GENOME-WIDE PATHWAY ANALYSIS AND VISUALIZATION USING GENE EXPRESSION DATA

M. P. KURHEKAR, S. ADAK
Bioinformatics Group,
IBM India Research Lab, Block 1, Indian Institute of Technology, Hauz Khas,
New Delhi, India 110016
{mkurhekar, asudeshn} @in.ibm.com

S. JHUNJHUNWALA
Department of Biochemical Engineering and Biotechnology, Indian Institute of Technology,
Hauz Khas, New Delhi, India 110016

K. RAGHUPATHY
Department of Electrical Engineering, Indian Institute of Technology,
Chennai, India

Visualization of results for high performance computing pose special problems due to the complexity and the volume of data these systems manipulate. We present an approach for visualization of c-DNA microarray gene expression data in metabolic and regulatory pathways using multi-resolution animation at different levels of detail. We describe three scoring functions to characterize pathways at the transcriptional level based on gene expression, coregulation and cascade effect. We also assess the significance of each pathway score on the basis of their biological connotation.

1 Introduction

1.1. Microarrays and Gene Expression

DNA and other microarray technologies are on their way to becoming standard tools of modern life sciences research[6]. A DNA microarray experiment shows the expression levels of thousands of genes at a single time point. This has given scientists the ability to get a snapshot view of genome-wide expression patterns. The vast quantity of data generated by genomic expression arrays affords researchers a significant opportunity to transform biology, medicine, and pharmacology using systematic computational methods. The availability of genomic (and eventually proteomic) expression data promises to have a profound impact on the understanding of basic cellular processes, the diagnosis and treatment of disease, and the efficacy of designing and delivering targeted therapeutics. Particularly relevant to these objectives is the development of a deeper understanding of the various mechanisms by which cells control and regulate the transcription of their genes.

1.2. Pathways

Living organisms behave as complex systems that are flexible and adaptive to their surroundings. At the cellular level, organisms function through intricate networks of chemical reactions and interacting molecules. These networks or biochemical pathways may be considered as the wiring diagrams[12] for the complete biological system of an organism. The best characterized among them are metabolic pathways, the biological networks that involve enzymatic reactions of chemical compounds. Regulatory pathways are another class of pathways that represent protein-protein interactions. Pathways are the key to understanding how an organism reacts to perturbations in its environment (e.g. heat shock, chemical or hormone stimulus) or internal changes (e.g. disease, development, etc.)

1.3. Pathways and Gene Expression

With the advent of microarrays, it is hoped that knowledge of genome-wide expression levels will speed up the understanding of biological systems. Microarrays can serve as an efficient tool to study the expression levels of genes involved in pathways. The effect of induction and repression of target genes on metabolic and regulatory pathways is particularly important in drug development. In this paper, we present a method for using genomic expression data to elucidate and visualize the effect of different stimuli on these genetic networks.

Typical automatic analysis of microarray expression data is performed by clustering the expression profiles: using pair-wise measures such as correlation[1,7] and mutual information[2]; using more multivariate methods like principal components[18] and Fourier analysis[20]. Clustering methods are based on the microarray expression data and subsequent efforts are made to correlate clusters with pathways[22]. Several authors have suggested methods for synthesizing pathways using gene expression data: linear models[5], Boolean networks[16] and Bayesian networks[9]. However, it is difficult to evaluate such reversed engineered pathways in view of known metabolic and regulatory pathways. This has led to efforts being made to map reconstructed pathways onto known pathways[8, 22].

In this paper, we present a method for scoring of putative pathways. The scores are defined to measure the impact of gene expression levels from a series of microarray experiments on metabolic and regulatory pathways. We also present an animated visualization technique that allows the user to observe the complex changes that occur in pathways as tracked by the changing expression levels in a series of microarray experiments. All methods have been implemented in a stand-alone JAVA application.

