Proc. of the Second Intl. Conf. on Advances In Bio-Informatics, Bio-Technology and Environmental Engineering- ABBE 2014.

ISBN: 978-1-63248-053-8 doi: 10.15224/ 978-1-63248-053-8-07

Inferring Genes Involved in Metabolic Pathways by Using Support Vector Machines

Shohei Maruyama, Yasuo Matsuyama, Sachiyo Aburatani

***Abstract*—The development of a method to annotate unknown gene functions is an important task in bioinformatics. The identification of the relevant genes to metabolic pathways is also helpful for understanding the genes. However, the relationships between metabolic pathways and genes are complicated. Thus, it is difficult to identify the relevant genes by linear models. In this study, we propose a new method based on the SVM approach, for inferring the genes involved in metabolic pathways from the gene expression profiles. To improve the classification performances of SVMs, we developed a method for finding the important interactions for classification, from a huge number of experiment combinations. The interactions selected by our method were added as new features to the training data set of the SVMs. Furthermore, feature selection by the Gini importance was applied, to avoid overlearning of the SVMs. To demonstrate the validity of our method, we trained SVMs with *Saccharomyces cerevisiae* gene expression profiles against eight metabolic pathways, and evaluated their classification performances. As a result, we achieved high performances with some metabolic pathways. Thus, our method is useful for inferring the relevant genes to metabolic pathways.**

Keywords*—metabolic pathways, gene involved in, gene expression profiles, microarray, interaction, support vector machines, SVM, Gini importance, random forests, machine learning, *Saccharomyces cerevisiae

I. Introduction

For understanding life system, it is important to identify the genes that are involved in metabolic pathways. Because gene expression profiles reflect various intracellular phenomena, gene expression profiles are useful for revealing the relevant genes to metabolic pathways. Pearson product-moment correlation coefficient has been utilized to gene expression profiles for revealing the relevant genes [1-3]. The method is based on the idea that the coexpression genes have similar function. However, the correlation coefficient can express only linear relationship between genes. Thus, we cannot utilize the method to infer the relevant genes which 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 profiles 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.

In this report, we propose a method based on the SVM approach, for inferring the relevant genes to metabolic pathways from the gene expression profiles. Since the relationships between metabolic pathways and genes are complicated, various gene expression profiles are needed to infer the relationships. However, it is costly to measure the expression values under numerous new experimental conditions. Instead of adding new experiments to the gene expression profiles, we introduced the interaction between experiments to the SVM training data set. Since there were too many experiment combinations to add to the data set, we developed a new method to find the important interactions for classification from a huge number of them. Before we trained the SVMs, we performed feature selection by the Gini importance [7-8], to avoid overlearning of the SVMs. To show the validity of our method, we then trained the SVMs with the gene expression profiles of *Saccharomyces cerevisiae* against eight metabolic pathways defined by KEGG, and evaluated their classification performances.

II. METHODS

A. *Gene Expression Profiles*

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 [9]. The raw CEL files were processed by MAS5.0 [10-11]. Each experiment was normalized with mean 0 and variance 1.

B. *Metabolic Pathway*

We utilized eight metabolic pathways which are classified at the KEGG PATHWAY database [12]. Table 1 shows the list of metabolic pathways and the number of genes involved in each metabolic pathway.

C. *Support Vector Machines*

To infer whether a gene is involved in a certain metabolic pathway, we trained the SVM classifiers from the gene expression profiles, where the profiles 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

Shohei Maruyama *(Author)*, Yasuo Matsuyama
Dept. of Computer Science and Engineering, Waseda Univ.
Tokyo, Japan.

Sachiyo Aburatani
CBRC, National Institute of AIST.
Tokyo, Japan.

Proc. of the Second Intl. Conf. on Advances In Bio-Informatics, Bio-Technology and Environmental Engineering- ABBE 2014.

