Mining Biologically Active Patterns in Metabolic Pathways using Microarray Expression Profiles

Hiroshi Mamitsuka
Institute for Chemical Research
Kyoto University
Gokasho, Uji, 611-0011, JAPAN
mami@kuicr.kyoto-u.ac.jp

Yasushi Okuno
Graduate School of Pharmaceutical Sciences
Kyoto University
Sakyo-ku, Kyoto, 606-8501, JAPAN
okuno@pharm.kyoto-u.ac.jp

Atsuko Yamaguchi
Institute for Chemical Research
Kyoto University
Gokasho, Uji, 611-0011, JAPAN
atsuko@kuicr.kyoto-u.ac.jp

ABSTRACT

We present a new probabilistic framework for analyzing a metabolic pathway with microarray expression profiles. Our purpose is to find biologically significant paths and patterns in a given metabolic pathway. Our approach first builds a Markov model using a graph structure of a known metabolic pathway, and then estimates parameters of a mixture of the Markov models using microarray data, based on an EM algorithm. In our experiments, we used a main pathway of glycolysis to evaluate the effectiveness of our method. We first measured the performance of our method comparing with that of another method, in a supervised learning manner, and found that our method significantly outperformed another method, which was trained by microarray data only. We further analyzed the trained models and obtained a number of new biological findings on frequent patterns (paths) and long-range correlations in a metabolic pathway.

Keywords

Metabolic Pathways, Microarray Profiles, Model-based Clustering, Markov Models, Finite Mixture Models

1. INTRODUCTION

With the advent of a variety of high-throughput experimental techniques in molecular biology, various different types of genomic and proteomic data have been obtained in recent years. New biological insights, which would be missed when using one type of data only, might be obtained by mining multiple different types of data. Mining such multiple data sources has been regarded as more important for new biological discoveries.

We present a new approach for combining a known metabolic pathway and microarray data. A metabolic pathway is represented by a graph in a number of web sites on pathways. These sites include KEGG [16], UM-BBD [11], EcoCyc [17], EMP [4], etc. Figure 1 shows a part of the glycolysis pathway, which is obtained from the KEGG database. In a pathway graph, a chemical compound corresponds to a vertex, and a chemical reaction corresponds to a directed edge labeled by a protein, which catalyzes the reaction. More precisely, a directed edge and its two (head and tail) vertices correspond to a chemical reaction, catalyzed by a protein, and a compound (product) corresponding to the head vertex is generated from another compound (substrate) corresponding to the tail vertex. A pathway graph in the currently available databases is drawn by simply gathering known reactions, i.e. directed edges, and so by connecting multiple reactions (directed edges), a long path of a series of connected edges can be generated, but in fact it is sometimes biologically inactive or useless, although each reaction in such a path truly exists. Thus an important issue in a metabolic pathway graph is mining biologically active paths from a large number of plausible paths in the graph of a given metabolic pathway.

We present a new systematic method for this issue. In our approach, we use microarray expression profiles to find such biologically active paths. In the microarray data, a set of proteins (genes) are co-expressed under a certain experimental condition. Thus, we can infer that reactions corresponding to edges labeled by co-expressed proteins are activated, and a path connected by the activated edges is biologically active. In order to combine a pathway graph and microarray data, we propose to apply a Markov model mixture to modeling a metabolic pathway. We focus on a biologically significant path in a metabolic pathway, and the path can be a sequence of reactions. Thus we are convinced that a Markov chain is reasonable to modeling such biologically significant paths in a metabolic pathway. As is well known, microarray expression data is noise-abundant, and then we use a probabilistic Markov model, which can be considered as a noise-robust model, as an component of the mixture. We estimate probability parameter values of the Markov model mixture using microarray expression data, based on an EM algorithm [7].

We evaluated our method using two different types of data sets: a main metabolic pathway of glycolysis/gluconeogenesis starting with β-D-Fructose 6-Phosphate (F6P) and finishing with Phosphoenolpyruvate (PEP) and a data set of microarray gene expressions of yeast. We first evaluated our method in a supervised learning manner. That is, we randomly generated negative examples and examined the performance of discriminating positive examples from negatives. The performance of our method was compared with that of a support vector machine called one-class SVM, since it is an un-

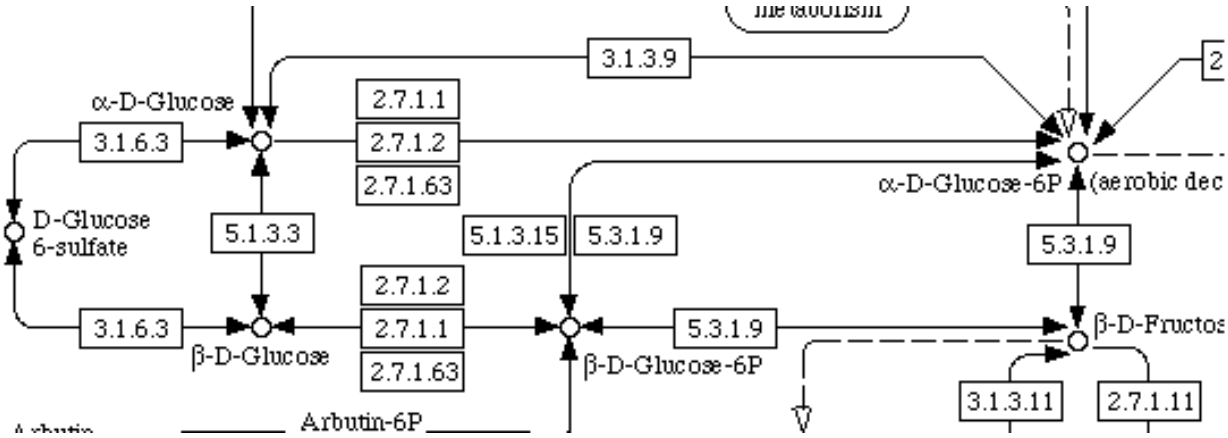


Figure 1: A part of the glycolysis pathway in KEGG.

supervised learning approach like our method and is a modification of a general SVM, which can be regarded as one of the most powerful learning methods in the current literature. The results obtained showed that our method significantly outperformed one-class SVM, which was trained by microarray data only. This result indicates that our approach of using a mixture of Markov models is highly effective for combining two different types of data sets, a metabolic pathway and a microarray data set.

