Genome-scale Gene Expression Analysis and Pathway Reconstruction in KEGG

Mitsuteru Nakao [1]
nakao@kuicr.kyoto-u.ac.jp

Hidemasa Bono [1]
bono@kuicr.kyoto-u.ac.jp

Shuichi Kawashima [1]
shuichi@kuicr.kyoto-u.ac.jp

Tomomi Kamiya [1]
tomomi@scl.kyoto-u.ac.jp

Kazushige Sato [2]
kazs@scl.kyoto-u.ac.jp

Susumu Goto [1]
goto@kuicr.kyoto-u.ac.jp

Minoru Kanehisa [1]
kanehisa@kuicr.kyoto-u.ac.jp

[1] Institute for Chemical Research, Kyoto University, Uji, Kyoto, 611-0011, Japan
[2] SGI Japan, Ltd, Ebisu Shibuya-ku, Tokyo 150-6031, Japan

Abstract

The massively parallel hybridization technologies by DNA chips and microarrays make it possible to monitor expression patterns of the whole set of genes in a genome under various conditions. The vast amount of data generated by such technologies necessitates the development of a new database management system that integrates expression data with other molecular biology databases and various analysis tools. We report here an extension of our KEGG (Kyoto Encyclopedia of Genes and Genomes) and DBGET/LinkDB systems for analyzing gene expression data in conjunction with pathway information and genomic information. It is now possible to make use of expression data for the reconstruction of pathways from the complete genome sequences.

1 Introduction

The emerging technology of DNA chips and microarrays makes it possible to simultaneously analyze the expression of a number of genes, such as the whole set of genes in the completely sequenced genome [1, 2, 3, 4]. To effectively analyze the new type of data derived from the expression profiles we should integrate them with the functional data such as pathways and assemblies, as well as with the traditional molecular biology data such as nucleotide and amino acid sequences. It is also a high priority task to integrate them with the statistical analysis and visualization tools [5, 6].

We have been developing the KEGG (Kyoto Encyclopedia of Genes and Genomes) system [7, 8] for computer representation and utilization of functional data, which are related to networks of interacting molecules such as metabolic pathways, various regulatory pathways, and molecular assemblies. In addition, we maintain the DBGET/LinkDB system [9] which was originally developed for an integrated retrieval of a number of molecular biology databases. It is also the backbone database management system for KEGG. One of the main features of KEGG is a collection of pathway maps, which computerizes the network information of molecular interactions such as for metabolism and signal transduction. Another feature of KEGG is a collection of genome maps for completely sequenced organisms, as well as for the fruit fly, mouse, and human. Furthermore, KEGG provides many analysis tools, such as for searching similar sequences in the pathway maps, identifying gene clusters that are conserved in two genomes, and reconstructing pathways from a whole set of genes by a genome-wide homology search. Because KEGG and DBGET/LinkDB are tightly coupled, once the expression data is incorporated in the DBGET/LinkDB system it can be applied to various computational analysis tools in KEGG.

In this paper, we report an integration of gene expression data into the DBGET/LinkDB and KEGG systems and show how we can make use of the integrated system for analysis of expression data. The integration includes a visualization of genome-scale gene expression data not only by the

standard array view but also by the genome map and pathway map views. The analysis includes a metabolic pathway reconstruction by differential gene expression patterns obtained by comparison of the reference state and the perturbed state (e.g. the wild-type and a disruptant, or the control and an environmental shift).

