

where v_{tj} is the tth coefficient of the jth principal component. a_{it} is the expression measurement for gene i under tth condition. A′ is the data in terms of principal components and V is the set of ortho-normal eigenvectors.

2.4.5 Statistical analysis using Significance Analysis of Microarrays (SAM):

SAM is a statistical method to identify the genes that are undergoing considerable change in expression between two sets of microarray data [Tusher et al., 2001]. SAM is a hypothesis testing based on student t test. Suppose n_1 observations of x_i and n_2 observations of y_i are given. It is assumed that x_i and y_i are normally distributed. Then a hypothesis is created that the population means are equal. Then it can be found out if the observations are consistent with the hypothesis [Meyer, 1975]. For unpaired SAM, a statistic is defined [Tusher et al., 2001] based on the ratio of change in gene expression to standard deviation in data for that gene.

$$d(i) = \frac{r(i)}{s(i) + s_o} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(2.20)}$$

$$r(i) = \overline{x}(i) - \overline{y}(i) \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(2.21)}$$

where $\overline{x}(i)$ and $\overline{y}(i)$ are defined as the average levels of expression for gene i in two different sets. *s(i)* is the standard deviation of repeated expression measurements.

$$s(i) = \sqrt{a\left\{\sum_m [x_m(i) - \overline{x}(i)]^2 + \sum_n [y_n(i) - \overline{y}(i)]^2\right\}} \qquad \ldots\ldots\ldots\ldots \text{(2.22)}$$

where,

$$a = \left(\frac{1}{n_1} + \frac{1}{n_2}\right) * \frac{1}{n_1 + n_2 - 2} \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(2.23)}$$

s_o is a positive constant which ensures the variance of *d(i)* is independent of gene expression.

Genes are ranked according to the magnitude of their *d(i)* values, therefore *d(1)* has the largest relative difference, *d(2)* has the second largest and *d(i)* has ith largest difference.

A large number of surrogate data is generated by permutation of the data used for analysis. For each of the permutations relative differences $d_p(i)$ were also calculated and the genes were ranked in the same way, so that $d_p(i)$ has the ith largest relative difference for pth permutation. Expected relative difference was calculated by

$$d_E(i) = \frac{\sum_p d_p(i)}{N} \qquad \text{............................} \; (2.24)$$

Where N is the total number of permutations. To identify the significant changes in expressions, observed relative difference $d(i)$ is plotted against the expected relative difference $d_E(i)$. For vast majority of the genes $d(i)$ and $d_E(i)$ values are expected to be same, hence they should be close to $d(i) = d_E(i)$ line. Some genes can also be far from the line. If the distance of a gene from the line is greater than a threshold value, say delta (Δ), that gene can be called significant [Tusher et al., 2001].

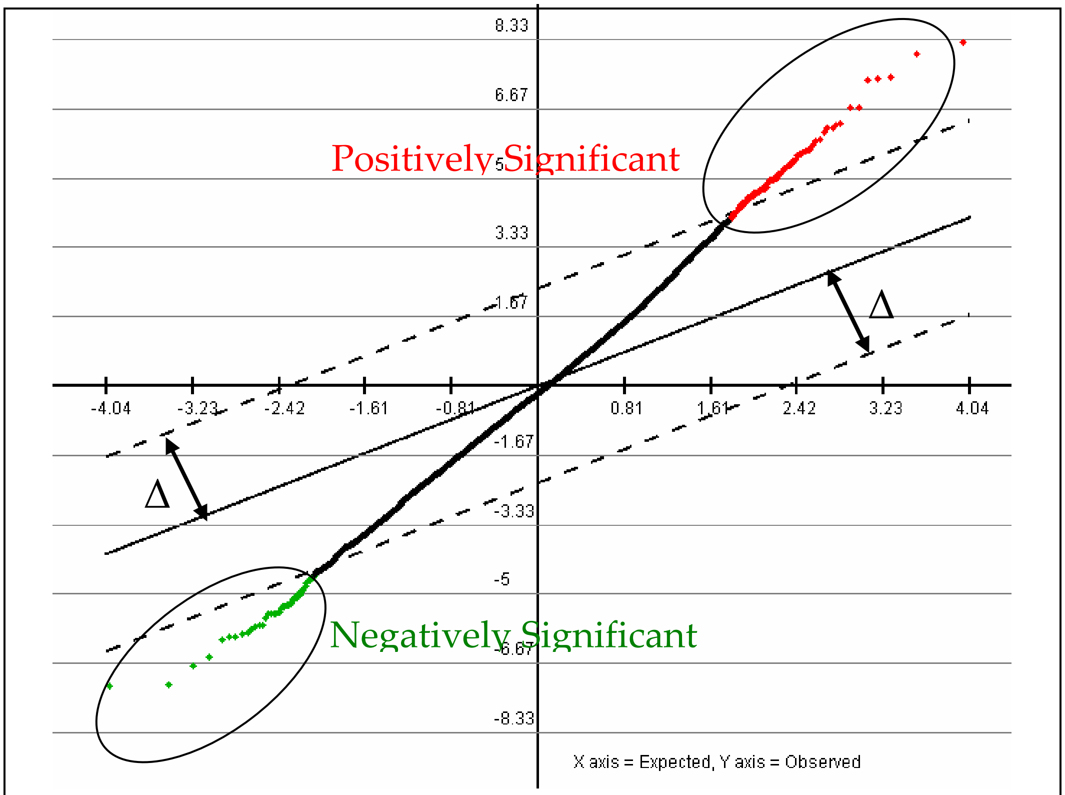


Figure 2.7: SAM graph.

SAM can also give a measure of false discovery rate (FDR). It's a measure of percentage of genes identified as significant by chance. To determine the number of falsely significant genes generated by SAM, two parallel cutoffs were defined.

Cutoffs are lines on both sides of d(i) = d_E(i) and parallel to it. The distance of the parallel lines from the line d(i) = d_E(i) is given by the threshold value. The genes that are above the upper line can be called significantly induced and the genes which are lying below the lower line are called significantly repressed. The number of falsely significant genes corresponding to each permutation was computed by counting the number of genes that exceeded the horizontal cutoffs for that permutation. The estimated number of falsely significant genes is the average of the significant genes found in all the permutations.

2.4.5 Paired SAM:

In control and perturbed experiments plants were harvested at same time points. So the difference in expression level of the perturbed and control samples should be compared for each time points separately. If unpaired SAM (explained in 2.4.5) is used, then it calculates the average expression level of the control and perturbed sets separately and finds the genes that are differentially expressed based on the averages calculated. Here we lose the information of individual time points by taking the average. Paired SAM computes the difference in expression of a gene between controlled and perturbed at each time point and calculates the statistic based on that. If there are K time points [1, 2, 3,… k] and x_{ij} of control pairs with y_{ij} of perturbed, r_i and s_i are calculated from the following equations [Stanford SAM manual]:

$$z_{ij} = x_{ij} - y_{ij} \quad \text{..(2.25)}$$

$$r_i = \sum_j z_{ij} / K \quad \text{..(2.26)}$$

$$s_i = \left[\sum_j (z_{ij} - r_i)^2 / \{K(K-1)\} \right]^{1/2} \quad \text{...(2.27)}$$

Paired SAM can only be used if there is equal number of observations (time points) in the two sets to be compared and the samples are collected at the same time points.

2.4.5.2. One class SAM:

Both two class paired and unpaired SAM are used when there are two sets of data and the objective is to find out what are the genes that make the two sets different. But when there is only one set of data, then the objective is to find out variables (here genes) that are most important (undergoing huge change in expression). For calculating the SAM statistic d_i, for ith gene, the variables r_i and s_i are computed as follows [Stanford SAM manual]:

$$r_i = \bar{x}_i = \sum_j x_{ij} \qquad \text{…………………………………………..(2.27)}$$

$$s_i = \left\{ \sum_j (x_{ij} - \bar{x}_i)^2 / n(n-1) \right\}^{1/2} \qquad \text{……………………………………………(2.28)}$$

Paired SAM and one class SAM differ in the way r_i and s_i are calculated from that data, but calculation of d_i and finding the significant genes is similar to that of unpaired SAM.

3. Plant Physiology Under Conditions of Elevated CO_2.

3.1 Photosynthesis and CO_2 fixation in plants:

Plants are the central link in the transformation of the inorganic CO_2 of the atmosphere to the organic carbon of the biosphere by photosynthesis. Photosynthesis is the process of converting light energy to chemical energy and storing it in the form of sugar, carbohydrate and lipids [Dey et al., 1996].

C3 and C4 are two different types of photosynthesis techniques. They are called C3 and C4 because in these two techniques CO_2 is first incorporated into a 3-carbon and 4-carbon compound respectively. The vast majority of plants we see around us assimilate carbon dioxide via C3 photosynthetic pathway. In brief, CO_2 enters the leaf through the stomata, and diffuses into the mesophyll cells where ribulose bisphosphate carboxylase (RuBisCo) catalyzes the carboxylation (addition of CO_2) of ribulose bisphosphate (RuBP) to form two PGA (Phosphoglycerate – a three carbon compound) molecules (fig 3.1).

Although the functional essence of C4 type of CO_2 assimilation is identical to the C3 pathway, the primary mode of CO_2 capture is substantially more efficient. By contrast to C3 systems where the carboxylating reactions are sequestered only in the mesophyll, C4 photosynthesis employs two tissue types, the mesophyll and the bundle sheath cells to achieve the same result.

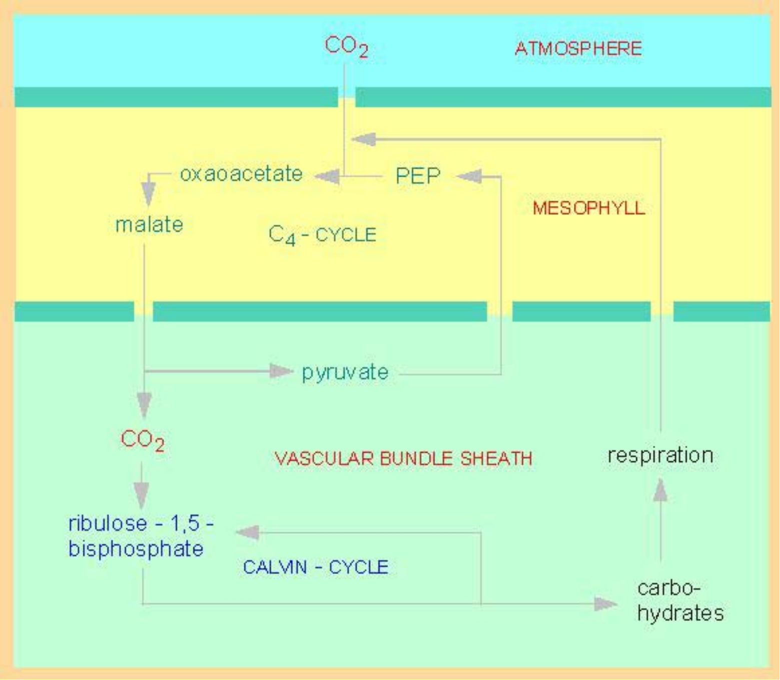


Figure 3.1: Diagrammatic representation of C4 photosynthesis. Figure was obtained from the website http://www.biologie.uni-hamburg.de/b-online/e24/24b.htm

C4 enzymes are located in the mesophyll, while the C3 enzymes involved in the Calvin cycle are specific to the bundle sheath. In short, CO_2 enters through the stomata and diffuses into the mesophyll tissue where it is fixed by Phosphoenolpyruvate Carboxylase to form oxaloacetate (Fig 3.1) which is then converted into malate (a 4-carbon molecule), and transported into the bundle sheath cells. Here, the C-4 acid is decarboxylated and the released CO_2 re-fixed by Rubisco and assimilated through the enzymes of the photosynthetic carbon reduction cycle to form sucrose and starch. Because the C4 pump is highly efficient at PEP (phosphoenolpyruvate) carboxylation, the Rubisco in the bundle sheath is super-saturated with CO_2 such that photorespiration is virtually eliminated.

Photosynthesis comprises of two parts [Lehninger, 2002]: light and dark reactions. In light reaction plant uses light energy to produce the energy storing compounds, the ATP and NADPH. Subsequently in the dark reactions, ATP and NADPH are used in the CO_2 fixation in Calvin cycle.

In plants CO_2 fixation takes place in Calvin cycle (fig 3.2). The cycle spends ATP as an energy source and consumes NADPH as reducing power to produce the sugar. Calvin cycle comprises of three separate phases.

- In phase 1 (Carbon Fixation), CO_2 reacts with five-carbon sugar named ribulose bisphosphate (RuBP) to produce 3-phosphoglycerate. The enzyme that catalyzes CO_2 fixation is called ribulose bisphosphate carboxylase (RuBisCo). It is the most abundant protein in chloroplasts and probably the most abundant protein on Earth [Lehninger, 2002].
- In phase 2 (Reduction), ATP and NADPH produced from the light reactions of photosynthesis are used to convert 3-phosphoglycerate to glyceraldehyde 3-phosphate, the three-carbon carbohydrate precursor to glucose and other sugars.

- In phase 3 (Regeneration), more ATP is used to convert part of the glyceraldehyde 3-phosphate pool back to RuBP, thereby completing the cycle.

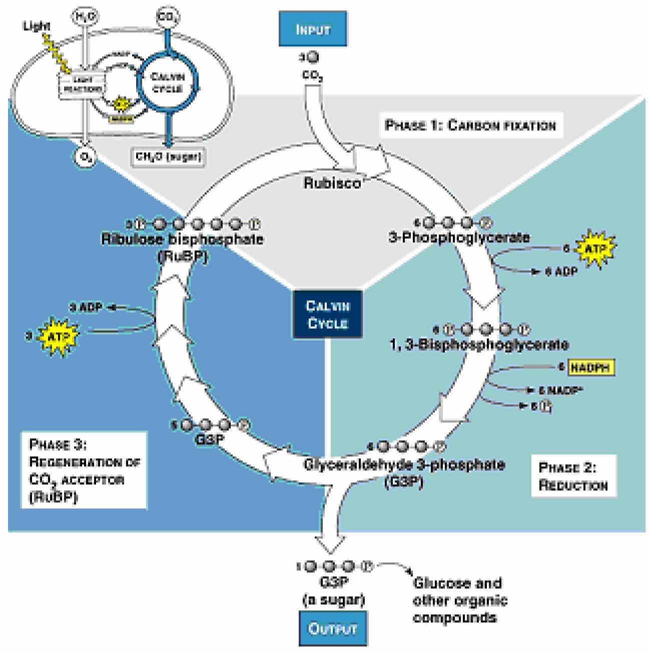


Figure 3.2: Calving cycle reaction

Figure was obtained from the website http://www.msu.edu/~smithe44/calvin_cycle_process.htm

For every three molecules of CO_2 that enter the cycle, the net output is one molecule of glyceraldehyde 3-phosphate (G3P). For each G3P synthesized, the cycle spends nine molecules of ATP and six molecules of NADPH [Lehninger, 2002]. The light reactions sustain the Calvin cycle by regenerating the ATP and NADPH.

The enzyme rubisco, that catalyses the carboxylation reaction of RuBP in Calvin cycle, also catalyses the condensation of O_2 with RuBP to form 3-

phosphoglycerate and phosphoglycolate (fig 3.3). This process is called photorespiration. In photorespiration oxygen competes with CO_2 for active sites of rubisco. If CO_2 concentration increases carboxylation of RuBP will be preferred over oxygenation, leading to increase in net photosynthesis rate.

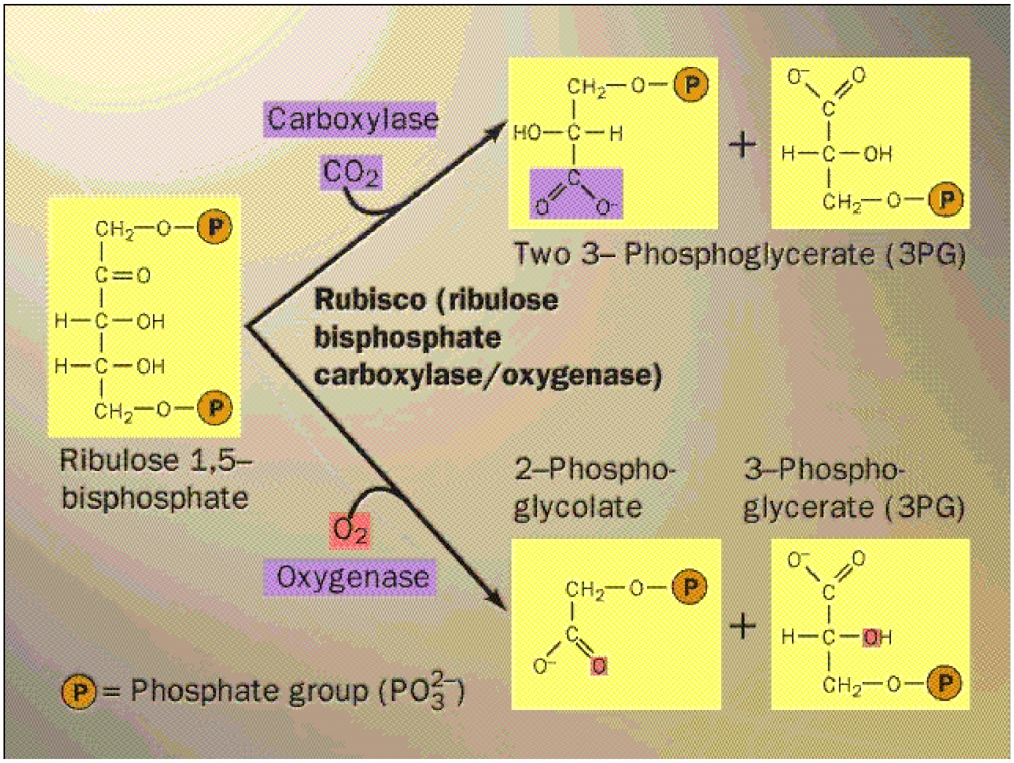

Carboxylase
CO2
Two 3– Phosphoglycerate (3PG)
Rubisco (ribulose bisphosphate carboxylase/oxygenase)
Ribulose 1,5–bisphosphate
2–Phospho-glycolate
3–Phospho-glycerate (3PG)
O2
Oxygenase
P = Phosphate group (PO3 2–)

Figure 3.3: Carbon fixation reactions.
Figure was obtained from the website http://www.usd.edu/biol/courses/Principles/Biol164/photorespiration.gif.

3.2 Plant response to elevated CO_2 levels:

CO_2 is the main source of carbon for the production of all the organic compounds in the plant. Therefore any changes in the ambient CO_2 concentration are expected to directly affect physiology of the plant. In light of the expected increase in the ambient CO_2 concentration by the end of the

21st century, it is important to study the response of the plants to such an increase. Many studies have been carried out concerning this topic. Photosynthetic CO_2 uptake and primary production increase in higher plants transferred to an elevated CO_2 atmosphere for two reasons [Webber et al., 1994] 1) Increased CO_2 concentration will lead to relative increase in the rate of carboxylation reaction compared to the RuBP oxygenation reaction. This will lead to increase in the photosynthesis rate with respect to utilization of other resources like light, water and nitrogen. 2) At normal conditions Rubisco is not saturated, thus, a further increase in carboxylation rate can be achieved by increased CO_2 concentration. But this effect can only operate when RuBP is in excess. First effect is more prominent when light is strictly limiting [Webber et al., 1994].

Elevated CO_2 will stimulate the carboxylation reaction catalyzed by Rubisco [Stitt et al., 1991]. However the rest of the plant may, for various reasons, be unable to utilize or store this additional carbohydrate. In this case, it is possible that long term and indirect effects become prominent, in which feedback regulation leads to an inhibition of photosynthesis [Stitt et al., 1991]. Therefore, the planning and interpretation of the experiments on the effect of increased CO_2 levels on plant physiology clearly distinguish between the short and long term effects of CO_2. Within each timeframe, it is important to relate the results of CO_2-enhancement to the basic understanding of photosynthetic metabolism, and how it is regulated.

Short term effect: As mentioned earlier RuBP be used by two parallel reactions, i.e. carboxylation and oxygenation. It is also important to consider how the relative changes in the rates of carboxylation and oxygenation will interact with other reactions involved in photosynthesis. Here two general aspects need to be considered. Firstly, continued catalysis by Rubisco requires that a molecule of RuBP is generated for every molecule used in carboxylation or oxygenation. Therefore, an increased net rate of carboxylation will require an increased Calvin cycle activity, and increased supply of NADPH and ATP from the light reactions [Stitt et al., 1991]. Secondly an increased net rate of carboxylation will lead to an increased rate of carbohydrate (the end product of photosynthesis) production.

Long term effect: The long term response of photosynthesis and carbohydrate content to elevated CO_2 can be related to source sink status of the plant. Enhanced CO_2 frequently leads to a larger stimulation of photosynthesis in young seedlings than in older plants [Stitt et al., 1991].

Continued catalysis of Rubisco requires that RuBP is regenerated at a higher rate to cope up with the increased rate of reaction. An increased rate of carboxylation will also require an increased rate of end-product synthesis. So the limitation of the rate of the reaction can be classified in the following three cases:

RuBisCo Limitation: This situation appears if photosynthesis is limited by Rubisco. "Rubisco has a very high affinity for RuBP [Andrews et al., 1987] and is often fully saturated in-vivo" (i.e. every active site contains a molecule of RuBP). When the reaction is not limited by RuBP regeneration, an increased rate of RuBP consumption can be matched by an increased rate of RuBP production, Rubisco will remain RuBP saturated. So increased reaction rate will require more RubisCo production.

RuBP regeneration limitation of Photosynthesis: If carboxylation is occurring at a faster rate than RuBP regeneration, then the RuBP concentration will decrease and Rubisco activity will be restricted. This is referred to as RuBP regeneration limitation of photosynthesis. Enhanced CO_2 will still lead to an increased net rate of photosynthesis when RuBP regeneration limits the rate of photosynthesis for the following reason. Normally, a considerable portion of available ATP, NADPH and Calvin cycle activity is involved in regeneration of RuBP which is subsequently oxygenated. In enhanced CO_2 the rate of oxygenation will be reduced leading to a more efficient use of RuBP.