2 Methods

2.1 Expression Data

In this paper, we consider series of microarray experiments, which measure genome-wide expression levels, as observed over time or over increasing levels of different stimuli like temperature, radiation, drug dosage etc. Well-known examples of such series of microarray experiments include:

- *Yeast* – diauxic shift microarray data[4], yeast sporulation data[3], yeast response to various environmental changes[7], yeast cell cycle data[20]
- *E.coli* – heat shock microarray time series[19]
- *Human* – response of human fibroblasts to serum[11]

For each microarray experiment series, let G be the set of genes investigated in a series of T experiments. For each gene $g\chi\ G$, we regard the expression data as a mapping from g to an ordered series of numbers, $X_{t,g}$ ($t=1,2,\ldots,T$). In this, $X_{t,g}$ denotes the expression ratio of the gene g in the t-th microarray experiment of the series. The expression ratio is calculated as:

$$X_{t,g} = L_{t,g} / L_{0,g} \quad (1)$$

where $L_{t,g}$ is the actual expression level for the gene g in the t-th microarray experiment and $L_{0,g}$ is the expression level of the gene g in the reference sample. That means $L_{0,g}$ is the unperturbed case.

2.2 KEGG: The Pathway Database

The KEGG (Kyoto Encyclopedia of Genes and Genomes) database provides a catalog of metabolic and regulatory pathways that may be considered wiring-diagrams of genes and molecules[18]. In addition, it provides up-to-date links from the gene catalogs generated by genome sequencing projects. More than diagrams, the KEGG database also provides direct links from the genes to the gene products (enzymes and other proteins) involved in the biochemical pathways. This feature of KEGG is particularly useful in mapping gene expression data to known metabolic and regulatory pathways.

In the case of a series of microarray experiments, visualizing the course of a pathway is highly informative and essential in understanding how the pathway is affected over the sequence of experiments. While visualization is essential to understanding each individual pathway, it is also necessary to provide the user some indication of the relative importance of the more than 100 different pathways in the KEGG database. In this paper, we describe three kinds of pathway scores, which are based on “activity”, “coregulation” and “cascade” effects in pathways.

Method Outline

The methods described in this paper allow scoring and visualization of the putative pathways in the KEGG database according to the gene expression levels in a microarray experiment series. The method can be summarized as follows:

- Given the input
  - Gene expression data from a microarray experiment series
  - Putative pathways of the KEGG database
- Answer the questions
  - Which pathways are most affected during the course of the experiments?
  - What is the nature of the effect? (Details such as which genes in a pathway are most affected, are the genes over-expressed or under-expressed, which reactions are disrupted etc.)
- By providing the output
  - Pathway scores – these quantify "activity", "coregulation", and "cascade" effects in pathways as measured by the gene expression levels from the microarray experimental data.
  - Pathway animated view – these show the effects on individual pathways over the course of a microarray experiment series.

3 Pathway Scoring

The pathway scoring methods described below measure the changes in metabolic and regulatory pathways as indicated by genome-wide gene expression levels. A high level of gene expression indicates that the cell required the particular protein coded by the gene and hence the expression of the gene has been induced. Thus, significant induction in the genes of a pathway shows that the pathway is being used more extensively than at the reference time point. Similarly, significant repression in the genes involved in a pathway shows that the pathway has been de-activated. By measuring the gene expression through a series of microarray experiments, it is thus possible to measure the effect on biochemical pathways as the cell is subjected to different stimuli. In this paper, we describe three kinds of pathway scores which progressively try to capture the complexity of biochemical pathways in living cells:

- The *Activity score* for a pathway gives a summary measure of the extent to which a pathway is perturbed from the reference state. This score will rank those pathways higher in which more genes were over-expressed or under-expressed with reference to reference state.
- The *Coregulation score* gives an indication of co-expression of the genes in a pathway under the given experimental conditions. It assigns higher scores to pathways whose genes show similar patterns of expression.

- The *Cascade score* takes into account the structure of a pathway as well as measuring *activity* and *coregulation*. It gives a measure of the extent to which a metabolic pathway is affected by analyzing the microarray data along reaction chains. If the first enzyme in a series of reactions is, say, over produced, this should be accompanied by an increase in production of the subsequent enzymes in the reaction chain. A high score is given to such over-expressed or under-expressed chains of reactions.

