Multiple kernel support vector machine can generate weights of feature matrices for toxicity prediction

HIROKI TAKAHASHI[†1] JUNKO YAMANE[†1]
WATARU FUJIBUCHI[†1]

Abstract: In order to reduce drug discovery period and costs, development of prediction system for toxicities and effects of medicine by artificial intelligence (AI) is expected. In our laboratory, a toxicity prediction system with multiple kernel support vector machine (MK-SVM) was constructed using SHOGUN library. In this study, sub-kernel matrices are created from three feature matrices (qRT-PCR expression values, correlations between genes by Bayesian network, and structure-activity relationships of each compound quantitated by E-Dragon) and a kernel matrix generated by the linear sum of these sub-kernel matrices was used for SVM prediction. Weights of each sub-kernel matrix in the linear sum indicate the contribution degree of each feature matrix in the classification. Therefore, focusing on the weights, we discuss whether each feature matrix can correctly predict toxicity or not.

Keywords: Artificial intelligence, Toxicity prediction, Support vector machine, Multiple kernel learning

1. Introduction

In recent years, drug discovery period and cost are increasing. The decline in the success rate of new drug development leads to worsening business of pharmaceutical industry, which leads to increase in the price of drugs as a countermeasure. In fact, Nivolumab, marketed drug for cancer immunotherapy recently, has become a problem for its high price [1]. Too high treatment cost significantly limits the patient's treatment options. In order to overcome the problem, drug discovery support systems by a computer have been developed. By predicting drug efficacy and toxicity before experiments, it is expected to reject unnecessary experiments and reduce drug discovery period and cost.

We aim to develop a prediction system for compound toxicity to human by artificial intelligence (AI), but it is difficult to predict toxicity to humans. The conventional method, toxicity tests by animal experiments, has very low success rate. If any these tests are passed, unidentified side effects are sometimes found later. However, it is impossible to conduct toxicity tests on humans. Therefore, we are developing a toxicity prediction system using human embryonic stem (hES) cells [2]. Of course, there is still ethical problem, so we paid maximum attention according to the manual of the ethics committees.

In this toxicity prediction system, the differential pattern of the gene expression level in hES cells between before and after compound exposure is regarded as feature values, and the presence or absence of an interested toxicity of the compound is predicted. In addition, relationship information data between genes are also prepared by the Bayesian network method as feature values. These feature values are input to support vector machine (SVM) classifier and are used as indicators of two-class classification of toxicity. We predicted three types of toxicity; neurotoxin (NT), genotoxic carcinogen (GC), and non-genotoxic carcinogen (NGC) to hES cells. Our toxicity prediction system succeeded in predicting these toxicities with accuracy rates of 95%, 100%, 95%, respectively.

†1 Theoretical Cell Science Lab, Center for iPS Cell Research and Application, Kyoto University, Japan

Based on these results, we are developing new toxicity prediction method with multiple kernel support vector machine (MK-SVM) in this study. SVM is an algorithm for finding a data-separating hyperplane that maximizes the margin between the hyperplane to the nearest neighbor vectors, what are called "support vector". Instead of solving the main problem of maximizing this margin, we can obtain the separation hyperplane by solving the dual problem with Lagrangian function [3].

$$\max_{\boldsymbol{\alpha}} \tilde{L}(\boldsymbol{\alpha}) = \max_{\boldsymbol{\alpha}} \left\{ \sum_{i=0}^{N-1} \alpha_i - \sum_{i=0}^{N-1} \sum_{j=0}^{N-1} \alpha_i y_i \alpha_j y_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \right\}$$

$$s.t.\ \ 0 \le \alpha_i \le C, \sum_{i=0}^{N-1} \alpha_i y_i = 0$$

Above equation means that each element $k(\boldsymbol{x}_i,\boldsymbol{x}_j)$ of the kernel matrix is used for optimization of the hyperplain. These element $k(\boldsymbol{x}_i,\boldsymbol{x}_j)$ is the product of two basis function $\varphi(\boldsymbol{x}_i)$ and $\varphi(\boldsymbol{x}_j)$. There exist various kernel functions for creating a kernel matrix, such as linear kernel, polynomial kernel, and Gaussian kernel.