Limitation of photosynthesis by end product synthesis: The rate of photosynthesis can also be limited by the rate at which the immediate products of CO_2 fixation (phosphorylated intermediates) are converted into non-phosphorylated end products (i.e. carbohydrate, amino acids and lipids). The major end products are sucrose and starch. Sucrose is synthesized from triose-phosphate in cytosol and starch is synthesized in the chloroplast stroma. If these reactions occur too slowly, phosphorylated intermediated will accumulate and the pool of P_i in the cytosol and chloroplast will be depleted. This eventually leads to an inhibition of photosynthesis because P_i is required in chloroplast for ATP synthesis. In this kind of conditions, increased rate of carboxylation or suppression of oxygenation and photorespiration merely generate excess electron transport and Calvin cycle capacity.

Exposure of plants (especially C3) to elevated CO_2 frequently results in an immediate increase in the rate of CO_2 assimilation; however, a reduction in photosynthetic capacity often occurs after a prolonged period (days to weeks) at elevated CO_2. This down regulation, also called acclimation of photosynthesis is accompanied by a large increase in leaf carbohydrates. On average, large soluble sugar increase by 52% and starch content increases by 160% [Cheng et al., 1998]. It was found that sugars can influence many metabolic and cellular processes in both prokaryotes and eukaryotes through modulation of gene expression [Sheen et al., 1994]. It is well established that increased sugar levels can trigger repression of photosynthetic gene transcription. It was shown [Sheen et al., 1990] that transcription of seven photosynthetic genes, including *rbcS*, is repressed by glucose and fructose. Cheng et. al. [Cheng et. al., 1998] conducted experiments to study the effect of short and long term elevated CO_2 on expression of Rubisco gene. For long term experiment, plants were grown continuously for 40 day at ambient or high CO_2. High CO_2 grown plants on average was found to have 2 day more advanced developmentally. This long term growth resulted in a 2-fold or greater increase in Glucose and Fructose and 3.5 fold increase in starch, whereas Sucrose amount remained relatively constant. It was also found Rubisco decreased by 34%, *rbcL* mRNA decreased by 38% and *rbcS* transcript decreased by 60%. In another experiment, 30 day-old ambient CO_2 grown plants were transferred to high CO_2 for up to 12 days, and leaves were collected at the beginning of the light period on each sampling day of transcript measurement. Rubisco protein, *rbcL* and *rbcS* transcripts were measured on 3, 6, 9 and 12th day from transfer and was found to be less in elevated CO_2 compared to ambient CO_2. Abundance of *rbcS-1A, rbcS-1B, rbcS-*

2B, rbcS-3B transcripts were measured through 24 hour period on day 6 of exposure [Cheng et al., 1998].

One of the most prominent consequences of elevated CO_2 enrichment is decrease in Nitrogen concentration [Sherwood, 2001]. The reduction of nitrogen concentration caused by elevated CO_2 is the result of increase in plant total carbohydrate amount resulting from enhanced growth in CO_2 enriched air. On exposing wheat to elevated CO_2 also leads to higher rate of CO_2 fixation. Increased CO_2 fixation requires more reducing power for reduction of 1,3-bisphosphoglycerate to glyceraldehyde-3-phosphate in Calvin cycle reaction. Reduction of nitrate or nitrite is required for nitrogen assimilation in plants. At elevated CO_2 as more reducing agent is used in Calvin cycle less reducing agent is available for reduction of nitrate or nitrite leading to a decrease in nitrogen assimilation [Bloom, 2002].

Atmospheric CO_2 enrichment may also alter the abilities of plants to combat diseases that periodically afflict them [Sherwood, 2001]. A good example of this phenomenon comes from the experiment of Malmstrom and Field (1997), who studied CO_2-induced growth responses of healthy oat plants and oat plants infected with the barley yellow dwarf virus (BYDV). In response to a doubling of the air's CO_2 concentration, they found that after only sixty days, total biomass in CO_2-enriched healthy plants had increased by 12% over total biomass in non-CO_2-enriched healthy plants. In BYDV-infected plants, however, biomass had increased by 36%, a response that was three times greater than the 12% increase observed in healthy plants.

Masle [2000] studied effect of elevated CO_2 concentrations on wheat plants. He has grown two sets of wheat plants one at 350 ± 10 ppm and the other at 900 ± 12 ppm over 4 weeks. He observed plants under 900 ppm CO_2 grew significantly more than those under 350 ppm, showing a 52% to 93% increase in total dry weight at the end of the experiment and a 39% to 82% increase in leaf area. These differences mostly caused by an increase in C content per unit leaf area. There was an increase in sugar contents and the 25% increase in rate of leaf photosynthesis per unit leaf area at elevated CO_2. Another effect of elevated CO_2 was to increase the cell division rate, achieved by reducing the time interval between successive divisions. A similar result was obtained by Kinsman et al. (1997) of enhanced cell division rates in the shoot apex of Dactylis under 700 ppm CO_2 compared with 350 ppm.

Chen et al. studied response of potato tuber cell division and growth to elevated CO_2. They observed that, elevated CO_2 increased accumulation of total net biomass, and increased tuber growth rate by about 36 %, but did not increase the number of tubers. They also found elevated CO_2 increased glucose concentration and soluble invertase activity.

Microarray analysis can study expression of thousands of genes simultaneously (detailed information is given in the transcriptional profiling chapter) and has become an important technology for genomic analysis [explained in more detail in chapter 2]. Compared to validating expression of annotated genes, confirming functional role assignments for putative genes and determining functions for hypothetical and unknown genes is significantly more difficult. It is not easy to find the proper condition under which those genes are significantly regulated. Precise functional assignments generally require series of biochemical and genetic analyses to confirm a gene product's action. However, microarray data provides information on pattern of gene expression that can give some insight of plausible function which can be eventually tested. Another advantage of microarray data is that, they provide support for the genes coding proteins for putative functions.

Microarray was used to study transcriptional profiling for *Arabidopsis thaliana* by lot of scientists. Kim et al. used microarray to study the gene expression of chromosome 2 of Arabidopsis Thaliana [Kim et al., 2003] under different biotic (4 different bacterial infection) and abiotic stresses (three different abiotic stress heat, cold and salt). A total of 4437 genes were used for analysis and 334 genes were found to be differentially expressed in response to at least one biotic treatment. They found 497 genes that were differentially expressed at 95% confidence under one or more of the conditions, which included 43 genes that have been previously characterized and 247 genes coding for putative functions. A gene encoding Glutathione-S-transferase (GST, At2g29450) was found to be up regulated in response to all stresses. Genes coding for cold-regulated protein cor15a precursor and 15b precursor were up regulated.

3.3 *Arabidopsis thaliana*: A Model Plant

The *Arabidopsis Thaliana* plant has become a model system to study C3 plant physiology because of the following reasons: *Arabidopsis* has a small genome of 125 Mb and 5 chromosomes and was completely sequenced in the year 2000. It is a well studied organism as most of the genes are well annotated and most of the pathways (metabolic and signaling) are well characterized. Hence it is easy to analyze and compare the results obtained with the results in literature. It has a rapid life cycle (about 4/6 weeks from germination to mature seed) for liquid cultures. It also has advantages like prolific seed production and easy cultivation in restricted space.

4. GENE EXPRESSION PROFILING of *A. thaliana* under conditions of elevated CO_2

4.1 Overview of the Experiment:

Two sets of *Arabidopsis thaliana* (Columbia strain) plants were grown for 12 days in Gamborg media at 23^0C under constant light. At the beginning of the 13th day, 3 and 4 plant liquid cultures, respectively, were harvested from the control and perturbed sets and they were used as reference of the plant growth up to that stage. On the 13th day one of the plant sets was fed with air of ambient composition and the other set was fed with air of 1% CO_2 , in the rest of the text it will be termed as control and perturbed set respectively. Two plants from each set were harvested at the time points 0.5hr, 1hr, 1.5hr, 2hr, 3hr, 6hr, 12hr and 23hr of the 13th day. Each harvested plant was immediately weighed, frozen in liquid nitrogen and stored at –80oC for further analysis.

Experimental setup was shown in Fig 4.1A. It consists of a shaker with 20 flasks containing liquid culture. A manifold was used to supply air to the flasks uniformly during the experiment. For the first 12 days manifold was not connected and the plants were grown in ambient air condition. On the

13th day the flasks were connected to the manifold to supply air of 1% CO_2 for perturbed and air of ambient composition for control.

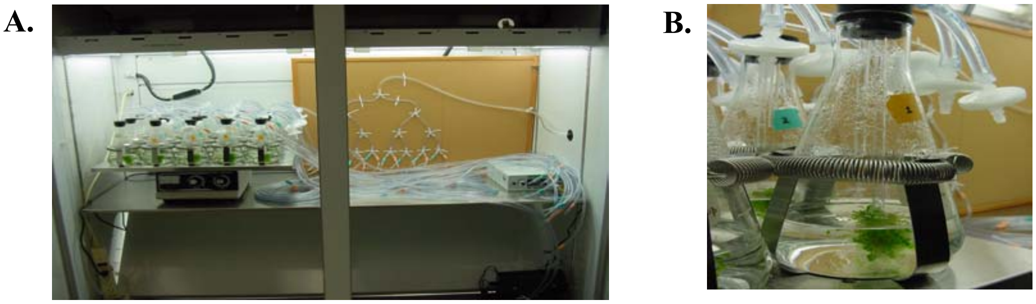


Figure 4.1: **A**. Picture of the experimental setup in the growth chamber. **B**. Picture of a shake-flask in this setup.

Each of the plants harvested were ground to a state of paste. 3 gm of the ground plant was used to extract RNA by Trizol extraction [protocol attached] and used for the hybridization on cDNA microarray slide.

At the phase of slide hybridization, a pool of equal amounts of mRNA from each sample (from both control and perturbed) was used as reference. The relative expression of each sample compared to the reference was measured by 2 DNA microarrays in which the dyes (Cy3 and Cy5) of the reference and the plant sample were swapped. Therefore, the total number of DNA microarray slides to be analyzed was: 2 x 19 (control set) + 2 x 20 (perturbed set) = 78.

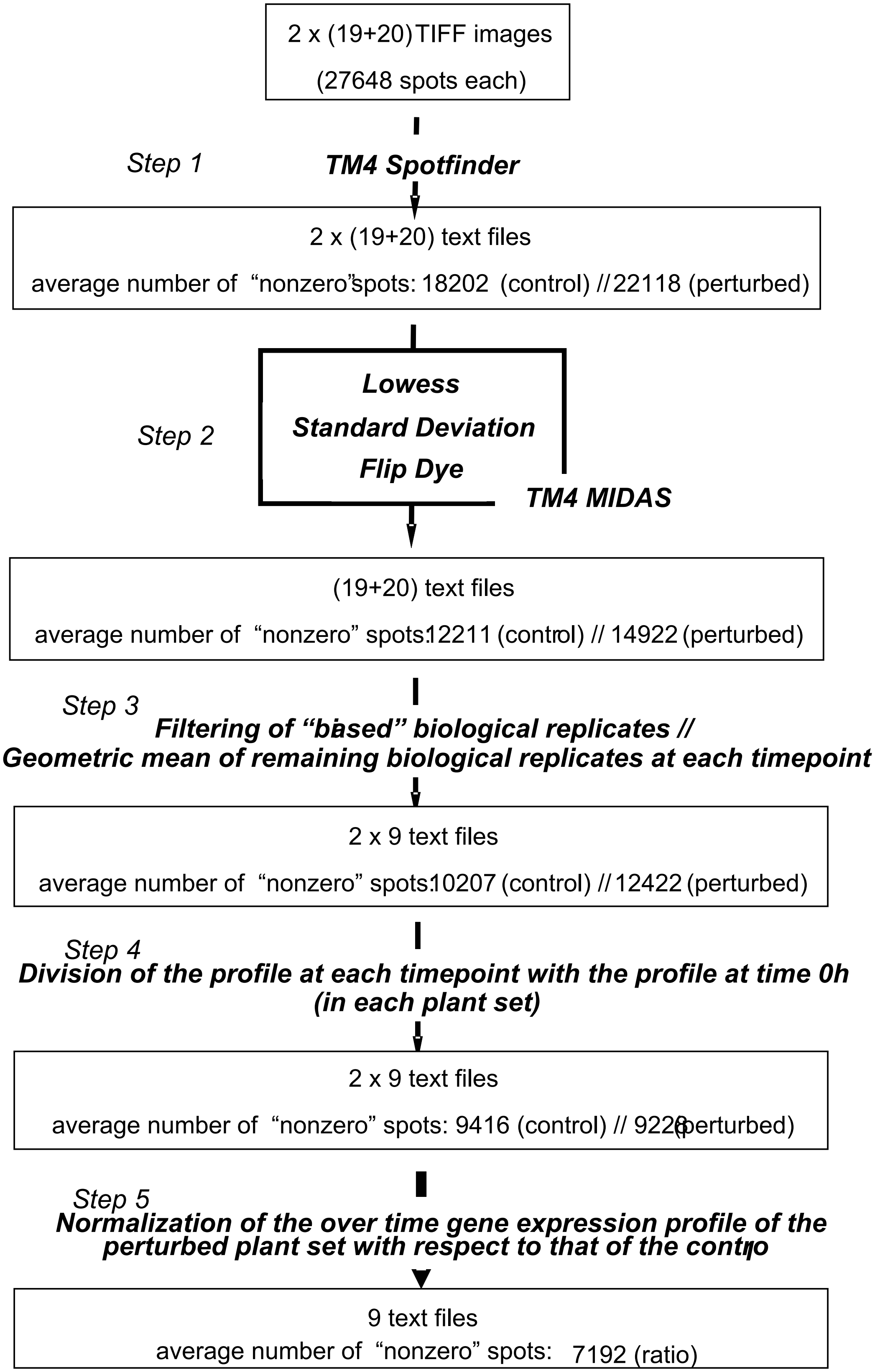

2 x (19+20) TIFF images
(27648 spots each)
Step 1
TM4 Spotfinder
2 x (19+20) text files
average number of “nonzero”spots: 18202 (control) // 22118 (perturbed)
Step 2
Lowess
Standard Deviation
Flip Dye
TM4 MIDAS
(19+20) text files
average number of “nonzero” spots: 12211 (control) // 14922 (perturbed)
Step 3
Filtering of “biased” biological replicates //
Geometric mean of remaining biological replicates at each timepoint
2 x 9 text files
average number of “nonzero” spots: 10207 (control) // 12422 (perturbed)
Step 4
Division of the profile at each timepoint with the profile at time 0h
(in each plant set)
2 x 9 text files
average number of “nonzero” spots: 9416 (control) // 9228 (perturbed)
Step 5
Normalization of the over time gene expression profile of the
perturbed plant set with respect to that of the control
9 text files
average number of “nonzero” spots: 7192 (ratio)

Figure 4.2: Flow chart of the steps that were followed for data analysis

4.2 Image Processing:

From the 78 DNA microarrays, 2x 78 TIFF images were generated. These were processed using the TM4 Spotfinder image processing software (version 2.2.1_NoDB) as explained in Chapter 2. In the image processing the default values of the software parameters were used except of the following:

- No QC filter was used.
- Flagged values were generated.
- Minimum spot size used = 7

The "flags" generated for each spot in the analyzed microarrays provide a measure of the "quality" of the spot based on the underlying image processing algorithm. Since only the spots flagged either B or C in both channels (Cy3 and Cy5) are used by the normalization software (TM4 MIDAS) for further analysis, the "flagging" of spots is equivalent to filtering out those that contain gross errors and any potential inclusion in the rest of the analysis could "skew" the results. Table 4.1 shows the number of "acceptable" spots for each of the plant sample, which will be considered in the rest of the analysis (at the bottom the average number of "acceptable" spots for each plant set after the image processing is also depicted).

Table 4.1: The table shows the list of acceptable spots of all the plants after image processing (step 1 of fig 4.2)

Control:

time point	biological replicate 1		biological replicate 2	
	flip dye		flip dye	
0 hr	20945	20973	19561	20313
	20329	21320		
.5hr	15048	16315	18713	13981
1 hr	15281	14289	19110	16424
1.5 hr	19275	16834	18815	19556
2 hr	16115	20789	17642	14238
3 hr	18606	19940	17544	17615
6 hr	19589	21809	20156	18841
12 hr	13594	13481	19510	17658
23 hr	16954	19728	20992	19519
Average	**17573.6**	**18547.8**	**19115.88889**	**17571.66667**

Perturbed:

time point	biological replicate 1		biological replicate 2	
	flip dye		flip dye	
0 hr	20366	23824	20506	20783
	20317	24146	22504	20272
.5hr	23644	19110	22751	18797
1 hr	21961	19568	21445	21124
1.5 hr	23190	18074	23507	23894
2 hr	23831	24377	23547	21442
3 hr	19641	20780	24092	23940
6 hr	24080	22704	23041	20406
12 hr	25040	24949	21267	21109
23 hr	21821	23495	22730	22655
Average	**22389.1**	**22102.7**	**22539**	**21442.2**

4.3 Data Normalization and Filtering:

4.3.1 Normalization using TM4 MIDAS:

The 78 textfiles (TAV format) generated from Spotfinder were normalized using the TIGR TM4 MIDAS (version V2.16) software as discussed in chapter 2. The following sequence of normalization methods was applied (if non-default values were used for some of the parameters in each method, they are provided next to the method's name):

i) *lowess*: Applied block-wise

ii) *standard deviation (SD)*: Applied block-wise

iii) flip dye: Cross log ratio data keep range: ± 2SD

Table 4.2 shows the number of spots in each plant sample, which had non-zero intensity after the end of the normalization process, i.e. step 2 in fig 4.2 (at the bottom the average number of "acceptable" spots for each plant set after the image processing is also depicted).

Table 4.2 Number of spots with non-zero intensity values after normalization

Control:

time points (hr)	Two biological replicates	
0hr	15778	13948
	14625	
0.5 hr	9892	7422
1hr	7179	10460
1.5hr	10661	13700
2hr	12710	12254
3hr	12250	12482
6hr	15121	15729
12hr	7997	14380
23hr	11334	14088

Perturbed:

time points (hr)	Two biological replicates	
0hr	14285	13825
	13593	11856
0.5 hr	12982	13488
1hr	11232	13519
1.5hr	11591	18652
2hr	16359	17243
3hr	11835	17991
6hr	18295	15950
12hr	21860	13946
23hr	17242	12712

4.3.2 Outlier detection

In each plant set, two liquid cultures were harvested at each time point except of 0hr at which 3 and 4 cultures, respectively, were harvested for the control and the perturbed systems. Even though there is intra species biodiversity among the plants, it is expected that the plants harvested at the same time point for each set would be at the same physiological level and would, thereby, have similar expression profiles. It is expected then that the biological replicates at each time point will statistically form separate populations, allowing for the phenotypic differentiation between the various timepoints. This implies that in the phylogenic tree generated by HCL the biological replicates at the same time point should be the closest..

The limitation in the present study was the acquisition of only 2 replicates at most of the time points but 0hr. In this case, even though some replicates did not cluster with the closest distance in the hierarchy tree, it is not easy to decide which of the two might contain gross errors and should be excluded from the rest of the analysis without the possibility of making the wrong decision. In the present analysis, all replicates at all timepoints were considered in the rest of the processing.

4.3.3 Averaging of the Biological Replicates:

The gene expression profile of the plants at each time point of the sampling period was represented by the geometric mean of the biological replicates harvested at the time point. In this way, genes that are not consistently present in all replicates are excluded from the rest of the analysis. The number of non-zero spots at each time-point after this normalizaton step is shown in Table 4.3 (at the bottom the average number of “acceptable” spots for each plant set after the image processing is also depicted).

Table 4.3 Number of spots with non-zero intensity values after averaging (after step 3 in fig 4.2)

Control	
time point (hr)	number of genes that have a nonzero intensity value
0	12892
0.5	6802
1	6925
1.5	10265
2	11067
3	11253
6	13948
12	7840
23	10876
Average	10207.56

perturbed	
time point (hr)	number of genes that have a nonzero intensity value
0	10041
0.5	11782
1	10942
1.5	11201
2	15248
3	11520
6	15139
12	13564
23	12366
Average	12422.56

4.3.4 Normalization with respect to 0 h time point:

To correctly compare the control and perturbed sets and identify their difference due to the applied perturbation, the plants should be at the same growth level at the initiation of the perturbation. Taking into consideration that the experiments of control and perturbed sets took place on different days, the gene expression profile of a culture at a particular time point was divided by that at 0h in the same plant set. It is clear that only the genes that have nonzero expression at the particular timepoint and time 0h will still have nonzero expression after this normalization step.. The spots that have nonzero expression after this normalization procedure are shown in Table 4.4 for each of the acquired time points.This again leads to loss for information. At each time point number of genes that have nonzero expression value after division is tabulated in table 4.3(at the bottom the average number of "acceptable" spots for each plant set after the image processing is also depicted).

Table 4.4 Number of spots with non-zero intensity values after normalization with respect to 0h time point (after step 4 in fig 4.2)

Control	
time point (h)	number of genes that have a nonzero intensity value
0	12892
0.5	6401
1	6535
1.5	9494
2	10133
3	10275
6	11506
12	7393
23	10117
Average	9416

Perturbed	
time point (hr)	number of genes that have a nonzero intensity value
0	10041
0.5	8576
1	8825
1.5	8939
2	9351
3	9016
6	9475
12	9417
23	9412
Average	9228

4.3.5 Elimination of initial time points:

There was huge oscillation in the gene expression both in perturbed and control in the initial period (first 3 hours) of the experiment. So first few time points show only transient effect and doesn't show any permanent effect in the gene expression. The initial time points also show a huge oscillation both in the genomic and metabolomic level (Fig 4.3). This oscillation was not due to CO_2 perturbation, but because of the physical perturbation of the system. Therefore only the last 4 time points (3, 6, 12 and 23hr) are considered for this analysis.