We further analyzed the parameter values of our trained model for the glycolysis pathway. This pathway is a simple left-to-right type, and one or more edges, each of which is labeled by a different protein, exist for each reaction step. From the analysis of the trained parameter values, we obtained a number of biologically significant findings. First, we found that biologically active reaction paths (patterns) were much fewer than all possible paths from F6P to PEP, and these patterns could be characterized by only two reaction steps. We further found that these two steps had a very strong long-range correlation and patterns found in the long-range correlation could be gradually replaced alternatively. Furthermore, biologically active patterns at some condition could be inactive under another condition, and the reverse was also found.

These empirical results indicate that our approach of using a Markov model mixture is a very powerful tool for mining biologically significant patterns and paths in a metabolic pathway, using microarray data.

2. MARKOV MODEL MIXTURE AND RELATED WORK

Various types of probabilistic models such as Bayesian networks [13; 23] and probabilistic boolean networks [28; 25] have been already applied to obtain gene regulatory networks from microarray data. We emphasize that the properties of a metabolic pathway are different from those of such a regulatory network, and then these probabilistic models are not suitable for modeling metabolic pathways. Figure 2 illustrates a simple metabolic pathway in which compound A can be generated from compound B by enzyme (protein) X or from C by Y. If we model this pathway by a Bayesian network, the Bayesian network indicates that compound A

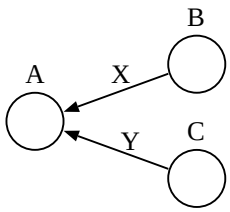


Figure 2: A simple pathway example.

depends on both compounds B and C. However, we note that when compound A is synthesized, it does not need both B and C, because compound A can be generated from B if enzyme X exists, regardless of how much compound C and enzyme Y exist. In other words, the reaction of generating compound A from compound B is independent of that from compound C, and the reverse is also true. This indicates that a Bayesian network is not suitable for modeling a metabolic pathway, and this fact is also true of a probabilistic boolean network. More generally, we can say that a metabolic path of generating a compound from another is a sequence of reactions, namely a type of time-series sequence, and so our model has to be trained using a set of time-series sequences. Thus we use a (first-order) Markov model for modeling a metabolic pathway.

We generate such a sequence of reactions from microarray expression profiles. That is, we focus on a set of proteins (enzymes), which are co-expressed under a certain experimental condition, and then obtain a series of reactions, which can be catalyzed by the co-expressed enzymes. A series of reaction steps in a metabolic pathway can proceed very fast, and hundreds of thousands of molecules can be generated within a second [31]. Thus it would be reasonable to use a sequence of reactions, which can be generated in the above procedure, as a training example to find biologically active paths in a given pathway.

Markov model-based approaches for time-series sequences have been successfully applied to a number of applications, such as speech recognition [24], natural language processing [19], biological sequence analysis [10] and web mining [2;

3]. The closest approach to our method is Cadez et al.'s analysis of web access patterns [5]. We use a mixture of Markov models as they used, but our component Markov model is different from that used by them, in terms that a state transition probability is specified not only by two states but by the transition itself. This is because in a metabolic pathway graph, multiple edges exist between same two vertices and these edges are labeled by different proteins.

Another item of note is that we already have a large amount of accumulated knowledge of pathways, and the accumulated knowledge would be very effective as background knowledge for building pathway graphs. Particularly, the accumulated knowledge of metabolic pathways is very abundant [18]. This fact implies that we may possibly skip the process of building a structure of an objective model, i.e. the most laborious process of learning a model, by using a known graph structure of a given metabolic pathway. If the graph structure of our model is given to us, all we have to do is just to estimate the parameter values of the model. Therefore, this abundant knowledge of metabolic pathways supports our approach. In addition, we do not need to apply a hidden Markov model (HMM) to our issue, because we can assume that our model structure is already given. This situation is the same as that of analyzing web access patterns, as pointed out in [5]. Instead of HMM, we use a mixture of Markov models, and estimate the probability parameters of the model from given time-series sequences [20; 6]. We note that the representation power of a mixture of Markov models is equal to that of an HMM. One advantage of the mixture of Markov models is that we can obtain frequent patterns in given sequences as in separate components, while the patterns are represented in only one model (network) in HMM. Thus we believe that our framework is very appropriate (and effective) for the problem of mining biologically active paths from a currently available metabolic pathway graph.

Tons of work has already been done on analyzing microarray data in the context of metabolic pathways. The following would be the tip of the iceberg: manually mapping co-expressed proteins onto pathways [8; 15; 22], simulating a kinetic model for fitting it to microarray expression profiles [1], analyzing microarray data by scoring pathways [32], clustering proteins (genes) using microarray data to build pathways [26; 9], co-clustering proteins (genes) using microarray data and metabolic pathways [14] and so on. We emphasize that the purposes of those approaches are different from that of this work, in which we focus on 'paths' and 'patterns', biologically activated. In addition, we note that a mixture of probabilistic Markov models has never been applied to analyzing metabolic pathways, using microarray data. Furthermore, some of the above work has used primitive and unsystematic approaches [8; 22]. We then emphasize that our approach is a general and systematic framework for automatically finding biologically significant paths and patterns in a given metabolic pathway.

