Comparison of Random Forest and SVM for Raw Data in Drug Discovery: Prediction of Radiation Protection and Toxicity Case Study

Atsushi Matsumoto, Shin Aoki, and Hayato Ohwada

***Abstract*—This paper compares random forest and SVM for raw data in drug discovery. Both machine-learning methods are often applied in drug discovery. We should select our methods depending on the problem. This is very important. SVM is suitable for virtual screening when the target protein is known. In contrast, random forest is suitable for virtual screening when the target protein is not decided uniquely or unknown, because random forest can find good combinations of features from many features. Therefore, random forest is thus more effective for problems including many unknown parts. Incidentally, selecting the good features is important in both methods. In particular, we must narrow the features using importance calculations if we lack sufficient biochemical knowledge. In this study, we predicted the radiation protection function and toxicity for radioprotectors targeting p53 as a case study. When predicting the radiation protection function the target protein is known. In contrast, when predicting toxicity, the target protein is not decided uniquely or is unknown. We evaluated each experiment based on its AUC score. As a result, we found that when predicting the radiation protection function, SVM was better than random forest. By contrast, when predicting toxicity, random forest was better than SVM.**

***Index Terms*—SVM, random forest, drug discovery, radioprotector.**

I. Introduction

In recent years, machine learning has become attractive in the drug discovery field [1]-[3]. The objective of using machine learning is to efficiently search for candidate compounds for new drugs based on the result of feature prediction, classification, etc. SVM [4] and random forest [5] are often used in the drug discovery field [6], [7].

In this study, we compared random forest and SVM for the raw data in drug discovery.

We took up two types of problems in drug discovery. First, there is the problem in which the target protein is known. In this case, we search for a compound that bonds to the target protein. If the compound bonds to the target protein, the function of the protein is activated or inhibited. As a result, various symptoms are addressed. The second type is the problem in which the target protein is not decided uniquely or is unknown, for example when we do not know about the drug mechanisms or we cannot uniquely determine the drug mechanisms. Predicting toxicity has the same problem.

Manuscript received December 5, 2015. Revised February 19, 2016.

Atsushi Matsumoto, Shin Aoki, and Hayato Ohwada are with the Tokyo University of Science, Noda-shi, Chiba-ken, Japan (e-mail: 7415617@ed.tus.ac.jp, shinaoki@rs.tus.ac.jp, ohwada@rs.tus.ac.jp).

SVM has high performance and is easy to apply to various problems. It is necessary to provide optimal features if one wishes to bring about the appropriate learning. In contrast, random forest is an ensemble algorithm using many weak discriminators. A weak discriminator consists of a random combination of features. Therefore, random forest is an effective method if we cannot determine which features are important. In this study, we predicted the radiation protection function and toxicity for radioprotectors targeting p53 as a case study. Predicting the radiation protection function is a problem in which the target protein is known. In contrast, predicting toxicity is a problem in which the target protein is not decided uniquely or is unknown.

II. Radioprotector

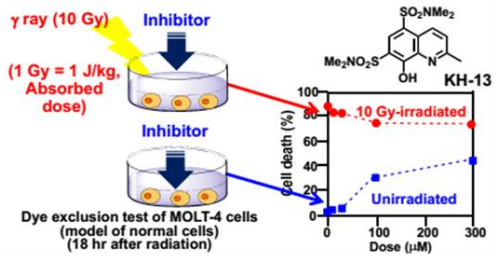
This chapter describes the radioprotector. Cancer is one of the most common causes of death in the world. There were an estimated 367,000 cancer deaths in Japan in 2014 (218,000 males and 150,000 females) [8]. Radiation therapy is one of the main approaches against cancer cells, although this therapy has adverse side effects, including p53-induced apoptosis of normal tissues and cells [9]. In brief, radiation kills the normal cells around the cancer cells. It is considered that p53 would be a target for therapeutic and mitigative radioprotection to avoid the apoptotic fate.