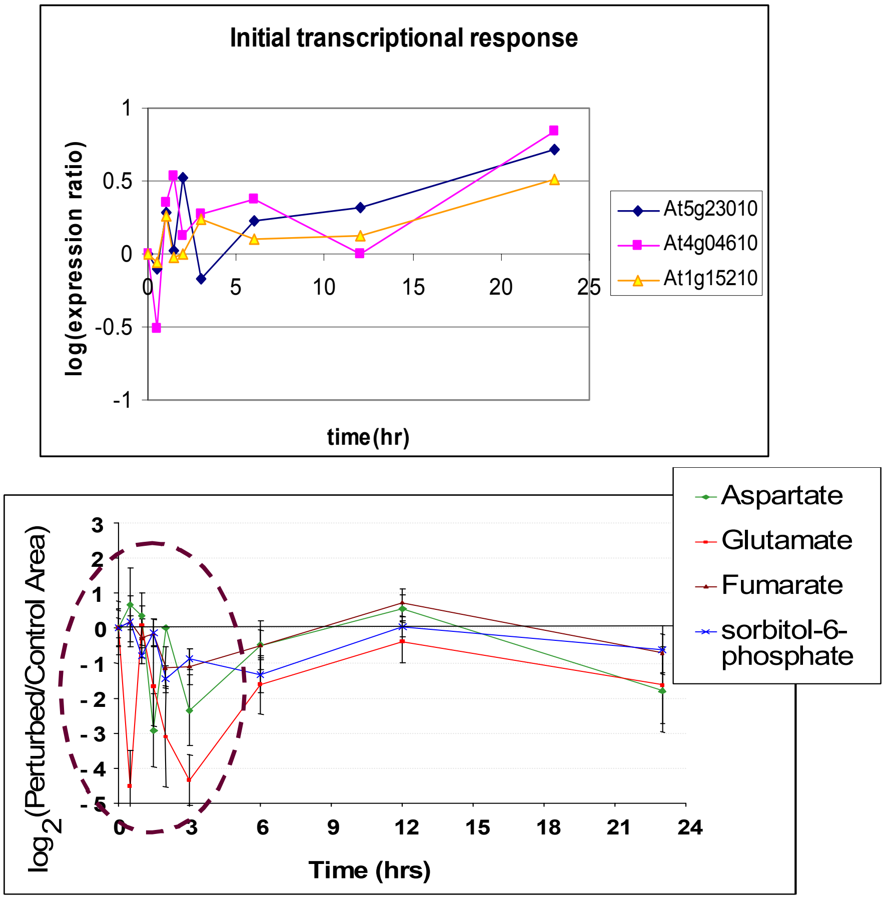


Figure 4.3: Initial response at transcriptional and metabolic level.

4.4 Clustering of Experiments:

In this experiment five time points from control and perturbed were used for analysis. It is believed that the gene expression of the perturbed set is going to be different from that of control set due to environmental stress. To verify this hypothesis, time-points of both control and perturbed were clustered. If there is a huge change in gene expression due to elevated CO_2, then control and perturbed time points are expected to cluster separately. When the time points of control and perturbed are clustered together using hierarchical clustering following phylogenic tree was observed (Fig 4.4).

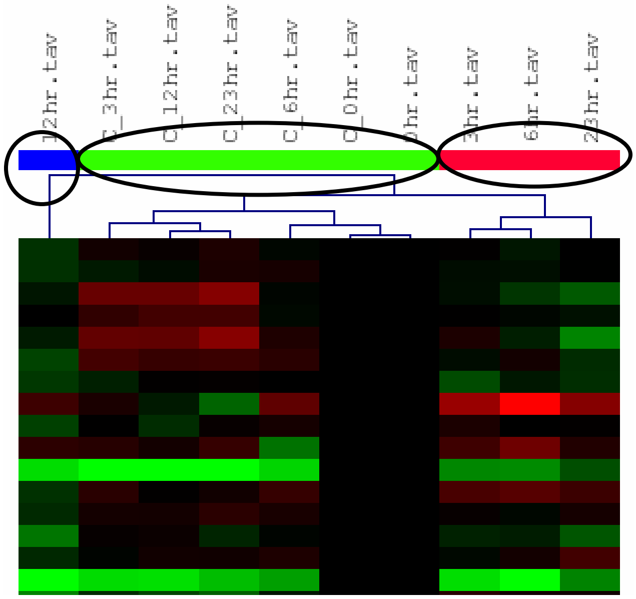


Figure 4.4: Clustering of the experiments.

Optimum number of clusters was found to be 3 from FOM. K-means clustering was used with optimum number of clusters as 3. One cluster was found to contain all the control time points and 0th time-point of perturbed. The second cluster was found to contain 3, 6 and 23 hr time points pf perturbed. 12 hr time point of perturbed was clustered separately in the third cluster. These three clusters obtained are marked with green, red and blue color in Fig 4.4. Hierarchical clustering shows the three different clusters obtained from k-means clustering forms three different branches of its phylogenic tree.

Principal component analysis (PCA) could separate the control and perturbed experiments effectively (fig 4.5). PCA of the experiments show that first three components can capture 95% of the information. So each time points of the experiment can be represented as a single spot in three dimensional space. Three dimensional view of the experimental time points are plotted in fig 4.5. Each point in fig 4.5 is a time-point of the experiment and its color is same as the color it is assigned in fig 4.4. PCA three dimensional diagram shows that all the green and red points are clustered separately. The blue point which represents time point 12hr of perturbed stands apart, showing similarity with hierarchical and k-means clustering.

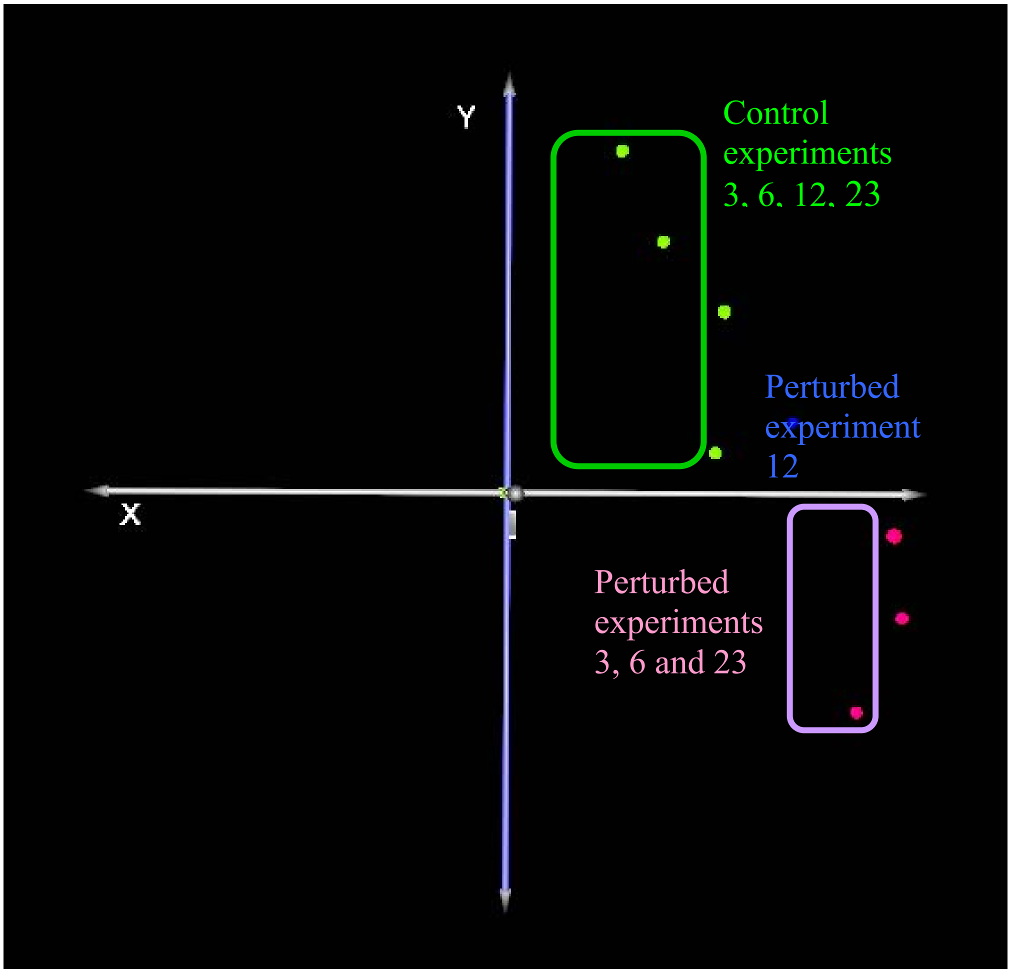


Figure 4.5: PCA of the experiments show that control and perturbed experiments cluster separately.

4.5 Clustering of Genes:

TIGR TM4 software MultiExperiment Viewer (MeV) (version 2.1) was used for clustering analysis as discussed in chapter 2.

4.5.1 Analysis using Paired SAM:

TAV files generated from normalization were loaded in MeV. As only last 4 time points are considered there are 5 TAV files (4 time points and 0th time point) from each of control and perturbed set. There was lot of genes whose data was missing in some of the time points. If there is some missing data then the insufficient information can lead to erroneous result. A criterion of 90% cutoff was used for selecting the genes which are used for analysis. This means a gene should have its expression value present in at least 9 of the 10 files. With 90% cutoff it was found that 7321 genes are used for analysis.

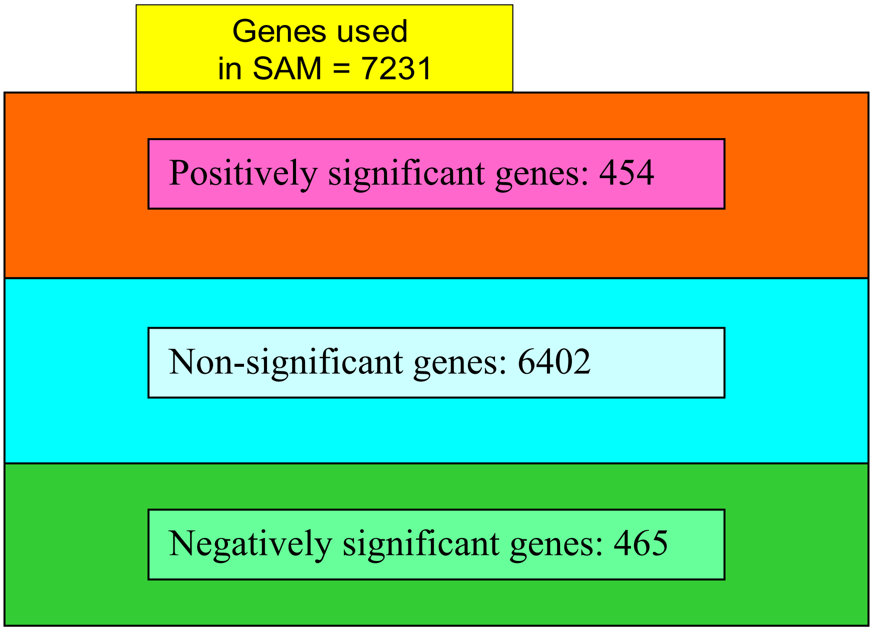

Genes used
in SAM = 7231
Positively significant genes: 454
Non-significant genes: 6402
Negatively significant genes: 465

Figure 4.6: Results obtained from paired SAM.

SAM analysis was performed for different delta values to check what would be the optimum value. It is a subjective question and the judgment may vary from person to person. Analysis was performed for delta values 0.9, 0.925, 0.95, 0.975 and 1 and found that a delta value 0.95 is optimum. For rest of the analysis positively, negatively and nonsignificant genes refer to the clusters obtained from paired SAM with delta = 0.95.

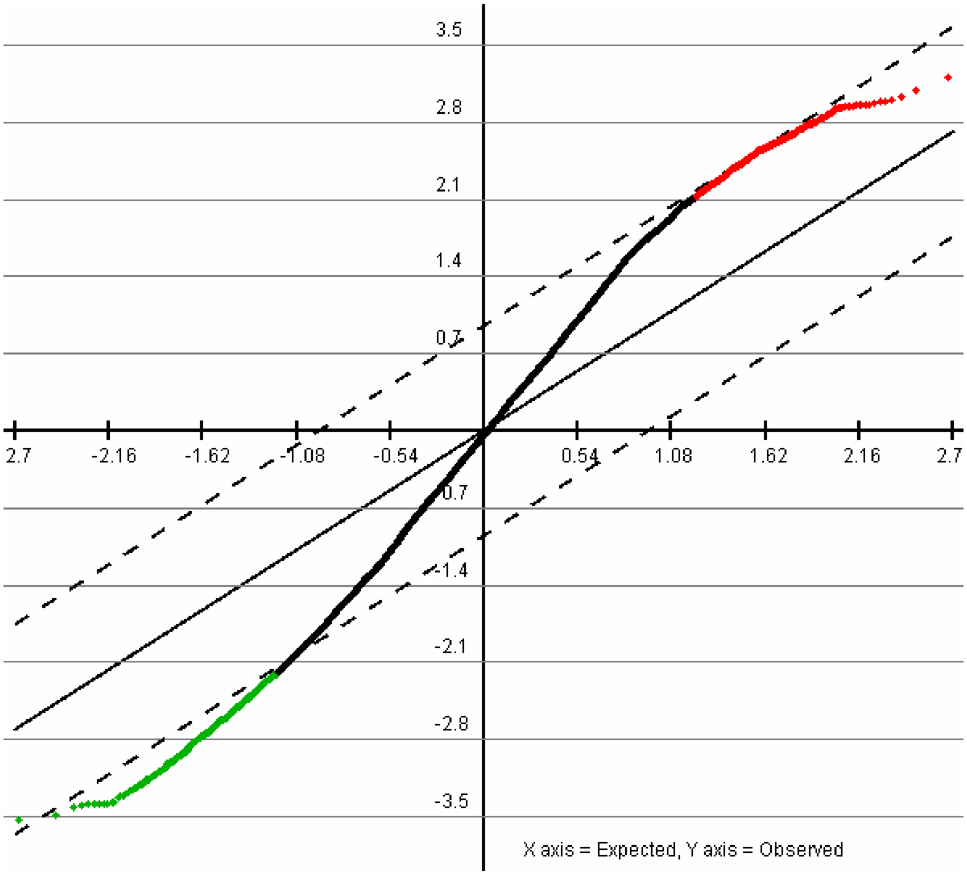


Figure 4.7: The SAM curve with delta = 0.95.

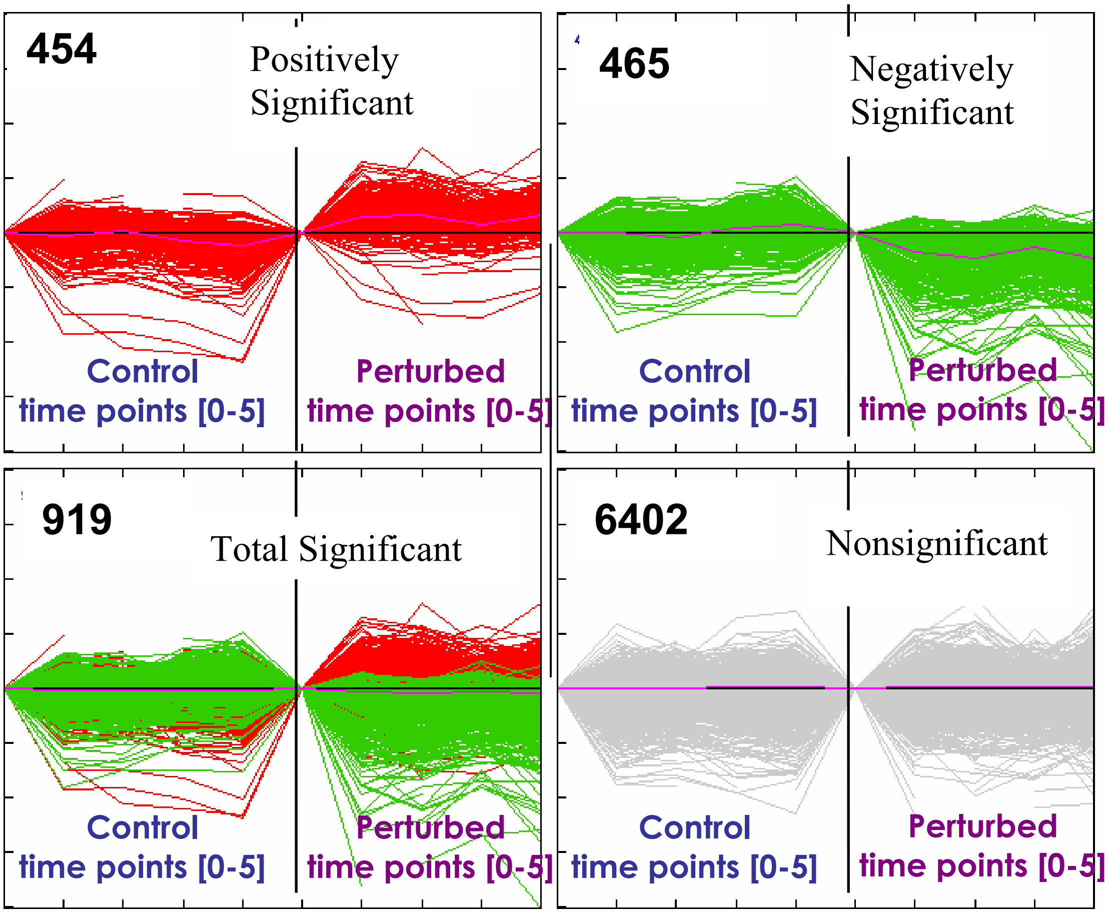


Figure 4.8: Expression profiles of positively, negatively and non-significant genes from paired SAM.

4.5.2 Analysis using one class SAM:

The ratio of the expression of genes in perturbed and control were calculated by dividing the expression of the perturbed set by control set and used for k-means clustering. A total of 5 files were obtained from the ratio. These data were used for one class SAM analysis. The TIGR TM4 software doesn't have one class SAM function, so Stanford SAM software was downloaded from http://www-stat.stanford.edu/~tibs/SAM/ and was used for one class SAM analysis. 100% cutoff was used, i.e. the genes that are present at all the time points were considered for analysis. There were 4703 genes that were present in all the time points, and hence were used for one class SAM analysis. Delta value 1.65 was found optimum and it gives low FDR like 0.57% (which is comparable to the FDR obtained from paired SAM i.e. 0.648%). Number of significant genes found was 589, this is comparable to the number of significant genes obtained from paired SAM that has no missing value. So the criterion of calling a gene significant remains same in both paired and one class SAM. 318 and 271 genes were found positively and negatively significant from one class SAM (Fig 4.9).

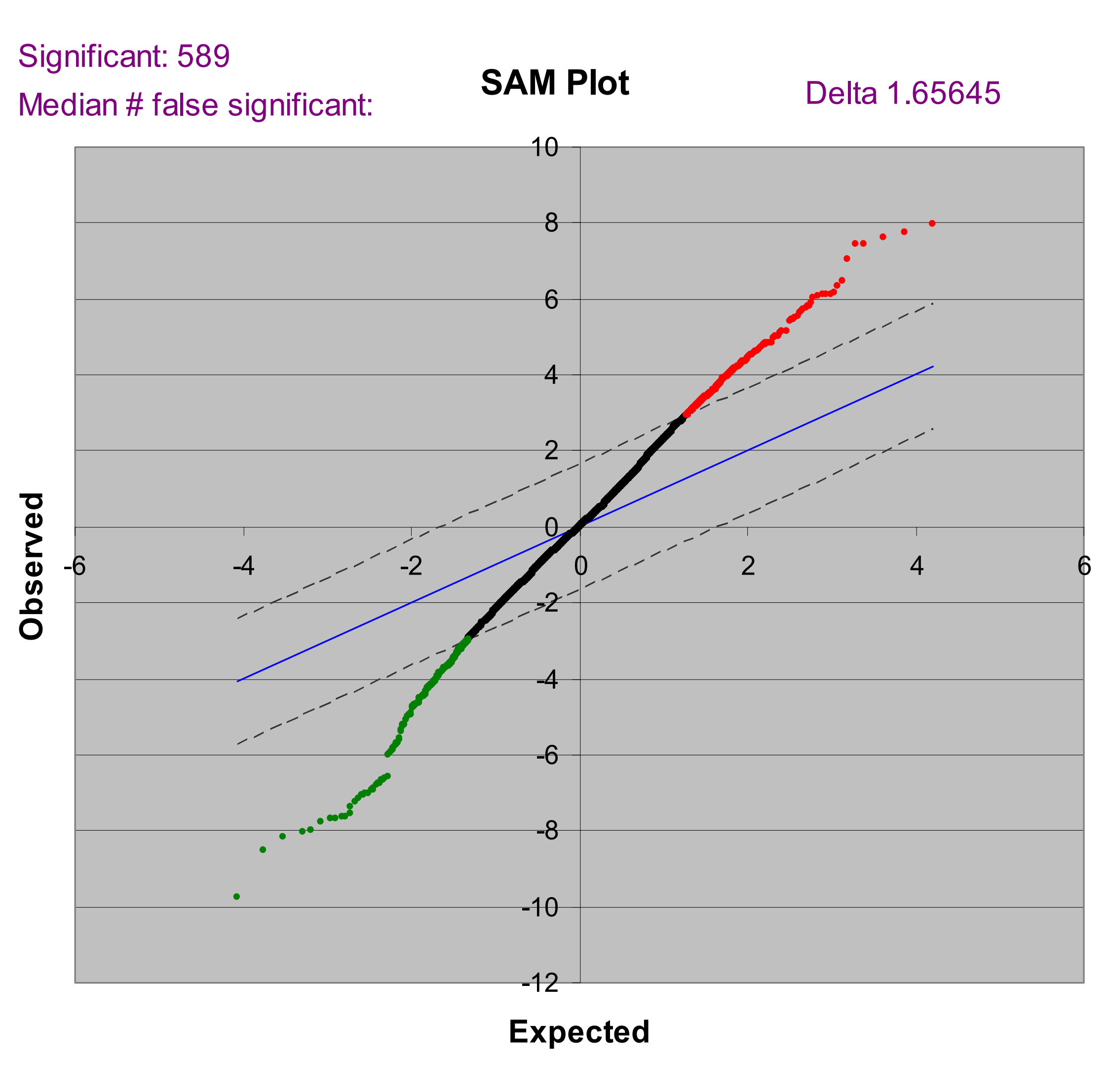


Figure 4.9: The SAM plot of one class SAM.