Since it is important to assess the relative importance of pathways rather than the absolute scores, each type of score is further normalized on a scale of 0-100 as follows: Relative Score = [Score/Max. Score]x100, where Max. Score is the maximum score (of the same type) among all putative pathways in KEGG.

Another important normalization required for the scoring functions is based on the number of enzymes in a pathway for a given organism. For example, in case of a pathway like prostaglandin and leukotriene metabolism is probably defunct in yeast as only two of the enzymes in this pathway are known to be present in yeast and they are entirely disconnected. Such defunct pathways should be differentiated from valid pathways while scoring, and their score should not be given importance. For a metabolic pathway P, the "validity factor normalization" with respect to the organism under investigation is defined as follows:

$$VF_{org}(P) = 1, \text{ if } P_{org}/P_{ref} >= 0.3 \qquad (2)$$
$$= P_{org}/P_{ref}, \text{ if } P_{org}/P_{ref} < 0.3, \quad \text{where}$$

P_{org} : number of enzymatic reactions in the organism specific version of P

P_{ref} : number of enzymatic reactions present in the reference version of P as provided by the KEGG database, it is the unperturbed case.

(Enzymatic reactions are uniquely identified by the substrate-product-enzyme combination). Thus, if only a few enzymes in a particular metabolic pathway are known to exist in an organism, the pathway will be given a low score by discounting the original score by the "validity factor". The threshold of 0.3 was used for P_{org}/P_{ref} in defining the validity factor, as it was found to be empirically suitable.

3.1 Activity Score

Consider a pathway P and let the set of genes involved in the pathway be denoted by G_P. The activity score for the pathway P with respect to a user-defined threshold $\hbar$ is defined as follows using (1) and (2):

$$\text{Activity Score}(P,\hbar) = \text{VF} \times \sum_{g \in G_p} \sum_{t=1}^{T} I(g,t), \text{ where} \qquad (3)$$

$$I(g,t) = 1 \text{ if } X_{t,g} > \hbar_1 \text{ or } X_{t,g} < 1/\hbar_2; \text{ 0 otherwise.}$$

Thus, according to activity score, pathways will be scored higher if there are more genes that are over-expressed above a given threshold value $\hbar_1$ or under-expressed below a given threshold value $1/\hbar_2$. The thresholds represent the minimum fold-deviation, with respect to the reference sample, that is considered meaningful. Similar activity scores have been previously used[21] where the threshold is determined based on the data. However, due to inherent noise in the experimental data, it is difficult to validate and interpret the resulting scores when the threshold is data dependent.

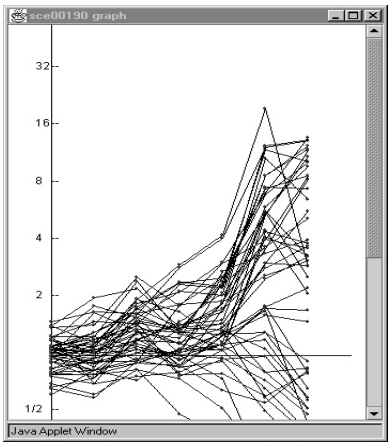

Fig 1a. Time progress of the Oxidative Phosphorylation pathway showing high activity is verified by the activity score. (Diauxic shift)

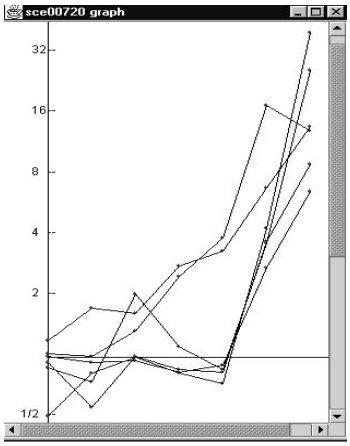

Fig 1b. Time progress of CO_2 fixation pathway shows high coregulation is verified by the coregulation score. (Diauxic Shift)

3.2 Coregulation Score

The activity score merely considers the number of genes that are over-expressed or under-expressed in a pathway, but it does not capture similarity in expression patterns among the genes of a pathway. Coregulation score ranks those pathways higher in which genes show greater similarity in their expression pattern.