3. DATA PREPARATION

3.1 Generating Reaction Sequences from Microarray Expression Profiles

A microarray data set is a table in which a column and a row correspond to a gene and an experimental condition, respectively. Figure 3 shows a schematic example of the data set,

		Genes		
		YGR043c	YLR354c	...
Experimental conditions	α-f #1	-0.23	0.08	
		0.25	-0.45	
	Cdc .	.	.	
	. .	.	.	
	. .	.	.	
	.			
	Dia #80	2.66	-1.64	

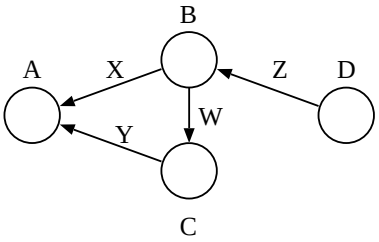
Figure 3: A schematic table of the microarray data set used in our experiment.

which will be actually used in our experiments. A numerical value in the table indicates how much a corresponding gene is expressed under a corresponding experimental condition. Usually if the value is high, a protein corresponding to this gene is highly produced and can be active in a metabolic pathway under the experimental condition. However each numerical value shows a relative value to that obtained under a reference (normal) condition. Thus, even when a value of a gene (protein) is zero, this protein may be produced and active if it is active under the reference condition. In particular, proteins appeared in a metabolic pathway are considered to be possibly always active. Thus, in this paper, to convert each original numerical value into a binary, i.e. active or inactive, we use a cut-off value around zero.

From each row of a microarray data set, we generate training examples, each of which is a sequence of reactions. The procedure to generate the sequences from a row is as follows: We first fix the beginning and ending vertices in a given pathway. We then convert each value of the row into a binary, i.e. one or zero, depending on whether or not it is larger than a pre-fixed cut-off value. For each row, using all proteins, each of whose binary value is one, we examine whether a path from the beginning to the ending vertices can be generated by the edges labeled by these proteins. If we find one or more paths, we store all of them.

We can use a dynamic programming approach for this process of generating paths starting from the beginning to the ending vertices. We can utilize a fact that a sequence can be generated only when its prefix subsequence is already generated, and then we generate a sequence of length t by using its prefix subsequence of length $t-1$. More concretely, for each t starting from 1 to T, we generate a sequence of length t from its prefix subsequence of length $t-1$, when the final compounds of the former and latter sequences are respectively the head and tail of a directed edge in a given pathway and the subsequence is already generated. The total computation time for this generation process is $O(n^2T)$, where n is the number of genes and T is the maximum sequence length. Thus, we note that sequences can be generated efficiently, although the number of possible sequences of length T reaches n^T.

Finally, if either of the beginning and ending vertices is more than one, we repeat the above procedure with varying the beginning and ending vertices. All of the paths generated are used to train our model. Figure 4 illustrates a simple



	W	X	Y	Z
I	1	0	1	1
II	0	1	0	1
III	1	1	1	0

	Reaction sequences
I	D - Z - B - W - C - Y - A
II	D - Z - B - X - A
III	-

Figure 4: An example of (a) a pathway graph, (b) binary microarray data and (c) generated sequences.

example of generating reaction sequences from a given pathway. As shown in Figure 4 (a), this example has four compounds, A, B, C and D, and four proteins X, Y, Z and W. First, when all binary values of four proteins are one except X as shown in (I) of Figure 4 (b), path 'D - Z - B - W - C - Y - A' is generated as shown in (I) of Figure 4 (c). Similarly a possible path is 'D - Z - B - X -A' for (II) of Figure 4 (b), and no paths are found for (III) of Figure 4 (b).

4. MODEL-BASED CLUSTERING

4.1 Preliminaries for Model Description

We use the following notation in this paper. We denote a variable by a capitalized letter, e.g. X, and the value of a corresponding variable by that same letter in lower case, e.g. x. Let $\mathbf{X}$ be a multivariate random variable taking on values corresponding to a sequence of reactions. Let Z be a discrete-valued latent variable taking on values $z_1, ..., z_K$, each of which corresponds to a latent class. Let $D = \{\mathbf{x}^1 \cdots \mathbf{x}^N\}$ be a set of (reaction) sequences, where each sequence $\mathbf{x}$ consists of observed T states and $T-1$ transitions: $\mathbf{x} = x_1 \cdot y_2, \cdots, y_T, x_T$, where state x_t takes discrete alphabets. Transition y_t also takes discrete alphabets, but we note that the number of alphabets taken by y_t depends on $t-1$ (more precisely x_{t-1}). We note that the Markov model can be interpreted as a pathway graph, and a state corresponds to a vertex labeled by a chemical compound, a transition corresponds to an edge labeled by an enzyme (protein), and a transition from a state to another state corresponds to a chemical reaction.

Our component Markov model has two types of probability parameters, i.e. the initial state probability $\theta_i (= p(x_1 = i; \theta))$, which is the probability that the initial state is i, and a state transition probability $\theta_{j,m|i}$ $(= p(x_t = j, y_t = m | x_{t-1} = i; \theta))$, which is the probability that the state and the transition at time t are j and m, respectively, given that the state at time $t-1$ is i. Here, $\sum_i \theta_i = 1$ and $\sum_{j,m} \theta_{j,m|i} = 1$.

Out mixture model has three types of probability parameters, i.e. the latent class probability $\pi_k (= p(z_k; \theta))$, which is the probability that the latent class is k, the initial state probability $\theta_{i|k} (= p(x_1 = i | z_k; \theta))$, which is the probability that the initial state is i given that the latent class is k, and a state transition probability $\theta_{j,m|i,k}$ $(= p(x_t = j, y_t = m | x_{t-1} = i, z_k; \theta))$, which is the probability that the state and the transition at time t are j and m, respectively, given that the state and the latent class at time $t-1$ are i and k, respectively. Here, $\sum_k \pi_k = 1$, $\sum_i \theta_{i|k} = 1$ and $\sum_{j,m} \theta_{j,m|i,k} = 1$. We note that the state transition probability includes transition y_t, since in a pathway graph, two vertices can be connected by more than one different edges and we need to distinguish these edges by their labels.