4.5.3 K-means clustering:

The ratio of the expression of genes in perturbed and control were used for k-means analysis.

Finding optimum number of clusters for k-means is inherent problem of k-means, please refer to chapter 2 of thesis. "Right" number of clusters were found by plotting FOM with number of clusters.

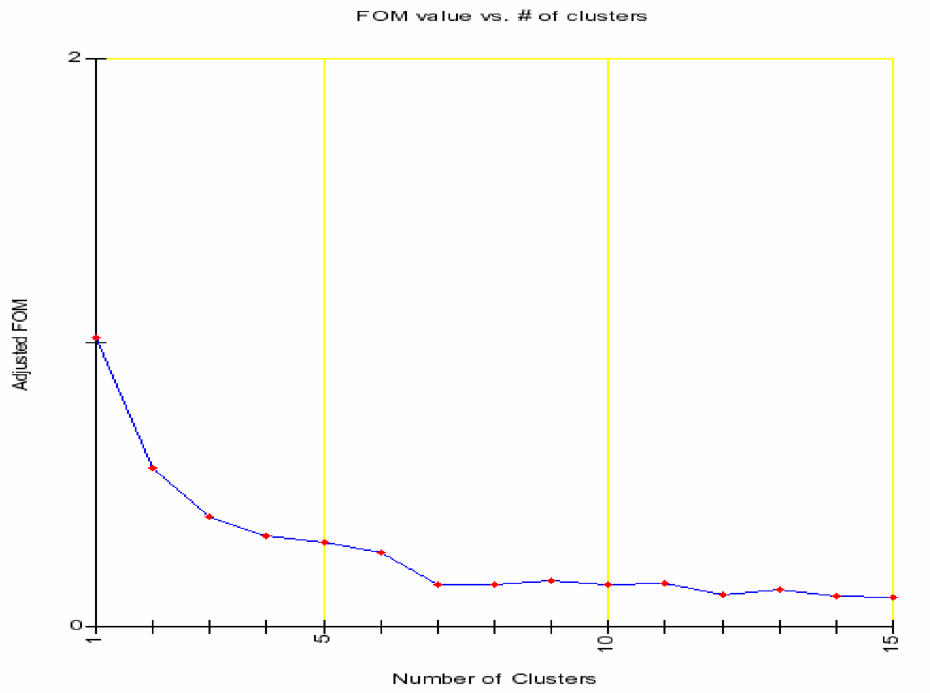


Figure 4.10: FOM values plotted with number of clusters to find optimum number of cluster

From the Fig 4.10 it is seen that FOM value falls with number of clusters sharply till cluster 4. From cluster 4 to 6 FOM value remains almost static. It again falls at cluster 7. With number of clusters as 7 following clustering was observed:

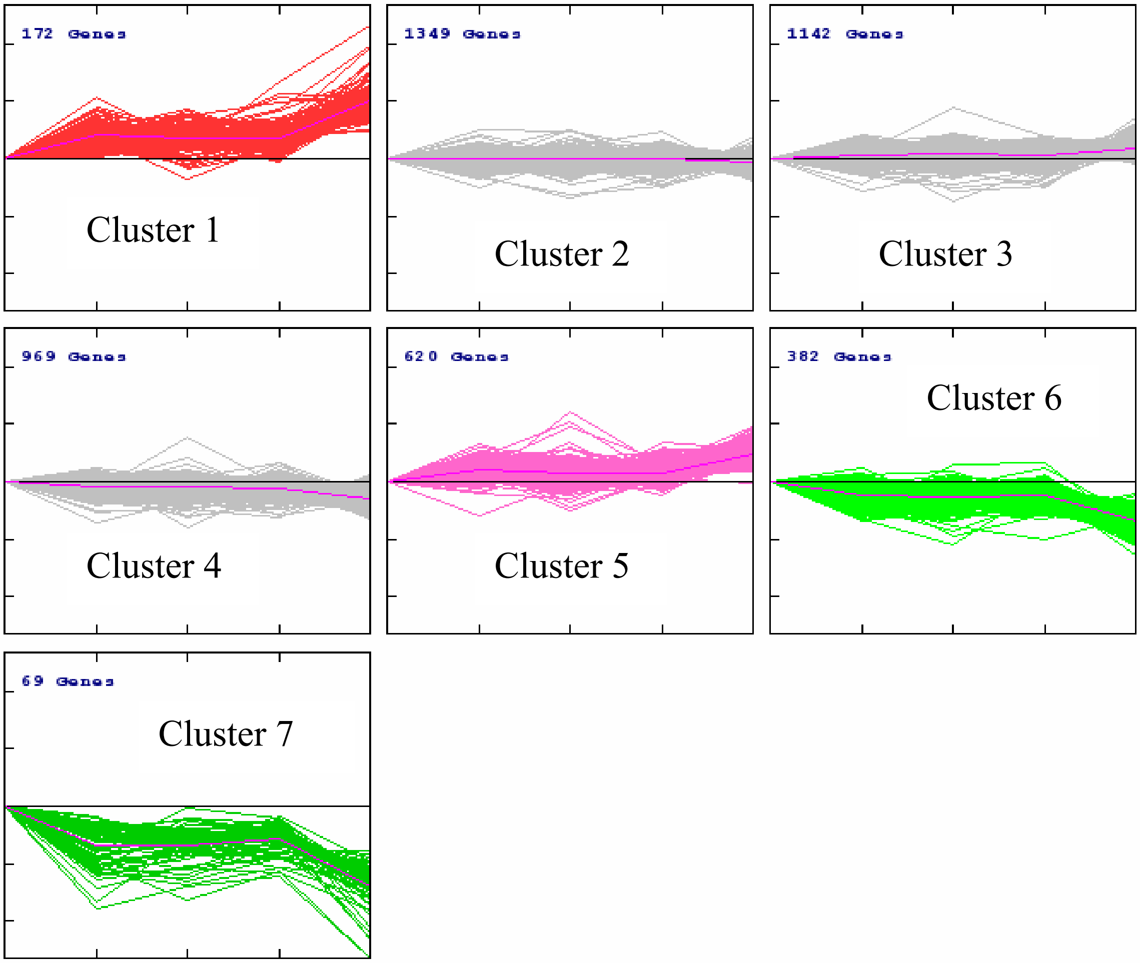


Figure 4.11: 7-Clusters obtained from k-means clustering with ratio of expressions.

Fig 4.11 shows that clusters obtained from k-means analysis. Clusters were colored accordingly:

Red: Strongly over-expressed

Pink: moderately over-expressed

Dark green: Strongly under-expressed

Light green: Moderately under-expressed

The rest of the clusters are not colored and indicates genes that are not undergoing considerable change in expression between perturbed and control.

K-means clustering of the same data set with number of clusters as 4 gives following clustering pattern (Fig 4.12). Coloring of the genes is based on the coloring done in Fig 4.11. Coloring shows that cluster 1 (strongly over-expressed) and part of cluster 5 (moderately over-expressed) of Fig 4.11 constitutes the cluster 4 of Fig. 4.12. Similarly cluster 7 (strongly under-expressed) and part of cluster 6 (moderately under-expressed) of Fig 4.11 constitute the cluster 1 of Fig 4.12.

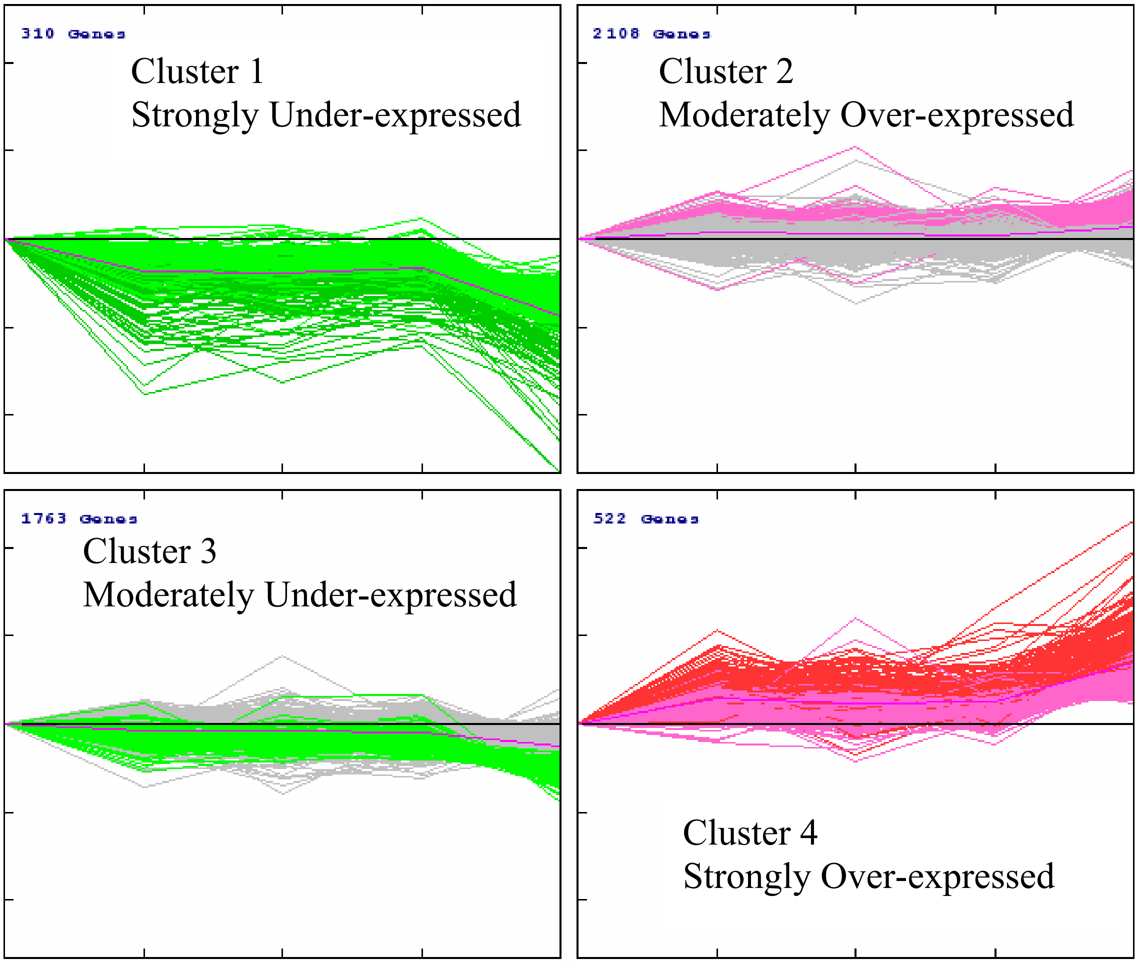


Figure 4.12: 4-Clusters obtained from k-means clustering with ratio of expressions. (Coloring of the genes is based on the coloring of clusters in Fig 4.11)

4.5.4 Comparison of clustering results:

4.5.4.1 Comparison of paired SAM and k-means:

A study was conducted to compare the results obtained from k-means, one class and paired SAM as measure of separating genes that show differential expression from control to perturb. As different cutoff were used for paired

SAM (90%) and one class SAM or k-means(100%) clustering, number of genes used for analysis were different. Number of genes used in paired SAM was 7321 whereas in k-means one class SAM was 4703. So the genes used for k-means clustering are a subset of genes used in paired SAM.

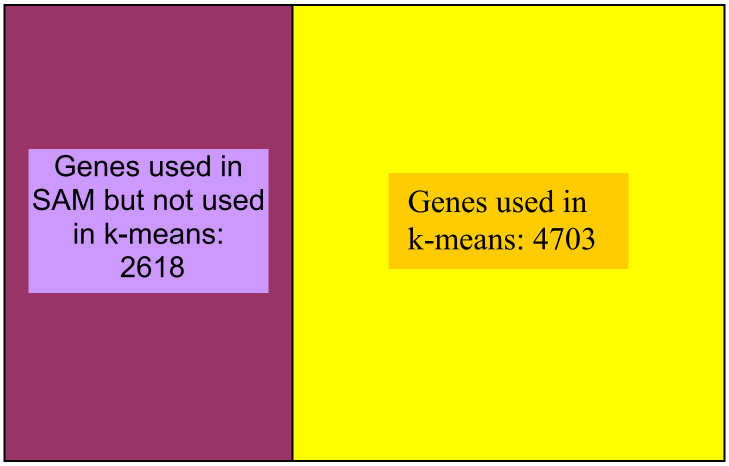

Genes used in SAM but not used in k-means: 2618
Genes used in k-means: 4703

Figure 4.13: Number of genes used in k-means (4703), is a subset of number of genes used in SAM (7321). (The yellow box represents the genes used for k-means analysis and the violet box represents the genes used in SAM but not used in k-means.)

.

Fig 4.14 gives pictorial representation of clusters obtained from k-means analysis. Clusters colored red represents over-expressed genes and clusters colored green represents under-expressed genes. Colored part of the rectangle represents the genes used for k-means clustering. The white part of the rectangle represents the genes used for paired SAM but not used for k-means clustering. 4 clusters obtained from k-means are shown in 4 different colors. Dark red and orange colored squares represent clusters of strongly and moderately over-expressed genes. Similarly dark green and light green squares represent clusters of strongly and moderately under-expressed genes respectively.

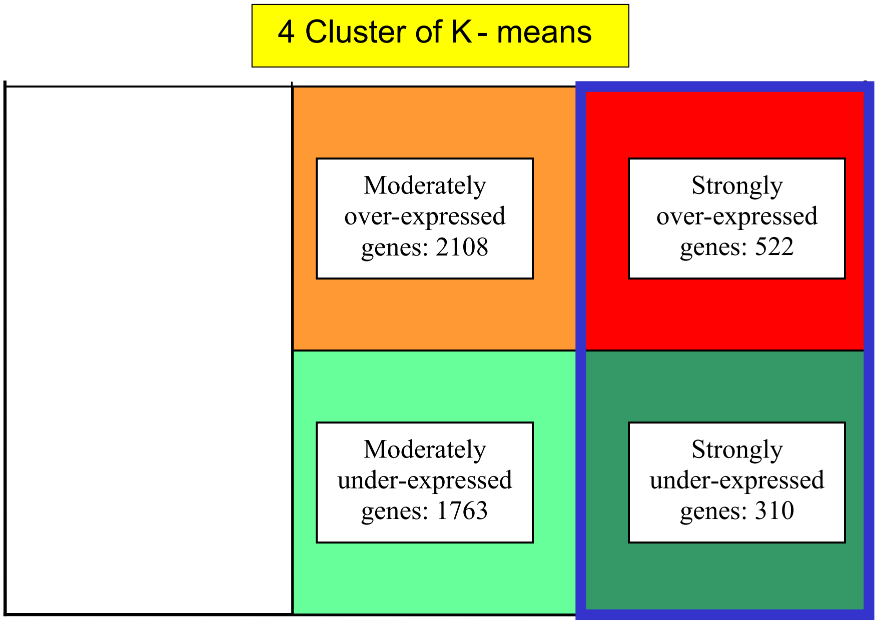

4 Cluster of K - means
Moderately over-expressed genes: 2108
Strongly over-expressed genes: 522
Moderately under-expressed genes: 1763
Strongly under-expressed genes: 310

Figure 4.14: Pictorial representation of clusters obtained from k-means analysis.

A pictorial representation of overlap of clusters obtained from SAM and k-means. Numbers of genes contained in each of the small rectangles are written in the white box inside

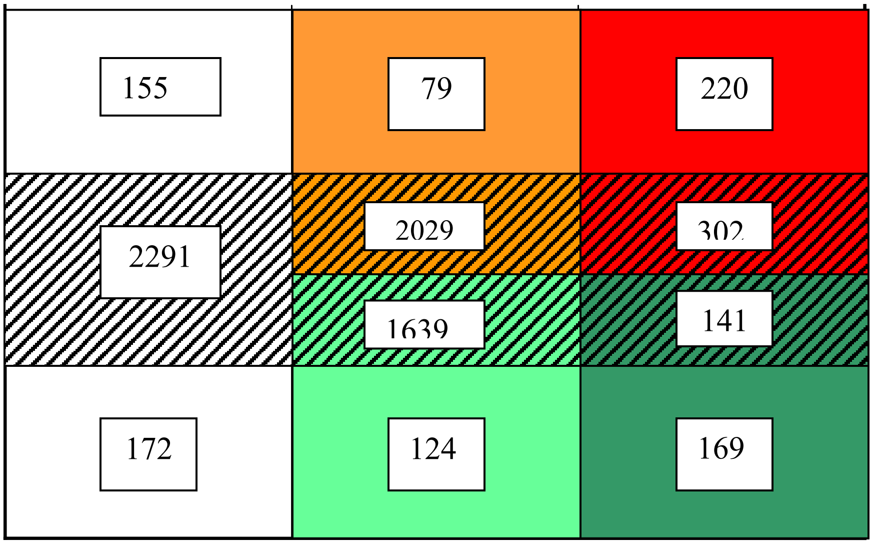


Figure 4.15: Shows the overlap of different clusters obtained from SAM and k-means clustering..

This figure 4.15 represents following things:

- Overlap of significant genes of SAM with cluster 1 and 4 of K-means (representing strongly over or under expressed genes) is much more than that of cluster 2 and 3 of k-means. This is expected, because cluster 1 and 4 of k-means represents "significant" genes from that analysis.
- There are 302 and 141 genes shows "significantly" over and under-expressed from k-means but are found non-significant by SAM.
- 155 and 172 positively and negatively significant genes respectively obtained from SAM were not used for k-means clustering.

4.5.4.2 Comparison of paired SAM and one class SAM:

Results obtained from paired SAM (using TIGR TM4 software) and one class SAM (using Stanford SAM software) were compared to study the similarity and differences between their results. Unlike k-means clustering, objective of both the techniques is to find out genes that are differentially expressed, whereas, the objective of k-means clustering is to group the genes that have similar expression profiles. As explained earlier, one class SAM uses only one set of data. The expression ratio of perturbed to control was used for one class SAM analysis, whereas control and perturbed sets were separately used for paired SAM. Though the question asked is same, one class and paired SAM uses different data sets and the algorithms to find out the significant genes are slightly different (refer to section 4.5.4). So it is expected that some of the genes will be identified as significant by one of these methods, not by the other. 90% cutoff (7321 genes) was used for paired SAM and 100% cutoff (4703 genes) was used for one class SAM. Following figure (Fig 4.18) gives a pictorial diagram of the overlap of the genes between one class and paired SAM. Overlap of one class SAM with paired SAM is much is much more than overlap between k-means and paired SAM. This is expected as paired and one class SAM has same objective, but they act on different type of data.

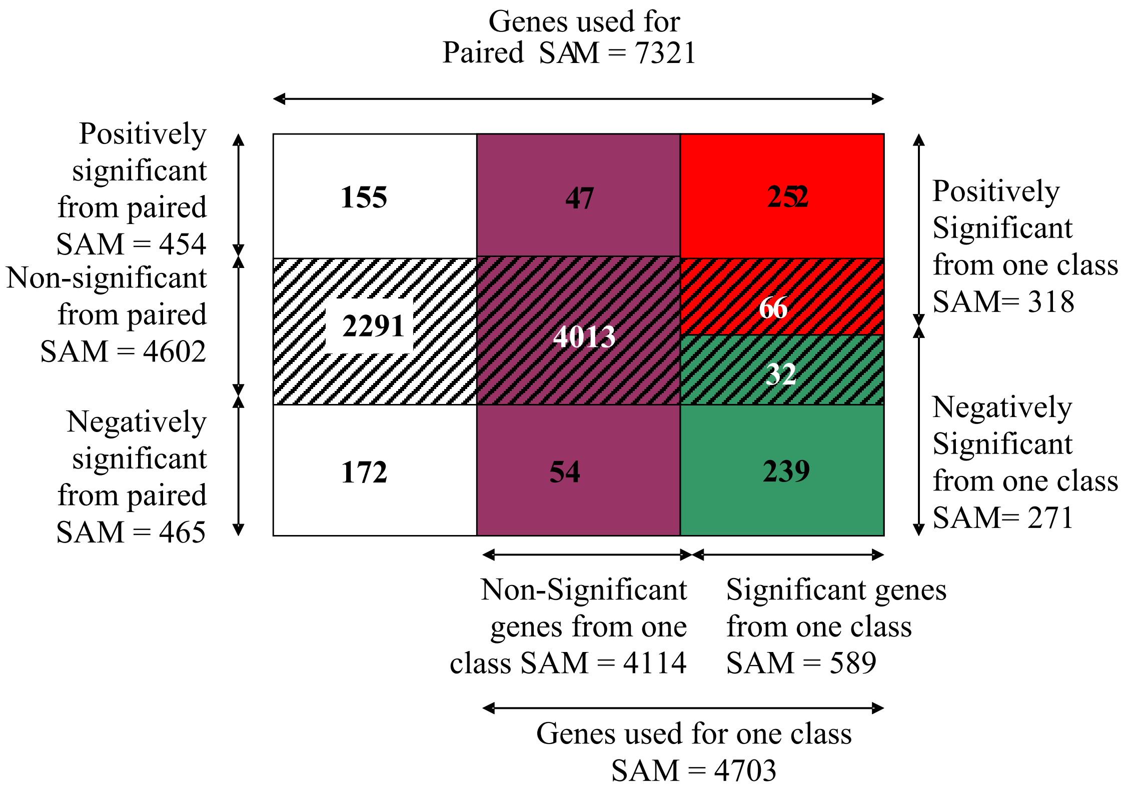

Genes used for
Paired SAM = 7321
Positively significant from paired SAM = 454
Non-significant from paired SAM = 4602
Negatively significant from paired SAM = 465
155
47
252
2291
4013
66
32
172
54
239
Positively Significant from one class SAM= 318
Negatively Significant from one class SAM= 271
Non-Significant genes from one class SAM = 4114
Significant genes from one class SAM = 589
Genes used for one class SAM = 4703

Figure 4.16: Overlap of the positively, negatively and non-significant genes from one class and paired SAM.