2 System Design

2.1 KEGG/EXPRESSION Database

We created a new database for KEGG, called EXPRESSION for expression data, and integrated it in the DBGET/LinkDB system. The information stored in each entry is currently as follows:

- Information about the array experiment:
  - ID of an expression profile
  - Authors and references
  - Name of the organism
  - Conditions of the experiment including the time-course information if any
  - Information on the related experiments
  - Information on the array including spot coordinates and probes
- Information about each spot:
  - ID of the spot
  - Name of the gene associated with the spot
  - Information on the data normalization method
  - Foreground and background intensities for the spot from channel 1
  - Foreground and background intensities for the spot from channel 2
  - Ratio of the intensity of the channel 1 to the intensity of the channel 2

In the DBGET/LinkDB system, the EXPRESSION database is organized as a flat-file database which is a collection of entries containing just the text information mentioned above. Each entry starts with a line for the ENTRY field and ends with a line of triple slashes (///). The original information is stored and managed in a relational database and the flat-file version for DBGET are automatically generated from the relational database.

In addition to the numerical and text information of each experiment, we provide a graphical view of the array data by a Java applet (Fig. 1). This particular view is created by taking the ratios of the two channels, namely Cy5- and Cy3-labeled spots, which are useful for identifying clusters of co-expressed genes that are also clustered in the pathways or in the chromosomal locations by mapping to the KEGG/PATHWAY and KEGG/GENOME databases, respectively (Fig. 2). We are also working on another Java applet which handles the time-course information stored in the condition field, where we can see and analyze the changes of the expression levels for individual genes through a series of hybridization experiments.

2.2 Integration of KEGG/EXPRESSION into DBGET/LinkDB

Fig. 2 summarizes how the KEGG/EXPRESSION database is integrated in the DBGET/LinkDB system. The databases available from the KEGG project include PATHWAY for diagrams of metabolic and regulatory pathways, GENES for sequence and other information of genes for all the completely

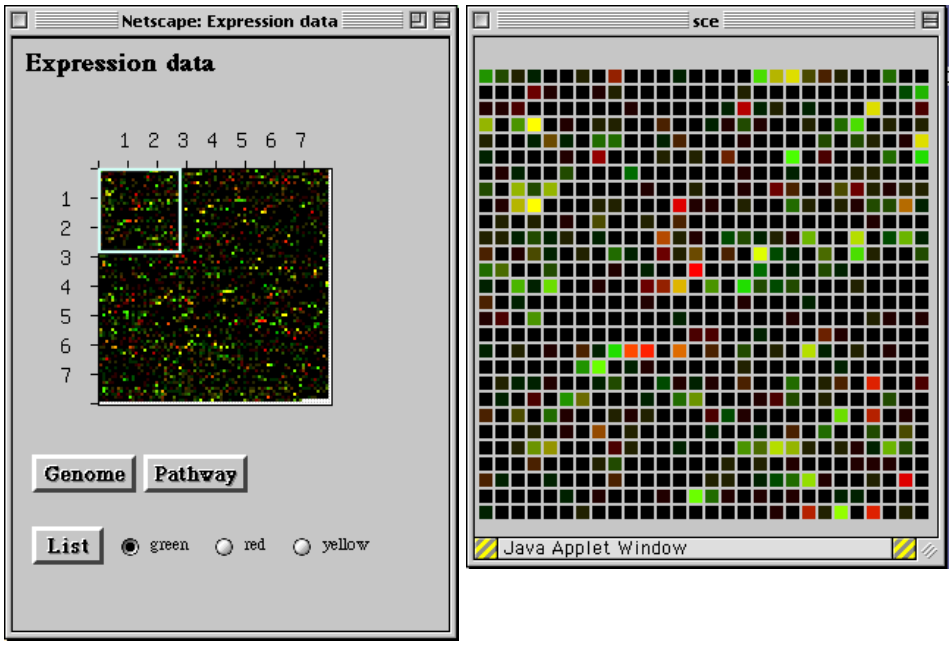

Figure 1: Graphical Java applet windows for handling array data stored in the EXPRESSION database: the overall view for the entire array (left) and an expanded view of the selected part (right).

sequenced organisms and some partially sequenced organisms, and GENOME for graphical manipulation of chromosomal information in the completely sequenced organisms as well as in the mouse and human. Because each spot of an array corresponds to a gene, there are direct links from the EXPRESSION database to the GENES database. The standard databases such as GenBank, SWISS-PROT, and LIGAND/ENZYME can be retrieved via the GENES database or directly using the LinkDB search.

