International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:9, No:3, 2015

Application of KL Divergence for Estimation of Each Metabolic Pathway Genes

Shohei Maruyama, Yasuo Matsuyama, Sachiyo Aburatani

***Abstract*—**Development of a method to estimate gene functions is an important task in bioinformatics. One of the approaches for the annotation is the identification of the metabolic pathway that genes are involved in. Since gene expression data reflect various intracellular phenomena, those data are considered to be related with genes' functions. However, it has been difficult to estimate the gene function with high accuracy. It is considered that the low accuracy of the estimation is caused by the difficulty of accurately measuring a gene expression. Even though they are measured under the same condition, the gene expressions will vary usually. In this study, we proposed a feature extraction method focusing on the variability of gene expressions to estimate the genes' metabolic pathway accurately. First, we estimated the distribution of each gene expression from replicate data. Next, we calculated the similarity between all gene pairs by KL divergence, which is a method for calculating the similarity between distributions. Finally, we utilized the similarity vectors as feature vectors and trained the multiclass SVM for identifying the genes' metabolic pathway. To evaluate our developed method, we applied the method to budding yeast and trained the multiclass SVM for identifying the seven metabolic pathways. As a result, the accuracy that calculated by our developed method was higher than the one that calculated from the raw gene expression data. Thus, our developed method combined with KL divergence is useful for identifying the genes' metabolic pathway.

***Keywords*—**Metabolic pathways, gene expression data, microarray, Kullback–Leibler divergence, KL divergence, support vector machines, SVM, machine learning.

I. Introduction

FOR understanding life system, it is important to identify the genes that are involved in metabolic pathways. Because gene expression data reflect various intracellular phenomena, gene expression data are useful for revealing the genes that are involved in metabolic pathways.

The Pearson product-moment correlation coefficient has been utilized to gene expression data for revealing the relevant genes [1]-[3]. The method is based on the idea that co-expression genes have similar function. However, the Pearson product-moment correlation coefficient can express only linear relationship between genes. Thus, we cannot utilize the method to infer the relevant genes that have non-linear relationship with other relevant genes.

Support vector machines (SVMs) [4], [5] are useful to treat this problem. SVMs are a supervised machine learning method for classification. SVMs can treat non-linear relationships between genes by the kernel trick. Brown et al. [6] utilized SVMs with gene expression data to recognize six functional classes of genes: tricarboxylic acid (TCA) cycle, respiration, cytoplasmic ribosomes, proteasome, histones, and helix-turn-helix proteins. They compared the classification performances of the SVMs with those of four machine learning algorithms (Parzen windows, Fisher's linear discriminant, C4.5, and MOC1), and showed that the SVMs achieved the best classification performance among them.

Brown et al. [6] applied raw gene expression data to SVMs. However, typical gene expression data includes the replicates for estimation of the variability associated with gene expressions. Thus, it is considered that gene expression data have to be applied to SVMs with the method reflecting the variability associated with gene expressions.

In this report, we propose a method based on the SVM approach, for identifying the genes' metabolic pathways from the gene expression data. To reflect the variability of gene expressions, we calculated the similarities between genes by KL divergence, and utilized the similarities as the feature vectors of SVMs. Then, we applied our developed method to the gene expression data of *Saccharomyces cerevisiae* against seven metabolic pathways defined by KEGG, and evaluated their classification performances.

II. Methods

To estimate the genes' metabolic pathway accurately by reflecting the variability of gene expressions, first, we estimated the distribution of each gene expression from replicate data. Next, we calculated the similarity between all gene pairs by KL divergence, which is a method for calculating the similarity between distributions. Finally, we utilized the similarity vectors as feature vectors and trained the multiclass SVM for identifying the genes' metabolic pathway.

A. Estimation of Distribution of Gene Expression

To reflect the variability of gene expressions, we estimated the distribution of the gene profiles. We assumed that

1. The distribution of each gene profile follows multivariate normal distribution.
2. Experiments are statistically independent.

Because of the independence of experiments, the covariance matrix of the multivariate normal distribution can be written by only the variance of each experiment. The mean and the variance, which are the parameters of the multivariate normal distribution, were estimated by calculating the mean and the unbiased variance from the replicates of each experiment.

Shohei Maruyama and Yasuo Matsuyama are with the Dept. of Computer Science and Engineering, Waseda Univ. Tokyo, Japan (e-mail: maru371@ruri.waseda.jp).

Sachiyo Aburatani is with the CBRC, National Institute of AIST, Tokyo, Japan.