The figure 4.16 represents following things:

- Out of 318 genes that are found positively significant from one class SAM (the red box), 252 genes (79%) were also found significant from paired SAM, whereas 42% genes from the cluster of over-expressed genes from k-means were found to overlap with paired SAM.
- Similarly 239 genes out of 271 (the green box) negatively significant genes (88%) overlap with that of paired SAM, whereas 54% of the genes overlap between under-expressed cluster of k-means and negatively significant of paired SAM.

- The violet box represents the non-significant genes from one class SAM.
- There are 101 (47 + 54) genes that are found non-significant from one class SAM but significant from paired SAM.
- There are 98 (66 + 32) genes that are found significant by one class SAM but non-significant by paired SAM.

4.5.4.3 Comparison of k-means and one class SAM:

For both the analysis k-means and one class SAM same data set was used, which is the ratio of perturbed to control expression. Same cutoff (100%) was also used. As explained earlier, objective of one class SAM is to find out the genes that are differentially expressed, whereas the objective of k-means is to group the genes that have similar expression profiles.

Figure 4.19 shows a pictorial diagram of the overlap of the positively and negatively significant genes of one class SAM with 4 clusters obtained from k-means.

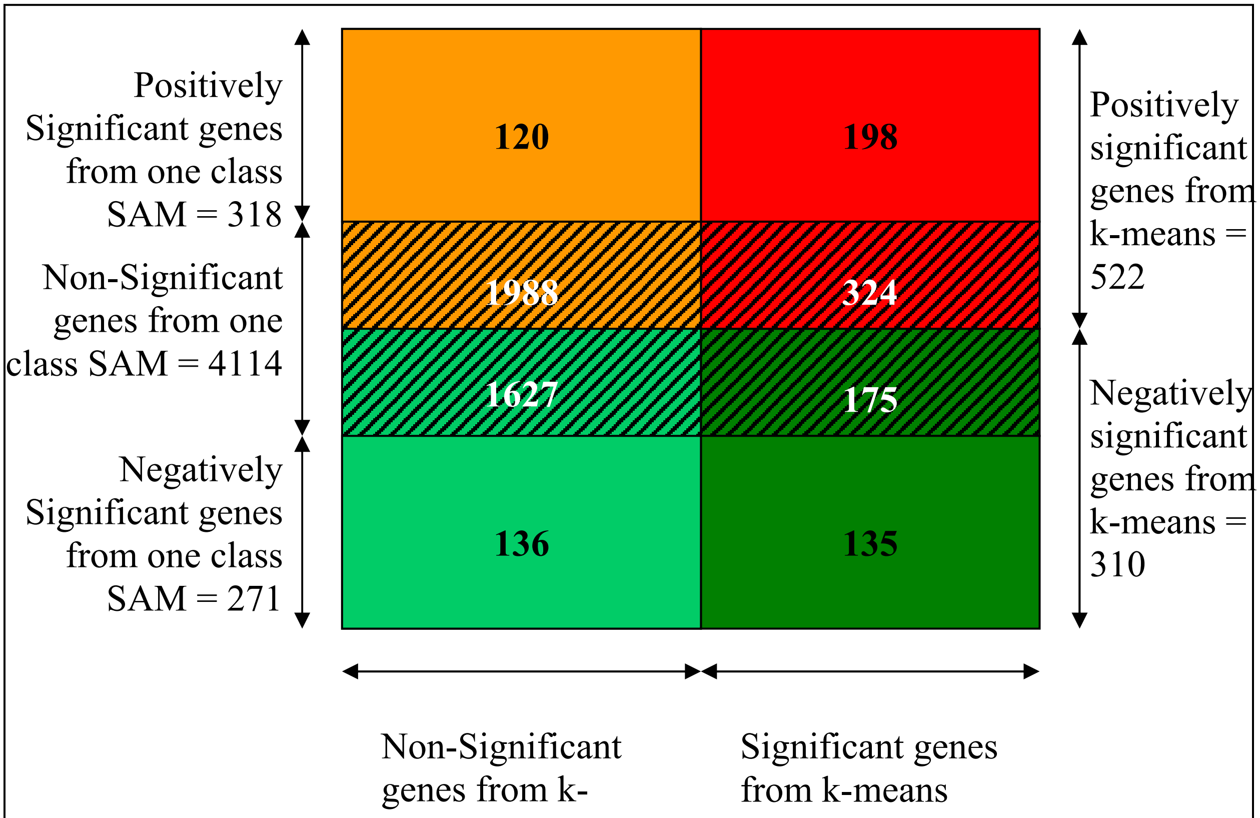

Positively Significant genes from one class SAM = 318
Non-Significant genes from one class SAM = 4114
Negatively Significant genes from one class SAM = 271
120
198
1988
324
1627
175
136
135
Positively significant genes from k-means = 522
Negatively significant genes from k-means = 310
Non-Significant genes from k-
Significant genes from k-means

Figure 4.17: A pictorial diagram of the overlap of the positively and negatively significant genes of one class SAM with 4 clusters obtained from k-means.

The figure 4.17 represents following things:

- Out of 522 genes found as strongly over-expressed from k-means analysis, only 198 genes are found positively significant from one class SAM. Overlap between over-expressed cluster of k-means with positively significant of one class SAM (38%) is less than paired SAM (42%).
- Out of 310 genes found as strongly under-expressed from k-means analysis, only 135 genes are found negatively significant from one class

SAM. Similar to the positively significant, overlap of k-means with one class SAM (43%) is less than paired SAM (54%).

- The shaded part on the diagram represents the non-significant genes (4114) from one class SAM.
- 120 genes were found positively significant from paired SAM that are not part of cluster of over-expressed genes from k-means.
- 136 genes were found negatively significant from paired SAM that are not part of cluster of under-expressed genes from k-means.

4.5.5 Functional classification of significant genes:

Genes that are found significant from paired SAM were broadly classified based on their function. Different functional categories were enzymes, transcription factors, ribosomal protein, transport related proteins, disease related proteins and proteins of unknown function. Classification was done separately for positively and negatively significant genes and were represented by different slice of the pie in figure 4.20. The genes that are encoding enzymes were further classified based on whether or not they are also found significant by k-means and one class SAM. There were four classes, i) genes that are significant by k-means, one class and paired SAM, ii) genes that are significant by both one class and paired SAM, iii) genes that are significant by paired SAM and over-expressed by k-means, iv) genes that are significant by paired SAM only.

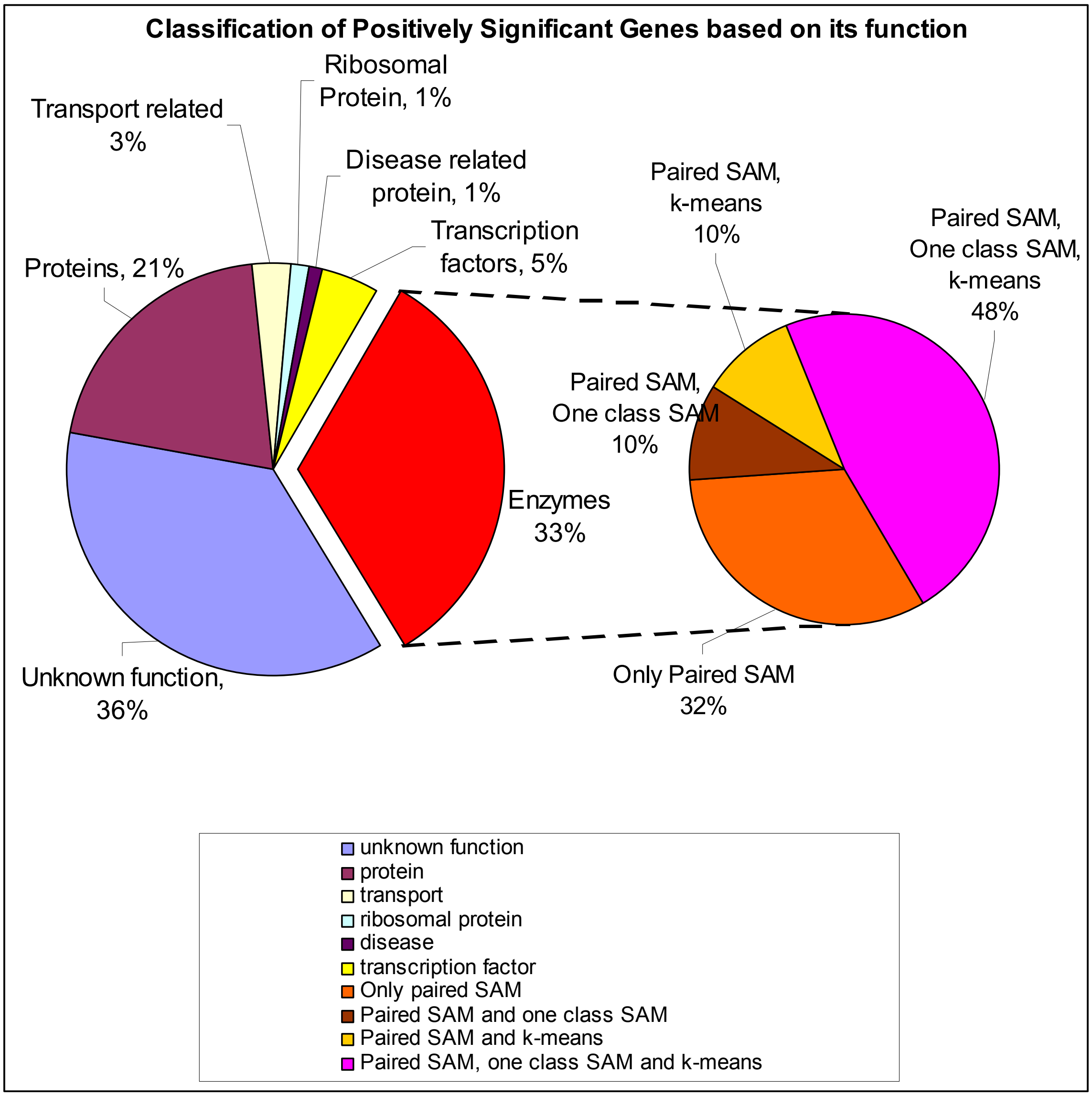


Figure 4.18A: Functional classification of positively significant genes from paired SAM.

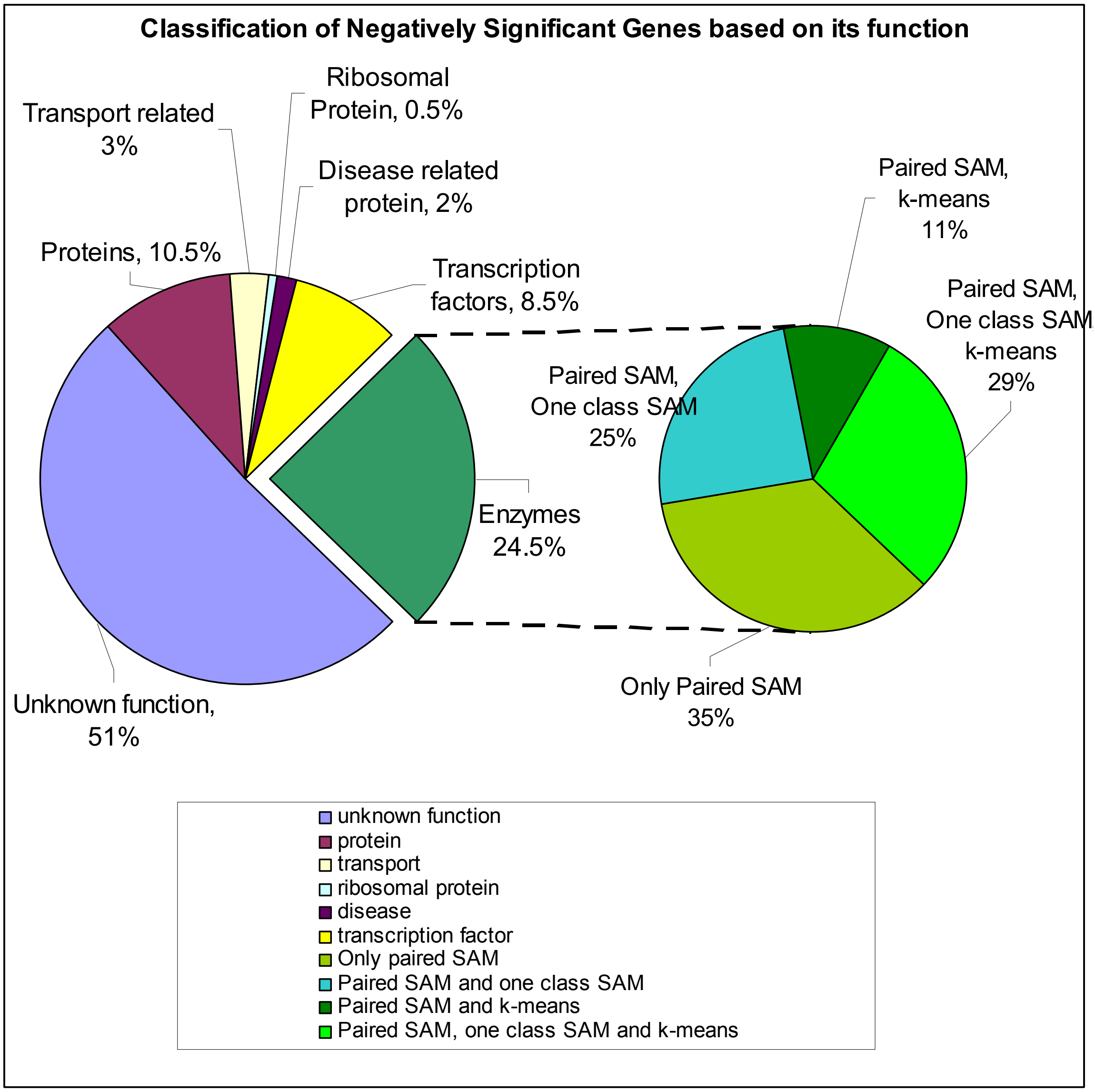


Figure 4.18B: Functional classification of negatively significant genes from paired SAM.

4.6 Use of metabolic information in the analysis of gene expression profiles:

The transcriptional activity of the following pathways has been studied. In all the figures, the enzymes encoded by the measured genes are colored red, green, blue and black. Colors represent the following gene expression level:

Red: the gene found positively significant from SAM.

Green: the gene found negatively significant from SAM.

Blue: the gene found non-significant from SAM.

Black: the gene is missing and not used in SAM analysis.

(+): the gene is present in strongly over-expressed cluster of k-means

(-): the gene is present in strongly under-expressed cluster of k-means

(0): the gene is missing in k-means analysis but used in SAM.

4.6.1 Photosynthesis and Calvin Cycle

Photosynthesis is involved in fixation of atmospheric CO_2. As CO_2 is a substrate of the reaction catalyzed by Rubisco and increase in CO_2 concentration will directly affect the rate of carbon fixation and Calvin cycle activity. So this pathway was studied:

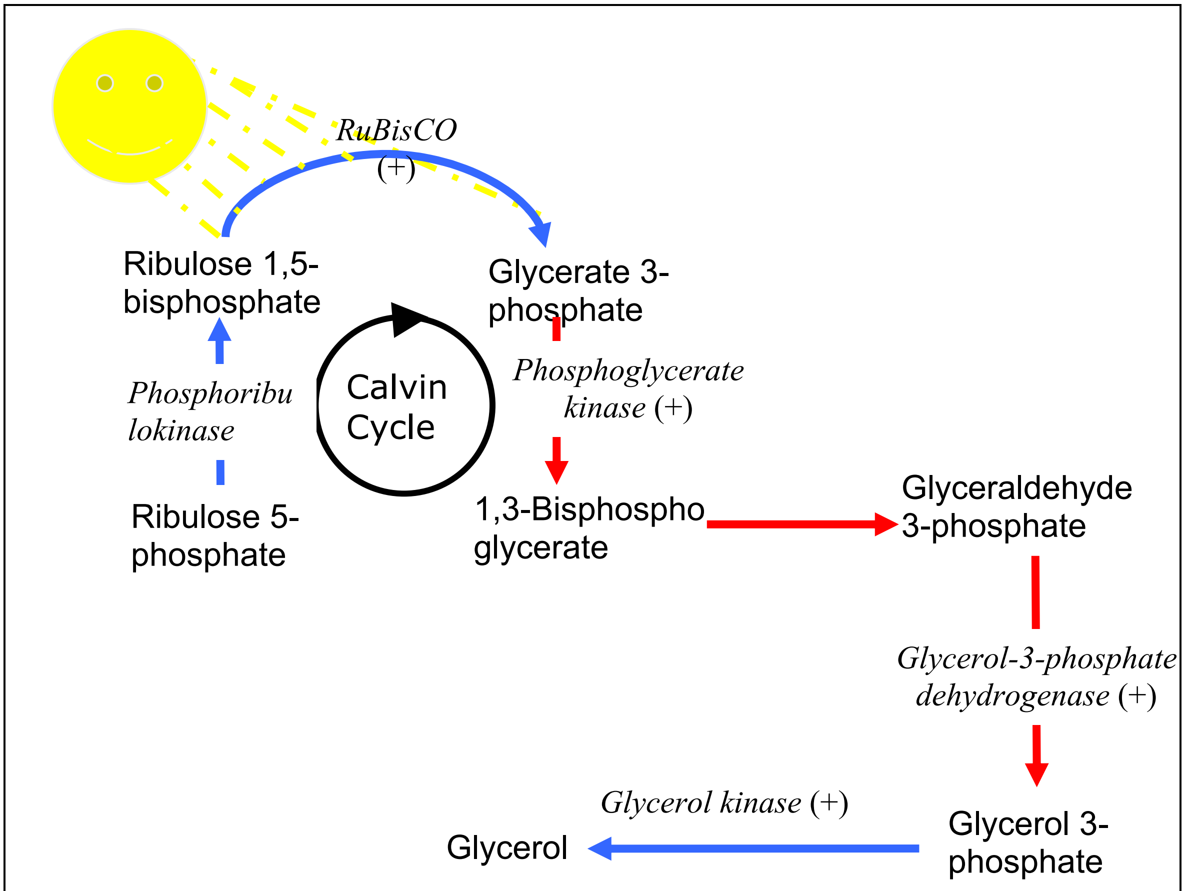


Figure 4.19: Carbon fixation reactions in Calvin cycle and subsequent production of glycerol from 1,3-bisphosphoglycerate