The multiple kernel learning is a learning method with one kernel produced by the linear sum of sub-kernel matrices [4]. In this study, we created three sub-kernel matrices from three feature data (qRT-PCR gene expression value, Bayesian network feature values and quantitative structure-activity relationship feature values) and classify compounds into two classes using a kernel matrix based on these sub-kernel matrices.

$$k(x_i, x_j) = \sum_{i=1}^{K} \beta_k k_i(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

$$s.t.\ \beta_k > 0, \sum_{k=1}^{K} \beta_k = 1$$

In the above equation, β_i means the weight of the ith sub-kernel. As β increases, the proportion in the generated kernel matrix increases. Therefore, β can be regarded as the index parameter of classification contribution of the subkernel matrix. That is, when calculate a maximum prediction rate, the sub-kernel

matrix with the largest β is capable of predicting the presence or absence of toxicity. According to the viewpoint, we consider the classification contribution of each sub-kernel matrices using weight β.

2. Method

2.1 Data preparation and normalization

Data preparation and normalization were carried out according to [2]. The data includes 9 neurotoxins (valproic acid, cyclopamine, phenytoin, methylmercury, acrylamide, 4-OH-2',3,3',4',5'-PCB, 2,5-hexanedione, warfarin, thalidomide), 5 genotoxic carcinogens (benzo[a]anthracene, 3-methylcholanthrene, benzo [a]pyrene, diethylnitrosamine, diethylstilbestrol), and 6 non-genotoxic carcinogens (2,3,8-tetrachlorodibenzodioxin (TCDD), lithocholic acid, thioacetamide, butylated hydroxyanisole, methapyrilene hydrochloride, phenobarbital) to predict its toxicity. The dataset can be downloaded from our website (http://stemcellinformatics.org/toxicology/). The qRT-PCR gene expression data was fitted to an empirical Bayesian linear model to exclude batch effects, and normalized by dividing them by the expression level of ACTB as internal standard.

We can use the normalized data as feature amount data, or use intergenic correlation data created by RX-TAOgen, which can construct Bayesian network (BN) based on the Gibbs sampling [5]. As a conventional feature dataset, 1,665 feature values of quantitative structure-activity relationship (QSAR) were prepared by E-Dragon website (http://www.vcclab.org/lab/edragon/) [6].

2.2 Feature extraction based on the result of previous study

The qRT-PCR gene expression data is composed of 10 genes, 5 doses (1, 1/2, 1/4, 1/8 and 1/16 with the maximum non-toxic dose as 1), and 4 time points (24, 48, 72 and 96 hours after exposure), so it has $10 \times 5 \times 4 = 200$ features per one data point. In the BN dataset, there exists 10 genes × 10 genes = 100 relationship network information values as features. Since we performed twice for each compound, there are a total of 40 data points. If we use all features, the number of features is far exceeded than the number of data points, which leads to decline of prediction accuracy. In order to optimize the number of features, we calculated the accuracy rates for each number (from 1 to N; N is the total number of features). Based on the results of the previous study [2], we extracted feature values used for the subsequent analysis with the optimum numbers (Figure 1).

2.3 Running multiple kernel support vector machine

Sub-kernel matrices are created from the above three features, and one combined kernel matrix produced by the linear sum of them is used to perform two-class classification. The subsequent procedures were performed by the SHOGUN library (version 6.0.0) [7]. We used three kernels (Linear, Polynomial, and Gaussian kernel), and various parameters; seven regularization parameters ($C = 10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^{2}, 10^{3}$), seven degree parameters of polynomial kernel (D = 1, 2, 3, 4, 5, 6, 7), and seven width parameters of Gaussian kernel (W = 1, 5, 10, 50, 100, 500, 1,000). Cross validation was performed on each of 20 compounds and we calculated the prediction accuracy rate and weights of sub-kernel matrices for each compound from its