ISBN: 978-1-63248-053-8 doi: 10.15224/ 978-1-63248-053-8-07

TABLE 1. LIST OF METABOLIC PATHWAYS AND THE NUMBER OF GENES INVOLVED IN EACH PATHWAY

Metabolic pathway	# of genes
Carbohydrate metabolism (Crb.)	205
Energy metabolism (Enr.)	103
Lipid metabolism (Lpd.)	116
Nucleotide metabolism (Ncl.)	116
Amino acid metabolism (Amn.)	157
Metabolism of other amino acids (Oth.)	49
Glycan biosynthesis and metabolism (Gly.)	75
Metabolism of cofactors and vitamins (Vtm.)	107

involved in a certain pathway. Given a profile $\mathbf{x}$ of a gene, the SVM method constructs the model as follows:

$$\begin{cases} \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}) + b > 0, & \text{The gene is positive.} \\ \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}) + b < 0, & \text{The gene is negative.} \end{cases}, \quad (1)$$

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 profiles $\mathbf{x}_1,\ldots,\mathbf{x}_N$ with the corresponding target values $t_1, \ldots, t_N$, 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

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

subject to

$$t_n(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_n) + b) \geq 1 - \xi_n,\ \xi_n > 0,\ n = 1, \ldots, N. \quad (3)$$

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 profiles to a higher dimension space by the kernel function.

D. *Feature Selection by the Gini Importance*

The compiled gene expression profiles included irrelevant experiments, which lead to overfitting. We utilized the Gini importance to select only the relevant experiments before model construction. The Gini importance [7-8] gives a relevance score to each experiment. The Gini importance is based on the classification algorithm, random forests [7]. Random forest classifiers are composed of decision trees.

CART: Classification and regression trees (CART) [13] are a type of decision trees used by random forests. The CART method builds binary decision trees based on the impurity of each node. The Gini impurity, which is one of the node impurity measures used in CART, is defined as

$$i(\tau) = 1 - \left(\frac{n_\tau^{(+)}}{n_\tau}\right)^2 - \left(\frac{n_\tau^{(-)}}{n_\tau}\right)^2, \quad (4)$$

where $n_\tau^{(+)}$ and $n_\tau^{(-)}$ are the numbers of positive genes and negative genes at a node τ, and n_τ is the total number of genes at node τ. The Gini impurity $i(\tau)$ reaches 0 when the ratio of classes of genes at the node τ inclines to one class. Thus, we can measure how well a potential split is separating the samples of the two classes at the node τ by calculating $i(\tau)$.

The decrease $\Delta i(\tau)$ of the Gini impurity is calculated from splitting the node τ to two sub nodes, τ_l and τ_r, by a threshold t_θ on a variable θ, as follows:

$$\Delta i(\tau) = i(\tau) - \left\{\frac{n_{\tau_l}}{n_\tau} i(\tau_l) + \frac{n_{\tau_r}}{n_\tau} i(\tau_r)\right\}, \quad (5)$$

where n_{τ_l} and n_{τ_r} are the numbers of genes at nodes τ_l and τ_r, respectively. At each node, the decrease $\Delta i(\tau)$ is calculated over all variable θ and all available threshold t_θ. Then, we determine $\theta^\star$ and $t_\theta^\star$, which maximize $\Delta i(\tau)$, and we split each node by $t_\theta^\star$ on $\theta^\star$ and make the decision tree grow.

Random Forests: Random forests are an ensemble learning method for classification. The main idea of random forests is to obtain better predictive performance by combining many weak decision trees. The training algorithm of random forests repeatedly builds CART on bootstrap samples with random subsets of the experiments.

Gini Importance: Important experiments for building decision trees decrease the Gini impurity greatly. The total decrease in Gini impurity of an experiment yields the importance of the experiment. The Gini importance $I_G(\theta)$ of an experiment θ is defined as the total decrease in the Gini impurity of the experiment for all nodes τ in all trees T*:*

$$I_G(\theta) = \sum_T \sum_\tau \Delta i_\theta(\tau, T), \quad (6)$$

where $\Delta i_\theta(\tau, T)$ is the decrease in the Gini impurity that results from a split on experiment θ.

E. *Selection of Interaction Terms*

To select the important interaction terms from the huge number of experiment combinations, before we apply feature selection by the Gini importance, we developed a new method for improving the SVM classifiers, by adding the interaction terms between experiments to the profiles as additional dimensions. In our method, there are four steps.