Coregulation scores have been previously used[21] which were based on the averages of all pairwise correlations among all genes in a pathway. However, the coregulation score as defined in [21] did not perform very well with experimental data. This is probably because pairwise correlations fail to capture the simultaneous co-expression of all genes in a pathway. We define a slope coregulation score that captures simultaneous coregulation by looking at the variation in the "slopes" among all genes in a pathway.

Consider a pathway P and let the set of genes involved in the pathway be denoted by G_P and $N_P = |G_P|$. The slope coregulation score for a pathway P is defined as follows using (1) and (2):

$$\text{Slope Coregulation Score}(P) = \text{VF} \times \sum_{t=1}^{T} \text{SlopeScore}_{\text{t,p}} \text{ , where}$$

$$\text{SlopeScore}_{\text{t,P}} = N_p \Big/ \left[\sum_{g \in G_p} \left(\text{Trend}(g,t) - \text{Mean}[\text{Trend}(g,t)]\right)^2 \right],$$

$$\text{where Trend}(g,t) = \left(\log_2\left(X_{t,g}\right) - \log_2\left(X_{t\text{-}1,g}\right)\right) \Big/ \left(A_t - A_{t-1}\right) \text{ and}$$

$A_t - A_{t\text{-}1}$ represents the fold change in the experimental condition for the microarray experiment series. For example, in case of microarray time series data, this represents the time lapse between experiments t and t-1; in case of microarray experiments performed over increasing temperature levels, this represents the change in temperature between experiments t and t-1, etc.

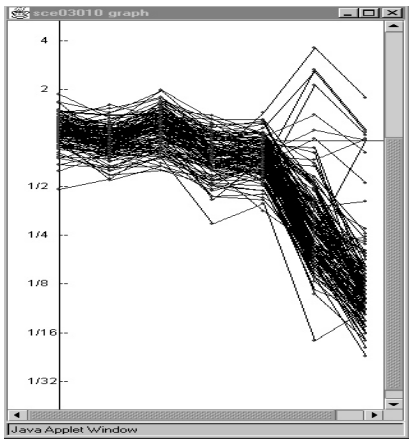

Fig 2a. Time progress of Ribosomal proteins which have a coregulation score of 100 - highest among all pathways. (Diauxic shift).

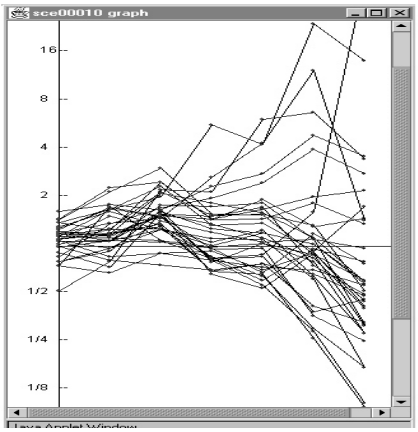

Fig2b. Time progress of the glycolysis pathway which have a coregulation score of 25 - highest among metabolic pathways (Diauxic shift).

3.3 Cascade Score

Using coregulation scores, pathways in which the genes have similar expression patterns will be ranked higher. However, this does not account for a) genes whose expression levels do not show much deviation from the expression level at the reference time point; b) the structure of the pathway. The first problem can be

resolved by combining activity and coregulation scores, as was done in a combined scoring function described in [21]. The combined scoring function still does not take into account the structure and ordering of reactions in metabolic pathways.