III. Dataset and Labeling

This chapter describes the dataset used for learning in this study. This study used compounds in connection with a p53 inhibitor. The Aoki group had a variety of successes in their study of radioprotectors [10]. Their experiment data was used in this study. Eighty-four compounds obtained in experiments were selected for learning. Two experiments were performed for each compound. First, experiments administering the compounds to normal cells were performed. These experiments were able to measure the toxicity. Next, experiments administering the compounds to gamma irradiated cells were performed. These experiments were able to measure the radiation-protection function. Their experiment (Fig. 1) measured the cell death rate for each concentration case.

The death rate of the cells was used as an indicator in both experiments. If the death rate in the case of the unirradiated experiment was low and the death rate in the case of the irradiated experiment was greatly reduced, that compound is useful as a radioprotector.

doi: 10.18178/ijmlc.2016.6.2.589



Ariyasu, S.; Sawa, A.; Morita, A.; et al. *Bioorg. Med. Chem.* 2014, 22, 3891-3905

Fig. 1. Example of experiments.

Now, Define C and R as a threshold of toxicity and radioation protection.

$$\mathrm{C} = \text{Cell death\%(maximum)} - \text{Cell death\%(minimum)} \quad (1)$$

$$\mathrm{R} = \text{Cell death\%}(0\mu\mathrm{M}) - \text{Cell death\%}(\text{Concentration when cell death due to toxicity is 20\%}) \quad (2)$$

With regard to the radiation protection function, we only used range of concentration unaffected toxic, in order to reduce the effects of toxicity (see Fig. 2). We set thresholds for labeling compounds based on the death rate of the cells (see Table I).

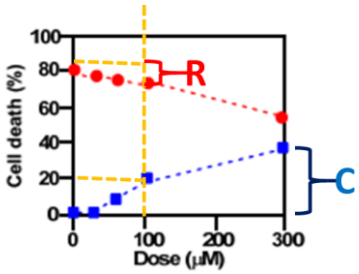


Fig. 2. Definition of threshold.

TABLE I: LABELING OF COMPOUNDS

	Threshold based on the death rate of the cells [%]	
	Toxicity	*Radiation protect*
class0	***~ 20***	***~ 10***
class1	***21 ~ 100***	***11 ~ 100***

We performed classifications based on these labels. These were binary classifications. For toxicity, class1 has 42 compounds. For the radiation protection function, class1 has 39 compounds. The other compounds were defined as being in class0.

IV. METHOD

This chapter describes our method. We used SVM and random forest as machine-learning methods. Our method consists of four steps, including the importance calculation, normalization, the selection of features, and finally the machine-learning step. First, the importance of the various features was calculated based on Gini in random forest because we should prioritize in selecting compound features. It is not a good idea to use a lot of chemical characteristics of unknown relevance to radioprotection. There were 489 calculated features in all. Roughly divided, the features consist of five types related to the structure, ALogP, size or weight, energy, or others. ALogP is an indicator of lipid solubility. These features were calculated by Discovery Studio, 3D modeling software that can calculate the chemical properties. Non-numeric data were excluded in this step. The dataset was normalized to a standard normal distribution for each feature. For compound i, the parameter p_i is transformed by

$$p_i = \frac{p_i - \mathrm{mean}(p)}{sd(p)} \quad (3)$$

where mean(p) indicates the mean value, and $sd(p)$ indicates the standard deviation of parameter p.

Features were selected based on their importance. The top 5%, 10%, 15%, 20%, 25%, 30% and all features were used in this study. Next, we performed a classification based on a defined label using random forest and SVM. We used the scikit-learn package for Python3.

We now describe the parameter settings. For random forest, there were 5000 trees, and the depth of search was determined using a grid search from 3 to 6. For SVM, the kernel was used as a radial basis function (RBF). This RBF is a non-linear kernel. If RBF is selected for use with SVM, it will require a cost parameter and a gamma parameter. Cost parameters and gamma parameters were determined using a grid search for 20 split from 0.0001 to 10,000.

Finally, we describe the evaluation. We used the Area Under Curve (AUC) score. AUC is part of the performance metric for binary classification. It is less affected by sample balance than the accuracy. As a reality, biological data are often unbalanced. The application of machine learning for drug discovery research often uses the AUC score for evaluation. When using raw data, there are few compounds . Thus, we used 10-fold cross-validation in all scenarios. The data was divided into 10 classes. The ratio between classes was the same as in the original data. Part of the divided data was used for testing, and the remaining data was used for training. To calculate the average, all of the divided data was evaluated.