Step 1. We discretize the gene expression profiles to focus on whether the genes are actually significantly expressed. Let $\mathbf{X} = (x_{nd})$ be an $N \times D$ matrix, which includes the gene expression profiles of N genes measured

Proc. of the Second Intl. Conf. on Advances In Bio-Informatics, Bio-Technology and Environmental Engineering- ABBE 2014.

ISBN: 978-1-63248-053-8 doi: 10.15224/ 978-1-63248-053-8-07

in D experiments. The gene expression value of gene n and experiment d is discretized by the threshold t_x, as follows:

$$x'_{nd} = \begin{cases} -1 & x_{nd} < -t_x \\ 0 & -t_x \le x_{nd} \le +t_x \\ +1 & +t_x < x_{nd} \end{cases}, \quad n = 1, \dots, N, \;\; d = 1, \dots, D. \tag{7}$$

Step 2. We find the target vectors. The target vector of experiment d is defined as the vector that has a perfect interaction with the discretized expression vector $(x'_{1d}, \cdots, x'_{Nd})^{\mathrm{T}}$. Thus, when an experiment vector is similar to the target vector, it is assumed that the experiment has an interaction with the experiment d. The target vector $\mathbf{X}^{\star} = (x^{\star}_{nd})$ is defined as satisfying

$$x'_{nd} \times x^{\star}_{nd} = y_n, \;\; n = 1, \dots, N, \;\; d = 1, \dots, D. \tag{8}$$

Step 3. We measure the similarities between the target vectors and the discretized gene expression profiles. Inner products were utilized as the method to measure the similarities. A smaller inner product of two experiments indicates more similarity between the experiments. The similarity **S** is then calculated by

$$\mathbf{S} = \mathbf{X}^{\star\mathrm{T}}\mathbf{X}'. \tag{9}$$

We replace the diagonal elements of **S** with 0.

Step 4. We obtain the value $v_n^{dd'}$ of the interaction term between the experiments d and d' of gene n, as follows:

$$v_n^{dd'} = x_{nd} \times x_{nd'}, \quad n = 1, \dots, N, \quad d = 1, \dots, D, \;\; d' \in \{d' \mid t_s > |s_{dd'}|\}, \tag{10}$$

where d' is the number of the experiment that is similar to the target vector of the experiment d, and the parameter t_s is the threshold for determining whether experiments d and d' have an interaction.

F. *Experimental Design*

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

$$k(\mathbf{x}_n, \mathbf{x}_m) = \exp(-\gamma \|\mathbf{x}_n - \mathbf{x}_m\|^2), \;\; \gamma > 0. \tag{11}$$

We set the parameter γ to $1/D'$, where D' is the number of elements of the SVM input vector. The parameter C of SVM is set to 1.

Since the number of negative genes is larger than the number of positive genes, we performed down-sampling against the negative genes, to equalize the numbers of negative genes positive genes. We performed the down-sampling 3 times, and calculated the average of the accuracies by 10-fold cross validation against each negative gene sample (that is, we constructed 30 SVM classifiers with each metabolic pathway and calculated 10 accuracies). The performance of each metabolic pathway's classifier was reported by the average of the 10 accuracies.

We repeatedly constructed 500 CARTs in the random forest learning. The $\sqrt{D}$ experiments were selected randomly in each construction, where D is the number of experiments before the random selection. We reported the classification performance of each metabolic pathway in the case where we set the numbers of selected experiments from 10 to 3,000 against the original gene expression profiles. We also reported the performance in the case where we set the numbers of selected experiments from 10 to 1,000 against the gene expression profiles that included the interaction terms.

We set t_x to 0.0, 0.4, 0.6 and 1.0, and selected the interaction terms as the top 1,000 d', instead of choosing the threshold t_s. The original experiment vectors and the selected interaction term vectors were combined into one matrix, as the new gene expression profiles. When we combined them, we used only the original experiment vectors with the Gini importances that were within the top 500. We utilized the feature selection by the Gini importance to assess the new gene expression profiles.

