DOI 10.5012/bkcs.2011.32.4.1237

The Predictive QSAR Model for hERG Inhibitors Using Bayesian and Random Forest Classification Method

Jun Hyoung Kim, Chong Hak Chae, Shin Myung Kang, Joo Yon Lee, Gil Nam Lee, Soon Hee Hwang, and Nam Sook Kang*

Drug Discovery Platform Technology Team, Korea Research Institute of Chemical Technology, Daejeon 305-600, Korea
**E-mail: nskang@krict.re.kr*
Received November 20, 2010, Accepted February 12, 2011

In this study, we have developed a ligand-based *in-silico* prediction model to classify chemical structures into hERG blockers using Bayesian and random forest modeling methods. These models were built based on patch clamp experimental results. The findings presented in this work indicate that Laplacian-modified naïve Bayesian classification with diverse selection is useful for predicting hERG inhibitors when a large data set is not obtained.

Key Words : hERG, Classification, Bayesian, Random forest, *in-silico* prediction

Introduction

The human ether-a-go-go related gene (hERG) channel is a key cardiac ion channel that is crucial for the regulation of cardiac action potential.[1] Blockage of this potassium channel extends the repolarization phase, leading to a prolonged QT interval, which is now well understood as the root cause of the cardio-toxicity of numerous approved drugs.[2] Therefore, the hERG channel is a general anti-target in the pharmaceutical industry, and the prediction of hERG activity of new drug candidates has become increasingly important in drug discovery and development. The typical high-throughput hERG screening comprises a radioligand binding assay, a patch clamp assay, a cell-based fluorescence assay, and a rubidium efflux assay.[3] However, the highest throughput methods are too expensive, technically demanding, labor-intensive, and time consuming for the treatment of ever increasing hit compounds in early-stage drug discovery projects. Therefore, in efforts to find a more economic and reliable approach and circumvent these problems, a number of computational methods have been explored for the prioritization of compounds according to their potential to cause cardiotoxic side effects. Many computational prediction models[4-9] have been reported recently for hERG channel blockers, and they can be broadly divided into three categories: structure-based, 3D-QSAR, and classification models. Most studies on hERG blockade have been performed based on 2-dimensional ligand structures using classification-based approaches,[10] including naïve Bayesian,[11] decision tree,[12] random forest,[13] and support-vector machines.[14-16] Bayes's rule of conditional probability[17] is a widely used method of statistical inference applied to many real-world problems that makes it possible to model uncertainty about the world and outcomes of interest by combining common-sense knowledge and observational evidence. The random forest modeling approach is a combinational classifier that consists of many decision tree predictors and outputs the class that is the mode of the output of individual trees.[18]

In this study, we have developed a ligand-based *in-silico* prediction model to classify chemical structures into hERG blockers using Bayesian and random forest modeling methods. These models were built based on patch clamp experimental data and were experimentally validated using in-house compounds. The models described here have comparable predictive powers to those yielded by the approaches noted above.

Experimental

Data Sets. The *in vitro* hERG inhibition data were collected from the literature, Prous Science Integrity,[19] and an in-house experiment. The compounds measured in Human Embryonic Kidney 293 (HEK293) or Chinese Hamster Ovary (CHO) cells in a whole-cell patch-clamp assay were collected with the experimental IC_{50} or pIC_{50} ($-\log IC_{50}$) values from the public domain. The hERG activities of in-house compounds were also measured by an automated planar patch clamp (PatchXpress 7000A) and HEK293 cells. To apply Bayesian[11] and random forest classifiers,[13] the collected compounds with $IC_{50} \geq 10$ μM or $pIC_{50}(-\log IC_{50}) \leq 5$ were assigned to class 0 (weak inhibitors), and the others were assigned to class 1 (strong inhibitors). We eliminated compounds reported to have both class 0 and class 1 activity for a given single compound as outliers. A principal component analysis (PCA) and structural clustering were performed to obtain the overall chemical diversity for all the finally selected 280 compounds. We used the property descriptors and functional class fingerprints with a maximum diameter of 4 (FCFP_4).[20] The data was additionally preprocessed by mean centering and unit-variance scaling. The 89.4% variation in the data indicates that the PCA gives three significant PCs. The plots of each data set in the space defined by the three principle components, PC1-PC2-PC3, are presented in Figure 1. Finally, we divided the collected data of 258