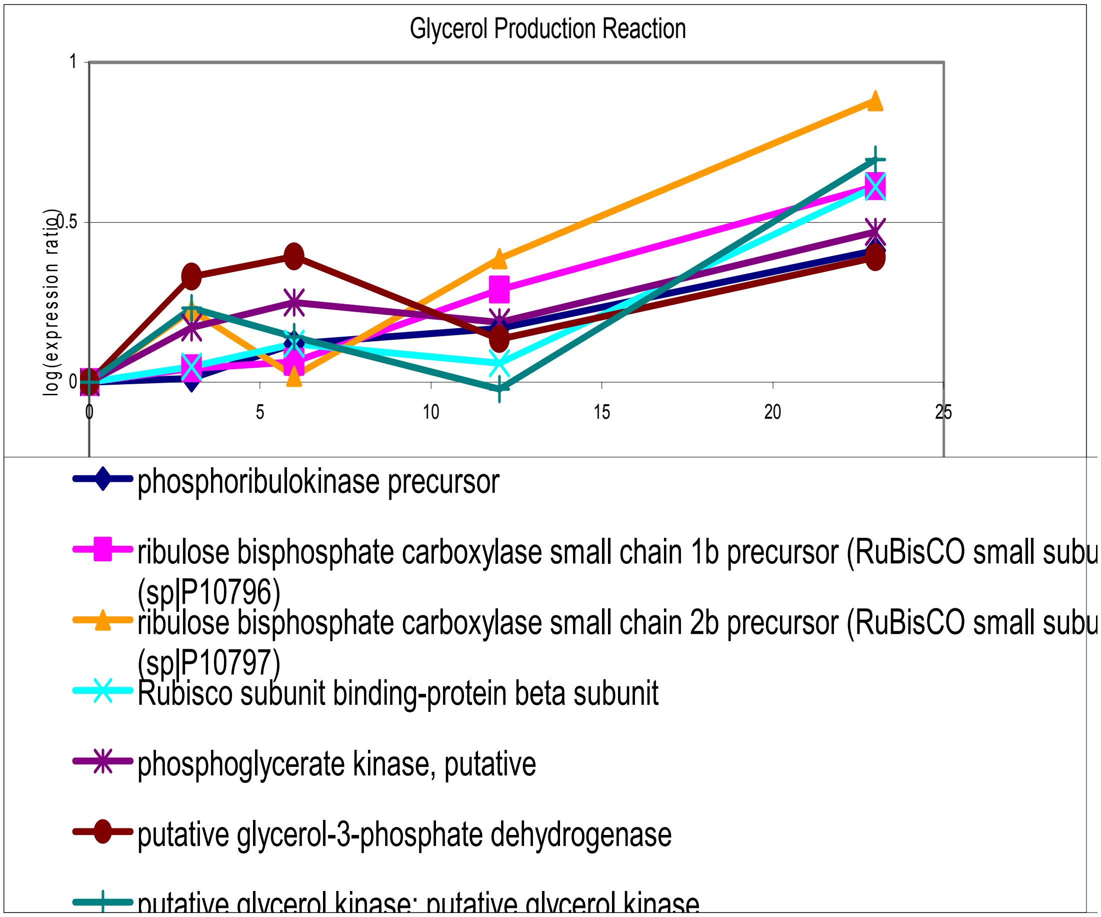
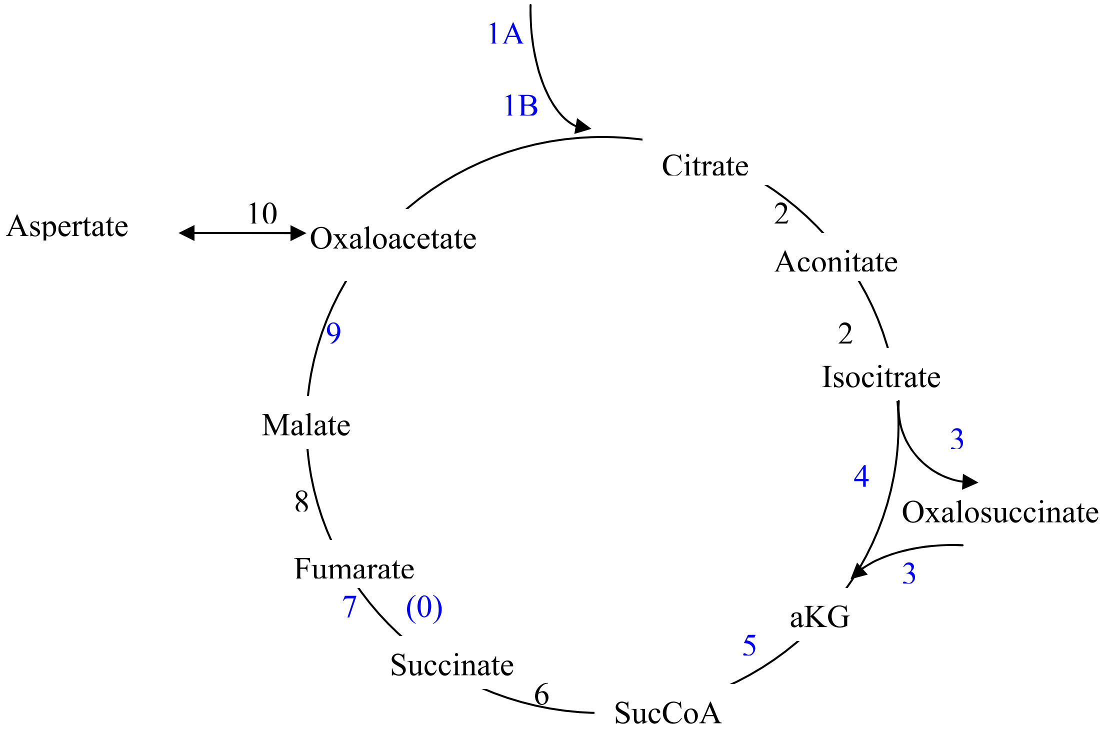


Figure 4.20: logarithm of expression ratio of perturbed to control were plotted for the genes encoding the enzymes in the figure 4.19

4.6.2 Tri Carboxylic Acid Cycle:

TCA cycle is one of the major pathways of central carbon metabolism. The rate of TCA cycle is directly related to respiration and amino acid biosynthesis.



Corresponding enzymes are:

1. A) ATP citrate lyase B) citrate synthase
2. aconitate hydratase
3. isocitrate dehydrogenase [NADP+],
4. isocitrate dehydrogenase [NAD+] subunit 2
5. 2-oxoglutarate dehydrogenase, E1 subunit
6. succinyl-CoA ligase beta subunit
7. succinate dehydrogenase [ubiquinone],
8. fumarase
9. malate dehydrogenase
10. aspartate aminotransferase

Figure 4.21: TCA cycle reactions with the metabolites and enzymes (The enzymes catalyzing these reactions are marked with numbers)

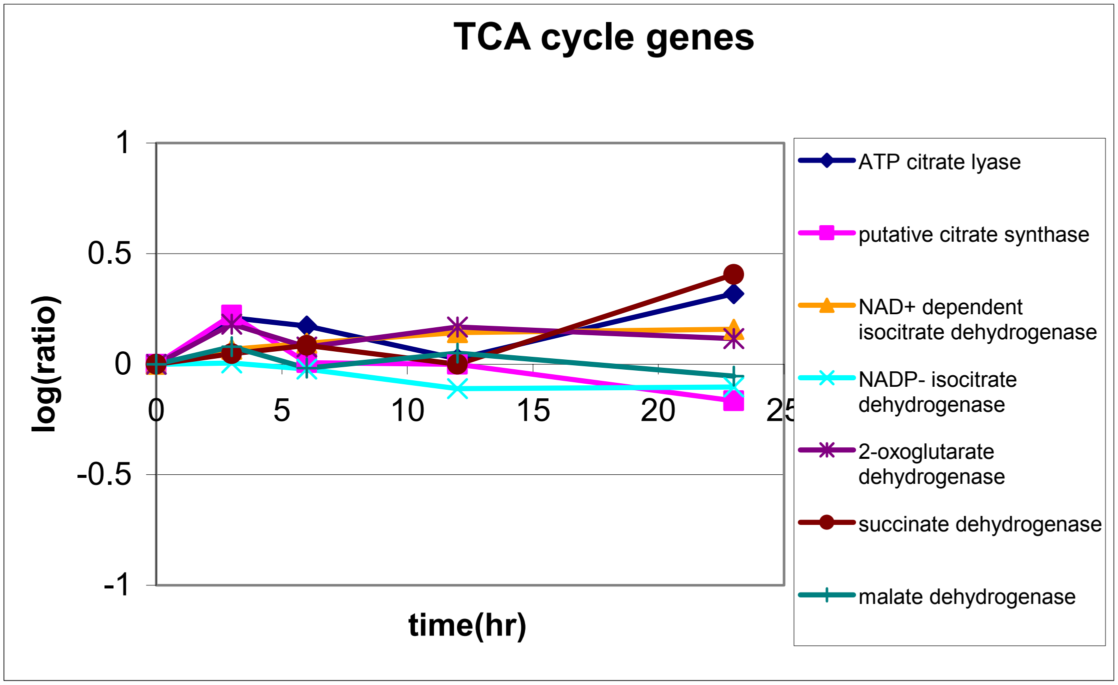
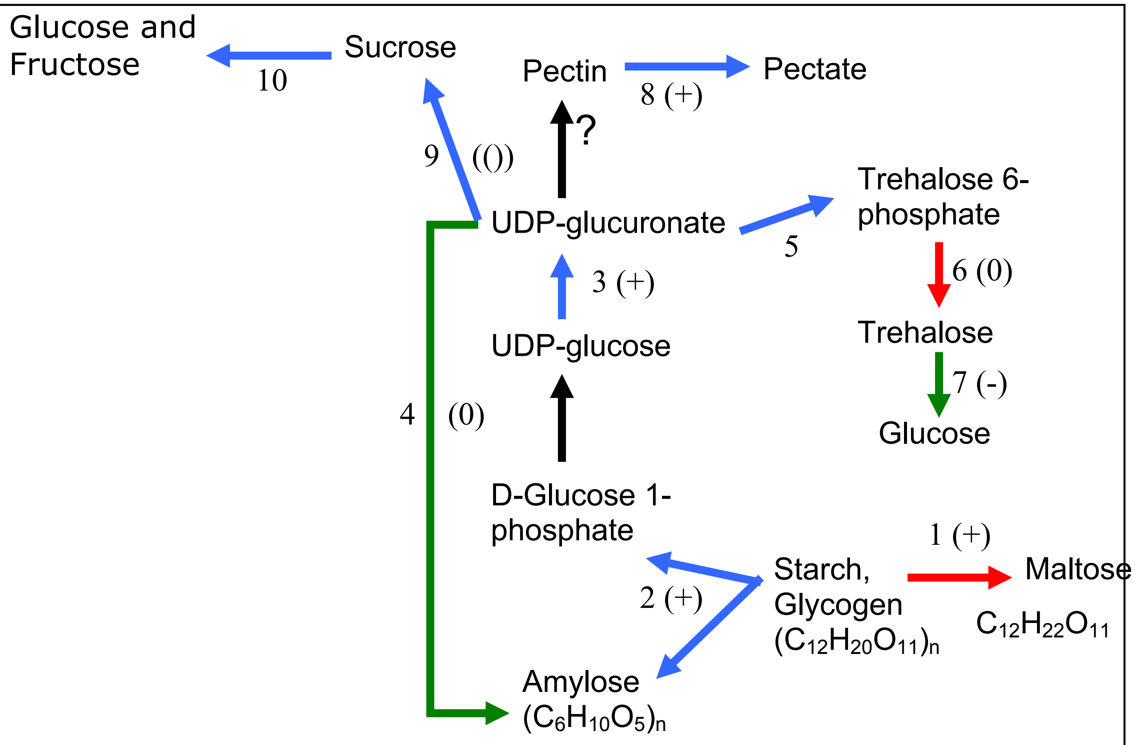


Figure 4.22: Graph shows the logarithm of the ratio of expression profile of perturbed and control.

4.6.3 Carbohydrate metabolism:

Figure 4.23 shows the reaction network of starch and carbohydrate metabolism



Corresponding enzymes are:
1. Beta-amaylase
2. Starch phosphorylase
3. UDP-glucose dehydrogenase
4. Starch synthase:
5. Trehalose 6 phosphate synthase:
6. Trehalose-6-phosphate phosphatase:
7. Trehalse:
8. Pectin methylesterase
9. Sucrose Synthase:
10. Alpha-glucosidase precursor

Figure 4.23: Reaction network of starch and carbohydrate metabolism is shown (Enzymes catalyzing the reactions are numbered and are also listed)

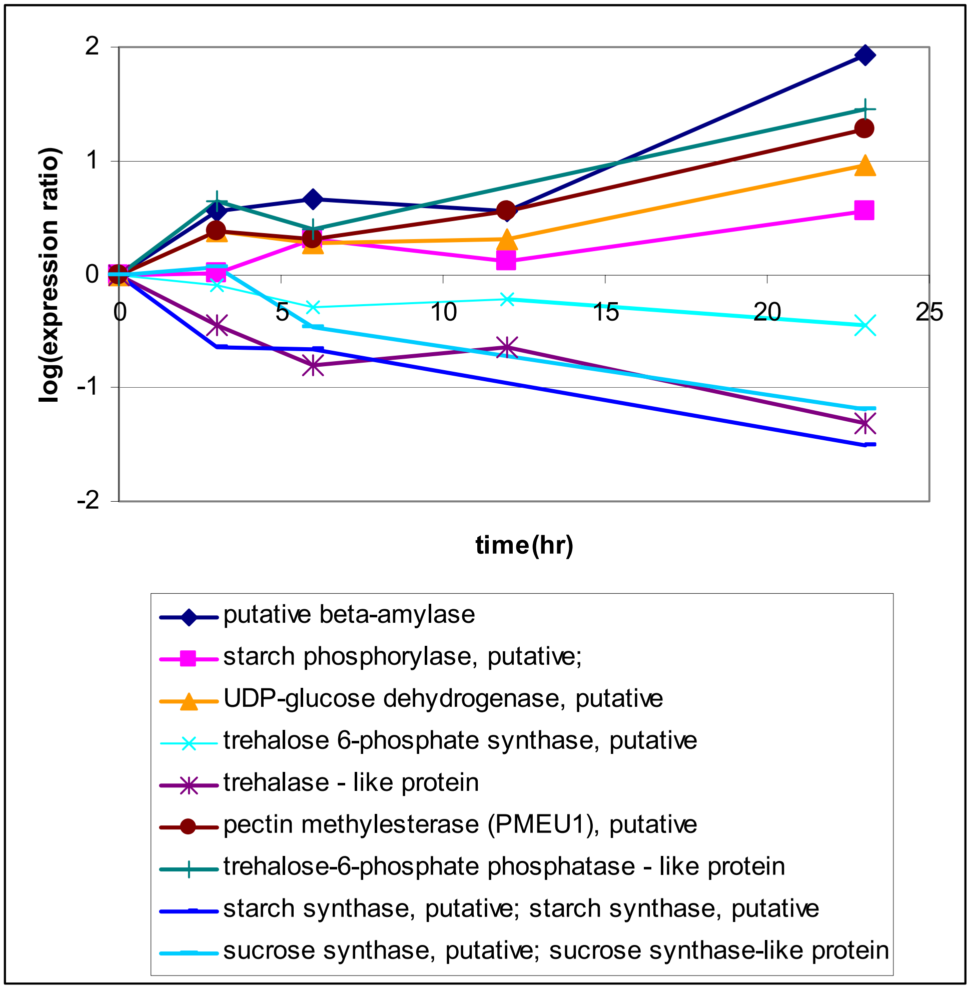


Figure 4.24 logarithm of the ratio of expression profile of the genes in carbohydrate metabolism are plotted

4.7 Plant Growth and Cell Wall Expansion:

Under condition of elevated CO_2 the rate of plant growth is expected to increase. In this context plant cell wall is expected to expand. To validate these assumptions and previous observations, the genes encoding proteins catalyzing the biosynthesis of cell wall components and/or the expansion of cell wall were studied.

4.7.1 Proteins Involved in Cell Wall Expansion:

Xyloglucan molecules are bound to the surface of cellulose microfibrils by hydrogen bonding and it helps cellulose microfibrils in cross binding to form the cell wall. Four proteins Xyloglucan Endotransglycosylase, expansin, cellulose synthase and endo-1,4-beta glucanase are found to be responsible for construction and modification of the cellulose xyloglucan framework [Kazuhiko et al., 2002]. All the five genes in fig 4.25 are present in the cluster of positively significant genes from SAM. But only three of them At2g01850, At4g03210 and At4g30280 are present in the cluster of strongly over-expressed genes from k-means. Other two genes have missing data at one of the time points, so are not used for k-means clustering.

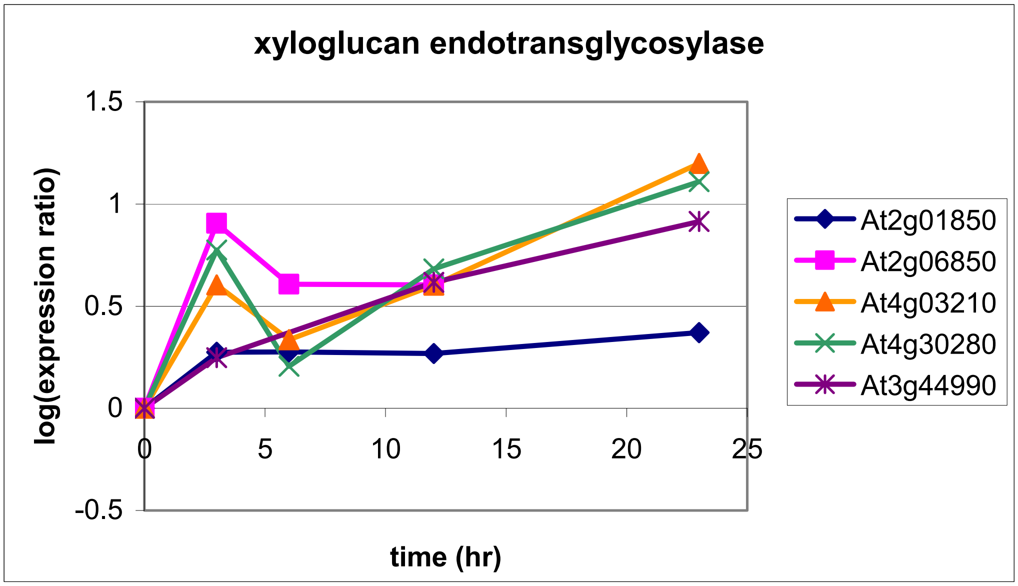


Figure 4.25: Different genes encoding xyloglucan endotransglycosylase are showing over-expression.

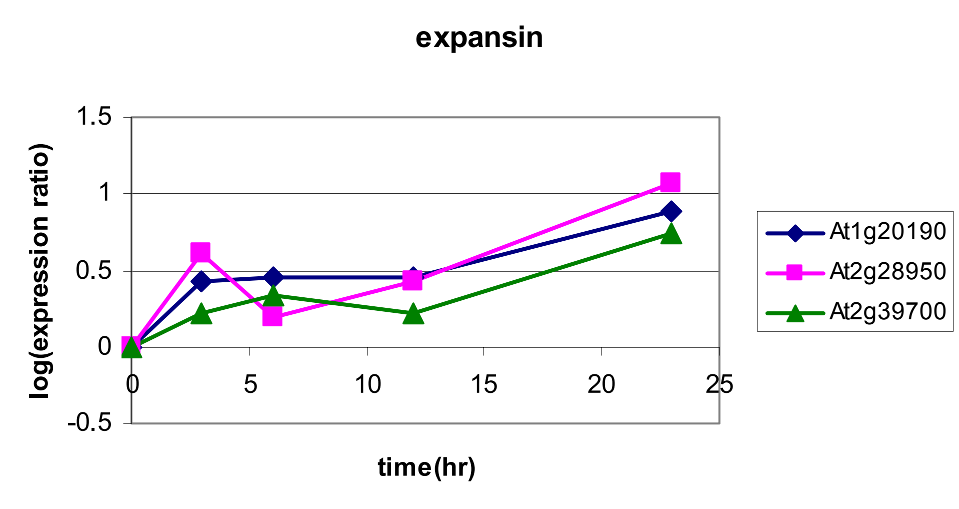


Figure 4.26: Three genes encoding expansin is found to be over-expressed at a longer time frame.

All three genes in fig 4.26 are present in cluster of strongly over-expressed genes from k-means. But the gene At2g39700 is found to be non-significant, while the other two are positively significant from SAM.

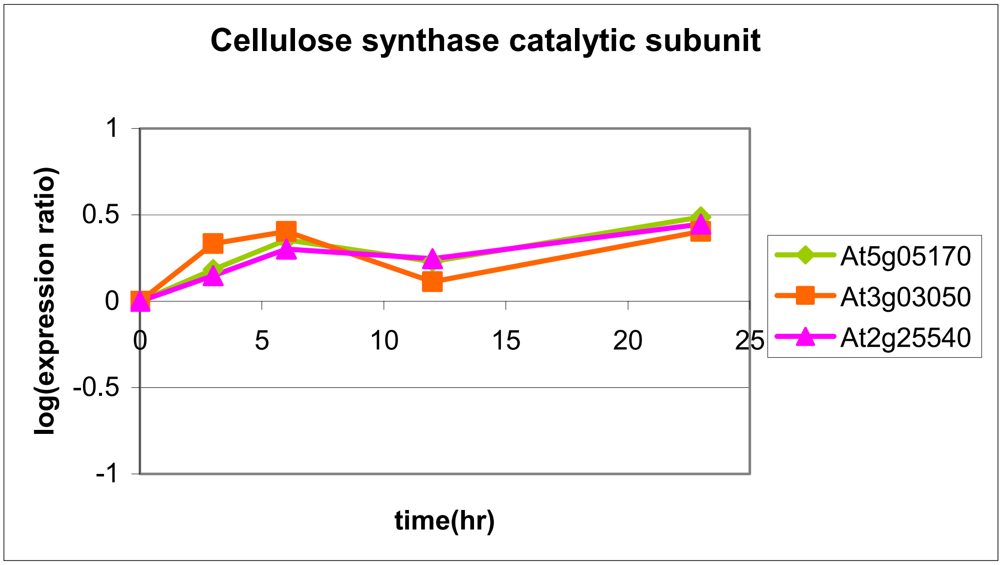


Figure 4.27: Genes encoding cellulose synthase show similar expression pattern.

All the three genes in fig 4.27 out here are present in the cluster of strongly over-expressed genes from k-means and positively significant genes from SAM.

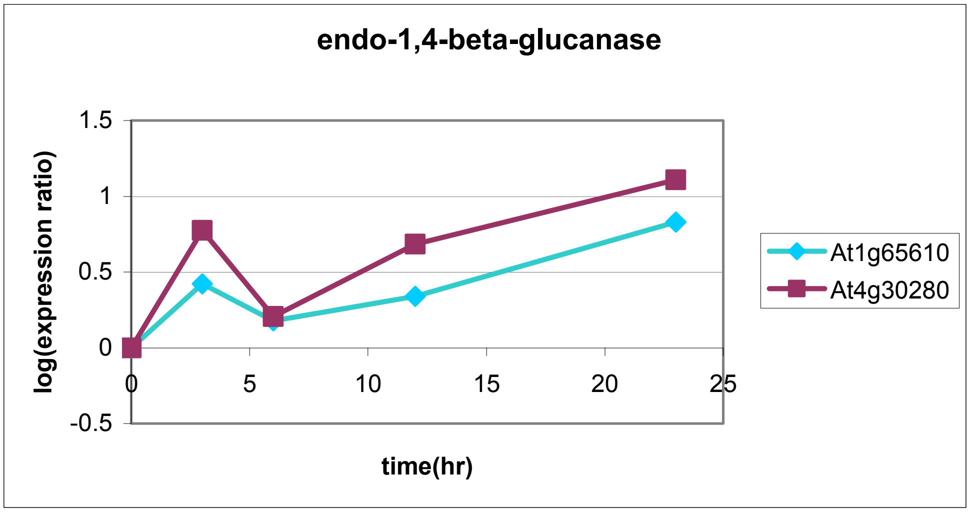


Figure 4.28: Genes encoding the enzyme endo-1,4-beta-glucanase are over-expressed and show similar expression.

Expression of all the four proteins involved in cell modification xyloglucan endotransglycosylase, expansin, cellulose synthase, endo-1,4-beta-glucanase are found to be over-produced in perturbed set compared to control.