Another type of integration is the linkage of the expression pattern with the pathway maps and genome maps (see below). Once the links are established, pathway maps and genome maps are colored according to the colors in the expression patterns. They can then be linked to the other databases for further analysis, e.g. by collecting regulatory signal sequence information for the genes with a similar expression pattern.

2.3 Mapping from KEGG/EXPRESSION to KEGG/PATHWAY

The KEGG/PATHWAY database is a collection of pathway diagrams that represent static networks of metabolism and regulation. An object in the diagram is a gene product such as an enzyme or a chemical compound such as a metabolic substrate. According to the constraint given, such as a set of genes in the genome or a compound found in the key word search, either the background color or the foreground color of the corresponding object(s) can be changed by the coloring tool that we have developed [7, 8].

Each gene product in the pathway diagram has the corresponding gene name that is linked to an entry in the GENES database. Because each spot in EXPRESSION is also associated with the gene name, it can be linked to the corresponding object in the pathway diagram if available. By converting the intensity ratio of the spot to the corresponding color and using the tool for coloring the objects in the pathway diagram, we can create a color-coded pathway diagram reflecting the patterns in the expression array (Fig. 2).

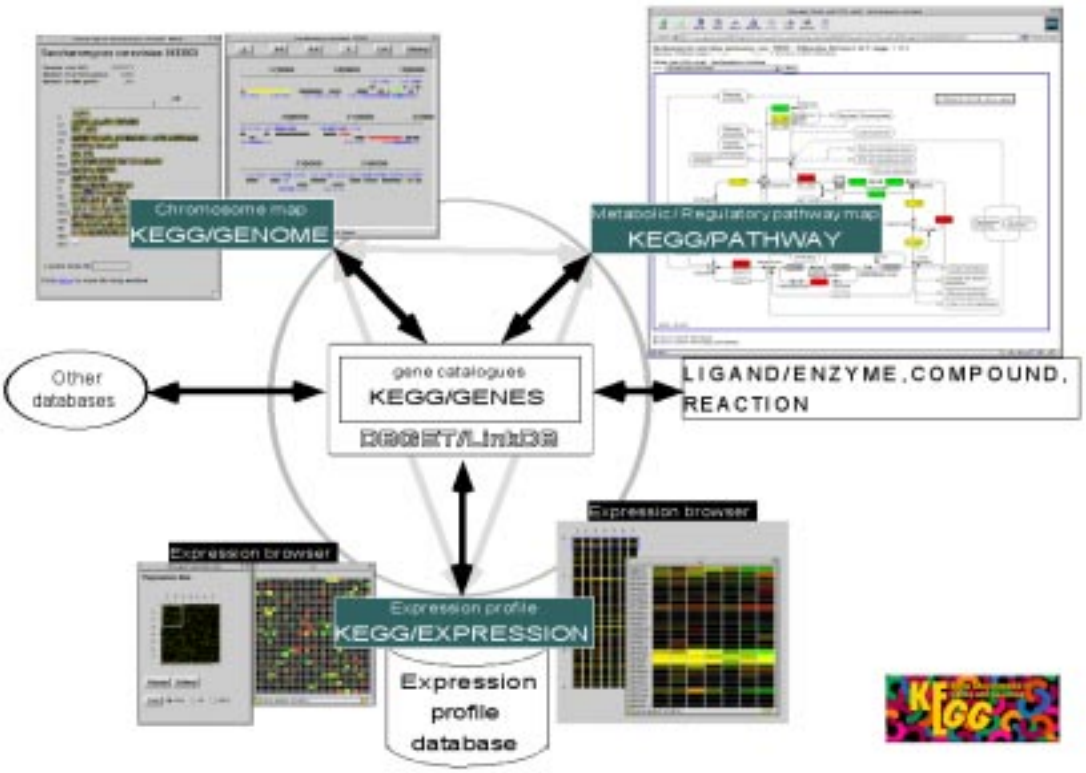

Figure 2: Integration of the KEGG/EXPRESSION database in the DBGET/LinkDB system, together with the GENES, GENOME, and PATHWAY databases in KEGG. All the databases are available via the Internet at the URL http://www.genome.ad.jp/kegg/.