V. RESULTS

TABLE II: LIST OF AUC SCORES

	feature_num[%]	Random Forest AUC	SVM AUC
Toxicity	5	0.778	0.681
	10	0.778	0.654
	15	0.737	0.674
	20	0.731	0.716
	25	0.727	0.636
	30	0.732	0.630
	100	0.718	0.663
Radio Protection	5	0.605	0.646
	10	0.619	0.494
	15	0.613	0.477
	20	0.593	0.520
	25	0.594	0.511
	30	0.583	0.569
	100	0.595	0.498

The results for random forest and SVM are summarized in Table II. There are 28 scenarios in all. feature_num[%] represents the use rate of the feature based on the importance calculation in machine learning. In addition, Fig. 3 and Fig. 4 indicate the importance of all the properties. The sum of importance values is 1.

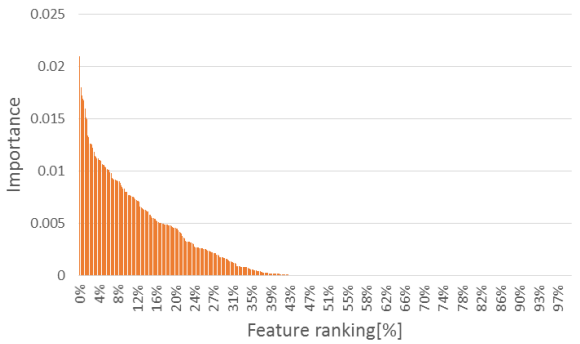


Fig. 3. Importance of the properties in toxicity.

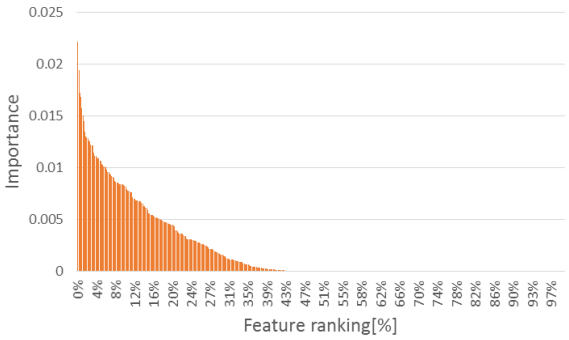
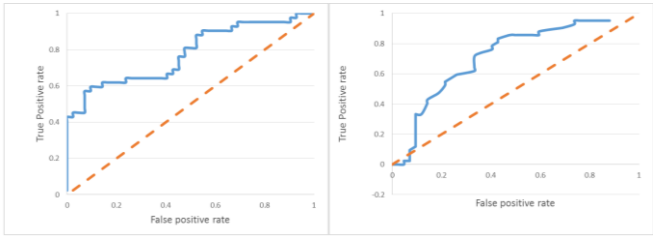


Fig. 4. Importance of the properties in radiation protection.

As a result, when predicting the radiation protection function, SVM is better than random forest as determined by the AUC score. In contrast, for predicting toxicity, random forest is better than SVM.

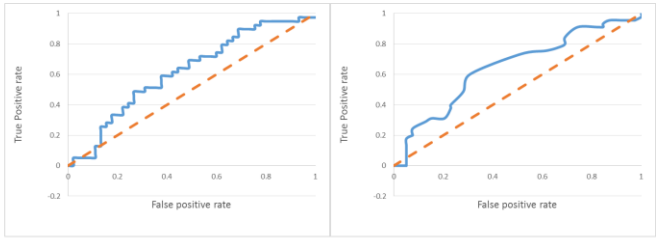
In radioprotection using SVM, the AUC score dropped significantly if numerous features were selected. Additionally, In spite of performing a 10-fold cross validation, the SVM AUC score is unstable. In general, a collective-learning algorithm such as random forest has more stable accuracy than SVM. However, SVM can outperform other methods if it is possible to adequately refine the feature selection and amount.

Given Fig. 3 and Fig. 4, using more than the top 30% of features would be inappropriate in this case study. The importance of more than half of these features was nearly zero.