III. Results

A. *Feature Selection by the Gini Importance*

We calculated the Gini importance of each experiment from the experiment vectors. Fig. 1 shows the histogram of the Gini importances. The details about the number of experiments in each metabolic pathway are shown in Table 2. In all of the metabolic pathways, the Gini importances of most experiments were less than 0.05. The number of unimportant experiments was the largest in "metabolism of other amino acids": there were 3,747 experiments with the Gini importances that were less than 0.05. On the other hand, the number of unimportant experiments was the smallest in carbohydrate metabolism: there were 2,840 experiments with the Gini importances that were less than 0.05. The correlation coefficient between the number of experiments with the Gini importances that were less than 0.05 and the number of genes in each metabolic pathway was -0.99.

In Fig. 2, we show the accuracy of each metabolic pathway in the case where we varied the number of selected

TABLE 2. THE NUMBER OF UNIMPORTANT EXPERIMENTS

Metabolic pathway	Gini importance	
	< 0.05	*0.05 ≤*
Carbohydrate metabolism	2,840	1,374
Energy metabolism	3,473	741
Lipid metabolism	3,327	887
Nucleotide metabolism	3,326	888
Amino acid metabolism	3,083	1,131
Metabolism of other amino acids	3,747	467
Glycan biosynthesis and metabolism	3,601	613
Metabolism of cofactors and vitamins	3,382	832

Proc. of the Second Intl. Conf. on Advances In Bio-Informatics, Bio-Technology and Environmental Engineering- ABBE 2014.

ISBN: 978-1-63248-053-8 doi: 10.15224/ 978-1-63248-053-8-07

experiments from 10 to 3,000. In all of the metabolic pathways, the highest accuracy was achieved when the number of experiments was set between 100 and 400. The accuracies decreased when the number of experiments exceeded 400. The largest decrease was measured in "glycan biosynthesis and metabolism". Its highest accuracy was 86.2%, and its lowest accuracy over 400 experiments was 76.4%. The difference in the accuracies was 9.8%. On the other hand, the smallest decrease was measured in "energy metabolism". Its highest accuracy was 78.7%, and its lowest accuracy over 400 experiments was 76.9%. The difference in the accuracies was 1.8%. The correlation coefficient between the difference and the number of genes in each metabolic pathway was -0.14.

To show the effect of feature selection, we compared the accuracy against all experiments and the highest accuracy in Fig. 2, for each metabolic pathway. The detailed comparison is shown in Fig. 3. In all of the metabolic pathways, the highest accuracy in Fig. 2 was higher than the accuracy against all experiments. The best improvement in the accuracy was achieved in "lipid metabolism": the accuracy of "lipid metabolism" against experiments selected by the Gini importance was 72.9%, which was 11.5% higher than its accuracy against all experiments. On the other hand, the worst improvement in the accuracy was measured in "carbohydrate metabolism": the accuracy of "carbohydrate metabolism" against experiments selected by the Gini importance was 76.0%, which was only 2.1% higher than its accuracy against all experiments.