Let $n_{i \to j,m}(\mathbf{x})$ be the value obtained from a given training data set as follows: If transition m from state i to state j is in $\mathbf{x}$ then $n_{i \to j,m}(\mathbf{x})$ is 1, otherwise it is 0. Similarly let $n_i(\mathbf{x})$ be the value obtained from a given training data set as follows: If state i is the initial state in $\mathbf{x}$ then $n_i(\mathbf{x})$ is 1, otherwise it is 0.

4.2 First-Order Markov Model

We first describe a probabilistic first order Markov model, which is used as a component of our mixture model. The model can be explained by using the initial state probabilities and the state transition probabilities as follows:

$$p(\mathbf{x}; \theta) = p(x_1; \theta) \prod_{t=2}^{T} p(x_t, y_t | x_{t-1}; \theta).$$

4.3 Mixture of First-Order Markov Models

We then describe our Markov model mixture, which we call 3M, using three types of probabilities: the latent class probabilities, the initial state probabilities and the state transition probabilities. 3M for $\mathbf{X}$ with K clusters has the following form:

$$\begin{aligned} p(\mathbf{x}; \theta) &= \sum_{k}^{K} p(z_k; \theta)\ p(\mathbf{x} | z_k; \theta) \\ &= \sum_{k}^{K} p(z_k; \theta)\ p(x_1 | z_k; \theta) \prod_{t=2}^{T} p(x_t, y_t | x_{t-1}, z_k; \theta). \end{aligned}$$

4.4 Estimating Probability Parameters from Given Reaction Sequences

A possible criterion for estimating probability parameters of 3M is the maximum likelihood, in which parameters are obtained to maximize the likelihood of given training data. In order to obtain the maximum likelihood parameters, we apply a general scheme, EM (Expectation-Maximization) algorithm [7], to 3M. The algorithm starts with initial parameter values and iterates both an expectation step (E-step) and a maximization step (M-step) alternately until a certain convergence criterion is satisfied.

In E-step, we estimate the latent classes using the complete

data log-likelihood as follows:

$$w(z_k|\mathbf{x}^d;\theta) \quad = \quad \frac{p(z_k;\theta)\ p(\mathbf{x}^d|z_k;\theta)}{\sum_{k'} p(z_{k'};\theta)\ p(\mathbf{x}^d|z_{k'};\theta)}.$$

In M-step, we sum over $w(z_k|\mathbf{x}^d;\theta)$ with $n_{i\to j,m}(\mathbf{x}^d)$ (or $n_i(\mathbf{x}^d)$), as follows:

$$\pi_k \quad \propto \quad \sum_d \ w(z_k|\mathbf{x}^d;\theta_{old}).$$

$$\theta_{i|k} \quad \propto \quad \sum_d n_i(\mathbf{x}^d)\ w(z_k|\mathbf{x}^d;\theta_{old}).$$

$$\theta_{j,m|i,k} \quad \propto \quad \sum_d n_{i\to j,m}(\mathbf{x}^d)\ w(z_k|\mathbf{x}^d;\theta_{old}).$$

5. EXPERIMENTAL RESULTS

5.1 Data and Parameters

We used the data sets of yeast (more precisely, *Saccharomyces cerevisiae*) in our experiments. The information of the pathway used in our experiments was downloaded from http://www.cpb.dtu.dk/models/yeastmodel.html [12]. The microarray data set of gene expressions on yeast proteins was downloaded from http://rana.lbl.gov/EisenData.htm [27], and we used original values of this publicly available data set, which can be seen at the above web site.

5.1.1 Metabolic Pathway

We used a main path of glycolysis/gluconeogenesis, because this pathway is the most fundamental pathway in metabolism. This glycolysis pathway has been analyzed by a number of groups, including DeRisi et al. [8], a pioneer group for analyzing pathways using microarray gene expressions. Figure 5 shows this pathway starting with F6P (β-D-Fructose 6-phosphate) and finishing with PEP (Phosphoenolpyruvate). As shown in this figure, we fix the beginning and ending compounds, but this pathway has two different edges from F6P to T3P1, three from T3P1 to 13PDG, three from 3PG to 2PG and five from 2PG to PEP. Totally there are 90 (= 2 $\times$ 3 $\times$ 3 $\times$ 5) possible different paths from F6P to PEP, and fourteen genes are used in these paths. So we used fourteen only among approximately six thousand columns (genes) of the data set of microarray expressions.

5.1.2 Microarray Expression Data

The data set of expression profiles has 80 rows, all of which differ in experimental conditions. These 80 experimental conditions are classified into five different classes, 'Cell-cycle α-factor (α-f)', 'Cell-cycle cdc15 (cdc)', 'Cell-cycle Elutriation (Elu)', 'Sporulation (Spo)' and 'Diauxic shift (Dia)', and three minor rows. In our experiments, we used the five classes of gene expression profiles. Figure 3 shows a schematic table of this microarray data set.

We generated reaction sequences from the microarray data set, following the procedure shown in the section of 'Data Preparation'. In this procedure, we used three different cut-off values, 0.1, 0 and -0.1, for each of the five classes of the microarray data set. Table 1 shows the original number of rows in the downloaded microarray data set and the number of sequences generated from the rows when we set a cut-off value at 0.1, 0 and -0.1.

Table 1: The number of sequences generated when we set the cut-off value at -0.1, 0 and 0.1.

		# reaction sequences		
Classes	#rows	0.1	0	-0.1
α-f	18	79	158	503
Cdc	25	56	100	228
Elu	14	66	129	183
Spo	13	60	69	85
Dia	7	15	51	108

5.1.3 Parameters

Throughout our experiments, we fixed the size of latent classes for 3M at ten.

5.2 Discriminating Positive Gene Expressions from Negatives

We first evaluated the performance of 3M comparing with that of another method in a supervised learning manner. We selected one-class SVM (hereafter OC.SVM) as a competing method, since it is an unsupervised learning approach like 3M and is a modification of a more general SVM, which is recognized as one of the most powerful learning methods. 3M and OC.SVM are both unsupervised learning methods, and so in our experiments, both were trained by positive examples only and predictions were done for both positive and negative examples. We note that 3M was trained by using both a microarray data set and a pathway, but OC.SVM was trained only by a microarray data set. Thus we evaluated these two methods by how accurately each of them discriminates positive rows of the microarray data set from the negative examples generated randomly.