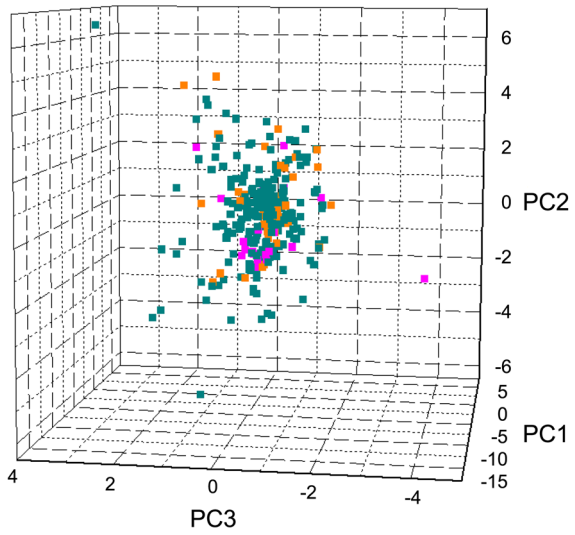

Figure 1. The PCA data plots of 280 hERG compounds for learning set (Blue), validation A set (Orange), and validation B set (Red).

Table 1. The number of compounds for data set

Class	Learning	Validation A	Validation B
0	88	16	12
1	136	18	10
Total	224	34	22

compounds among 280 compounds into a learning set (224 compounds) and a validation A set (34 compounds). The 224 compounds of the learning set were further divided into training and test sets by diverse or random selection. The in-house data of 22 compounds among the 280 compounds was assigned to a validation B set for a second evaluation. The information on the activity classes and the divided sets is listed in Table 1.

Classification Methods. Bayesian and random forest classification models were built from the training set using the molecular fingerprint and property descriptor sets in terms of the following: extended class fingerprint with a maximum diameter 6 (ECFP_6) in Pipeline Pilot 7.5,[20] number of rotatable bonds, number of hydrogen bond donors, number of hydrogen acceptors, AlogP, molecular weight, molecular fractional polar surface area, molecular fractional polar solvent-accessible surface area, and molecular solvent-accessible volume. The 224 compounds of the learning set were divided into training and test sets with a ratio ranging from 9/1 to 1/9, and classification models were then constructed based on each of the nine training sets. The training and test sets were generated by simple random selection or by diverse selection. For the diverse selection of training sets and test sets, the FCFP_4 fingerprint and property descriptors, respectively, were used.

Laplacian-modified naïve Bayesian classification models were built for each individual training set using the Learn Good Molecules component in Pipeline Pilot.[20] Laplacian-modified naïve Bayesian classifiers have proven useful in creating models that can work well even with noisy data.[21]

Random forest is essentially a collection of tree predictors, where each tree depends on the value of a randomly sampled parameter vector.[18] This means that a tree is trained only with randomly selected samples, which are called in-bag cases. In addition, a subset of descriptors is considered to be eligible to be divided at each node. To learn imbalanced data, the number of each class and the weight of each sample were set to be equal in 500 trees for the forest model. The square root of the number of descriptors was used as a splitting criterion within each tree.

Results and Discussion

In this study, predictions were carried out using the test set in the learning set and the validation A and validation B sets. Figure 2 shows the ability of the Laplacian-modified naïve Bayesian classification and random forest classification models to rank compounds according to their probability of being active, which varies according to the selection method, and the ratio of training sets to test sets. The area under the

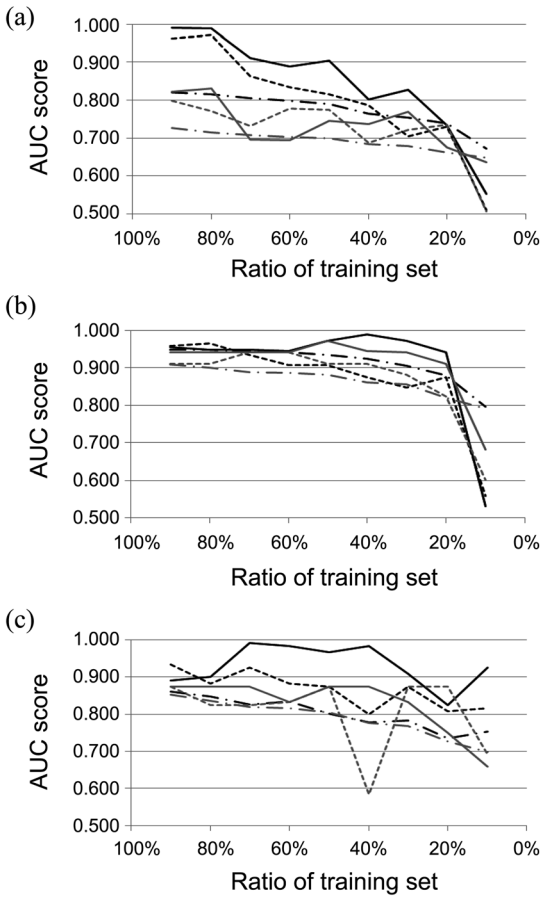

Figure 2. The plots of AUC scores for each test and validation sets: (a) test set, (b) validation A set, and (c) validation B set. The classification methods are distinguished between Laplacian-modified naïve Bayesian (black line) and random forest (gray line). The compound selection method of training sets are given by the line type: random (dot-dashed), fingerprint diverse (solid), and property diverse (dashed).