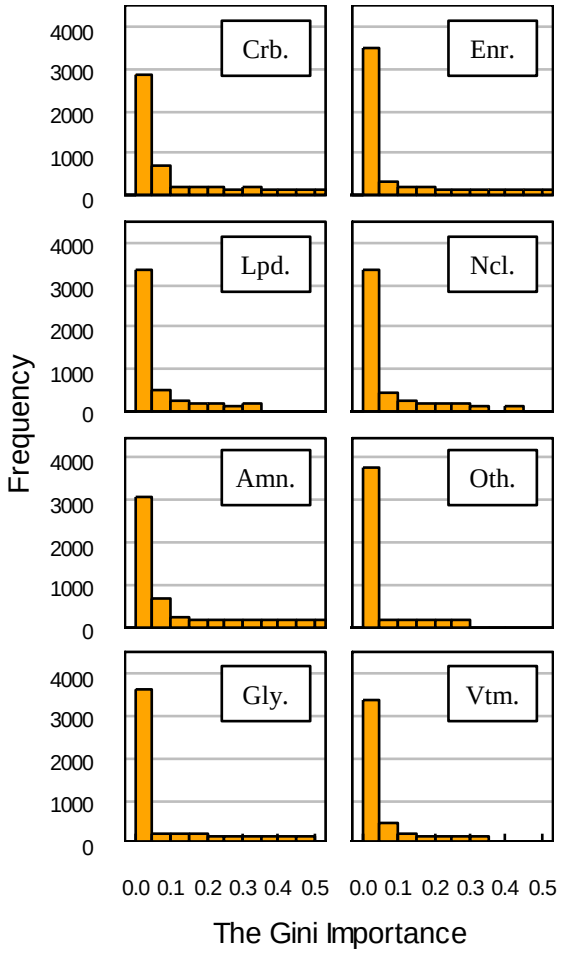

Fig. 1. Histograms of the Gini importances, calculated from only the experiment vectors in each metabolic pathway. The bin width was set to 0.05, which was selected as the majority of the widths suggested by Sturges' formula in each metabolic pathway.

B. *Model Construction Including Internal Terms*

The histogram of $|s_{dd'}|$ is shown in Fig. 4, where t_x was set to 0.0. The details of the ideal and actual max values of $|s_{dd'}|$ in each metabolic pathway are shown in Table 3. The ideal max value of $|s_{dd'}|$ is equal to the number of genes in each metabolic pathway. In "metabolism of other amino acids" and "glycan biosynthesis and metabolism", the actual max value is greater than half of the ideal max value. On the other hand, in the other metabolic pathways, the actual max value is less than half of the ideal max value. These results suggest that none of the experiment vectors were similar to the target vectors. Thus, there were no experiments that strongly interact with other experiments.

We calculated the Gini importance of each experiment from the new gene expression profiles, which consisted of experiment vectors and interaction vectors. Fig. 5 shows the histogram of the Gini importances. The details of the number of experiments in each metabolic pathway are shown in Table

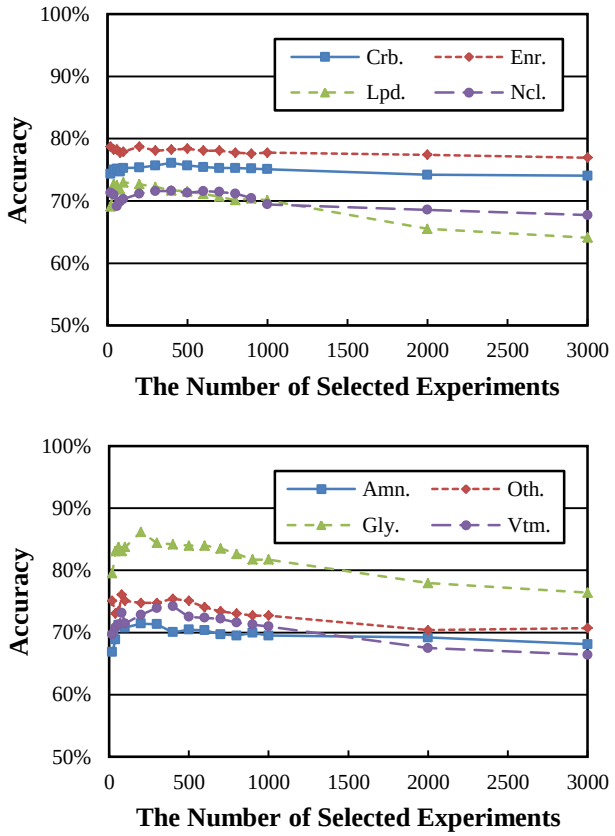

Fig. 2. Accuracy of each metabolic pathway when the number of selected experiments was set to 20, 40, 60, 80, 100, 200, 300, 400, 500, 600, 700, 800, 900, 1,000, 2,000, and 3,000.

Proc. of the Second Intl. Conf. on Advances In Bio-Informatics, Bio-Technology and Environmental Engineering- ABBE 2014.