Here, we define a cascade score that accounts for both these features in scoring pathways. This method of scoring is particularly useful to find out in which pathway a reaction chain is active or shutdown for the particular experiment. The cascade score is valid only for metabolic pathways, as it requires a network of linked reactions. It is computed as follows:

- Step 1: Create a list of all enzymatic reactions in the pathway. Discard all reactions in which the gene (corresponding to the enzyme) does not show any significant fold-deviations from the reference expression level.
- Step 2: Form all possible reaction chains (paths in a graph). Score each chain based on the coregulation of enzyme pairs as they occur in the chain. The edge weight *W* for an edge between two genes *h* and *g* is given by:

  $W(g \rightarrow h) = \#(h=1|g=1) / \#(g=1)$, where $g=1$ means *g* is active.
- Step 3: Find the chain with the highest score – assign the score for the chain as the cascade score for the pathway.

Details of the method for calculating cascade score are given in [15].

The table below gives the results for the three scoring methods for different yeast microarray experiment series.

Table 1. Top pathways based on scoring functions :

Microarray Experiment Series	Activity Score		Coregulation Score		Cascade Score
	Regulatory	Metabolic	Regulatory	Metabolic	
Diauxic Shift[4]	Ribosome	Oxidative Phosphoryl-ation	Electron Transport System - II	Reductive Carboxylate cycle	TCA cycle
Alpha Factor	Cell cycle	Riboflavin	Cell cycle	Riboflavin	Purine Metabolism
Elutriation	Cell cycle	Pentose Phosphate	Cell cycle	Porphyrin and chrolophyll metabolism	Glyoxylate and dicarbonate metabolism
Sporulation	Ribosome	Oxidative Phosphoryl-ation	Proteasome	Terpenoid Biosynthesis	Vitamin metabolism
Heat Shock	Ribosome	Purine Metabolism	Proteasome	Galactose	Fatty acid biosynthesis

The above table gives an idea of how the different analysis can give different results and how the scores point the affected pathways that are related with the experiment data. Consider the diauxic shift experiment, in [4] it is shown to be

related with ribosome and TCA cycle pathways. Diauxic shift is known to be related with Oxidative Phosphorylation and Electron Transport System-II. Co-regulation of the Electron Transport System-II becomes evident due to significant activity of the Oxidative Phosphorylation pathway. Figure 1b clearly shows reductive carboxylate (CO_2 fixation) pathway getting activated during diauxic shift response.

4 Multiresolution, Animated Visualization of Pathways

As mentioned earlier, KEGG contains information on a large number of putative pathways. The pathway scores are useful in directing the user to the "right" pathway in the context of a microarray experiment series. However, visualization of the pathways is necessary to show a user the details of pathway effects as measured by changes in gene expression levels in response to stimuli. The visualization technique of [14] requires a single absolute level for each member of the set of genes, which is a severe limitation. Also identification of affected pathways based on color rather than numerical value is prone to errors. Our technique removes these inadequacies.

4.1 Pointing the User in the Right Direction – Multiresolution Viewing

The metabolic pathways in KEGG are classified hierarchically at three levels of detail i.e. three resolutions. Resolution 1 is a coarse grained representation of the complete network of metabolic pathways. Resolution 2 provides medium grade resolution is terms of functionality like carbohydrate metabolism, nucleotide metabolism, etc. and contains pathways related to that function. The finest resolution is resolution 3, which shows the reaction network as well as the compounds and the enzymes involved. KEGG also organizes its regulatory pathways into groups at resolution 2 based on broad functionality.

The (activity, coregulation, cascade) score for a pathway group at the resolution 2 level is simply the average of the corresponding scores of the pathways belonging to that group. The scores at the resolution 2 level are normalized on a scale of 0-100, with the highest pathway being given a score of 100.

The user is directed using the relevant summary pathway scores at each of the coarser resolutions (resolution 1 and 2). Clickable maps allow the user to navigate easily through the pathways. Example of this multiresolution view is shown in figure 3 and figure 4. Figure 3 shows the resolution 2 view for energy metabolism that had the highest activity score among all pathway groups in resolution 1.

4.2 Directing the User to Impact - Animated Visualization

At any resolution 3 pathway, the user is presented with several choices for viewing the expression data for all the genes involved in the pathway. One such choice is an animated view:

- For a single microarray experiment, the organism specific pathway map from KEGG is colored. The enzyme boxes are colored based on their expression level (red indicating induction and blue indicating repression).
- For a microarray experiment series, the user can use a "next" button to view the experiments in sequence, allowing a visual monitoring of the pathway changes.