B. Kullback-Leibker Divergence (KL divergence)

The KL divergence [7] is a measure of the difference between two probability distributions. To calculate the similarities between genes reflecting the variability of gene expressions, we utilized the KL divergence.

For genes A and B, KL divergence is defined to be

$$D_{\mathrm{KL}}(A\|B) = \int_{-\infty}^{\infty} p_A(\mathbf{x}) \ln \frac{p_A(\mathbf{x})}{p_B(\mathbf{x})} d\mathbf{x}, \quad (1)$$

where $\mathbf{x}$ is a gene profile, and p_A and p_B are the probability density functions of each gene's profile.

We calculated the similarities between all gene pairs and created the similarity matrix, whose rows mean gene A and columns mean gene B. Then, we utilized the similarity vectors, which are the rows of the similarity matrix, as the feature vectors of the SVM classifiers.

C. Support Vector Machines (SVM)

To infer whether a gene is involved in a certain metabolic pathway, we trained the SVM classifiers from the similarity vectors, where the similarity vectors were mapped to a higher dimension space by the kernel trick. We define the positive genes as the genes that are involved in the certain pathway, and the negative genes as the genes that are not involved in a certain pathway. Given a similarity vector $\mathbf{x}$ of a certain gene, the SVM method constructs the model as follows:

$$\begin{cases} \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}) + b > 0, & \text{The gene is positive.} \\ \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}) + b < 0, & \text{The gene is negative.} \end{cases} \quad (2)$$

where $\mathbf{w}$ is the vector of coefficients, b is a bias parameter and $\phi(\mathbf{x})$ denotes a feature-space transformation.

Let us suppose that we have a training data set, which consists of N similarity vectors $\mathbf{x}_1$,…, $\mathbf{x}_N$ with the corresponding target values t_1, …, t_N, where t_n is $+1$ when the gene n is positive and t_n is -1 when the gene n is negative. The training algorithm of the soft margin SVMs [4] solves the optimization problem

$$\arg\min_{\mathbf{w},\xi,b} \left\{ \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{n=0}^{N} \xi_n \right\} \quad (3)$$

subject to

$$t_n(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n) + b) \geq 1 - \xi_n,\ \xi_n > 0,\ n = 1, \ldots, N. \quad (4)$$

where C is a constant that controls the error penalties.

The optimization problem (2) can be expressed only in terms of a kernel function $k(\mathbf{x}_n, \mathbf{x}_m) = \phi(\mathbf{x}_n)^T\phi(\mathbf{x}_m)$. Thus, we implicitly mapped the similarity vectors to a higher dimension space by the kernel function.

We utilized the radial basis function (RBF) kernel for SVMs. The RBF kernel is defined as

$$k(\mathbf{x}_m, \mathbf{x}_n) = \exp(-\gamma\|\mathbf{x}_m - \mathbf{x}_n\|^2),\ \gamma > 0. \quad (5)$$

where $\mathbf{x}_m$ and $\mathbf{x}_n$ are the similarity vectors.

To solve the problem of multiclass classification by SVM, we utilized a one versus one classifier. One versus one classifier constructs one classifier per pair of pathways. At prediction time, the pathway that received the most votes is selected.

III. Experimental Design

To evaluate our developed method, we applied the method to *Saccharomyces cerevisiae* and trained the multiclass SVM for identifying the seven metabolic pathways.

A. Gene expression data

We compiled the profiles of 4,783 *Saccharomyces cerevisiae* genes, which were measured in 4,214 experiments by Affymetrix arrays. They were downloaded as raw CEL files from the Gene Expression Omnibus (GEO) database [8]. The raw CEL files were processed by MAS5.0 [9], [10]. Each experiment was normalized with mean 0 and variance 1.

B. Metabolic Pathways

We utilized seven metabolic pathways which are classified at the KEGG PATHWAY database [11]. Both of "Amino acid metabolism" and "Metabolism of other amino acids" are the pathways that are related amino acids. Thus, we merged the two pathways as one "Amino acid metabolisms". Table I shows the list of metabolic pathways and the number of genes that are involved in each metabolic pathway.