Table 2. The predictive performance obtained from all of the learned models by Laplacian-modified naïve Bayesian classification

Data sets	Ratio	9/1	8/2	7/3	6/4	5/5	4/6	3/7	2/8	1/9
Test	Accuracy (%)	90.5	95.5	78.8	79.8	75.9	75.9	46.2	61.8	61.7
	Sensitivity (%)	84.6	92.6	75.0	79.6	66.2	80.2	11.6	50.9	100.0
	Specificity (%)	100.0	100.0	84.6	80.0	90.9	69.2	100.0	78.6	2.5
Validation A	Accuracy (%)	85.3	82.4	88.2	88.2	91.2	94.1	61.8	79.4	55.9
	Sensitivity (%)	83.3	77.8	77.8	77.8	83.3	88.9	27.8	66.7	100.0
	Specificity (%)	87.5	87.5	100.0	100.0	100.0	100.0	100.0	93.8	6.3
Validation B	Accuracy (%)	90.9	86.4	86.4	90.9	72.7	90.9	54.5	59.1	45.5
	Sensitivity (%)	80.0	80.0	70.0	80.0	40.0	80.0	0.0	10.0	100.0
	Specificity (%)	100.0	91.7	100.0	100.0	100.0	100.0	100.0	100.0	0.0

receiver operating characteristic curve (AUC) scores for Laplacian-modified naïve Bayesian and random forest classification are given in Figure 2 for the test, validation A, and validation B sets. (The ROC plots are shown in supplementary materials.) The AUC provides a simple quality assessment for a classification model. The closer the AUC score is to 1.0, the better the model is at distinguishing samples in the good class from samples in the bad class. As shown in Figure 2, our results generally indicated that Laplacian-modified naïve Bayesian classification in combination with diverse selection by fingerprint or property descriptors could achieve a better predictive model, although a large data set was not fulfilled. As for the test set, random forest classification models showed significantly lower enhancement, as seen in Figure 2(a). On the other hand, for validation A set, as shown in Figure 2(b), models built on ratios from 9/1 to 6/4 of training sets to test sets yielded high AUC values of more than 0.9. However, as shown in Figure 2(c), for the in-house data set assigned to the validation B set, Laplacian-modified naïve Bayesian classification in combination with diverse selection by the fingerprint descriptor achieved better performance than any other models. We also estimated the predictive performance by various statistics such as accuracy, sensitivity, and specificity (%) on each of the test sets in the learning set and validation A and validation B sets from all learned models by Laplacian-modified naïve Bayesian classification, as presented in Table 2. As also can be seen from the ratios ranging from 9/1 to 8/2 in Table 2, the prediction accuracies for each set of compounds range from 95.5% to 82.4%, the sensitivities range from 92.6% to 77.8%, and the specificities range from 100.0% to 87.5% for the Laplacian-modified naïve Bayesian classification with fingerprint diverse selection models.

To identify the important features that contributed to class 0 or class 1, the ECFP_6 descriptors were extracted from each learned model of which the normalized probability values are less than −0.70 or more than 0.35. The normalized probability is the final contribution of the feature to the total relative estimate. If the value of the normalized probability is positive, the presence of the feature increases the likelihood that the molecule is a member of the 'class 1' subset. On the other hand, if the value of the normalized probability is negative, it decreases the likelihood that the molecule is a member of the 'class 1' subset. Subsets of the most important features corresponding to class 0 and class 1 are shown in Figure 3. These features were selected for a substructure search of the entire test and validation A and validation B data sets, and the number of compounds retrieved in each of these sets are 47, 11, and 13 compounds, respectively.