ISBN: 978-1-63248-053-8 doi: 10.15224/ 978-1-63248-053-8-07

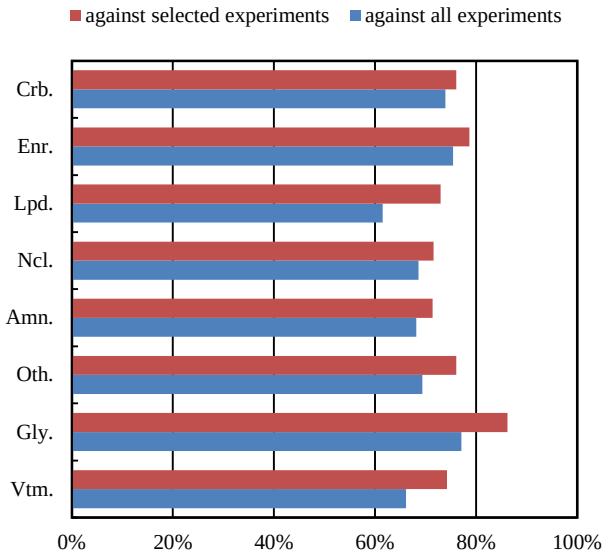

Fig. 3. Comparison of accuracy against all experiments with that against selected experiments. For each metabolic pathway, the highest accuracy in Fig. 2 is shown as the accuracy against selected experiments.

4. Half of the Gini importances were less than 0.05 in four metabolic pathways: "carbohydrate metabolism", "energy metabolism", "metabolism of other amino acids" and "glycan biosynthesis and metabolism". The number of unimportant experiments was the largest in "metabolism of other amino acids": there were 1,200 experiments with the Gini importances that were less than 0.05. On the other hand, the number of unimportant experiments was the smallest in "amino acid metabolism": there were 380 experiments with the Gini importances that were less than 0.05.

The two types of accuracy were compared, as shown in Fig. 6: one is the accuracy calculated from only the expression vectors, and the other is the accuracy calculated from the expression vectors and interaction vectors. The accuracy of each metabolic pathway indicates the maximum value of the accuracies which were calculated in the case where we varied the number of selected experiments and interaction terms from 10 to 1,000. The accuracy was improved in six metabolic pathways: "energy metabolism", "lipid metabolism", "nucleotide metabolism", "amino acid metabolism", "metabolism of other amino acids" and "metabolism of cofactors and vitamins". On the other hand, the accuracy was not improved in "carbohydrate metabolism" and "glycan biosynthesis and metabolism". In all metabolic pathways with improved accuracies, the highest accuracy was achieved when the threshold t_s was set to either 0.4 or 0.6.

TABLE 3. MAX VALUE OF SIMILARITY $|s_{dd'}|$ IN EACH METABOLIC PATHWAY

Metabolic pathway	Max Value		Actual / Ideal
	Ideal	*Actual*	
Carbohydrate metabolism	205	85	0.41
Energy metabolism	103	46	0.45
Lipid metabolism	116	46	0.40
Nucleotide metabolism	116	46	0.40
Amino acid metabolism	157	65	0.41
Metabolism of other amino acids	49	38	0.78
Glycan biosynthesis and metabolism	75	44	0.59
Metabolism of cofactors and vitamins	107	44	0.41

IV. Discussion

Most experiments were not important for classification, since their Gini importances were less than 0.05, as shown in Fig. 1. The number of unimportant experiments was negatively correlated with the number of genes, because the correlation coefficient between them was -0.99. From this result, we found that the experiments were selected depending on the number of genes in the training data set, rather than the metabolic pathways.

The accuracies decreased when unimportant experiments were added to the gene expression profiles, as shown in Fig. 2. This result suggests that the unimportant experiments caused overlearning of the SVM classifiers. The decrease in the