TABLE I
List of Metabolic Pathways and the Number of Genes Involved in Each Pathway

Metabolic pathway	# of genes
Carbohydrate metabolism (Crb.)	213
Energy metabolism (Enr.)	106
Lipid metabolism (Lpd.)	117
Nucleotide metabolism (Ncl.)	116
Amino acid metabolism (Amn.)	188
Glycan biosynthesis and metabolism (Gly.)	76
Metabolism of cofactors and vitamins (Vtm.)	110

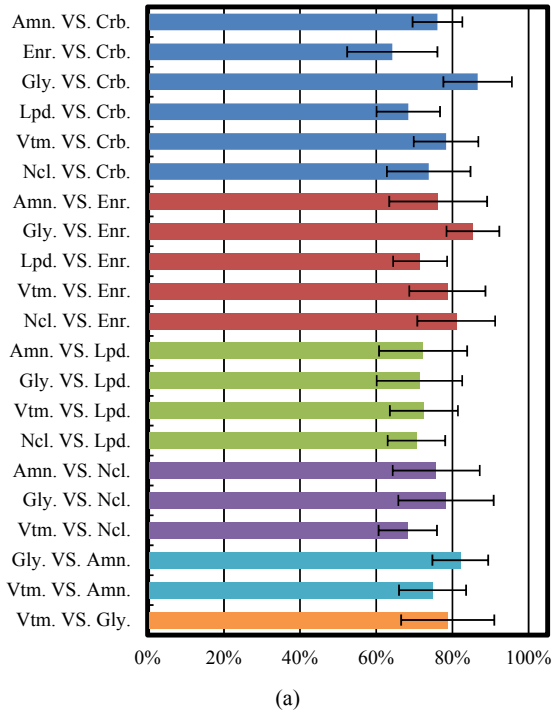
C. Evaluation Measure

To evaluate the classification performances of the binary classifiers, which compose the one versus one classifiers for identification of the genes' metabolic pathway, we utilized accuracy. Similarly, to evaluate the classification performances of the multiclass classifier, we utilized recall. Accuracy and recall are defined as

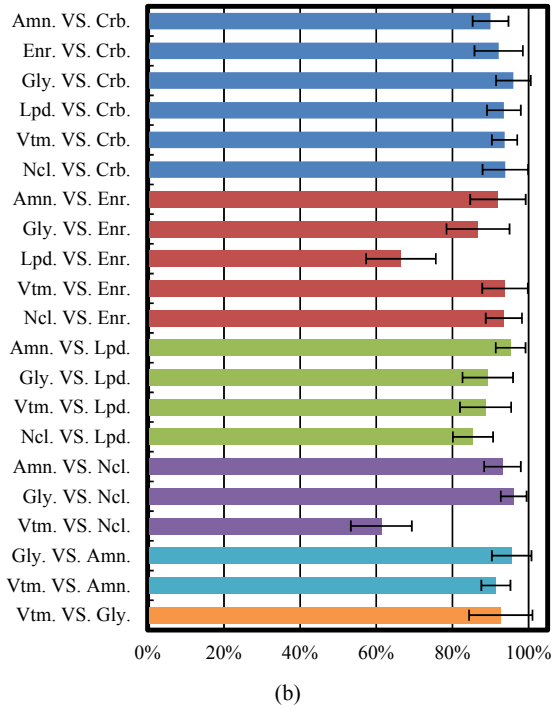
$$\text{accuracy} = \frac{(\text{relevant genes}) \cap (\text{retrieved genes})}{(\text{all genes})}, \quad (6)$$

$$\text{recall} = \frac{(\text{relevant genes}) \cap (\text{retrieved genes})}{(\text{relevant genes})}, \quad (7)$$

where the relevant genes are the genes that are involved in a certain pathway, and the retrieved genes are the genes that are identified as the genes that involved in the pathway.



(a)



(b)

Fig. 1 Accuracies of the binary classifiers that compose the one versus one classifier for identification of the genes' metabolic pathway (a) the accuracies with the raw gene expression data (b) the accuracies with the similarities that are calculated by KL divergence (our method). The vertical axis means the mean of accuracies of 10-fold cross validation. The horizontal axis means the pair of metabolic pathway that each binary classifier identifies. The error bar means the range of (mean) ± (standard deviation)

IV. Results

A. Classification Performances of Binary Classifiers

In Fig. 1, we show the accuracies of the binary classifiers, which compose the one versus one classifiers for identification of the genes' metabolic pathway. Fig. 1 (a) shows the accuracies with the raw gene expression data, and Fig 1 (b) shows the accuracies with the similarities that are calculated by KL divergence. When we trained SVMs using the raw gene expression data, the accuracies of the SVMs are around 70%. The best classifier in (a) is the pair of "Carbohydrate metabolism" and "Glycan biosynthesis and metabolism", whose accuracy is 86.61%. The worst classifier in (a) is the pair of "Carbohydrate metabolism" and "Energy metabolism", whose accuracy is 64.20%.