4.7.2 Nucleotide and Histone production:

Nucleotides are building blocks of DNA. Ribonucleotide reductase catalyses the following reaction in nucleotide production pathway [www.kegg.com]:

ADP → dADP

GDP → dGDP

CDP → dCDP

UDP → Dudp

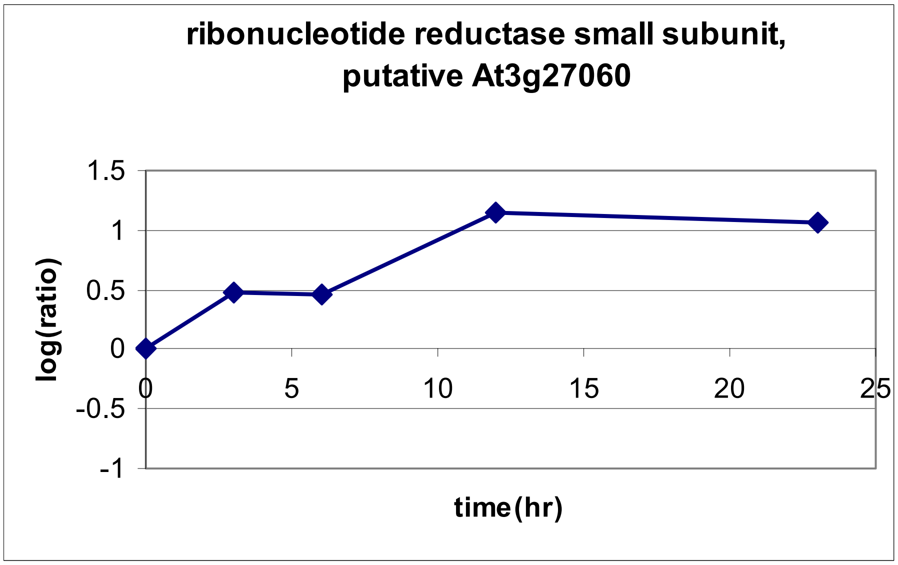


Figure 4.29: Ribonucleotide reductase catalyses the reaction in nucleotide production pathway found over-expressed.

The gene encoding this enzyme is found to be strongly over-expressed (fig 4.29).

DNA is wrapped around the Histone proteins. Histone has 4 subunits H2A, H2B, H3 and H4 [Albert et al., 2002]. Production of all the different subunits of Histone proteins are found to be increased (Fig 4.30) in perturbed set compared to control set. In fig 4.30 only the two genes encoding H2A are positively significant from SAM but the rest of the genes are non-significant.

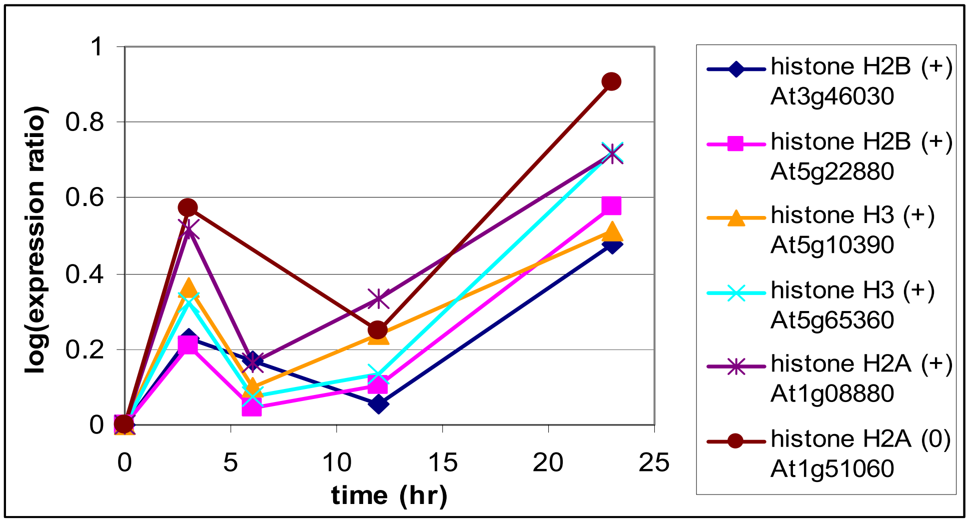


Figure 4.30: Genes encoding all the four subunits H2A, H2B, H3 and H4 of Histone protein are over-expressed.

As DNA is wrapped around Histone protein, increase in both the nucleotide production and Histone protein production is believed to be in agreement with each other.

Mei2 [Hirayama et al., 1997] has been thought to be a key protein for switching the mitotic cell cycle to meiosis. Mei2 has three putative RNA-recognition motifs (RRM) and actual RNA binding activity that is necessary for Mei2 function. The activity of Mei2 is thought to be regulated at the transcriptional and post-translational level [Hirayama et al., 1997]. Expression of Mei2 and RRM are decreasing (Fig 4.31). All the genes (in Fig 4.31) except At2g42890 are found negatively significant from SAM.

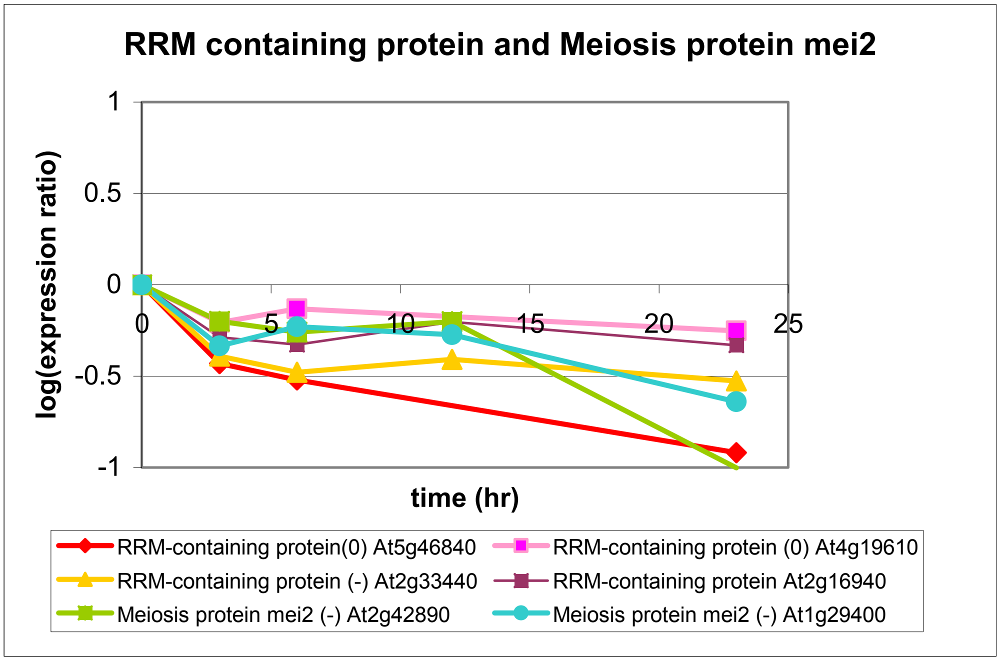


Figure 4.31: Genes encoding meiosis protein Mei2 and RRM containing proteins are under-expressed.

Hexose like D-Glucose, D-mannose, D-fructose, sorbitol are converted to their corresponding phosphate by hexokinase [Lehninger et al., 2002]. Both the genes producing hexokinase (At2g19860) and fructokinase (At1g06030) in fig

4.32 was found to be present in the cluster of strongly under-expressed genes from k-means and negatively significant genes from SAM

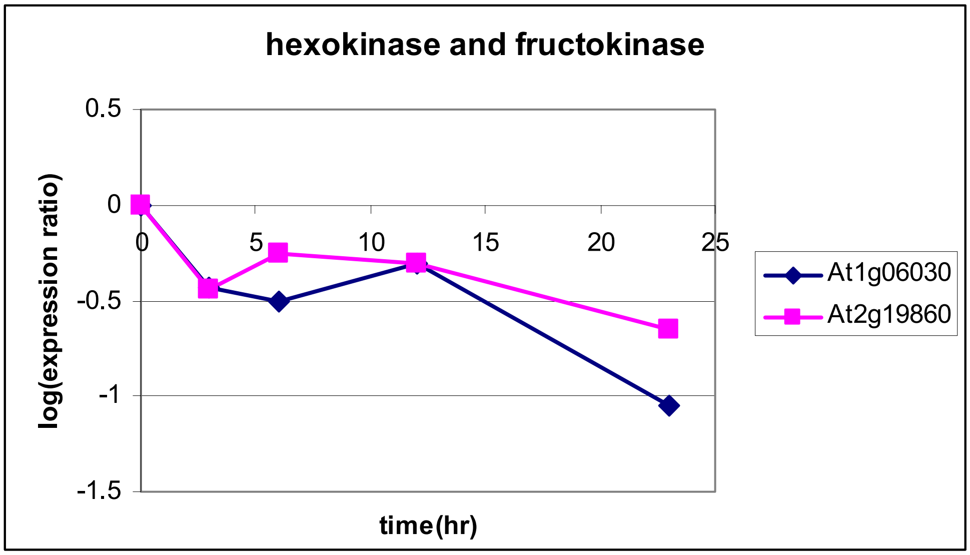


Figure 4.32:. Hexokinase and fructokinase are found under-expressed

Phosphoenolpyruvate carboxylase catalyses the Anaplerotic pathway reaction from phosphoenolpyruvate to oxaloacetate.. In Mesophyll cells carbon fixation reaction takes place, which produces oxaloacetate from pyruvate is also catalyzed by phosphoenolpyruvate carboxylase [Lehninger et al., 2002]. Both the genes in Fig 4.33 belong to the cluster of strongly over-expressed genes from k-means, but they were found non-significant from SAM.

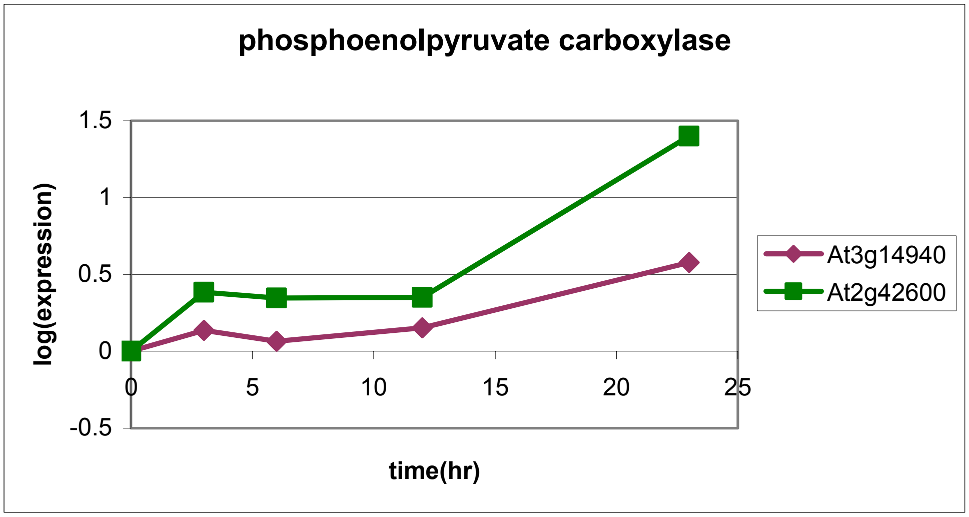


Figure 4.33: Two genes encoding phosphoenolpyruvate carboxylase were found over-expressed in perturbed system.

Pyruvate decarboxylase catalyses the reaction from pyruvate to AcetylCoA in glycolysis pathway. It also catalyses the reduction of private to acetaldehyde [www.kegg.com]. The pyruvate decarboxylase gene belongs to the cluster of strongly under-expressed genes from k-means, but it was found non-significant from SAM

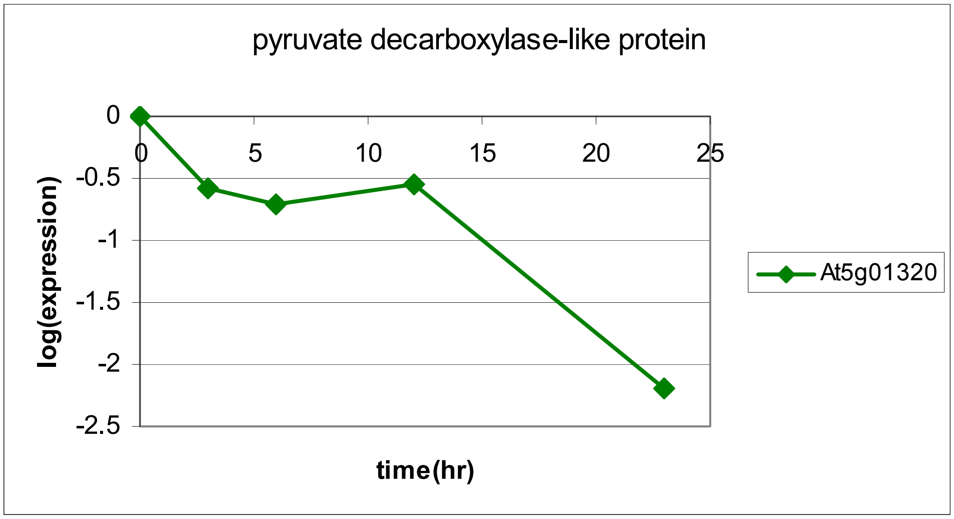


Figure 4.34: Gene encoding pyruvate decarboxylase-like protein shows a strong under-expression.

Nitrate reductase catalyses the reduction of nitrate to nitrite in the nitrogen assimilation pathway [www.kegg.com]. Nitrate reductase gene in fig 4.35 belongs to the cluster of strongly under-expressed genes from k-means and negatively-significant from paired SAM.

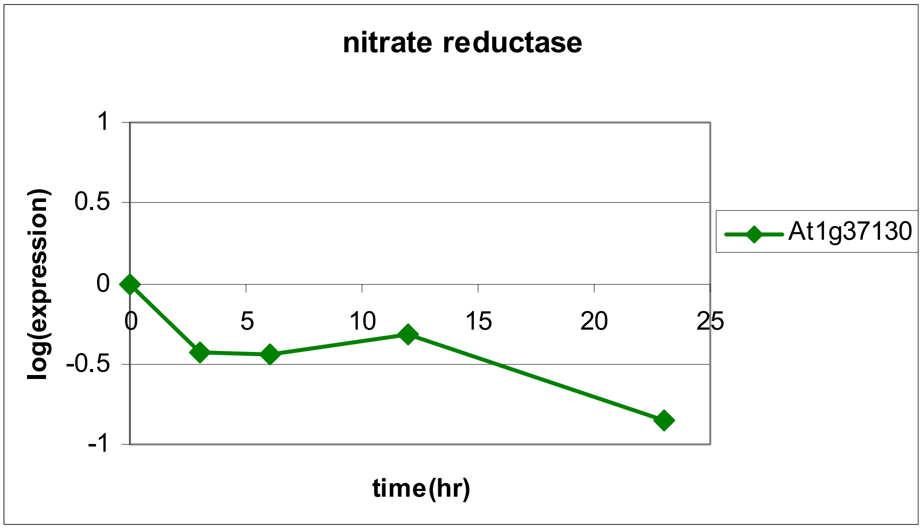


Figure 4.35: The gene encoding nitrate reductase is under-expressed.

Four genes encoding heat shock proteins are over-expressed (Fig 4.38) and belong to cluster of positively significant genes from SAM analysis. However, expression at 6hr is missing for the gene At5g56010. So this gene was not considered for k-means clustering. The remaining three genes are present in cluster of strongly over-expressed genes from k-means clustering.

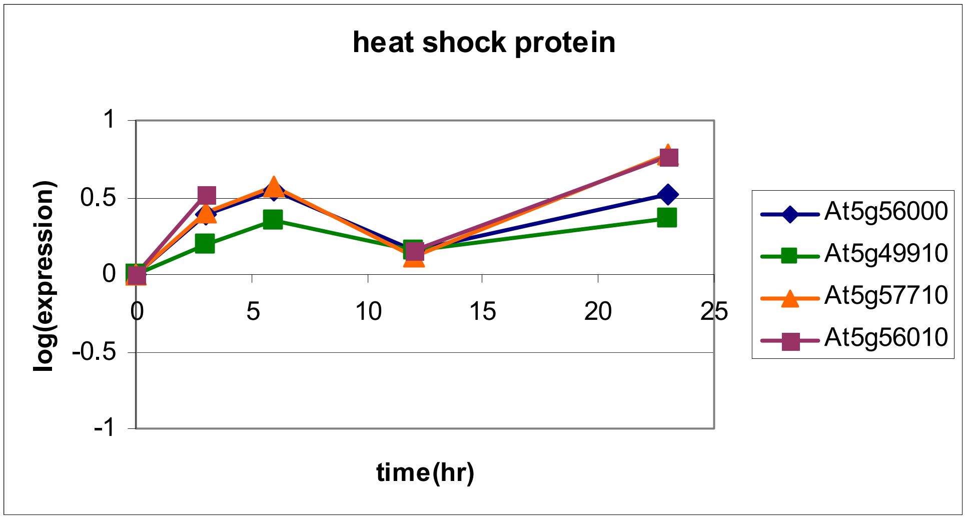


Figure 4.36: Heat shock proteins are over-expressed and show similar expression profiles

In each cluster from paired SAM and k-means there was considerable number of genes whose function can not be characterized. In the TIGR annotation either they are marked as expressed protein, hypothetical protein, putative protein, unknown protein or no function is assigned at all. For lot of genes putative function is assigned. It is debatable how sure one can be about the function of putative functions.

In the following chapter biological significance of the results obtained in this chapter will be discussed.

5. Discussion:

Transcriptional profiling analysis of two sets of data indicates that even during the short term treatment of the plants gene expression responds. Using full genome DNA microarray analysis of *Arabidopsis thaliana* physiology allowed us to study genomic response of the plants to the applied perturbation.

5.1 Data Normalization and Filtering:

Tables 4.2-4 depict the number of spots with non-zero intensity after normalization, multiplication of the biological replicates and division by 0th time point respectively. After normalization, average number non-zero spots are 12211 for control and 14922 for perturbed set. After the geometric mean of the biological replicates are taken, average number of spots for control and perturbed set are 10207 and 12422 respectively. Around 16% of the spots obtained after normalization were lost in both control and perturbed when geometric mean was taken. Due to division by 0th hr time point control and perturbed incur 7% and 26% loss of spot respectively.

Outlier analysis conducted in section 4.4 was not used to eliminate any experiment, when biological replicates were not clustering together. In this experiment there were two biological replicates at each time points. When the biological replicates are found inconsistent, then it can not be concluded which of the replicates are outlier. But if there are more than two bio-replicates, then if two replicates are inconsistent, then the third replicate can provide information to determine which one of these two is an outlier.

5.2 Data Analysis:

Results obtained using the algorithms k-means, paired and one class SAM was compared (section 4.5.4). Cluster of strongly over and under expressed genes obtained from k-means clustering is equivalent to significant genes of k-means (marked with blue box in Fig 4.14). Dataset used for k-means analysis is a subset of dataset used for SAM (Fig 4.13). There are 302 AND 141 genes that were found over and under-expressed by k-means but non-significant by paired SAM. Though these genes were used for SAM analysis but SAM algorithm do not consider them as significant. Fig 5.1 shows gene encoding phosphoenolpyruvate carboxylate is much strongly over-expressed compared to expression of the gene encoding 40S ribosomal protein –like, but the later is positively significant whereas the gene coding for phosphoenolpyruvate carboxylate is non-significant from SAM. In another similar example, Fig 5.1 shows gene encoding pyruvate decarboxylase is much strongly under-expressed compared to expression of the gene encoding 3-hydroxyisobutyryl-coenzyme A hydrolase, but the later is negatively significant whereas gene coding for pyruvate decarboxylase is non-significant from SAM. Both phosphoenolpyruvate carboxylate and pyruvate decarboxylase are present in the over and under-expressed cluster of k-means. Apparently, the genes that have sudden increase in expression at the last time point, is likely to be considered as non-significant by SAM. So analysis using only one clustering technique is not enough to find out the genes that are differentially expressed. k-means clustering helps to get lot of gene which are actually "significant" but found non-significant by SAM. SAM only calculates the difference in expression between control and perturbed set, but it does not take into account how the expression changes with time. as k-means compares expression profiles it inherently considers how the time sequence of the expression.

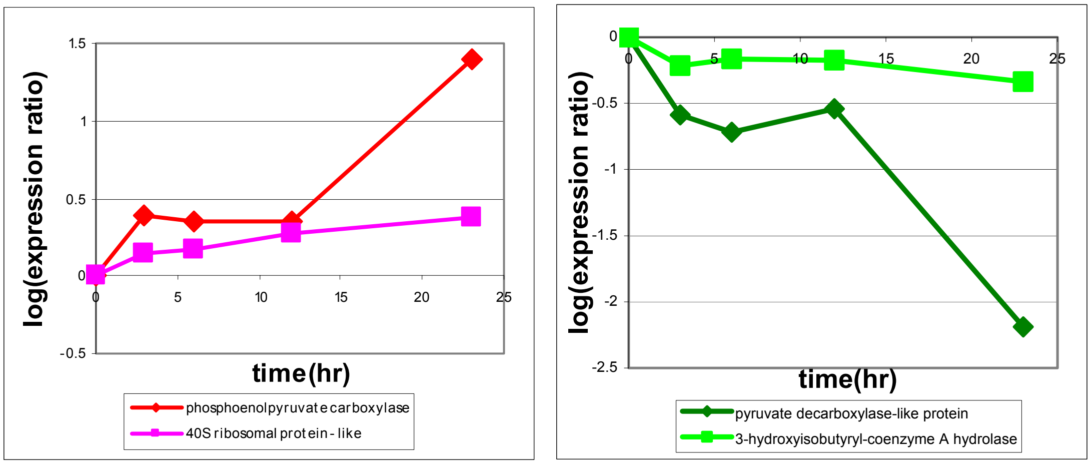


Figure 5.1 Limitation of SAM.