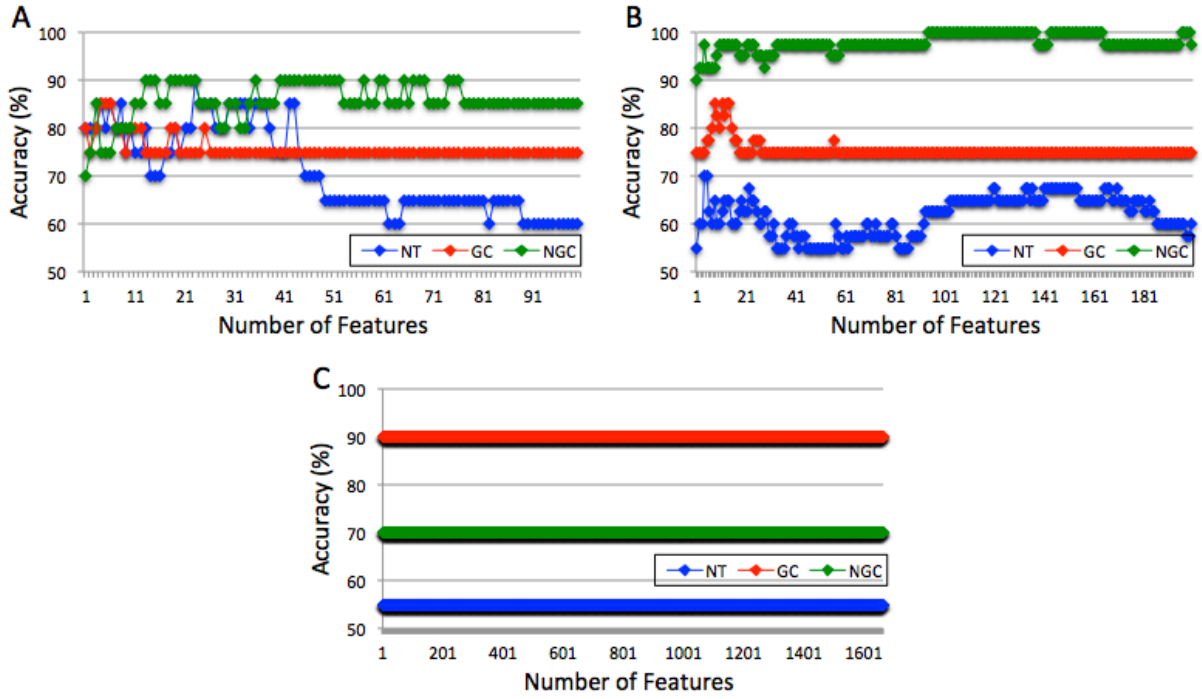

Figure 1. SVM prediction accuracies with *vectorial* kernels
The figure shows the highest prediction accuracies in each number of features using *vectorial* kernel (Linear, Polynomial, Gaussian). For each feature, p-value by Student's t-test is calculated, and those with a small p-value are preferentially used. Result of toxicity prediction with (A) Bayesian network values, (B) qRT-PCR expression values, and (C) E-Dragon features.

results.

3. Result

3.1 Weights of sub-kernel matrices by MK-SVM

The weights used for the linear sum of sub-kernel matrices were calculated with *CombinedKernel* class of SHOGUN library. We extracted weights with the highest accuracy in the various parameters. In this study, the highest accuracies were 85.0% for NT, 95.0% for GC and 100.0% for NGC. Figure 2 shows weights of sub-kernel matrices with the maximum accuracy.

3.2 Frequency of giving the maximum weight among the three weights

However, such weight is not always only one case. Since the accuracy rate is calculated using 40 data points, there are only 41 possible patterns of prediction accuracy (i / 40 × 100; i is from 0 to 40). On the other hand, there are $(1 + 7 + 7)^3 \times 7 =$ 23,625 parameter cases, so there is high possibility that two or more weights exist with the maximum accuracy and other parameters. In fact, in this analysis, the weights with various parameters were extracted (39 for NT, 51 for GC, and 20 for NGC). In order to evaluate the difference of these weights, we used the frequency of giving the maximum weight among the three ones. The results are shown in Table 1. In the case of NT, 39 weights were extracted, and the weights of BN were the largest in all cases. Likewise, in the case of NGC, the weights of PCR were the largest in all 20 cases. However, in the case of GC, the weight of BN is largely large, but depending on the compound, QSAR was larger than BN. Therefore, calculating weights for each compound may give high accuracy, but it may be inappropriate to obtain the classification contributions of each sub-kernel matrices.

4. Discussion

In this study, we regarded the weights of linear sum in multiple kernel learning as contribution of toxicity prediction of each sub-kernel matrix. However, in the prediction of GC toxicity, 3.2 result shows that the required features are different by the compound. In terms of obtaining classification contribution, it is not best to use the weight calculated in this study. In future, we would like to analyze all compounds with the same weight.