On the other hand, when we trained SVMs using the similarities that are calculated by KL divergence, the accuracies of the SVMs are around 90%, except "Energy metabolism" versus "Lipid metabolism" and "Nucleotide metabolism" versus "Metabolism of cofactors and vitamins". The best classifier in (b) is the pair of "Carbohydrate metabolism" and "Glycan biosynthesis and metabolism", whose accuracy is 95.98%. The worst classifier in (b) is the pair of "Nucleotide metabolism" and "Metabolism of cofactors and vitamins", whose accuracy is 61.32%.

B. Classification Performances of Multiclass Classifiers

Fig. 2 shows the comparison of the recall of the one versus one classifier for identification of the genes' metabolic pathway. In the entire metabolic pathway except "Energy metabolism", the recall of each metabolic pathway was improved by using KL divergence. The most improved metabolic pathway is "Amino acid metabolism", whose difference of recall is 50.30%. The worst improved metabolic pathway is "Nucleotide metabolism", whose difference of recall is 3.61%. The recall of "Energy metabolism" was not improved; there is no difference between the recall with the raw gene expression data and that with the similarities that are calculated by KL divergence.

The recalls of "Amino acid metabolism", "Carbohydrate metabolism" and "Glycan biosynthesis and metabolism" are higher than 80%. On the other hand, the recall of "Energy metabolism", "Lipid metabolism", "Metabolism of cofactors and vitamins" and "Nucleotide metabolism" are lower than 80%. Thus, the recalls of the four metabolic pathways are low compared to that of the others.

C. Breakdown of Estimated Genes' Metabolic Pathways

Fig. 3 illustrates the breakdown of the metabolic pathways when we estimated the relevant metabolic pathways of the relevant genes of "Energy metabolism". Fig. 3 (a) shows the breakdown with the raw gene expression data, and Fig. 3 (b) shows the breakdown with the similarities that are calculated by KL divergence. The ratio, which "Energy metabolism" genes are identified as "Energy metabolism" genes, is same between the raw gene expression data and the similarities calculated by KL divergence.

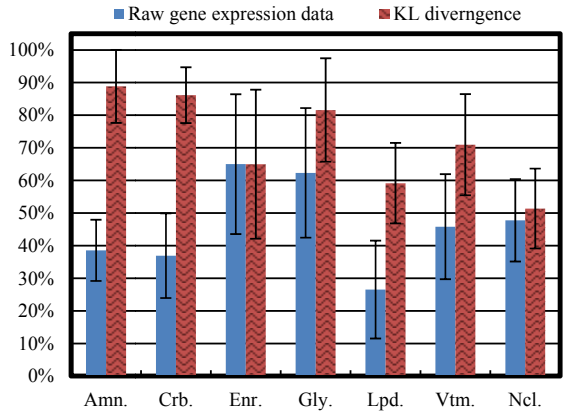


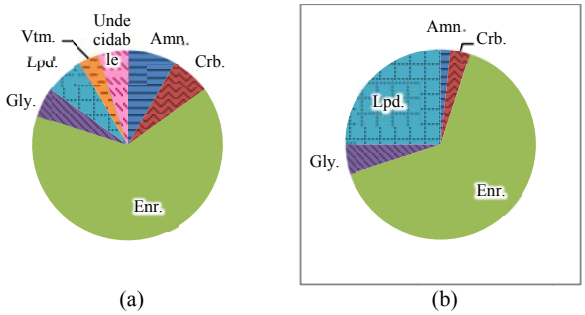
Fig. 2 Comparison of the recall of the one versus one classifier for identification of the genes' metabolic pathway. The blue bars are the recalls with the raw gene expression data. The red bars are the recalls with the similarities that are calculated by KL divergence (our method). The vertical axis means the mean of recalls of 10-fold cross validation. The horizontal axis means the metabolic pathway that genes are involved in. The error bar means the range of (mean) ± (standard deviation)

On the other hand, the ratio that "Energy metabolism" genes are identified as the other metabolism genes is different. When we used the raw gene expression data, genes were identified as various metabolic pathways. When we used the similarities calculated by KL divergence, genes were basically identified as "Lipid metabolism".

V. Discussion

Comparing Figs. 1 (a) and (b), the classification performances of most metabolic pathway pairs were improved by using KL divergence. On the average, the accuracies increased about 14%. This result suggests that it is useful for identifying genes' metabolic pathway to reflect the variability associated with gene expressions in the feature vectors of SVMs by KL divergence.