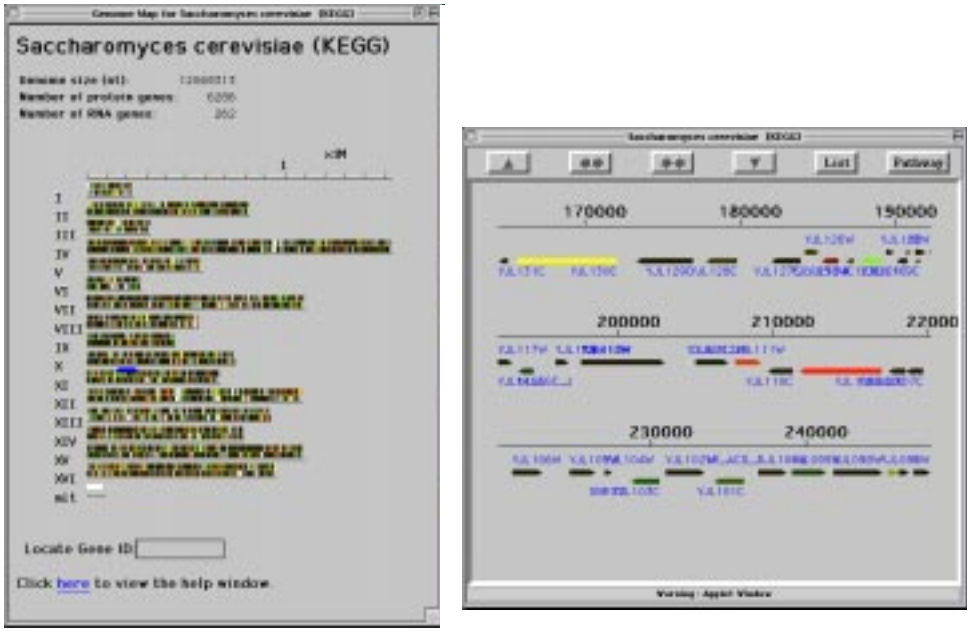

Figure 3: A genome browser view showing the expression data of *Saccharomyces cerevisiae* for the entire chromosomes (left) and for a selected portion in the expanded window (right). This browser is implemented as a Java applet.

By considering the time course information, we can color-code the pathway map for each time point in turn. Even though there is in general no relationship between enzymatic reaction kinetics and dynamic changes of enzyme gene expression levels, a dynamic behavior of the pathway may be observed in terms of the appearance and disappearance of the objects (genes). Such a dynamic behavior can be visualized by GIF animation.

Another way of visualizing the dynamic changes of the expression levels is to calculate the ratio of the expression levels between two time points and then to encode the values to different colors; for example red when the expression level increases, green when the level decreases, and yellow when no change is observed above a given threshold.

2.4 Mapping from KEGG/EXPRESSION to KEGG/GENOME

Fig. 3 shows a genome map browser that is integrated with the EXPRESSION database. This visualization provides facilities to analyze the relationship between the expression patterns and the chromosomal positions of genes.

This view is especially useful for the analysis of bacterial genomes, where operon structures are prevalent. A set of co-regulated genes in the expression pattern will appear as a set of positionally coupled genes in the genome. For eukaryotic genomes the genome map browser can be applied to the analysis of regulatory signal sites, for example by color-coding common sequence patterns upstream of the genes that are co-expressed according to the expression profiling. Furthermore, it may be worth exploring gene distribution profiles in each chromosome, for example analyzing the relationship between the expression patterns of house-keeping genes and the centromere/telomere structure, or even the chromatin structure.

