

# Drug clearance pathway prediction based on semi-supervised learning

YANAGISAWA KEISUKE<sup>1,a)</sup> ISHIDA TAKASHI<sup>1</sup> SUGIYAMA YUICHI<sup>2</sup> AKIYAMA YUTAKA<sup>1,b)</sup>

**Abstract:** It is necessary to confirm that a new drug can be appropriately cleared from the human body. However, checking the clearance pathway of a drug in the human body requires clinical trials, and therefore requires large cost. Thus, computational methods for drug clearance pathway prediction have been studied. The proposed prediction methods developed previously were based on a supervised learning algorithm, which requires clearance pathway information for all drugs in a training set as input labels. However, these data are often insufficient because of the high cost of their acquisition. In this paper, we propose a new drug clearance pathway prediction method based on semi-supervised learning, which can use not only labeled data but also unlabeled data. We evaluated the effectiveness of our method, focusing on the cytochrome  $P_{450}$  2C19 enzyme, which is involved in one of the major clearance pathways.

**Keywords:** drug clearance pathway prediction, drug discovery assistance, machine learning, semi-supervised learning

## 1. Introduction

Drug development is a time-consuming and expensive process. More than 10 years and tens of billion dollars are required for approval of a new drug [1]. One of the reasons for this time lag and large cost is that many candidate drug compounds are retracted in the later stage of development because of safety issues, including side effects, insufficient clearance, and so on. Thus, to reduce the cost of drug development, the appropriate selection of compounds based on the safety of a drug in the early stage of development is a very effective strategy, especially if determined before the compound synthesis stage.

Drug clearance pathway prediction is one of the main computational methods used to determine a drug's safety. In general, the method predicts whether a chemical compound is cleared from the human body by a target clearance pathway. Common clearance pathways include metabolism and excretion pathways such as those involving cytochrome  $P_{450}$  (CYP), organic anion transporting polypeptide (OATP), and others. Several drug clearance pathway prediction methods have been established to date. For example, Sorich *et al.* proposed a prediction method for UDP-glucuronosyl transferase (UGT) excretion pathway [2], and Hammann *et al.* proposed a prediction method for 3 types of CYP metabolism pathways [3]. Furthermore, Hotta *et al.* proposed a prediction method for multiple pathway categories, including the CYP, OATP, and renal excretion (Renal) pathways, simultaneously [4].

These drug clearance pathway prediction methods are based on supervised learning techniques, which is currently the most popular method for drug clearance pathway prediction. In supervised machine learning, the algorithm constructs a prediction model using labeled data, for which correct values are already known, and then predicts the labels of unknown data using the prediction model. Various algorithms have been applied for clearance pathway prediction, such as the rectangular boundary method [5], support vector machine (SVM) [6], [7], and Boosting [8] algorithms. According to previous studies, the SVM appears to be the best algorithm for this prediction problem [6]. However, the prediction accuracies are still insufficient for several pathways. One of the clear reasons for this insufficient prediction accuracy is insufficiency in the training data. It is difficult to increase the amount of labeled data because expensive wet experiments and clinical trials are required for determining the clearance pathways of a drug.

The semi-supervised learning method can be used for both labeled and unlabeled data, for which correct values are unknown in the training process. The effectiveness of the semi-supervised learning method has been confirmed in several fields, including for resolving the text-classification problem [9] and the substrate specificity prediction problem [10]. In clearance pathway prediction, unlabeled data are easily obtained because information for a vast amount of compounds can be gathered from public databases such as ZINC [11], PubChem [12], and DrugBank [13]. Therefore, the semi-supervised learning method would be suitable for drug clearance pathway prediction.

Here, we propose a novel drug clearance pathway prediction method based on the semi-supervised learning algorithm. We focus on CYP2C19, one of the major clearance pathways, as a prediction target. CYP2C19 has genotypes of two mutations, and

<sup>1</sup> Graduate School of Information Science and Engineering, Tokyo Institute of Technology, Meguro, Tokyo 152-8552, Japan

<sup>2</sup> RIKEN Innovation Center, Research Cluster for Innovation, RIKEN, Yokohama, Kanagawa 230-0045, Japan

<sup>a)</sup> yanagisawa@bi.cs.titech.ac.jp

<sup>b)</sup> akiyama@cs.titech.ac.jp

approximately 25% of the Japanese population are genotypically identified as poor metabolizers [14]. This means that if these individuals take a drug that is metabolized by CYP2C19, the drug will not be appropriately cleared. Therefore, prediction of this clearance pathway is important.