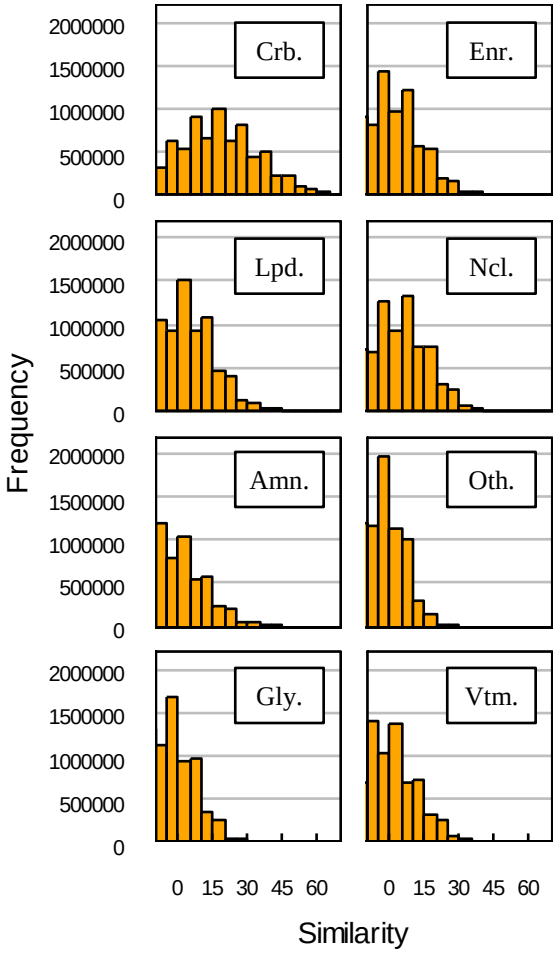

Fig. 4. Histograms of $|s_{dd'}|$ in the case where t_x was set to 0.0. The bin width was set to 5, which was selected as the majority of the widths suggested by Sturges' formula in each metabolic pathway.

Proc. of the Second Intl. Conf. on Advances In Bio-Informatics, Bio-Technology and Environmental Engineering- ABBE 2014.

ISBN: 978-1-63248-053-8 doi: 10.15224/ 978-1-63248-053-8-07

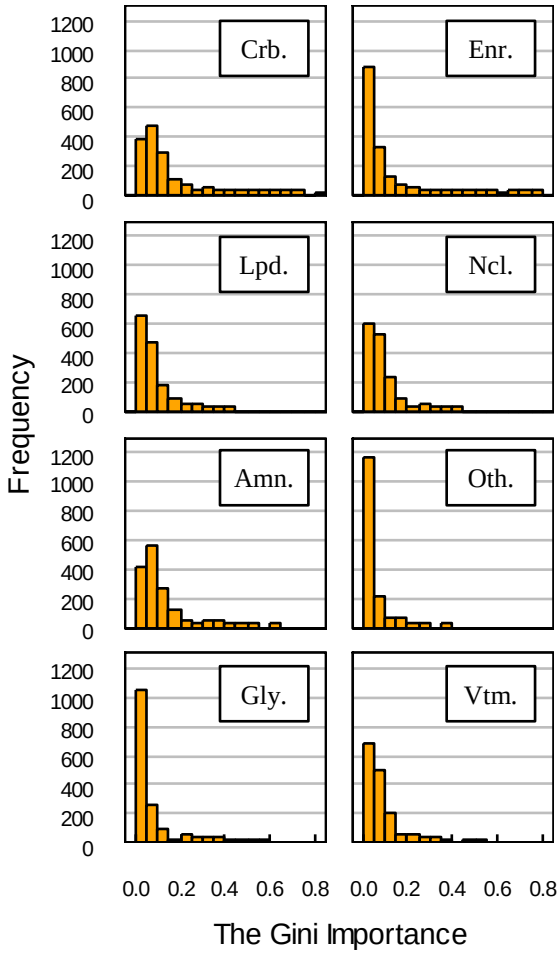

Fig. 5. Histograms of the Gini importances calculated from experiment vectors and interaction vectors in each metabolic pathway. The bin width was set to 0.05, which was selected as the majority of the widths suggested by Sturges' formula in each metabolic pathway.

accuracy was not correlated with the number of genes, because the correlation coefficient between them was -0.14. Thus, we considered that the overlearning was caused depending on the metabolic pathways, rather than on the number of genes in the training data set. Since the accuracies did not increase when we added experiments, except when the number of experiments was small, the Gini importance exactly expressed the importance of the experiment for classification. As shown in Fig. 3, the accuracies were improved in all of the metabolic pathways. Therefore, the feature selection by the Gini importance improved the classification performance of the SVM classifiers.