5.2.1 Data Preparation for Supervised Learning

We converted each value of a row of the original microarray data set into a binary, using a cut-off value, and a set of binaries for each row of the original microarray data set was regarded as a positive example. We randomly split the positive rows into training and test roughly equally for each class of the microarray data set. For the positive test data set, we randomly generated negative examples, whose number is the same as that of positive test examples. More precisely, we randomly generated a binary for each of the fourteen genes in Figure 5, under the constraint that from each set of the fourteen binaries we can generate at least one path from F6P to PEP in Figure 5. This set of random fourteen binaries corresponds to a negative test example, and we then repeated this random generation until the number of negative examples reaches that of positive rows. We note that we used each set of fourteen binaries as an input for OC.SVM, whereas sequences generated from each binary set were used for training 3M. That is, pathway dependency information is not given to OC.SVM.

We further repeated this procedure of preparing a pair of training and test data sets (and a negative test data set) ten times, and experimental results obtained were averaged over these ten runs.

5.2.2 Competing Method: One-Class Support Vector Machine (OC.SVM)

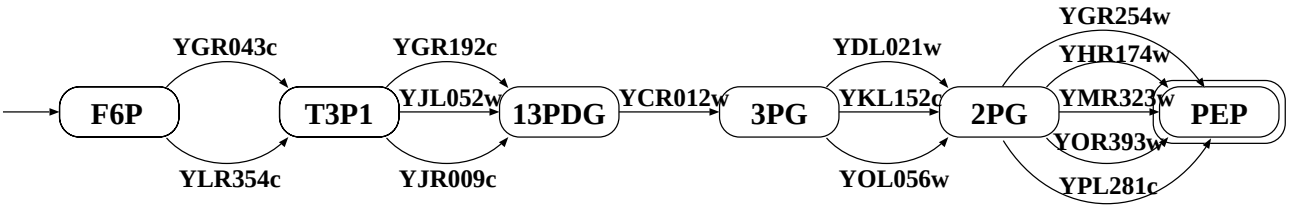


Figure 5: Glycolysis pathway used in our experiments: F6P→PEP.

We here briefly explain the learning scheme of OC.SVM. Let $\mathbf{g}_i$ $(i = 1, ..., M)$ be an input example, i.e. a row of the microarray data set. The goal of OC.SVM is to find a hyper plane h that separates given examples from the origin in a hyper space at a threshold ρ. For this goal, the following quadratic problem is solved:

$$\min \quad \frac{1}{2}\|h\|^2 + \frac{1}{\mu M}\sum_i \xi_i$$

subject to

$$(h \cdot \Phi(\mathbf{g}_i)) \geq \rho - \xi_i \quad (i = 1, \cdots, M), \quad \xi_i \geq 0,$$

where Φ be a kernel map.

In our experiments, we used LIBSVM ver.2.36, which has an option to run the OC.SVM and can be downloaded from http://www.csie.ntu.edu.tw/~ cjlin/libsvm/. We note that in prediction, for each example, LIBSVM gives only a binary output, indicating whether this example belongs to the class of training data or not (positive or negative). We will show the results obtained by using a linear kernel, but we emphasize that any significantly better performance was not obtained by other kernels.

5.2.3 *Average prediction accuracies and t-values of mean difference significance test*

We note that in both training and prediction, 3M uses reaction sequences generated from each row of the microarray data set. Thus we have to predict the class of each row using the predictions done for reaction sequences by 3M. For this purpose, we used the following procedure: First, we computed the likelihood of each reaction sequence by 3M, and then the likelihoods for all sequences (obtained from all rows in both positive and negative test data sets) were sorted. For the sorted sequences, we found a threshold for discriminating positives from negatives, satisfying to obtain the maximum discrimination accuracy for test sequences. We then assigned a class to each sequence, depending on whether the likelihood of the sequence is lower or higher than the threshold. Finally, for each row of test examples, we took a majority vote over the classes of reaction sequences, which are generated from the row.

We computed the prediction accuracy of each method by dividing the number of correctly predicted rows by the total number of rows. As mentioned earlier, we averaged the prediction accuracies obtained over ten runs.

We further used 't' values of the (pairwise) mean difference significance test for statistically comparing the prediction accuracy of our method with that of OC.SVM. The t values are calculated using the following formula: $t = \frac{|ave(D)|}{\sqrt{\frac{var(D)}{n}}}$, where we let D denote the accuracy difference between our method and OC.SVM for each of ten runs, $ave(W)$ the average of W, $var(W)$ the variance of W, and n the number of data sets (ten in our case). For $n = 10$, if t is larger than 3.250 then it is more than 99% statistically significant that 3M achieves a higher prediction accuracy than OC.SVM.

Table 2: Average prediction accuracies (%) and t-values in parenthesis.

Class	Cut-off	3M	OC.SVM
α-f	0.1	**70.0**	61.3 (**3.33**)
	0	**80.0**	51.7 (3.16)
	-0.1	62.1	**69.3** (2.04)
Cdc	0.1	**95.0**	66.7 (**5.97**)
	0	**90.0**	53.8 (**7.51**)
	-0.1	**85.8**	64.2 (**4.72**)
Elu	0.1	**90.0**	56.7 (**10.0**)
	0	**87.5**	71.3 (2.77)
	-0.1	**84.0**	72.0 (2.71)
Spo	0.1	**80.0**	50.0 (**3.87**)
	0	**95.0**	65.0 (**3.87**)
	-0.1	**85.0**	60.0 (**5.00**)
Dia	0.1	**95.0**	50.0 (**9.49**)
	0	**90.0**	67.5 (**5.28**)
	-0.1	**87.5**	70.0 (2.83)
# highest pred. acc.		**14**	1
# stat. sig. diff.		-	**10**

Table 2 shows the prediction accuracies (and t values in parenthesis) obtained by 3M and OC.SVM. We tested three cut-off values for each of the five classes, and totally fifteen different data sets were tested. The prediction accuracies obtained by 3M were better than those of OC.SVM in all of these fifteen cases except only one, being statistically significant in ten cases out of all of the advantageous cases. This result indicates that 3M, trained by using both a metabolic pathway and microarray gene expressions, significantly outperformed OC.SVM, which was trained by only one type of data, microarray data. In other words, we can say that only using the given microarray data set was not enough for discriminating positives from negatives by OC.SVM, but combining the information of the metabolic pathway and the microarray data set by our method significantly improved the predictive performance obtained by OC.SVM.