## 2. Drug clearance pathway prediction based on semi-supervised learning

### 2.1 Semi-supervised learning

Semi-supervised learning is a machine learning method originated from the self-training algorithm proposed by Scudder [15]. The supervised learning method uses only labeled data for the training stage, and cannot use unlabeled data. By contrast, the unsupervised learning method uses only unlabeled data for the training stage. Some semi-supervised learning algorithms make clusters based on the unlabeled data and simultaneously make predictions based on the labeled data. Thus, better prediction can be obtained by using both labeled and unlabeled data. There are various semi-supervised learning algorithms, such as co-training [16], transductive framework [17], and Universum SVM [18].

In this study, we used a transductive SVM (TSVM) algorithm, which is an extension of the SVM algorithm based on a transductive framework. The TSVM algorithm was proposed by Joachims [9], and details of the training algorithm are described below:

- (1) The support vector is constructed from only labeled data, as in the conventional SVM model.
- (2) Unlabeled data are classified based on the support vector, as in the conventional SVM model.
- (3) The labeled and unlabeled data are given a penalty parameter, which is a weight for misclassification. The penalty for unlabeled data is smaller than that for labeled data.
- (4) The support vector is constructed again from both the labeled and unlabeled data.
- (5) The unlabeled data are re-classified based on the new support vector.
- (6) Step (4) and (5) are repeated until there is no change in any of the classified results.
- (7) The penalty for unlabeled data is strengthened.
- (8) Step (4) to step (7) are repeated.
- (9) When the penalty for unlabeled data is equal to the penalty for labeled data, the training process is ended and the last support vector is outputted.

As this algorithm suggests, if the number of unlabeled data entries is equal to 0, the TSVM algorithm works in the same way as the conventional SVM algorithm. Therefore, the comparison between the SVM and TSVM algorithms is easier than among other semi-supervised learning algorithms.

In this experiment, we used SVM<sup>light</sup> v6.02 for implementation of both the SVM and TSVM [19] algorithms.

### 2.2 Dataset

In this study, we focused on a single clearance pathway, CYP2C19. Thus, positively labeled compounds are metabolized by CYP2C19, and negatively labeled compounds are not metabolized by CYP2C19. We adopted the dataset used by Toshimoto *et al.* [7] as the labeled data. This dataset contains 240 com-

pounds, 10 of which are positively labeled and 230 of which are negatively labeled.

In the TSVM algorithm, both unlabeled and labeled data are used in the training process. Ideally, both the labeled and unlabeled data should be sampled from the same population. However, the labeled data are often biased because the clearance pathways are generally determined only for drug candidates, and the distribution of labeled compounds might be different from that of whole chemical compounds. Thus, to reduce the influence of sampling from different populations, all of the unlabeled data were selected only from already approved drug compounds. We constructed an unlabeled dataset using the ZINC database, which is a popular public compounds database, and includes data for more than 35 million compounds. There are also data subsets in the ZINC database, and we used the ZINC drug database (zdd) subset, which consists of 2,924 FDA-approved drug compounds, including chiral compounds and duplicated compounds to the labeled data. We omitted these compounds because the chiral compounds cannot be distinguished using our method, and finally obtained 1390 compounds as the unlabeled dataset.

### 2.3 Features

Kusama *et al.* used four features for the prediction: molecular weight (MW), octanol-water distribution coefficient (logD), protein unbound fraction in plasma (fup), and category of charge at neutral pH [5]. In this study, we essentially used the same features as those used in Kusama *et al.*: MW, logD, and fup. In addition to these features, we used the number of positively charged atoms at neutral pH (charge+) and the number of negatively charged atoms at neutral pH (charge-) instead of the category of charge at neutral pH. These features were calculated using PreADMET v2.0 software (Bioinformatics & Molecular Design Research Center, South Korea).

## 3. Results

We constructed a prediction model using the SVM and TSVM algorithms, and performed evaluation experiments to check the performance of our proposed method.

### 3.1 Evaluation measure

In this study, the number of positive and negative labels were imbalanced, and therefore accuracy, which is the ratio of correctly classified data, is an inappropriate measure for this evaluation. Thus, we employed the f-measure as the evaluation measure. The f-measure is the harmonic mean of precision and recall, and is useful for evaluating the performance of prediction based on an imbalanced dataset. Previous related studies also used the f-measure [4], [7], therefore making it suitable for comparison to previous work.