We used only three kinds of kernels in this study (Linear, Polynomial and Gaussian). In the previous study [2], three more *structured* kernels (EKM, Saigo, ME) were used [8]. These kernels can be used with kemba-svm, which is the SVM analysis tool. It is known that the *structured* kernels are able to calculate with higher accuracies than the *vectorial* kernels for all toxicity. We would like to use these kernels by combining the SHOGUN library with kemba-svm *kern* mode.

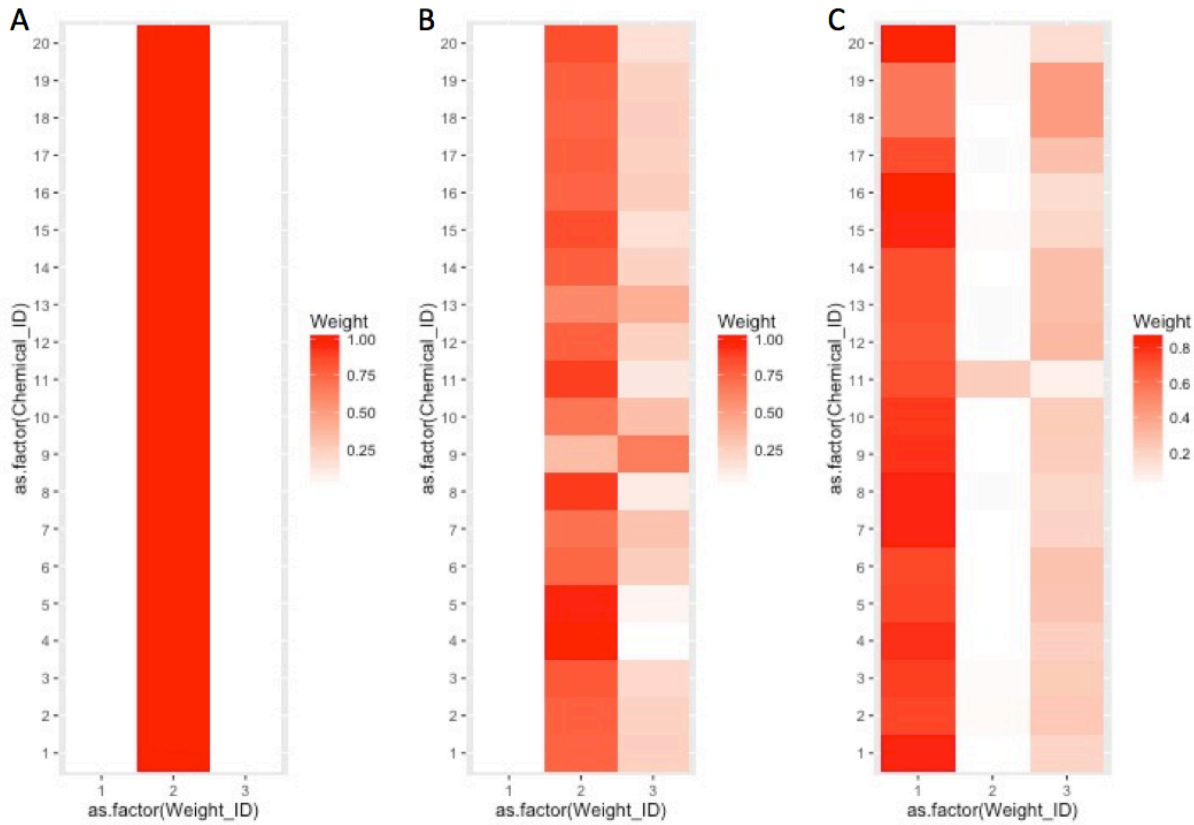

Figure 2. Weights of sub-kernel matrices

One of the weights of sub-kernel matrices with the highest accuracy rate is randomly selected and showed by the white-red heatmap. Red color indicates that the weight is larger. Weight_ID on the horizontal axis notes that 1 is PCR, 2 is BN, and 3 is QSAR. Chemical_ID of the vertical axis is equal to the row numbers of the compounds in Table 1. (A) Weights in NT toxicity prediction. (B) Weights in predicting GC toxicity. (C) Weights in NGC Toxicity Prediction.