3 Results

3.1 Data set

Here we report the result of our analysis on the publicly available microarray data for *Saccharomyces cerevisiae*, which is the same data set as the one used in the cluster analysis by Eisen *et al.* [4] The data include various experimental observations, such as a metabolic shift with genetic reprogramming, referred to as the diauxic shift from fermentation to respiration [1], and the developmental program of sporulation, consisting of meiosis and spore morphogenesis [2]. We obtained the data from their web site (http://rana.stanford.edu/clustering/).

3.2 Pathway Reconstruction using KEGG/EXPRESSION

One way of attempting to reconstruct a pathway is to just map the expression pattern onto the KEGG pathway diagrams as explained in section 2.3. However, this is obviously limited by the amount of data that already exists in the KEGG reference pathway diagrams. Therefore, it is necessary to develop methods for identifying new pathways by making use of the expression data. A general strategy is to first cluster genes from an experiment or several experiments and then to search possible functional connections of genes in the same cluster. This is based on the assumption that there is a tendency of co-regulated genes to have functional correlations in the pathway. If there is already a cluster in the pathway that is formed by partial members of a gene cluster in the expression pattern, then additional members may also belong to the same cluster in the pathway. Thus, this procedure would assist to assign functions to hypothetical genes and also possibly to extend the knowledge of pathways.

We developed a clustering framework for expression profiles. It is based on the distance between two genes, which is defined by computing the correlation of the expression patterns of two genes. We currently provide three clustering methods: single linkage, complete linkage and UPGMA (unweighted pair-group method with arithmetic means) [4]. The analysis of the *Saccharomyces cerevisiae* expression data [4] identified sets of co-regulated gene clusters; the most highly correlated ones were the clusters for ribosomal proteins and proteins in glycolysis.

Fig. 4 shows the result of mapping clustered genes coding for proteins in glycolysis onto the KEGG pathway diagram. The main pathway from D-glucose to ethanol was highly correlated. It is interesting to note that the enzymes responsible for this major pathway exhibited similar expression patterns over various environmental conditions as according to the data set we analyzed.

The pathway reconstruction by computation has mainly relied on the sequence similarity and on the possible functional coupling according to the conserved operon structures among prokaryotes [10, 11]. The gene expression data provides additional information for the pathway reconstruction, which will be most useful in eukaryotes where the simple correlation of positional coupling and functional coupling is not generally observed. Furthermore, the information of chemical compounds and chemical reactions can also be used, at least for the reconstruction of metabolic pathways. KEGG provides a pathway computation tool by using the reaction data stored in the LIGAND database [12]. For example, from a set of co-regulated enzyme genes all possible networks of chemical compounds can be computed from the corresponding list of substrate-product relations, which may lead to the discovery of a new pathway.

3.3 Expression Profile and Subcellular Locations

Organisms often have duplicate genes which apparently exhibit the same function. Aminoacyl-tRNA synthetase (ARS) coding genes in *Saccharomyces cerevisiae* contain such duplicate genes, one for cytoplasm and another for mitochondrion. We found that the expression patterns of the genes and the subcellular locations of the gene products are correlated in some cases.