TABLE 4. THE NUMBER OF UNIMPORTANT EXPERIMENTS AND INTERACTIONS

Metabolic pathway	Gini importance	
	< 0.05	*0.05 ≤*
Carbohydrate metabolism	383	1,117
Energy metabolism	886	614
Lipid metabolism	658	842
Nucleotide metabolism	594	906
Amino acid metabolism	411	1,089
Metabolism of other amino acids	1,171	329
Glycan biosynthesis and metabolism	1,057	443
Metabolism of cofactors and vitamins	674	826

The accuracies were enhanced by adding the interaction terms between experiments, as shown in Fig. 6. Since the highest accuracies were achieved when we set the threshold t_x to either 0.4 or 0.6, the pair of interacting experiments can be found by focusing on only the genes that are significantly expressed. On the other hand, the accuracies got worse when we set the threshold t_x to a large value, such as 1.0. This result means that we cannot find the interacting experiment pair when we focus on too few genes.

v. Conclusions

We have proposed a new method based on the SVM approach, for inferring the genes involved in metabolic pathways from the gene expression profiles. To improve classification performances of SVMs, we developed a method for finding the important interactions for classification, from a huge number of experiment combinations. The interactions selected by our method were added as new features to the training data set of SVMs.

We trained SVMs with the *Saccharomyces cerevisiae* gene expression profiles against eight metabolic pathways, and evaluated their classification performances. As a result, we achieved high performances in some metabolic pathways. Thus, our method is

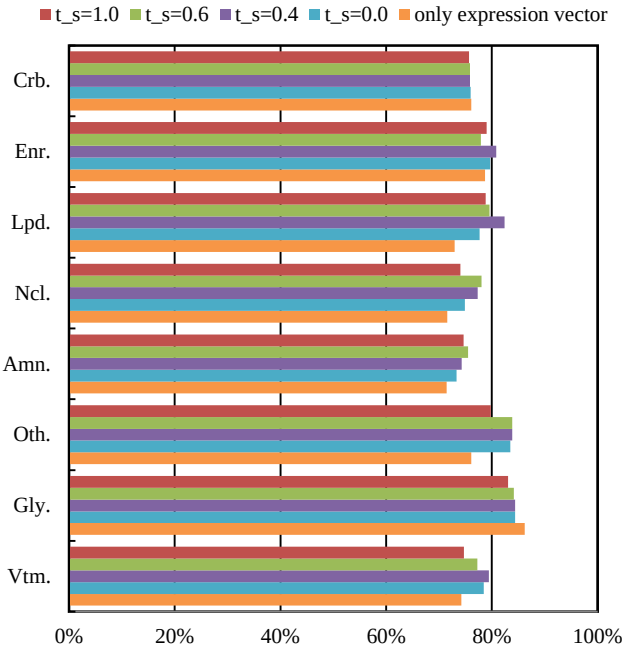

Fig. 6. Comparison of the accuracy calculated from the experiment vectors with the accuracy calculated from the experiment vectors and interaction vectors. The accuracy of each metabolic pathway indicates the maximum value of the accuracies, calculated in the case where we varied the number of selected experiments and interaction terms from 10 to 1,000.

Proc. of the Second Intl. Conf. on Advances In Bio-Informatics, Bio-Technology and Environmental Engineering- ABBE 2014.

ISBN: 978-1-63248-053-8 doi: 10.15224/ 978-1-63248-053-8-07

useful for inferring the relevant genes to metabolic pathways.

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] L. Breiman, "Random forests," Mach. Learn., vol. 45, pp. 5-32, 2001.

[8] B. H. Menze, B. M. Kelm, R. Masuch, U. Himmelreich, P. Bachert, W. Petrich, and F. A. Hamprecht, "A comparison of random forest and its Gini importance with standard chemometric methods for the feature selection and classification of spectral data," BMC Bioinformatics, vol. 10, p. 213, Jan. 2009.

[9] 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.

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

[11] 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.

[12] 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.

[13] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, Classification and Regression Trees, vol. 19. 1984, p. 368.

About Author:

Shohei Maruyama

Master student of Waseda University.

Research interest:
Machine learning, metabolic pathway, gene regulatory networks.