To calculate the f-measure, precision and recall values are required:

$$\text{Precision} = \frac{\#TP}{\#TP + \#FP} \quad (1)$$

$$\text{Recall} = \frac{\#TP}{\#TP + \#FN} \quad (2)$$

TP, FN, and FP represent the number of true positives, false nega-

tives, and false positives, respectively. Thus, precision is the ratio of the number of positives that correctly predict to the number of predicted positives, and recall is the ratio of the number of positives that correctly predict to the number of all positives, which are predicted as both positives and negatives.

$$F = 2 \cdot \frac{\text{recall} \cdot \text{precision}}{\text{recall} + \text{precision}} = \frac{2 \cdot \#TP}{\#FN + \#FP + 2 \cdot \#TP} \quad (3)$$

Because it is based on the harmonic mean, a better f-measure is obtained when both the recall and precision are relatively high.

To calculate the evaluation measure, we used leave-one-out cross validation. If the number of labeled data entries is  $N$ ,  $N - 1$  data entries are used as a training data and the remaining data entry is predicted. This training and prediction are repeated  $N$  times, and the results of each prediction are cumulated. The cumulated result is then used to calculate the evaluation measure.

### 3.2 Kernel selection and hyper-parameter optimization

The selection of a kernel function and hyper-parameter optimization highly influence the performance of prediction in an SVM-based algorithm. In this study, we employed the Gaussian kernel shown below.

$$K(\mathbf{x}, \mathbf{z}) = \exp\left(-\gamma \|\mathbf{x} - \mathbf{z}\|^2\right) \quad (4)$$

$\mathbf{x}$  and  $\mathbf{z}$  indicate the vectors of features. If these features are similar, the output of  $K(\mathbf{x}, \mathbf{z})$  is higher; therefore,  $K(\mathbf{x}, \mathbf{z})$  indicates the similarity of the two vectors.

When employing the Gaussian kernel, there are two hyper-parameters that need to be optimized: cost of the soft-margin parameter  $C$ , and the width parameter of the Gaussian kernel  $\gamma$ . The larger the cost parameter  $C$  and the width parameter  $\gamma$ , there is an increased likelihood of obtaining a more complicated hyperplane. For hyper-parameter optimization, we employed the following two steps.

**global optimization** First, we trained the TSVM and SVM for each combination of 20 patterns of parameter  $C$  and 20 patterns of parameter  $\gamma$  as follows:

$$\gamma = \{2^{-15}, 2^{-14}, \dots, 2^3, 2^4\} \quad (5)$$

$$C = \{2^{-5}, 2^{-4}, \dots, 2^{13}, 2^{14}\} \quad (6)$$

The best global parameters  $C_0$  and  $\gamma_0$  were obtained on the basis of the evaluation measure.

**local optimization** Second, we determined the best parameter using the results of global optimization. As for global optimization, we trained the TSVM and SVM for each combination of 24 patterns of parameters  $C$  and  $\gamma$  as follows:

$$\gamma = \{\gamma_0 \cdot 2^{-3}, \gamma_0 \cdot 2^{-2.75}, \dots, \gamma_0 \cdot 2^{2.5}, \gamma_0 \cdot 2^{2.75}\} \quad (7)$$

$$C = \{C_0 \cdot 2^{-3}, C_0 \cdot 2^{-2.75}, \dots, C_0 \cdot 2^{2.5}, C_0 \cdot 2^{2.75}\} \quad (8)$$

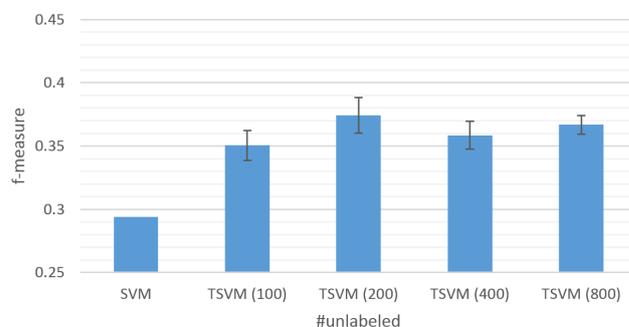
We obtained the best parameters  $C_{best}$  and  $\gamma_{best}$  from these combinations.