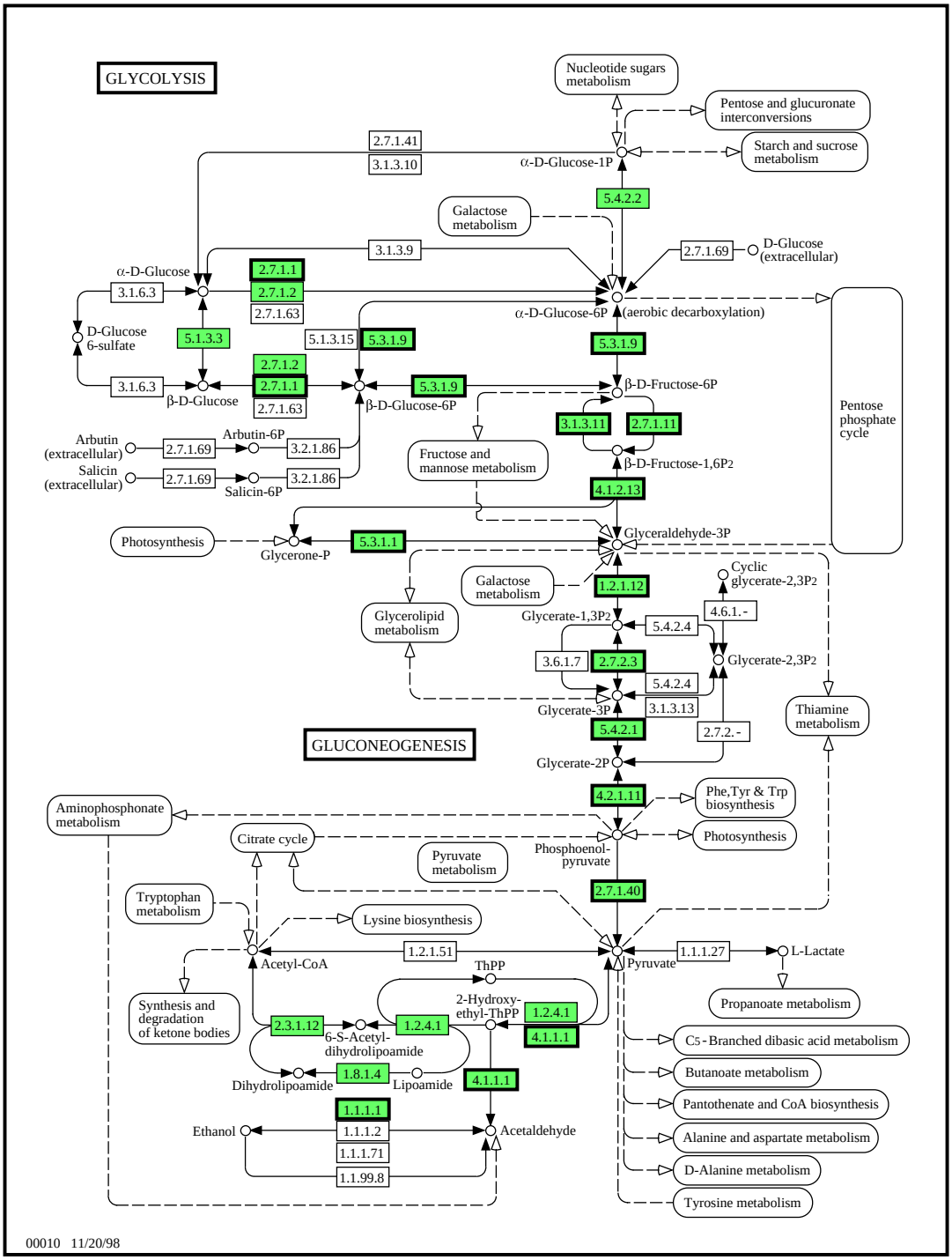

Figure 4: The enzymes whose genes share similar expression patterns are marked with bold lines in glycolysis for *Saccharomyces cerevisiae*. The enzymes whose genes are identified in the genome are colored in gray.