Figure 6 shows how the t-value varies with the three cut-off values for each of the five classes. The trend of the five curves in Figure 6 indicates that as with the larger cut-off value the t-value is larger. Particularly, when the cut-off value is -0.1, the performance difference is statistically

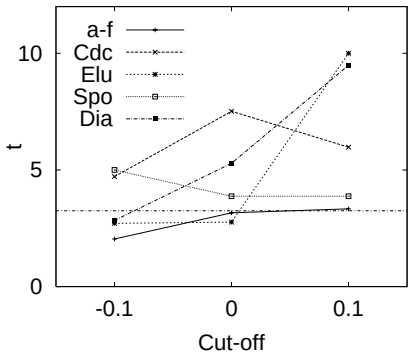


Figure 6: Variation of t-values for three cut-off values.

insignificant in three out of the five cases. As shown in Table 1, a larger cut-off value results in the smaller number of generated examples, and the reverse is also true. Thus, strangely enough, the statistical significance of our method decreases as the number of reaction sequences increases. A plausible explanation for this result is that the setting of a low cut-off value, namely -0.1, generates noisy sequences, which may lower the quality of our mining results. This implies that we may need to set the cut-off value at zero or larger for our experiments.

5.3 Analyzing Glycolysis Pathway

For analyzing this glycolysis pathway, we focused on Spo and Dia of the five classes, because α-f, Cdc and Elu, which are relating to cell-cycle, have only a relatively small influence on glycolysis (see e.g. [21]). We run our learning algorithm, using all training examples generated, and obtained ten component Markov models for each of Spo and Dia. We checked up parameter values of these totally twenty components obtained, when the cut-off value was set at zero. We found that these components could be all clearly clustered into extremely fewer 'patterns' in terms of parameter values, and surprisingly each of these patterns could be characterized by only two reaction steps, F6P→T3P1 and 3PG→2PG, in this pathway. We first note that for all of the twenty components, the state transition probability of YOL056c was always zero for 3PG→2PG. This means any path through '3PG - YOL056c - 2PG' is biologically inactive for both Spo and Dia. Thus YGR043c/YLR354c and YDL021w/YKL152w can be active for F6P→T3P1 and 3PG→2PG, respectively. We further explored how the state transition probability of each of the four enzymes, YGR043c, YLR354c, YDL021w and YKL152w, varies for twenty components.

We obtained only two and three patterns for Dia and Spo, respectively. Figure 7 shows two patterns (named as Dia-(a) and Dia-(b)) obtained for Dia. Dia-(a) was the major pattern, taken by nine components out of the ten obtained components of our mixture, and Dia-(b) was the minor pattern found in only one component. In Dia-(a), the state transition probability of YGR354c was 1.0 for F6P→T3P1, and the probabilities of YDL021w and YKL152c were approximately 0.3 and 0.7, respectively, for 3PG→2PG. On the other hand, in Dia-(b), the state transition probabilities of YGR043c and YDL021w were both 1.0 for F6P→T3P1 and 3PG→2PG, respectively. This result indicates that 'F6P - YLR354c - T3P1 - ··· - 3PG - YKL152c - 2PG - ···' was the biologically most significant path for Dia, and other paths shown in Figure 7 seemed to be also biologically active. However, the other paths, such as 'F6P - YGR043c - T3P1 - ··· - 3PG - YKL152c - 2PG - ···', seemed to be biologically inactive for Dia.

Figure 8 shows three patterns (named as Spo-(a), Spo-(b) and Spo-(c)) obtained for Spo, and a two-dimensional (2D) space obtained by plotting two state transition probabilities, i.e. YGR043c for F6P→T3P1 and YDL021w for 3PG→2PG, of ten trained components of our mixture. Spo-(a) was the major pattern, taken by eight components, and Spo-(b) and Spo-(c) were minor patterns, each of which was taken by only one component. Spo-(c) corresponds to the upper right point in the 2D space, and the state transition probabilities of YGR043c and YDL021w were both 1.0. Also, in Spo-(b), the state transition probabilities of YGR043c and YDL021w were both higher than those of YLR354c and YKL152c, respectively. Interestingly, the pattern of Spo-(c) was common to the pattern shown in Dia-(b) (Figure 7 (b)). That is, three minor components of Dia and Spo, namely Dia-(b), Spo-(b) and Spo-(c), had the same pattern. Spo-(a) had a different interesting pattern, which indicates anticorrelation between YGR043c and YDL021, as shown in the ellipse on the diagonal line from upper left to lower right in the 2D space. That is, as with increasing the state transition probability of YGR043c, that of YDL021w decreases, and the reverse is also true. This result indicates that our method found a pair of alternative paths, namely 'F6P - YGR043c - T3P1 - ··· - 3PG - YKL152c - 2PG -' and 'F6P - YLR354c - T3P1 - ··· - 3PG - YDL021w - 2PG -'. Furthermore, our biologically significant finding is that they were not just switched on and off, and could be gradually replaced as shown in the ellipse.