There are 155 and 172 genes that are found positively and negatively significant by SAM, but k-means do not find them significant because they were not used for k-means analysis. Only SAM provides information about these genes. There are 79 and 124 genes that are found positively and negatively significant by SAM but non-significant by k-means. These genes would have rejected if only k-means is used, but is able to find them significant.

5.3 Analysis of gene expression profiles in the context of *A. thaliana* physiology:

Elevated CO_2 will stimulate the carboxylation reaction catalyzed by Rubisco [Stitt et al., 1991]. Genes coding for Rubisco small chain 1b and 2b precursors were moderately over-expressed [Fig 4.20]. Increased carboxylation reaction also increases other reactions in Calvin cycle. Another enzyme of Calvin cycle Phosphoglycerate kinase was also found strongly over-produced from the results of both SAM and k-means [Fig 4.20]. Increased rate of carbon fixation causes increased rate of glycerol production which is used in cell wall. Gene coding for Glycerol-3-phosphate dehydrogenase and Glycerol kinase, two enzymes involved in glycerol production (Fig 4.19) are over-expressed (Fig 4.20). From metabolic analysis it is observed that glycerol is over produced.

Cheng et al.(1998) studied the short term effect of elevated CO_2. 30 day-old ambient CO_2 grown plants were transferred to high CO_2 for up to 12 days. A decrease in the transcription of Rubisco gene was observed. However, in our results we have found that the expressions of two small subunits of Rubisco are increasing. The plausible reason could be, we have studied the expression in first 23 hours, i.e. the immediate response of the plant to elevated CO_2 is to increase carbon fixation by increasing the transcription of Rubisco gene. But after a certain time, when substantial amount of carbon fixation has taken place, the reaction becomes limited by RuBP regeneration or end product synthesis. The time scale that Cheng et al. has considered, 12 day, is much larger than the time scale of our experiment. The contradiction shows how the time frame of the experiment should be taken into consideration when comparing the data from the literature.

Current view of plant cell wall model envisaged xyloglucan molecules to be bound to the surface of cellulose microfibrils by hydrogen bonding [Keegstra et al., 1973]. Some xyloglucan molecules are further linked covalently through certain cross linking poly-saccharides. Therefore, a cellulose microfibril coated with xyloglucan molecules is interconnected to two or more microfibrils, thereby forming a single super-molecular framework structure surrounding the cell [Keegstra et al., 1973]. Following proteins are found to be responsible for construction and modification of the cellulose xyloglucan framework [Kazuhiko et al., 2002]. These proteins include (1) xyloglucan endotransglucosylase/hydrolases (XTH) [Fry et al., 1992], [Nishitani et al., 1992] [Okazawa et at., 1993], (2) expansins [McQueen-Mason et al., 1992] [Shcherban et al., 1995] (3) cellulose synthases [Pear at al., 1996], [Arioli et al., 1998], [Taylor et al., 1999] (4) membrane-anchored endo-(1–4)glucanases [Brummell et al., 1998], [Nicol et al., 1998]. Genes encoding all four enzymes were found over-expressed (Fig 4.25, 4.26, 4.27, 4.28) leading to the conclusion that cell wall synthesis is going on. From the metabolic analysis it was also observed that cell wall material xylitol and arabinose are over produced at elevated CO_2.

It was observed at elevated CO_2 cell division increases and time interval between two successive division decreases Masle [2000]. Chen et al. [2003] observed at elevated CO_2 net rate of biomass accumulation increases. From this study, it was found nucleotide production (Fig 4.29) as well as different subunits of Histone production increases (Fig 4.30), which implies faster rate of DNA replication. On the basis of increased rate of cell wall production and nucleotide biosynthesis it is speculated that rate of cell division is increasing.

One of the most prominent consequences of elevated CO_2 enrichment is decrease in Nitrogen concentration [Sherwood, 2001]. Nitrate reductase catalyses the reduction of nitrate to nitrite in the nitrogen assimilation pathway [www.kegg.com]. In this study it was observed that gene coding for nitrate reductase (Fig 4.35) is under-expressed. So it can be concluded that nitrogen assimilation is reduced at elevated CO_2, which is in conjunction with the previous study [Sherwood, 2001].

Genes encoding meiosis protein Mei2 was found to be under-expressed (Fig 4.31). Four genes coding for RNA recognition motifs (RRM) (Mei2 has three putative RNA-recognition motifs) were also found to belong to cluster of strongly under-expressed genes obtained from k-means analysis. As Mei2 protein is an essential component of the switch from mitotic to meiotic growth, under-production of this protein implies that cells had prolonged mitotic growth. So the plants stayed in the vegetative growth phase for a longer period.

The enzyme UDP-glucose dehydrogenase, catalyzing the reaction from UDP-glucose to UDP-glucuronate is also overproduced (Fig 4.24). So the net flux in the pathway from starch to UDP-glucuronate is possibly increasing. UDP-glucuronate can produce trehalose 6-phosphate, pectin, sucrose or Amylose. All the enzymes catalyzing these reactions, except the enzyme catalyzing the reaction from UDP-glucuronate to pectin, are under-produced. So the flux through those reactions is expected to decrease. But the flux to UDP-glucuronate from UDP-glucose is increasing. Assuming that there is not much accumulation of UDP-glucuronate, it is speculated that the reaction from UDP-glucuronate to pectin will increase. Though it was not possible to verify this fact as the gene encoding this enzyme is missing. However pectin to pectate formation, the next reaction of that pathway is increasing as pectate methylesterase is over-produced (Fig 4.24). Pectin also has to be produced at a higher rate so that it can be converted to pectate. Which supports the hypothesis that, the flux from UDP-glucuronate to pectin is increasing. Pectin molecules are synthesized in the Golgi complex and secreted at the cell surface, where they cross-link the cellulose microfibrils into the matrix of the cell wall [Lodish et al.].

Hexokinase converts Hexose like D-Glucose, D-mannose, D-fructose, sorbitol to their corresponding phosphate. Fructokinase converts fructose to fructose-6-phosphate. Genes encoding both the enzymes were found to be under-expressed (Fig 4.32).

6. Future Work:

6.1 Integration of genomic and metabolomic data:

With the advent of the DNA microarray technology, it became possible to study the expression of entire cellular genomes. Tanscriptional profiling alone can not provide a comprehensive picture of the cellular physiological state and it should be complemented by other cellular fingerprints. One of the essential aspects of systems biology is integration of different fingerprints of cellular responses. In this project genomic and metabolomic profiling of a systematically perturbed system was studied. Even though these studies were conducted independently, but essentially they reflect the physiological condition and response of the same system.

6.1.1 Clustering Gene Expression and Metabolite Concentration Profiles:

Similarity in expression profile of a gene and concentration profile of a metabolite could be due to functional relation between the gene and the metabolite. Two examples were shown in Fig 6.1, where 3-phosphoglycerate mutase catalyses the reversible reaction from glycerate3-phosphate to glycerate-2-phosphate and Glucose-6P isomerase catalyses the reversible reaction from glucose-6P to fructose-6P.

Transcription and metabolic profiles can be clustered together using TIGR MeV software to find the profiles that have similar pattern. Magnitude change in metabolite concentration due to elevated CO_2 is larger [Kanani et al., 2004] than that of gene expression change. When metabolites and genes are clustered together using Euclidian distance, metabolites form a separate cluster. Use of Pearson correlation distance will cluster on the basis of expression profiles rather than the absolute expression value, which will find the genes and metabolites that have similar pattern.

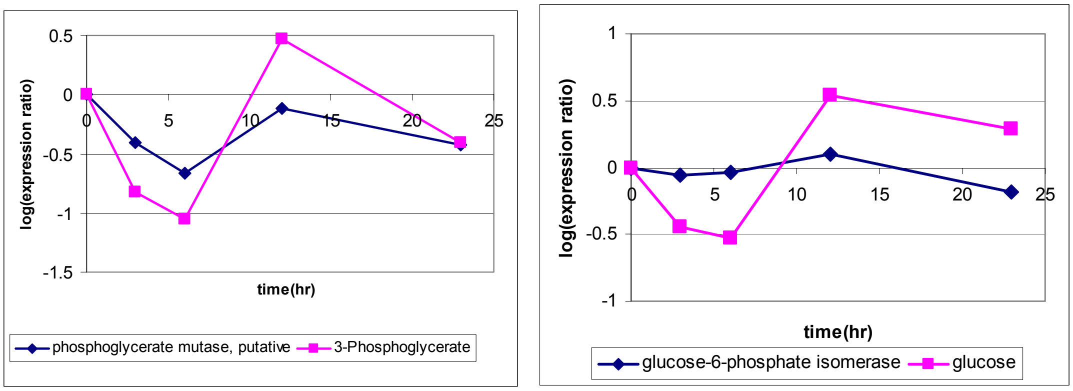


Figure 6.1 Transcription and metabolic profiles of genes and metabolites that are functionally related show similarity.

6.1.2 Model to Correlate Genomic and Metabolomic Data:

The experiment generated time series data of gene expression and metabolite concentration. Expression of genes control concentration of metabolites in the cell, metabolite concentration can also regulate gene expression. As the functional relationship between genes and metabolites is not completely understood, till date no mathematical model was proposed to correlate them. The wealth of data generated by DNA microarray and GC-MS can be used for developing a mathematical model. Gene expression and metabolic data can be viewed as input and output variables of a systematically perturbed system, where the functional relationship between the variables is not completely known. Partial Least Square (PLS) can be used for correlating gene and metabolic data.

6.2 Proposed Modifications in Experimental Design:

Following modifications are suggested for future experiments:

6.2.1 Increased number of biological replicates:

For the current experiment two biological replicates were used at each time point except time 0. It was observed quite often that expressions of genes in two different biological replicates are considerably different. As there were only two experiments it can not be concluded, which of the replicates is an outlier. At least three biological replicates should be considered for analysis, as the third replicate can help to conclude about the outlier.

6.2.2 Time point of sample collection:

In this project plants were not harvested at a constant time interval. At the initial period they were harvested at short time interval or 0.5hr and at longer period they were harvested in 12hr difference. Last two time points were 12 and 23 hrs, there should have at least one time point between them. If the plants are harvested at constant time interval modeling of the time series gene expression data will be easier.

6.3 Multiple Perturbations:

In this experiment sucrose was growth media of the plants. The same experiment can be conducted but with glucose or galactose as growth media in stead of sucrose. The response of the plant to CO_2 stress grown at different media can be compared at genomic level. This will provide a better understanding of gene regulation.

References:

Andrews, J. T. & Lorimer G. H. (1987) Rubisco: structure, mechanisms and prospects for improvement. Biochemistry of Plants Vol. 10, pp 132-207. Academic Press, Newyork

Arioli, T., Peng, L.C., Betzner, A.S., Burn, J., Wittke, W., Herth, W., Camilleri, C.H., Plazinski, J., Birch, R., Cork, A., Glover, J., Redmond, J. and Williamson, R.E. (1998) Science 279: 717–720

Bowes G (1991) Growth at elevated CO2: Photosynthetic responses mediated through Rubisco, Plant Cell Envisonment 14: 795-806

Brummell, D.A., Catala, C., Lashbrook, C.C. and Bennett, A.B. (1997) Proc. Natl. Acad. Sci. USA 94: 4794–4799,

Nicol, F., His, I., Jauneau, A., Vernhettes, S., Canut, H. and Höfte, H. (1998) EMBO J. 17: 5563–5576

Chen CT, Setter TL. Response of potato tuber cell division and growth to shade and elevated CO2, Ann Bot (Lond). 2003 Feb;91(3):373-81

Cheng S., Moore B., Seemann J. Effects of short and long term elevated CO2 on the expression of Ribulose-1,5-Bisphosphate Carboxylase/Oxygenase genes and carbohydrate accumulation in leaves of *Arabidopsis thaliana*, 1998, 116: 715-723

Cho R J. et al. A Genome wide transcriptional analysis of the mitotic cell cycle. Mol. Cell 2, 65-73 (1998)

Cleveland, W.S. Robust locally weighted regression and smoothing scatterplots.
J. Amer. Stat. Assoc. **74**, 829–836 (1979).

del Campillo E. Multiple endo-1,4-beta-D-glucanase (cellulase) genes in Arabidopsis Curr Top Dev Biol. 1999;46:39-61

Datta Susmita, Datta Somnath, Comparisons and validation of statistical clustering techniques for microarray gene expression data. Bioinformatics, 459-466, Vol 19, No 4, 2003

Delp G, Palva ET, A novel flower-specific Arabidopsis gene related to both pathogen-induced and developmentally regulated plant beta-1,3-glucanase genes, Plant Mol Biol. 1999 Feb;39(3):565-75

Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA* **95**, 14863–14868 (1998).

Fry, S.C., Smith, R.C., Renwick, K.F., Martin, D.J., Hodge, S.K. and Matthews, K.J. (1992) Biochem. J. 282: 821–828

Fulton M and Cobbett C., Two a-L-arabinofuranosidase genes in *Arabidopsis thaliana* are differentially expressed during vegetative growth and flower development*

Golovkin Maxim and Anireddy S. N. Reddy A calmodulin-binding protein from *Arabidopsis* has an essential role in pollen germination Proc Natl Acad Sci U S A. 2003 September 2; 100 (18): 10558–10563

Harmer SL, Hogenesch JB, Straume M, Chang HS, Han B, Zhu T, Wang X, Kreps JA, Kay SA: Orchestrated transcription of a key pathway in Arabidopsis by the circadian clock. Science 2000, 290:2110-2113

Heyer, L. J., Kruglyak, L. & Yooseph, S. Exploring expression data: identification and analysis of coexpressed genes. *Genome Res.* **9**, 1106–1115 (1999).

Hirayama T, Chika Ishida, Takashi Kuromori, Shusei Obata, Chikashi Shimoda, Masayuki Yamamoto, Kazuo Shinozaki and Chikara Ohto, Functional cloning of a cDNA encoding Mei2-like protein from *Arabidopsis thaliana* using a fission yeast pheromone receptor deficient mutant, FEBS Letters Volume 413, Issue 1 , 11 August 1997, Pages 16-20

Jeremy Berg, John Tymoczko and Lubert Stryer, Biochemistry, 5th edition.

Kazuhiko Nishitani, New Directions to Post-genomic Cell Wall Research, Plant Cell Physiol. 43(12): 1397–1397 (2002)

Keegstra, K., Talmadge, K.T., Bauer, W.D. and Albersheim, P. (1973) Plant Physiol. 51: 188–196

Kim H Erik C S, Hass B. Cheung F. Town C., Quackenbush J (2003) Gene Expression Analyses of Arabidopsis Chromosome 2 Using a Genomic DNA Amplicon Microarray, Genome Research

Kinsman EA, Lewis C, Davies MS, Young JE, Francis D, Vilhar B, Ougham HJ (1997). Elevated CO_2 stimulates cells to divide in grass meristems: a differential effect in two natural populations of *Dactylis glomerata*. Plant Cell Environ **20**: 1309−1316

Lodish, Harvey; Berk, Arnold; Zipursky, S. Lawrence; Matsudaira, Paul; Baltimore, David; Darnell, James E, Molecular Cell Biology**.** 4th ed.

Masle Josette, The Effects of Elevated CO_2 Concentrations on Cell Division Rates, Growth Patterns, and Blade Anatomy in Young Wheat Plants Are Modulated by Factors Related to Leaf Position, Vernalization, and Genotype, Plant Physiol. 2000 April; 122 (4): 1399–1416

Meyer S L, Data Analysis for scientists and engineers, John Wiley and Sons, 1975

McQueen-Mason, S., Durachko, D.M. and Cosgrove, D.J. (1992) Plant Cell 4: 1425–1433

Michèle Rouleau, Frédéric Marsolais, Martine Richard, Ludovic Nicolle, Brunhilde Voigt, Günter Adam, and Luc Varin Inactivation of Brassinosteroid Biological Activity by a Salicylate-inducible Steroid Sulfotransferase from Brassica napus J Biol Chem, Vol. 274, Issue 30, 20925-20930, July 23, 1999

Nishitani, K. and Tominaga, R. (1992) J. Biol. Chem. 267: 21058–21064

Okazawa, K., Sato, Y., Nakagawa, T., Asada, K., Kato, I., Tomita, E. and Nishitani, K. (1993) J. Biol. Chem. 268: 25364–25368

Onouchi, H., Igeno, M.I., Perilleux, C., Graves, K. and Coupland, G. (2000) Mutagenesis of plants over-expressing CONSTANS demonstrates novel interactions among Arabidopsis flowering-time genes. Plant Cell, 12, 885-900.

Pear, J.R., Kawagoe, Y., Schreckengost, W.E., Delmer, D.P. and Stalker, D.M. (1996) Proc. Natl. Acad. Sci. USA 93: 12637–12642. Rose, J.K.C., Braam, J., Fry, S.C. and Nishitani, K. (2002) Plant Cell Physiol

Quackenbush J, 2001, Computational analysis of microarray data, Nature Genetics, Vol 2, 418-427

Quackenbush, John Microarray data normalization and transformation. Nature Genetics, Dec2002 Supplement 2, Vol. 32 Issue 4, p496,

Robson Frances, M. Manuela R. Costa, Shelley R. Hepworth, Igor Vizir, Manuel Pinƒ eiro, Paul H. Reeves1, Joanna Putterill, and George Coupland Functional importance of conserved domains in the flowering-time gene CONSTANS demonstrated by analysis of mutant alleles and transgenic plants The Plant Journal (2001) 28(6), 619- 631

Raychaudhuri, S., Stuart, J. M. & Altman, R. B. Principal components analysis to summarize microarray experiments: application to sporulation time series. *Pac. Symp. Biocomput.* **2000**, 455–466 (2000).

Saeed AI, Sharov V, White J, Li J, Liang W, Bhagabati N, Braisted J, Klapa M, Currier T, Thiagarajan M, Sturn A, Snuffin M, Rezantsev A, Popov D, Ryltsov A, Kostukovich E, Borisovsky I, Liu Z, Vinsavich A, Trush V, Quackenbush J. TM4: a free, open-source system for microarray data management and analysis

Schaffer R, Landgraf J, Accerbi M, Simon V, Larson M, Wisman E: Microarray analysis of diurnal and circadian-regulated genes in *Arabidopsis*. *Plant Cell* 2001, 13:113-123

Shcherban, T.Y., Shi, J., Durachko, D.M., Guiltinan, M.J., McQueen-Mason,S.J., Shieh, M. and Cosgrove, D.J. (1995) Proc. Natl. Acad. Sci. USA 92:9245–9249

Sheen J, Metabolic Repression of transcription in higher plants (1990) Plant Cell 2 : 10271038

Sheen J, Feedback control of gene expression. Photosynth Res 1994 39: 427-438

Stitt M. Rising CO_2 levels and their potential significance for carbon flow in photosynthetic cells, Plant Cell and Environment (1991) 14: 741-762

Sun N., Ma L., Pan D., Zhao H. Deng X. W. (2003) Evaluation of light regulatory potential of Calvin cycle steps based on large scale gene expression profiling data. Plant Molecular Biology 53: 467-478

Tamayo, P. *et al.* Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc. Natl Acad. Sci. USA* **96**, 2907–2912 (1999).

Tavazoie, S., Hughes, J. D., Campbell, M. J., Cho, R. J. & Church, G. M. Systematic determination of genetic network architecture. *Nature Genet.* **22**, 281–285 (1999).

Taylor, N.G., Scheible, W.R., Cutler, S., Somerville, C.R. and Turner, S.R. (1999) Plant Cell. 11: 769–779

Tepperman JM, Zhu T, Chang HS, Wang X, Quail PH: Multiple transcription-factor genes are early targets of phytochrome A signaling**.** *Proc Natl Acad Sci USA* 2001, **98:**9437-9442

Thum K., Shin M., Palenchar P., Kouranov A., Coruzzi G. (2004) Genome wide investigation of light and carbon signaling interactions in Arabidopsis, Genome Biology: 2004, 5:R10

Tusher Virginia Goss, Tibshirani Robert, and Chu Gilbert, Significance analysis of microarrays applied to the ionizing radiation response, PNAS, April 24, 2001**,** vol. 98, no. 9

Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. Serial analysis of gene expression. Science 270, 484–487 (1995).

Vivian G. Cheung, Michael Morley, Francisco Aguilar, Aldo Massimi, Raju Kucherlapati2 & Geoffrey Childs, Making and reading microarrys. Nature Genetics, Vol 21, 1999. 15-19

Wang R, Guegler K, LaBrie ST, Crawford NM: Genomic evidence sion patterns and novel metabolic and potential regulatory a nutrient response in *Arabidopsis* reveals diverse expres-genes induced by nitrate**.** *Plant Cell* 2000, 12**:**1491-1509

Wang YH, Garvin DF, Kochian LV: Rapid induction of regulatory and transporter genes in response to phosphorus, potassium, and iron deficiencies in tomato roots. Evidence for cross talk and root/rhizosphere-mediated signals. *Plant Physiol* 2002, 130:1361-1370

Webber N. Andrew, Nie G and Long S. P. (1994) Acclimation of photosynthetic proteins to rising atmospheric CO_2 , Photosynthesis Research 39: 413-425

Yang T, Poovaiah BW Calcium/calmodulin-mediated signal network in plants Trends Plant Sci. 2003 Oct;8(10):505-12

Yang, Y.H. et al. Normalization for cDNA microarray data: a robust composite method addressing single and multiple slide systematic variation. Nucleic Acids Res. 30, e15 (2002).

Yang, I.V. et al. Within the fold: assessing differential expression measures and reproducibility in microarray assays. Genome Biol. 3, research0062.1–0062.12 (2002).

Yang, Y.H. et al. Normalization for cDNA microarray data: a robust composite
method addressing single and multiple slide systematic variation. Nucleic Acids Res. 30, e15 (2002).