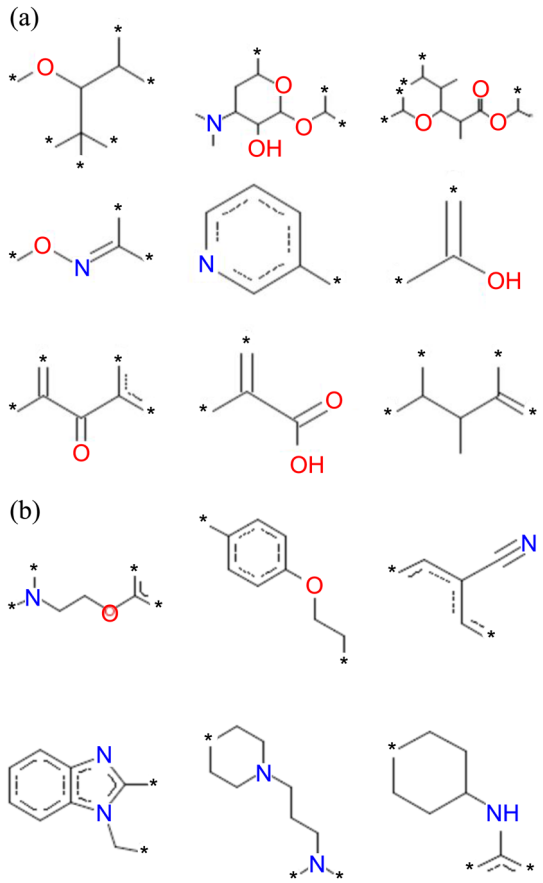

Figure 3. The important features for class 0 (a) and class 1 (b) which were obtained from each of nine Laplacian-modified naïve Bayesian classification models and then filtered by normalized probability and frequency.

Conclusion

In this study, we investigated hERG prediction with various training sets created by different classification methods. This work indicates that Laplacian-modified naïve Bayesian classi-

fication with diverse selection is useful for predicting hERG inhibitors when a large data set is not obtained. Introduction of a set of fingerprint or property descriptors for diverse selection also improves the prediction capability of machine learning methods such as Laplacian-modified naïve Bayesian and random forest classification.

Acknowledgments. This research was supported by the Center for Biological Modulators of the 21st Century Frontier R&D program, the Ministry of Science and Technology, Korea, and the Ministry of Knowledge Economy.

References

1. Vandenberg, J. I.; Walker, B. D.; Campbell, T. J. *Trends Pharmacol. Sci.* **2001**, *22*, 240.
2. Brown, A. M. *Cell Calcium* **2004**, *35*, 543.
3. Aronov, A. M. *Drug Discov. Today* **2005**, *10*, 149.
4. Mitcheson, J. S.; Chen, J.; Lin, M.; Culberson, C.; Sanguinetti, M. C. *PNAS.* **2000**, *97*, 12329.
5. Pearlstein, R. A.; Vaz, R. J.; Kang, J.; Chen, X.-L.; Preobrazhenskaya, M.; Shchekotikhin, A. E.; Korolev, A. M.; Lysenkova, L. N.; Miroshnikova, O. V.; Hendrix, J.; Rampe, D. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 1829.
6. Du, L.; Li, M.; You, Q.; Xia, L. *Biochem. Biophys. Res. Commun.* **2007**, *355*, 889.
7. Ekins, S.; Crumb, W. J.; Sarazan, R. D.; Wikel, J. H.; Wrighton, S. A. *J. Pharmacol. Exp. Ther.* **2002**, *301*, 427.
8. Cavalli, A.; Poluzzi, E.; De Ponti, F.; Recanatini, M. *J. Med. Chem.* **2002**, *45*, 3844.
9. Keserü, G. M. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 2773.
10. Thai, K. M.; Ecker, G. F. *Chem. Biol. Drug Des.* **2008**, *72*, 279.
11. Sun, H. *ChemMedChem* **2006**, *1*, 315.
12. Gepp, M. M.; Hutter, M. C. *Bioorg. Med. Chem.* **2006**, *14*, 5325.
13. Song, M.; Clark, M. *J. Chem. Inf. Model.* **2006**, *46*, 392.
14. Jia, L.; Sun, H. *Bioorg. Med. Chem.* **2008**, *16*, 6252.
15. Tobita, M.; Nishikawa, T.; Nagashima, R. *Bioorg. Med. Chem. Lett.* **2005**, *15*, 2886.
16. Leong, M. K. *Chem. Res. Toxicol.* **2007**, *20*, 217.
17. Jaynes, E. T. *Probability Theory: The Logic of Science*. Cambridge University Press: London, 2003.
18. Breiman, L. *Mach. Learn.* **2001**, *45*, 5.
19. Thomson Reuters Integrity^{SM}. Barcelona: Prous Science, S.A.U., a Thomson Reuters business. 2001. Available from: http://integrity.prous.com.
20. Accelrys Software Inc., Pipeline Pilot Release 7.5, San Diego: Accelrys Software Inc., 2007.
21. Glick, M.; Jenkins, J. L.; Nettles, J. H.; Hitchings, H.; Davies, J. W. *J. Chem. Inf. Model.* **2006**, *46*, 193.