From the three patterns obtained for Spo, three out of the four possible combinations of the four enzymes, YGR043c, YLR354c ,YDL021w and YKL152c, could be biologically active, and they were 'F6P - YGR043c - T3P1 - ··· - 3PG - YDL021w - 2PG -', 'F6P - YGR043c - T3P1 - ··· - 3PG - YKL152c - 2PG -' and 'F6P - YLR354c - T3P1 - ··· - 3PG - YDL021w - 2PG -'. However, the remaining one path, 'F6P - YLR354c - T3P1 - ··· - 3PG - YDL152c - 2PG -', seemed to be biologically inactive. Interestingly, this inactive path for Spo was active for Dia as the major pattern, as shown in Figure 7 (a). On the other hand, the major active pattern for Spo, 'F6P - YGR043c - T3P1 - ··· - 3PG - YKL152c - 2PG -' seemed to be biologically inactive for Dia, as we mentioned earlier. This result implies that yeast must have a strong selectivity for the glycolysis pathway, particularly the two steps on which we focused, depending on given conditions such as Dia and Spo. The yeast glycolysis has been well investigated [30; 29], but this type of obvious selectivity has not been clearly shown like this, and our method automatically found this selectivity in the glycolysis pathway.

We note that other reaction steps, T3P1→13PDG→3PG and 2PG→PEP, had no relation to the above five patterns. For example, the state transition probabilities of three proteins of T3P1→13PDG had an almost uniform distribution, i.e. $\frac{1}{3}$ for each of these three, in all twenty components of Spo and Dia. This result indicates that two steps, F6P→T3P1 and 3PG→2PG, have a strong correlation in metabolism.

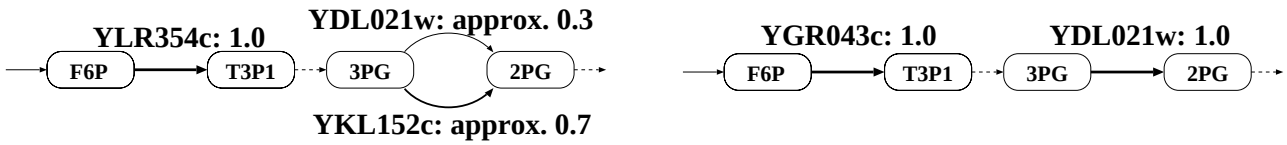


Figure 7: Two patterns obtained for Dia.

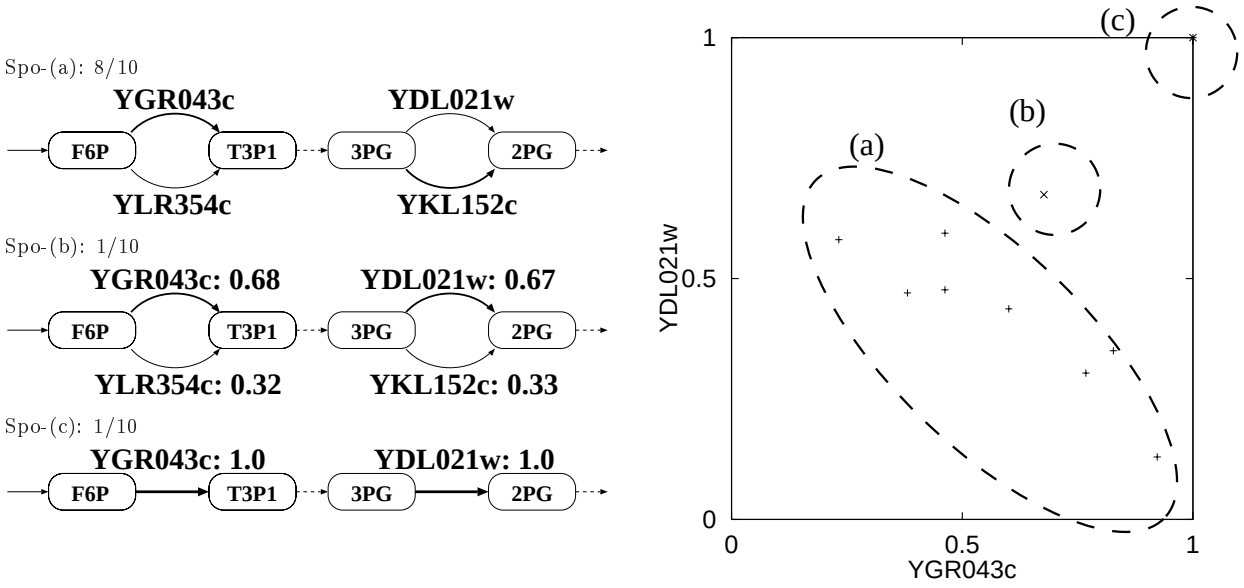


Figure 8: Three patterns obtained for Spo.

We emphasize that our approach of using a Markov model mixture could find this long-range correlation in given sequences.

Over all, these results obtained by analyzing trained components also confirm that our method is very effective for mining biologically significant patterns and paths from a known metabolic pathway and microarray data.

6. CONCLUDING REMARKS

We have presented a new method for mining biologically significant patterns from given metabolic pathways and microarray expression profiles. We have experimentally examined the effectiveness of our probabilistic approach using a real metabolic pathway and a microarray data set. Our method achieved a competitive performance for effectively combining two different types of data sets, a pathway graph and a microarray expression data set. Furthermore, various types of analysis on the trained models presented a variety of biologically significant findings and results, which also verify the effectiveness of our proposed method.

In our first experiment, we randomly generated negative examples. We note that these examples are synthetically generated just for examining the effectiveness of our approach of combining two different types of data. The problem of finding new biologically significant paths and patterns is a unsupervised learning issue in nature. As shown by the analysis of our trained models, our approach is very effective for this issue.

We further emphasize that our approach can be applied to any known metabolic pathway using a data set of microarray expressions. Future work includes to run our method on a variety of pathways and genome-wide large-scale pathways.

ACKNOWLEDGEMENTS

This work was partially supported by Grant-in-Aid for Scientific Research on Priority Areas (C) "Genome Information Science" from the Ministry of Education, Culture, Sports, Science and Technology of Japan.