### 3.3 Unlabeled dataset

In the evaluation, we generated subsets of sizes 100, 200, 400, and 800 by randomly selecting from the unlabeled dataset including 1390 compounds, and used these subsets in the training to

**Table 1** Prediction result of pathway CYP2C19

| CYP2C19    | f-measure |        |
|------------|-----------|--------|
|            | average   | S.E.   |
| SVM        | 0.2941    | —      |
| TSVM (100) | 0.3506    | 0.0118 |
| TSVM (200) | 0.3742    | 0.0140 |
| TSVM (400) | 0.3586    | 0.0109 |
| TSVM (800) | 0.3668    | 0.0072 |



**Fig. 1** Average and standard error of f-measure: CYP2C19

| p-value    | <b>Table 2</b> The result of the Student's $t$ -test |            |            |            |
|------------|--|------------|------------|------------|
|            | SVM  | TSVM (100) | TSVM (200) | TSVM (400) |
| TSVM (100) | <b>0.0010</b>  |            |            |            |
| TSVM (200) | <b>2.9e-4</b>  | 0.2430     |            |            |
| TSVM (400) | <b>2.2e-4</b>  | 0.5003     | 0.3934     |            |
| TSVM (800) | <b>3.5e-6</b>  | 0.2281     | 0.6916     | 0.5456     |

confirm the influence of the number of unlabeled data entries and how many unlabeled data entries are needed for this prediction problem. In addition, we performed the random sampling ten times and checked the performance for each sample.

### 3.4 Results of the evaluation test

**Table 1** and **Fig. 1** show the results of the evaluation test. Numbers in parentheses represent the numbers of unlabeled data entries. For the results of the TSVM, the averages and standard errors (S.E.) are shown for checking the influence of sampling bias of the unlabeled data. The results showed that the TSVM always performed much better than the SVM. The average f-measure was improved until reaching a size of 200 unlabeled data entries, and then appeared to be saturated. On the other hand, as the number of unlabeled data entries was increased, the S.E. of the f-measure was likely to decrease.

To assess the statistical significance of the improvement in prediction given by the TSVM and the effectiveness of the increase of the unlabeled dataset, we analyzed the f-measures using the Student's  $t$ -test and calculated p-values for all combinations. **Table 2** shows the results, and the values indicated in bold represent the values that are statistically significant ( $p < 0.05$ ). The results showed that all p-values for differences between the SVM and TSVM were less than 0.05. This indicates that the improvement conferred by the TSVM was statistically significant regardless of the number of unlabeled data entries. In contrast, the p-values for differences among TSVMs with different sizes of the unlabeled dataset were not less than 0.05. According to Joachims [9], there is a limitation of improvement in prediction even when using a vast amount of unlabeled data, which suggests that this small unlabeled dataset is sufficient for the improvement of the prediction,

**Table 3** precision & recall: CYP2C19

|            | precision |        | recall  |        | f-measure |        |
|------------|-----------|--------|---------|--------|-----------|--------|
|            | average   | S.E.   | average | S.E.   | average   | S.E.   |
| SVM        | 0.2083    | —      | 0.5000  | —      | 0.2941    | —      |
| TSVM (800) | 0.4088    | 0.0217 | 0.3500  | 0.0224 | 0.3668    | 0.0072 |

and increasing the number of unlabeled data entries is not effective for this prediction problem.

## 4. Discussion

### 4.1 Balance of precision and recall

If f-measure values are equivalent in two prediction methods, the precision and recall may nonetheless be different. Even if the performance is improved according to the f-measure, a large decrease of recall or precision can cause problems in some applications. Thus, we checked the precision and recall of the prediction results.

**Table 3** shows the precision and recall values for the TSVM (800). The precision of the TSVM was higher than that of the SVM, whereas the recall of the TSVM became lower than that of the SVM. Therefore, the TSVM can be suitable for predicting a few compounds that are more likely to metabolize. Because there are millions of candidate compounds, whereas only hundreds or thousands of compounds can reasonably be tested in wet experiments, the ability to obtain a relatively small amount of compounds that are predicted to be most likely to metabolize is a very important feature of this model. Thus, we think that this characteristic of the TSVM makes it more desirable than the SVM.