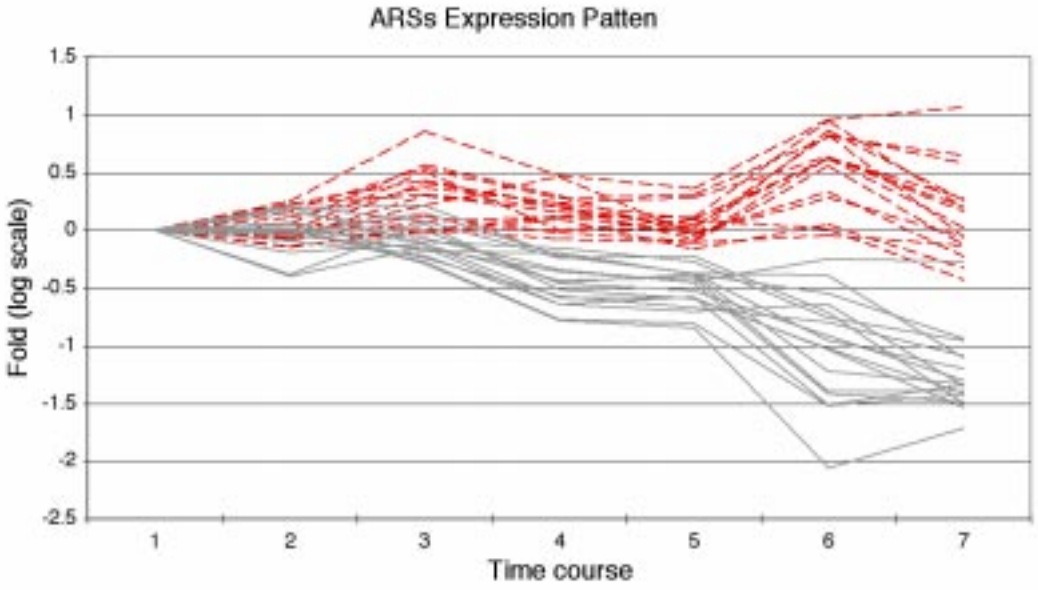

Figure 5: Expression patterns of aminoacyl-tRNA synthetase (ARS) genes in the diauxic shift. Cytoplasmic ARSs are represented by solid lines and mitochondrial ARSs are represented by dashed lines.

Fig. 5 shows the expression patterns of ARSs in the diauxic shift analysis of *Saccharomyces cerevisiae* [1]. It clearly shows the difference of the expression patterns between cytoplasmic ARSs and mitochondrial ARSs. While the expression level of cytoplasmic ARSs decreases as the concentration of glucose in the media decreases and the metabolism shifts to respiration, the expression level of mitochondrial ARSs are relatively stable.

When mapping the expression patterns to metabolic pathway diagrams the information on the subcellular locations is useful, because the objects in the current KEGG metabolic pathway map are categorized by functions or EC numbers only. As this ARS example showed, the distinction may not be detected if we use just the relationship between EC numbers and gene names.

4 Discussion

The publicly available expression data were stored in the KEGG/EXPRESSION database and used for the pathway reconstruction. Most of the currently available data are limited to eukaryotes– yeast, mouse and human. Therefore the analysis is focused on uncovering functional relationships among genes scattered on the genome rather than among those clustered in the chromosome.

We have initiated an experimental project to perform expression profile analysis of bacterial species, especially *Synechocystis* PCC6803 [13]. Among the many prokaryotes that have been sequenced the operon structure is least abundant in *Synechocystis*, which makes it difficult to assign gene functions because positional correlations cannot be utilized. Therefore it is expected that the direct measurement of co-expression is most useful in *Synechocystis*, identifying partners in the ABC transport system and the two-component signal transduction system, for example. The database we reported in this paper will be used to organize and analyze the data from the various research laboratories involved in the *Synechocystis* expression analysis.

Current difficulties in pathway reconstruction from expression data are summarized as follows.

The first one stems from the use of EC numbers for mapping genes to metabolic pathway diagrams as described in section 3.3. In addition to the information on the subcellular location, which has been useful in unicellular organisms and organisms with relatively small number of genes, the information on the tissue specificity should also be considered for organisms with a large number of genes, such as the mouse and human.