Feature_num[%] = 5, Random Forest Feature_num[%] = 20, SVM

Fig. 5. ROC curve of toxicity.



Feature_num[%] = 10, Random Forest Feature_num[%] = 5, SVM

Fig. 6. ROC curve of radiation protection.

Fig. 5 and Fig. 6 depict ROC curves representing the best AUC scores. The dotted line is a random line. The vertical axis is the true positive rate and the horizontal axis is the false positive rate. Each result corresponds to the best score in Table II.

In addition, Table III provides a ranking of features based on importance. Tables IV and V list the top 5% of features. The molecular properties conform to the Discovery Studio software.

TABLE III: TOP 5% FEATURES IN TOXICITY

Molecular properties	importance
Energy	0.0210
Molecular_3D_PolarSASA	0.0180
Jurs_TPSA	0.0172
Minimized_Energy	0.0169
Jurs_DPSA_3	0.0167
Jurs_TASA	0.0160
Molecular_Volume	0.0151
Strain_Energy	0.0150
Jurs_SASA	0.0134
Shadow_YZ	0.0132
Jurs_DPSA_1	0.0126
Shadow_XZ	0.0126
Jurs_PNSA_3	0.0126
Molecular_3D_SASA	0.0122
Jurs_PNSA_2	0.0118
Jurs_WNSA_1	0.0114
Molecular_SurfaceArea	0.0113
PMI_mag	0.0112
PMI_Z	0.0112
Jurs_DPSA_2	0.0111
Molecular_3D_SAVol	0.0111
Jurs_PPSA_1	0.0110
Shadow_XY	0.0106
Jurs_PPSA_2	0.0106

TABLE IV: TOP 5% FEATURES IN RADIATION PROTECTION

Molecular properties	importance
Energy	0.0221
Molecular_3D_PolarSASA	0.0194
Jurs_TPSA	0.0172
Jurs_DPSA_3	0.0168
Jurs_TASA	0.0157
Minimized_Energy	0.0150
Strain_Energy	0.0145
Molecular_Volume	0.0135
Molecular_SurfaceArea	0.0130
Jurs_SASA	0.0129
Shadow_XZ	0.0129
Shadow_YZ	0.0126
Jurs_DPSA_1	0.0125
Jurs_PNSA_2	0.0122
Molecular_3D_SASA	0.0122
Jurs_PNSA_3	0.0122
Jurs_DPSA_2	0.0114
PMI_Z	0.0111
Jurs_WNSA_1	0.0111
Jurs_WNSA_2	0.0109
ES_Count_sCH3	0.0109
Jurs_PNSA_1	0.0109
PMI_mag	0.0107
Molecular_3D_SAVol	0.0106

In Tables III and IV, similar elements occupied the high ranks.

VI. CONCLUSION

This paper compares random forest and SVM for raw data in drug discovery. We predicted the radiation protection

function and toxicity for radioprotectors targeting p53 as a case study. SVM is better than random forest for predicting the radiation protection function. However, random forest is better than SVM in predicting the toxicity. In predicting the radiation protection function, the target protein is known. In contrast, in predicting toxicity, the target protein is not decided uniquely or is unknown. In general, a collective-learning algorithm such as random forest had more stable accuracy than SVM. However, SVM can outperform other methods if it is possible to adequately refine the feature selection and amount.

There are often not enough suitable databases for machine learning to be applied to raw data. In such cases, selecting the appropriate features is important. We can obtain satisfactory results using existing machine learning techniques if it is possible to adequately refine the feature amount and quality.

REFERENCES

[1] M. Okada, T. Ito, H. Ohwada, and S. Aoki, "Docking score calculation using machine learning with an enhanced inhibitor database," *Journal of Medical Imaging and Health Informatics*, vol. 5, no. 5, pp. 1104-1107, September 2015.

[2] K. Y. Hsia, S. Ghosh, and H. Kitano, "Combining machine learning systems and multiple docking simulation packages to improve docking prediction reliability for network pharmacology," PLoS, vol. 8, no. 12, PMC3877102, December 2013.