### 4.2 ROC analysis

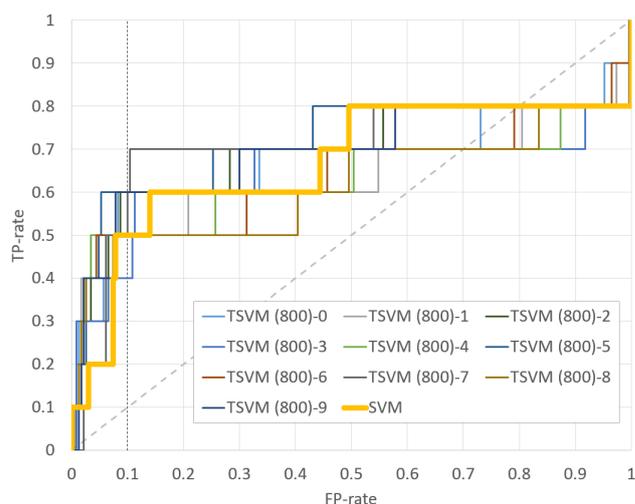
As previously discussed, positive compounds should be ranked among the top hundreds or thousands. This means that it is more important to determine the true positive rate in a couple of percent of data that are top-ranked by prediction. This determination is called early enrichment. Unfortunately, since there are only a few labeled data entries, especially for the positive data, we could not stably obtain the positive rate in the top few percent of the data. Due to this limitation, we drew a receiver operating characteristic (ROC) curve based on the positiveness of the data (in this case, the decision values of the SVM and TSVM), and calculated the area under the curve (AUC) for the overall data and the top 10%. For distinction, the overall AUC is referred to as AUC (100%) and the top 10% AUC is referred to as AUC (10%) hereafter. To draw the ROC curve, the FP-rate and TP-rate are required:

$$\text{FP-rate} = \frac{\#FP}{\#FP + \#TN} \quad (9)$$

$$\text{TP-rate} = \frac{\#TP}{\#TP + \#FN} \quad (10)$$

The ROC curve can be drawn once the positiveness of each tested data entry is obtained. To draw the curve, we added the data in order of their positiveness. If positive data are added, the TP-rate increases, whereas if negative data are added, the FP-rate increases. The ROC curve shows these changes as a line, and the curve describes the tradeoff between the TP-rate and FP-rate. Higher AUCs are obtained when the TP-rate is higher, even if the FP-rate is still low.

The ROC curves for the TSVM with 800 unlabeled data en-

**Fig. 2** ROC curve: CYP2C19**Table 4** AUC value: CYP2C19

|                      | AUC (100%) | AUC (10%) |
|----------------------|------------|-----------|
| SVM                  | 0.6665     | 0.0265    |
| TSVM (800) (average) | 0.6635     | 0.0393    |
| (random)             | (0.5000)   | (0.0050)  |

tries (TSVM (800)) and the SVM are shown in **Fig. 2**. There are 10 plots for the TSVM and one plot for the SVM. The diagonal dotted line shows the performance of random prediction, and the vertical dotted line shows the threshold for the top-ranked 10%. The results of the ROC analysis showed that the performance of the TSVM was comparable to that of the SVM in AUC 100% but was superior to the SVM in AUC (10%) (**Table 4**). This result suggests that the prediction improvement by the TSVM is more obvious in earlier enrichment, indicating it is effective for this type of prediction because of the importance of earlier enrichment described above.

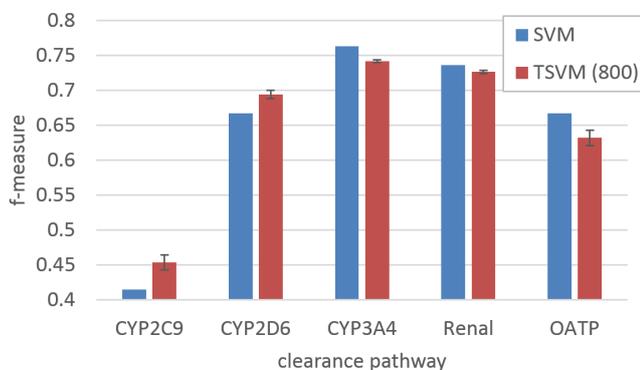
### 4.3 Application of the proposed method to other clearance pathways

Our results demonstrated that the TSVM is effective to predict the CYP2C19 clearance pathway. Thus, we also applied the method to other clearance pathways. We employed the same dataset used in Kusama *et al.* [5], which includes five pathways: CYP2C9, CYP2D6, CYP3A4, Renal, and OATP. The results with 800 unlabeled data entries are shown in **Table 5** and **Fig. 3**. The TSVM succeeded in improving the predictions for some pathways (CYP2C9, CYP2D6) compared with those of the SVM. However, the accuracy of prediction for the other pathways (CYP3A4, Renal, OATP) was equal to or worse than that of the SVM. The TSVM particularly improved the accuracy of the CYP2C9 pathway, which showed the worst f-measure among all of the tested pathways when the SVM was used. This result indicates that the TSVM is likely to improve the accuracy of prediction when that of the SVM is considerably insufficient.