7. REFERENCES

[1] A. Arkin, J. Ross, and H. McAdams. Stochastic kinetic analysis of developmental pathway bifurcation in phage λ-infected *escherichia coli* cells. *Genetics*, 149:1633–1648, 1998.

[2] P. Baldi, P. Frasconi, and P. Smyth. *Modeling the Internet and the Web: Probabilistic Methods and Algorithms.* John Wiley & Sons, 2003.

[3] D. J. Bertsimas, A. J. Mersereau, and N. R. Patel. Dynamic classification of online customers. In *Proceedings*

of the Third SIAM International Conference on Data Mining, pages 107–118, 2003.

[4] A. Burgard and C. Maranas. Review of the enzymes and metabolic pathways (EMP) database. *Metabolic Engineering*, 3:193–194, 2001.

[5] I. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White. Model-based clustering and visualization of navigation patterns. *Data Mining and Knowledge Discovery*, 7:399–404, 2003.

[6] J. Cornell. *Experiments with Mixtures: Design, Models, and the Analysis of Mixture Data*. John Wiley & Sons, NY, 2002.

[7] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. B*, 39:1–38, 1977.

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

[9] I. Dhillon, E. Markotte, and U. Roshan. Diametric clustering for identifying anti-correlated gene clusters. *Bioinformatics*, 19(13):1612–1619, 2003.

[10] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison. *Biological Sequence Analysis*. Cambridge University Press, Cambridge, UK, 1998.

[11] L. Ellis, B. Hou, W. Kang, and L. Wackett. The University of Minnesota biocatalysis/biodegradation database: Post-genomic data mining. *Nucleic Acids Research*, 31:262–265, 2003.

[12] J. Förster, I. Famili, P. Fu, B. Palsson, and J. Nielsen. Genome-scale reconstruction of the *saccharomyces cerevisiae* metabolic network. *Genome Reserach*, 13:244–253, 2003.

[13] N. Friedman, M.Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7:601–620, 2000.

[14] D. Hanisch, A. Zien, R. Zimmer, and T. Lengauer. Co-clustering of biological networks and gene expression data. *Bioinformatics*, 18(Suppl. 1):S145–S154, 2002.

[15] T. Ideker, V. Thorsson, J. Ranish, R. Christmas, J. Buhler, J. Eng, R. Bumgarner, D. Goodlett, R. Aebersold, and L. Hood. Integrated genomic and proteomic analysis of a systematically perturbed metabolic network. *Science*, 292:929–934, 2001.

[16] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori. The KEGG resource for deciphering the genome. *Nucleic Acids Research*, 32(1):D277–D280, 2004.

[17] P. Karp, M. Riley, M. Saier, I. Paulsen, J. Collado-Vides, S. Paley, A. Pellegrini-Toole, C. Bonavides, and S. Gama-Castroothers. The EcoCyc database. *Nucleic Acids Research*, 30(1):56–58, 2002.

[18] K. Koeller and C.-H. Wong. Enzymes for chemical synthesis. *Nature*, 409:232–240, 2001.

[19] C. Manning and H. Schütze. *Foundations of Statistical Natural Language Processing*. MIT Press, Cambridge, MA, 1999.

[20] G. McLachlan and D. Peel. *Finite Mixture Models*. John Wiley & Sons, NY, 2000.

[21] L. Newcomb, J. Diderich, M. Slattery, and W. Heideman. Glucose regulation of *saccharomyces cerevisiae* cell cycle genes. *Eukaryotic Cell*, 2:143–149, 2003.

[22] M.-K. Oh, L. Rohlin, K. Cao, and J. Liao. Global expression profiling of acetate-grown *escherichia coli*. *Journal of Biological Chemistry*, 277(15):13175–13183, 2002.

[23] I. Ong, J. Glasner, and D. Page. Modelling regulatory pathways in E.colli from time series expression profiles. *Bioinformatics*, 18(Suppl. 1):S241–S248, 2002.

[24] L. Rabiner. A tutorial on hidden Markov models and selected applications. *Proceedings of IEEE*, 77:257–286, 1989.

[25] I. Schmulevich, E. Dougherty, and W. Zhang. From boolean to probabilistic boolean networks as models of gene regulatory networks. *Proceedings of IEEE*, 90(11):1778–1792, 2002.

[26] E. Segal, H. Wang, and D. Koller. Discovering molecular pathways from protein interaction and gene expression data. *Bioinformatics*, 19:i264–i272, 2003.

[27] G. Sherlock, , T. Hernandez-Boussard, A. Kasarskis, G. Binkley, J. Matese, S. Dwight, M. Kaloper, S. Weng, H. Jin, C. Ball, M. Eisen, P. Spellman, P. Brown, D. Botstein, and J. Cherry. The Stanford microarray database. *Nucleic Acids Research*, 29:152–155, 2001.

[28] I. Shmulevich, E. Dougherty, S. Kim, and W. Zhang. Probabilistic boolean networks: a rule-based uncertainty model for regulatory networks. *Bioinformatics*, 18(2):261–274, 2002.

[29] B. ter Kuile and H. Westerhoff. Transcriptome meets metabolome: Hierarachical and metabolic regulation of the glycolytic pathway. *FEBS Letter*, 500(3):169–171, 2001.

[30] B. Teusink, J. Passarge, C. Reijenga, E. Esgalhado, C. van der Weijden, M. Schepper, M. Walsh, B. Bakker, K. van Dam, H. Westerhoff, and J. Snoep. Can yeast glycolysis by understood in terms of in vitro kinetics of the constituent enzymes? testing biochemistry. *European Journal of Biochemistry*, 267:5313–5329, 1998.

[31] D. Voet and J. Voet. *Biochemistry*. John Wiley & Sons, NY, 2003.

[32] A. Zien, R. Küffner, R. Zimmer, and T. Lengauer. Analysis of gene expression data with pathway scores. In *Proceedings of the Eighth International Conference on Intellingent Systems for Molecular Biology*, pages 407–417, 2000.