Table 1. Frequency of maximum value among three weights

	NT			GC			NGC		
	PCR	BN	QSAR	PCR	BN	QSAR	PCR	BN	QSAR
Valproic acid	0	39	0	0	40	11	20	0	0
Cyclopamine	0	39	0	0	49	2	20	0	0
Phenytoin	0	39	0	0	46	5	20	0	0
Benzo [a] anthracene	0	39	0	0	50	1	20	0	0
3-Methylcholanthrene	0	39	0	0	50	1	20	0	0
Methylmercury	0	39	0	0	46	5	20	0	0
Acrylamide	0	39	0	0	50	1	20	0	0
Benzo [a] pyrene	0	39	0	0	50	1	20	0	0
Diethylnitrosamine	0	39	0	0	0	51	20	0	0
Diethylstilbestrol	0	39	0	0	22	29	20	0	0
4-hydroxy PCB107	0	39	0	0	46	5	20	0	0
2,5-hexanedione	0	39	0	0	46	5	20	0	0
Warfarin	0	39	0	0	16	35	20	0	0
Thalidomide	0	39	0	0	46	5	20	0	0
TCDD	0	39	0	0	49	2	20	0	0
Lithocholic acid	0	39	0	0	51	0	20	0	0
Thioacetamide	0	39	0	0	46	5	20	0	0
Butylated hydroxyanisole	0	39	0	0	50	1	20	0	0
Methapyrilene hydrochloride	0	39	0	0	46	5	20	0	0
Phenobarbital	0	39	0	0	51	0	20	0	0
Rate (%)	0.0	100.0	0.0	0.0	83.3	16.7	100.0	0.0	0.0

The highest rates in this study were calculated to be roughly high, but NT prediction showed a lower accuracy than the result of the previous study [2]. It is thought that being too compatible with training data is caused. To prevent from such over learning may also be necessary depending on toxicity.

There is also room for improvement on the used data. ES cells are cultured through breaking the fertilized eggs, which remains some ethical problems. On the other hand, since induced pluripotent stem (iPS) cells can be prepared by reprogramming somatic cells such as skin cells or fibroblasts [9]. There are few ethical constraints. In the future, toxicity prediction by iPS cells will be carried out instead of ES cells.

We are also considering using public datasets. The National Institutes of Health provides screening results for thousands of compounds with cancer cell line NCI-60 [10]. Many gene expression level datasets by screening tests are also published in Connectivity Map [11] and LINCS [12]. By using these data, we will continue to improve our toxicity prediction systems.

Reference

[1] M. Sarfaty, M.Leshno, N. Gordon, A.Moore, V. Neiman, E.Rosenbaum, D.A. Goldstein. *Eur Urol*, S0302-2838(17)30670-X, 2017.

[2] J. Yamane, S. Aburatani, S. Imanishi, H. Akanuma, R. Nagano, T. Kato, H. Sone, S. Ohsako, W. Fujibuchi. *Nucleic Acids Res*, 44(12), 5515-5528, 2016.

[3] "The Shogun API cookbook/Kernel Support Vector Machine", Copyright 2015-2017, The Shogun developers.

[4] "The Shogun API cookbook/Multiple Kernel Learning", Copyright 2015-2017, The Shogun developers.

[5] T. Yamanaka, H.Toyoshiba, H. Sone, F.M. Parham and C.J. Portier. Environ Health Perspect, 112(16), 1614-1621, 2004.

[6] I.V. Tetko, J. Gasteiger, R. Todeschini, A. Mauri, D. Livingstone, P. Ertl *et al*. *J Comput Aided Mol Des*, 19(6), 453-463, 2005.

[7] SHOGUN website http://www.shogun-toolbox.org/

[8] W. Fujibuchi and T. Kato. *BMC Bioinformatics*, 8, 267, 2007.

[9] K. Takahashi and S. Yamanaka. *Cell*, 126(4), 663-676, 2006.

[10] W.C. Reinhold, M. Sunshine, H. Liu, S. Varma, K.W. Kohn, J. Morris, J. Doroshow and Y. Pommier. *Cancer Res*, 72(14), 3499-3511, 2012.

[11] J. Lamb, E.D. Crawford, D. Peck, J.W. Modell, I.C. Blat, M.J. Wrobel, J.Lerner et al. *Science*, 313(5795), 1929-1935, 2006.

[12] C. Liu, J. Su, F. Yang, K. Wei, J. Ma and X. ZhouX. *Mol Biosyst*, 11(3), 714-722, 2015.