To investigate the effect of the TSVM more deeply, we also drew an ROC curve for CYP3A4 prediction, which was worse when the TSVM was used compared to the SVM. ROC curves of CYP3A4 are shown in **Fig. 4** and the AUC values are shown in **Table 6**. According to the ROC curves and AUC values, the

**Table 5** Prediction result of other pathways

| f-measure | SVM    | TSVM (800) |        |
|-----------|--------|------------|--------|
|           |        | average    | S.E.   |
| CYP2C9    | 0.4151 | 0.4537     | 0.0110 |
| CYP2D6    | 0.6667 | 0.6940     | 0.0059 |
| CYP3A4    | 0.7630 | 0.7417     | 0.0018 |
| Renal     | 0.7361 | 0.7262     | 0.0021 |
| OATP      | 0.6667 | 0.6319     | 0.0110 |

**Fig. 3** Average and standard error of f-measure: other pathways

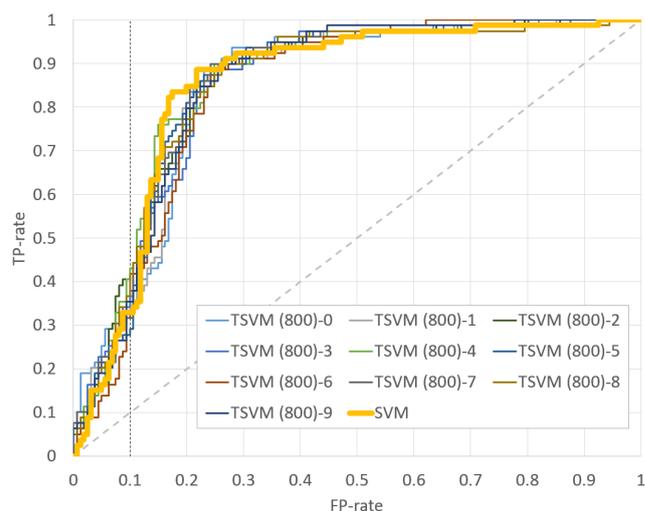
TSVM was worse than the SVM in terms of f-measure, whereas the TSVM was better than the SVM in terms of AUC (10%). The f-measure is calculated with respect to the points of the ROC curve lines, whereas the AUC is calculated using all or a particular (for example 10%) area of the ROC curve. Therefore, if one method is much better on one point of the ROC curve, the f-measure would be drastically improved, while the AUC would be only slightly improved. In this case, the calculation point of the f-measure for SVM prediction of CYP3A4 is the point (FP-rate, TP-rate) = (0.2547, 0.8987). This point is slightly above the other TSVM ROC curves, which explains why the CYP3A4 f-measure based on the SVM was better than that of the TSVM. On the other hand, because the distance of the ROC curves between the SVM and TSVM is very small, their AUC (100%) values were approximately equivalent.

## 5. Conclusion

In this study, we proposed a new drug clearance pathway prediction based on the TSVM. The TSVM improved the clearance pathway prediction for CYP2C19 compared with the SVM, and the improvement was statistically significant. By contrast, there was no significant improvement observed by increasing the number of unlabeled data entries in the TSVM. These results suggest that a dataset with only 100 unlabeled data entries is sufficient for the prediction problem with a dataset containing 240 labeled data entries.

Investigation of other semi-supervised learning methods is still required in future studies. We employed the TSVM as a semi-supervised learning method; however, there are still many other types of semi-supervised learning methods, even among those that are based on the SVM such as Universum SVM [18]. Therefore, we have to investigate which semi-supervised learning algorithm is the most suitable for drug clearance pathway prediction.

**Acknowledgments** We are grateful to Mr. Kota Toshimoto, RIKEN Innovation Center, Research Cluster for Innovation, RIKEN, for his suggestions and advice. This work was

**Fig. 4** ROC curve: CYP3A4**Table 6** AUC value: CYP3A4

|                      | AUC (100%) | AUC (10%) |
|----------------------|------------|-----------|
| SVM                  | 0.8499     | 0.0187    |
| TSVM (800) (average) | 0.8479     | 0.0215    |
| (random)             | (0.5000)   | (0.0050)  |

supported by the Education Academy of Computational Life Sciences (ACLS) at the Tokyo Institute of Technology.