Second, since the KEGG/PATHWAY database consists of about 100 diagrams, it is not easy to check all the diagrams to see any interesting coloring according to the expression patterns. A method for automatic extraction of interesting correlations should be developed by quantifying the importance of the clusters found in the pathways.

Third, in case where the relative expression levels are considered between two time points (section 2.3), it is necessary to distinguish between highly expressed genes and lowly expressed genes. A high ratio may simply be due to the small variations of very low expression levels. In such a case we should combine the intensity information.

Acknowledgments

The authors thank T. Katayama and Y. Okuji for their advice on visualization tools and for stimulating discussions. This work was supported in part by a Grant-in-Aid for Scientific Research on the Priority Area "Genome Science" from the Ministry of Education, Science, Sports and Culture of Japan. The computation time was provided by the Supercomputer Laboratory, Institute for Chemical Research, Kyoto University. H.B. was supported by the Research Fellowship of the Japan Society for Promotion of Science for Young Scientists.

References

[1] DeRisi, J. L., Iyer, V. R., and Brown, P. O., Exploring the metabolic and genetic control of gene expression on a genomic scale, *Science*, 278:680–686, 1997.

[2] Chu, S., DeRisi, J., Eisen, M., Mulholland, J., Botstein, D., Brown, P. O., and Herskowitz, I., The transcriptional program of sporulation in budding yeast, *Science*, 282:699–705, 1998.

[3] Spellman, P. T., Sherlock, G., Zhang, M. Q., Iyer, V. R., Ander, K., Eisen, M. B., Brown, P. O., Botstein, D., and Futcher, B., Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization, *Mol. Biol. Cell*, 9:3273–3297, 1998.

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

[5] Bassett, D. E. Jr, Eisen, M. B., and Boguski, M. S., Gene expression informatics – it's all in your mine, *Nature Genet.*, 21:51–55, 1999.

[6] Ermolaeva, O., Restogi, M., Pruitt, K. D., Schuler, G. D., Bittner, M. L., Chen, Y., Simon, R., Meltzer, P., Trent, J. M., and Boguski, M. S., Data management and analysis for gene expression arrays, *Nature Genet.*, 20:19–23, 1998.

[7] M. Kanehisa, A database for post-genome analysis., *Trends Genet.*, 13:375–376, 1997.

[8] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M., KEGG: Kyoto Encyclopedia of Genes and Genomes, *Nucleic. Acids Res.*, 27:29–34, 1999.

[9] Fujibuchi, W., Goto, S., Migimatsu, H., Uchiyama, I., Ogiwara, A., Akiyama, Y., and Kanehisa, M., DBGET/LinkDB: an Integrated Database Retrieval System., *Pac. Symp. Biocomput.*, 98:683–694, 1998.

[10] Ogata, H., Goto, S., Fujibuchi, W., and Kanehisa, M., Computation with the KEGG pathway database, *Biosystems*, 47:119–128, 1998.

[11] Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D., and Maltsev, N., The use of gene clusters to infer functional coupling, *Proc. Natl. Acad. Sci. USA*, 96:2896–2901, 1999.

[12] Goto, S., Nishioka, T., and Kanehisa, M., LIGAND database for enzymes, compounds, and reactions, *Nucleic. Acids Res.*, 27:377–379, 1999.

[13] Kaneko, T., Sato, S., Kotani, H., Tanaka, A., Asamizu, E., Nakamura, Y., Miyajima, N., Hirosawa, M., Sugiura, M., Sasamoto, S., Kimura, T., Hosouchi, T., Matsuno, A., Muraki, A., Nakazaki, N., Naruo, K., Okumura, S., Shimpo, S., Takeuchi, C., Wada, T., Watanabe, A., Yamada, M., Yasuda, M. and Tabata, S., Sequence analysis of the genome of the unicellular cyanobacterium *Synechocystis* sp. strain PCC6803. II. Sequence determination of the entire genome and assignment of potential protein-coding regions, *DNA Res.*, 3:109–136, 1996.