[3] S. Agarwal, D. Dugar, and S. Sengupta, "Article prev. article next article table of contents ranking chemical structures for drug discovery: A new machine learning approach," *Journal of Chemical Information and Modeling*, vol. 50, no. 5, pp. 716-731, April 2010.

[4] V. Vapnik, *The Nature of Statistical Learning Theory*, Spring-Verlag, NY, USA, 1995.

[5] L Breiman, "Random forests," *Machine Learning*, vol. 451, pp. 5-32, 2001.

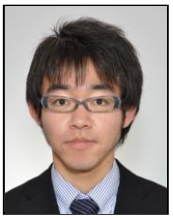
[6] T. Ito, H. Ohwada, and S. Aoki, "Combining two machine learning methods for predicting protein-ligand docking using structure and physiochemical properties," in *Proc. the 7th International Conference on Bioinformatics and Computational Biology*, March 2015, pp. 19-24.

[7] G. Riddick, H. Song, S. Ahn, J. Walling, D. Borges-Rivera, W. Zhang, and H. A. Fine, "Predicting in vitro drug sensitivity using Random Forests," *Bioinformatics*, vol. 27, no. 2, pp. 220-224, December 2010.

[8] The Editorial Board of the Cancer Statistics in Japan, "Cancer statistics in Japan 2014," Foundation for Promotion of Cancer Research (FPCR), March 2015.

[9] A. Morita, S. Ariyasu, B. Wang, T. Asamaru, T. Onoda, A. Sawa, K. Tanaka, I. Takahashi, S. Togami, M. Nenoi, T. Inaba, and S. Aoki, "AS-2, a novel inhibitor of p53-dependent apoptosis, prevents apoptotic mitochondrial dysfunction in a transcription-independent manner and protects mice from a lethal dose of ionizing radiation," *Biochemical and Biophysical Research Communications*, vol. 450, pp. 1498-1504, 2014.

[10] S. Ariyasu, A. Sawa, A. Morita, K. Hanaya, M. Hoshi, I. Takahashi, B. Wang, and S. Aoki, "Design and synthesis of 8-hydroxyquinoline-based radioprotective agents," *Bioorganic and Medicinal Chemistry*, vol. 22, issue 15, pp. 3891-3905, August 1, 2014.



Atsushi Matsumoto graduated from the Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science, Noda City, Japan, in 2015.

He is a student at Tokyo University of Science Graduate School, Division of Science and Engineering Industrial Administration Master's Course since 2015, Noda City, Japan. His research interests are in the field of machine learning and application, especially, bioinformatics.

Shin Aoki graduated from the University of Tokyo with B.S. (1986), M.S. (1988), and Ph.D. (1992) degrees in pharmaceutical sciences under the supervision of Prof. Kenji Koga. He started his academic carrier as an assistant professor at the University of Tokyo from 1990. Following postdoctoral positions with Professor Chi-Huey Wong at the Department of Chemistry, the Scripps Research Institute, USA, he joined Prof. Eiichi Kimura's research group in 1995 at the Faculty of Medicine, Hiroshima University, where he became an associate professor in 2001. In 2003, he has become a professor at the Faculty of Pharmaceutical Sciences, Tokyo University of Science. He is a recipient of the Award of Japan Society of Coordination Chemistry for Young Scientists (1999), the AJINOMOTO Award in Synthetic Organic Chemistry, Japan (2001), the Pharmaceutical Society of Japan Award for Young Scientists (2002), and so on. His major research interests are organic synthetic chemistry, bioinorganic chemistry, supramolecular chemistry, photochemistry, and medicinal chemistry, mainly using metal complexes in aqueous solution.

Hayato Ohwada graduated from the Department of Industrial Administration, Faculty of Science and Technology, Tokyo University of Science, Noda City, Japan, 1983. Then he graduated from Tokyo University of Science Graduate School, Division of Science and Engineering Industrial Administration Doctoral Course completed program with degree in 1988.

He was a research associate from 1988 to 1998, a lecturer from 1999 to 2000, and an associate professor (Tokyo University of Science) from 2001 to 2004. Then he is a professor at Tokyo University of Science Faculty of Science and Engineering from 2005. His research interests are in the fields of inductive logic programming and bioinformatics.