The only two accuracies, "Energy metabolism" versus "Lipid metabolism" and "Nucleotide metabolism" versus "Metabolism of cofactors and vitamins", were not improved by using KL divergence. The accuracies of these two pairs were low by both KL divergence and raw gene expression data. The Reductive citrate cycle and Phosphate acetyltransferase-acetate kinase pathway in "Energy metabolism" are known to be related with the main flow of Fatty acid biosynthesis in "Lipid metabolism". Those two different pathways were connected by a famous core component "acetyl-CoaA". Furthermore, Uridine monophosphate biosynthesis in "Nucleotide metabolism" shares the one of main flow with Pantothenate biosynthesis in "Metabolism of cofactors and vitamins". Thus, we considered that it is difficult to identify this two pairs from gene expression data. This two pairs need the classifiers based on the different ways from gene expression data.



(a) (b)

Fig. 3 Breakdown of the estimated "Energy metabolism" genes' metabolic pathways (a) the breakdown with the raw gene expression data (b) the breakdown with the similarities that are calculated by KL divergence. "Undecidable" means the ratio of the genes that cannot be identified

As shown in Fig. 2, most classification performances of the one versus one classifier for identification of the genes' metabolic pathway were improved by using KL divergence. Especially, the recalls of "Amino acid metabolism" and "Carbohydrate metabolism" increased about two fold or so. By using the raw gene expression data, the recalls of "Amino acid metabolism" and "Carbohydrate metabolism" were less than 40%. Utilizing KL divergence, the recalls of those two metabolisms became around 90%. Combined KL divergence with SVM is one of the most powerful methods to classify the genes into their metabolic functions.

The lower recalls of some metabolic pathways were considered to be occurred by low comparison with that of the others. This is caused by the low accuracy pairs in Fig. 1 (b). As shown in Fig. 3 (b) many genes of "Energy metabolism" were classified exactly, but identified some genes were identified as the "Lipid metabolism". It is difficult to identify the similar or shared metabolic pathways by our binary classifier. Thus, if we obtain the two exact classifiers ("Energy metabolism" versus "Lipid metabolism", "Nucleotide metabolism" versus "Metabolism of cofactors and vitamins"), we will achieve high classification performances in "Energy metabolism", "Lipid metabolism", "Metabolism of cofactors and vitamins" and "Nucleotide metabolism".

VI. Conclusions

We have proposed a new method based on the SVM approach, for identifying the genes' metabolic pathway from the gene expression data. To improve classification performances of SVMs, we calculated the similarities between genes by KL divergence, and utilized them as the feature vectors of SVMs. By using KL divergence, one can reflect the variability associated with gene expressions in the feature vectors of SVMs. Then, we applied our method to the *Saccharomyces cerevisiae* gene expression data against seven metabolic pathways, and evaluated its classification performances. As a result, the classification performances using our method were higher than that using the raw gene expression data in most metabolic pathways. Thus, our method is useful for identifying the metabolic pathway that genes are involved in.

REFERENCES

[1] T. Obayashi, Y. Okamura, S. Ito, S. Tadaka, Y. Aoki, M. Shirota, and K. Kinoshita, "ATTED-II in 2014: evaluation of gene coexpression in agriculturally important plants," *Plant Cell Physiol.*, vol. 55, no. 1, p. e6, Jan. 2014.

[2] K. Aoki, Y. Ogata, and D. Shibata, "Approaches for extracting practical information from gene co-expression networks in plant biology," *Plant Cell Physiol.*, vol. 48, no. 3, pp. 381-390, Mar. 2007.

[3] K. Saito, M. Y. Hirai, and K. Yonekura-Sakakibara, "Decoding genes with coexpression networks and metabolomics - 'majority report by precogs'," *Trends Plant Sci.*, vol. 13, no. 1, pp. 36-43, Jan. 2008.

[4] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, pp. 273-297, 1995.

[5] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machine," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1-27, Apr. 2011.

[6] M. P. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares, and D. Haussler, "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 97, no. 1, pp. 262-267, Jan. 2000.

[7] S. Kullback, and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 79-86, 1951.

[8] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Res.*, vol. 30, pp. 207-210, 2002.

[9] E. Hubbell, W. M. Liu, and R. Mei, "Robust estimators for expression analysis," *Bioinformatics*, vol. 18, pp. 1585-1592, 2002.

[10] S. D. Pepper, E. K. Saunders, L. E. Edwards, C. L. Wilson, and C. J. Miller, "The utility of MAS5 expression summary and detection call algorithms," *BMC Bioinformatics*, vol. 8, p. 273, 2007.

[11] M. Kanehisa, S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe, "Data, information, knowledge and principle: back to metabolism in KEGG," *Nucleic Acids Res.*, vol. 42, no. Database issue, pp. D199-205, Jan. 2014.