Fig 4 shows the resolution 3 view for the citrate cycle pathway at the last time point of the diauxic shift microarray experiment series[4]. The user can use the previous and next buttons to observe this pathway at each time point.

5 Results and Discussion

While the potential utility of expression data is immense, some obstacles will need to be overcome before significant progress can be realized. First, data from expression arrays is inherently noisy. Second, gene expression is regulated in a complex and seemingly combinatorial manner. Third, our knowledge regarding genetic regulatory networks is extremely limited. Never the less, gene expression data from microarrays are very useful for understanding biochemical pathways, their progress with time and their response to experimental stimuli.

The scoring and visualization methods used here give a natural way for using genome-wide expression data in understanding how biological systems function. However, current methods can be improved if protein microarrays become widely available. This is because current DNA microarray technology measures mRNA expression levels, which are only an indication of the level of activity of the final protein, also the mapping from gene to proteins/enzymes is many to many, which can result in misleading scores. Directions for future work include analysis and visualization for microarray experimental data that corresponds to two or more classes[1, 21].

References

1. A.A. Alizadeh et al, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling" *Nature* **403**, 503 (2000)
2. A.J. Butte et al, "Mutual information relevance networks: Functional genomic clustering using pairwise entropy measurements" *PSB* (2000)
3. S. Chu et al, "The transcriptional program of sporulation in budding yeast" *Science* **282**, 699 (1998)

4. J.L. DeRisi et al, "Exploring the metabolic and genetic control of gene expression on a genomic scale" *Science* **278**, 680 (1997)
5. P. D'haeseleer et al, "Linear modeling of mRNA expression levels during CNS development and injury" *PSB* **4**, 41 (1999)
6. D.J. Duggan et al, "Expression profiling using cDNA microarrays". *Nature genetics supplement* **21**, 10 (1999)
7. M.B. Eisen et al. "Cluster analysis and display of genome wide expression patterns" *PNAS* **95**, 14863 (1998).
8. M. Fellenberg et al, "Interpreting clusters of gene expression profiles in term of metabolic pathways" *German Conf. On Bioinformatics* Poster (1999)
9. N. Friedman et al, "Using bayesian networks to analyze expression data" *Proc. RECOMB* 127 (2000)
10. Goto et al, "Organizing and computing metabolic pathway data in terms of binary relations" *PSB* (1997)
11. V.R. Iyer et al, "The transcriptional program in the response of human fibroblasts to serum" *Science* **283**, 83 (1999)
12. M. Kanehisa in *Bioinfomatics: Databases and Systems*, "KEGG: From genes to biochemical pathways" Ed. S Letovsky (1999)
13. P.D. Karp et al, "Integrated access to metabolic and genomic data" Journal of Computational Biology (1996)
14. P.D. Karp et al, "Integrated pathway/genome databases and their role in drug discovery" Trends in Biotechnology (1999)
15. M.P. Kurhekar et al, "Analysis of pathways using cascade scores" *Technical Report*, IBM India Research Lab. (Draft under submission)
16. S. Liang et al, "REVEAL: A general reverse engineering algorithm for inference of genetic networks" *PSB* **3**, 18 (1998)
17. Nakao et al, "Genome-scale Gene Expression Analysis and Pathway Reconstruction in KEGG" *Genome Informatics* **10**, 94 (1999)
18. Raychauduri et al, "Prinicipal component analysis to summarize microarray experiments: application to sporulation time series" *PSB* (2000)
19. Richmond et al, "Genome-wide expression profiling in Escherichia coli K-12" *Nucleic Acids Research* **27**, 3821 (1999)
20. P.T. Spellman et al, "Comprehensive identification of cell cycle regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization" *Mol. Biol. Cell* **9**, 3273 (1998)
21. P. Tamayo et al, "Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation" *PNAS* **96** 2907 (1999)
22. A. zien et al, "Analysis of gene expression data with pathway scores" *Proc. ISMB'00* (2000)