## References

- [1] "PhRMA New Drug Approvals in 2011", <http://www.phrma.org/>
- [2] Sorich M.J., Miners J.O., McKinnon R.A. *et al.*: "Comparison of Linear and Nonlinear Classification Algorithms for Prediction of Drug and Chemical Metabolism by Human UDP-Glucuronosyltransferase Isoforms", *Journal of Chemical Information and Computer Sciences* 43(6), pp.2019-pp.2024, 2003
- [3] Hammann F., Gutmann H., Baumann U. *et al.*: "Classification of Cytochrome P450 Activities Using Machine Learning Methods", *Molecular Pharmaceutics* 6(6), pp.1920-pp.1926, 2009
- [4] Hotta S., Toshimoto K., Ikeda K. *et al.*: "Drug clearance pathway prediction on a web application", *IP SJ SIG Technical Report 2010-BIO-21(20)*, pp.1-pp.8, 2010
- [5] Kusama M., Toshimoto K., Maeda K. *et al.*: "In silico classification of major clearance pathways of drugs with their physicochemical parameters.", *Drug Metabolism and Disposition* 38(8), pp.1362-pp.1370, 2010
- [6] Toshimoto K., Kusama M., Ikeda K. *et al.*: "In silico Prediction System of Major Drug Clearance Pathways -Expansion for Multiple Pathway Prediction and External Validation-", *IP SJ SIG Technical Report 2010-BIO-20(8)*, pp.1-pp.8, 2009
- [7] Toshimoto K., Wakayama N., Hotta S. *et al.*: "Establishing an in silico prediction system of nine major in vivo drug clearance pathway using machine learning technique.", *The 27th Annual Meeting of Japanese Society for Alternatives to Animal Experiments*, P-43, 2014
- [8] Ikeda K., Toshimoto K., Kusama M. *et al.*: "Prediction of drug clearance pathway by boosting algorithm", *IP SJ SIG Technical Report 2009-BIO-17(10)*, pp.1-pp.8, 2009
- [9] Joachims T.: "Transductive inference for text classification using support vector machines", *Proceedings of the Sixteenth International Conference on Machine Learning*, pp.200-pp.209, 1999
- [10] Röttig M., Medema M.H., Blin K. *et al.*: "NRPSpredictor2 - a web server for predicting NRPS adenylation domain specificity", *Nucleic Acids Research* 39(suppl 2), W362-W367, 2011
- [11] John J.I., Teague S., Michael M.M. *et al.*: "ZINC: A Free Tool to Discover Chemistry for Biology.", *Journal of Chemical Information and Modeling* 52(7), pp.1757-pp.1768, 2012
- [12] Wang W., Xiao J., Suzek T.O. *et al.*: "PubChem: a public information system for analyzing bioactivities of small molecules.", *Nucleic Acids Research* 37(suppl 2), W623-W633, 2009
- [13] Law V., Knox C., Djoumbou Y. *et al.*: "DrugBank 4.0: shedding new light on drug metabolism", *Nucleic Acids Research*, 42(D1), D1091-D1097, 2014

- [14] Miyuki K., Ichiro I., Kohsuke M. *et al.*: "Genetic Polymorphism of Cytochrome P450x, CYP2C19, and CYP2C9 in a Japanese Population.", *Therapeutic Drug Monitoring* 20(3), pp.243-pp.247, 1998
- [15] Scudder III H.: "Probability of error of some adaptive pattern-recognition machines.", *Information Theory, IEEE Transactions* 11(3), pp.363-pp.371, 1965
- [16] Blum A., Mitchell T.: "Combining labeled and unlabeled data with co-training.", *Proceedings of the Eleventh Annual Conference on Computational Learning Theory*, pp.92-pp.100, 1998
- [17] Vapnik V.: "Statistical Learning Theory.", Wiley, 1999.
- [18] Weston J., Collobert R., Sinz F. *et al.*: "Inference with the Univer-sum.", *ICML '06 Proceedings of the 23rd international conference on Machine learning*, pp.1009-pp.1016, 2006
- [19] Joachims T.: "SVM<sup>light</sup> Support Vector Machine.", <http://svmlight.joachims.org/>