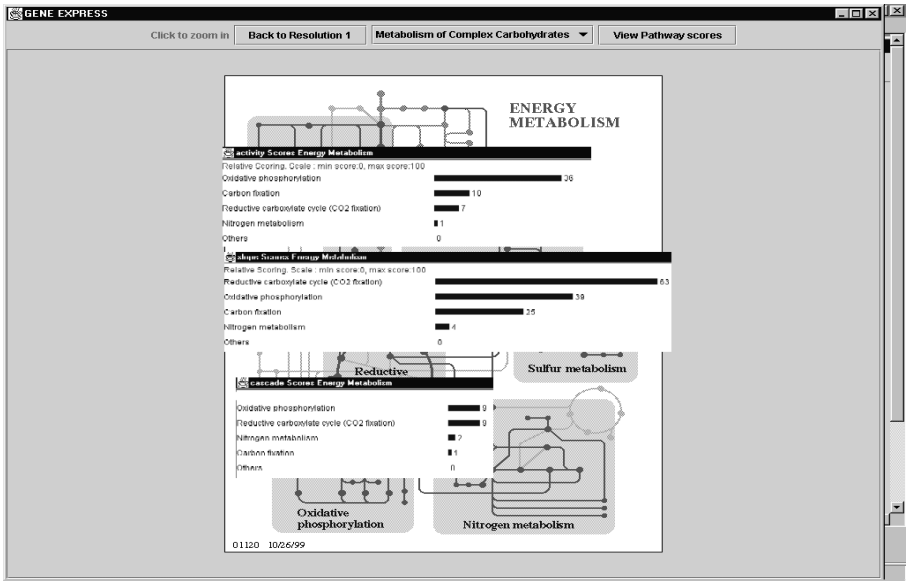

Fig 3. Second Resolution grouping of pathways belonging to Nucleotide Metabolism. The activity, co-regulation and cascade effect scores are also shown

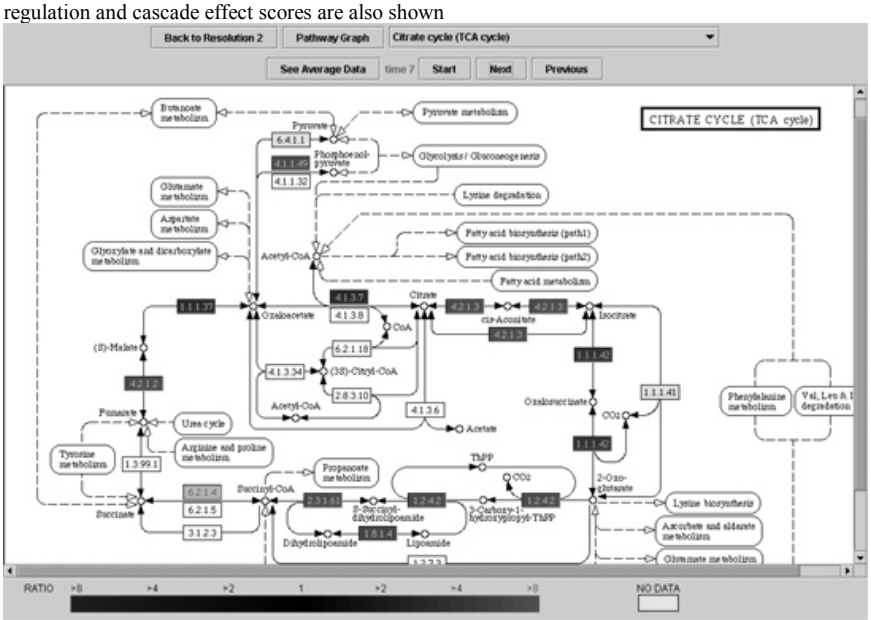

Fig 4. Finest Resolution (Resolution 3) for Citrate Cycle. The enzymes are colored according to their activity at time instant 7 of the Diauxic shift experiment [1].