Available online at www.sciencedirect.com
SCIENCE DIRECT®
Bioorganic & Medicinal Chemistry Letters 15 (2005) 2886–2890
Bioorganic & Medicinal Chemistry Letters

A discriminant model constructed by the support vector machine method for HERG potassium channel inhibitors

Motoi Tobita,* Tetsuo Nishikawa and Renpei Nagashima

Reverse proteomics research institute, 2-6-7 Kazusa-kamatari, Kisarazu-si, Chiba 292-0818, Japan

Received 31 January 2005; revised 17 March 2005; accepted 22 March 2005
Available online 19 April 2005

Abstract—HERG attracts attention as a risk factor for arrhythmia, which might trigger *torsade de pointes*. A highly accurate classifier of chemical compounds for inhibition of the HERG potassium channel is constructed using support vector machine. For two test sets, our discriminant models achieved 90% and 95% accuracy, respectively. The classifier is even applied for the prediction of cardio vascular adverse effects to achieve about 70% accuracy. While modest inhibitors are partly characterized by properties linked to global structure of a molecule including hydrophobicity and diameter, strong inhibitors are exclusively characterized by properties linked to substructures of a molecule.
© 2005 Elsevier Ltd. All rights reserved.

Inhibition of a *human ether-a-go-go* related gene (HERG) can lead to a prolongation of the QT interval which, in the worst case, triggers *torsade de pointes* arrhythmia. Many drugs and small molecules that are reported to inhibit HERG span wide range in their therapeutic categories and chemical structures.[1] Thus, a prediction of the HERG inhibiting potency of drug candidate molecules at an early stage of drug development process is important. Along this line, prior work includes the derivation of a pharmacophore model for HERG channel inhibitors using three-dimensional quantitative structure–activity relationship (3D-QSAR) approaches.[2–4] Ekins et al.[2] derived a pharmacophore model consisting of four hydrophobic features and one positive ionizable feature using an algorithm called catalyst. Cavalli et al.[3] also derived a pharmacophore model containing two aromatic moieties, a phenyl ring, and a basic nitrogen. Pearlstein et al.[4] used a comparative molecular similarity analysis (CoMSiA) 3D-QSAR approach together with the homology modeling of HERG using the MthK potassium channel structure as a template.[5] These authors suggest that (1) a hydrophobic feature, which optimally consists of an aromatic group, is capable of engaging in π-stacking with a Phe656 side chain, (2) the basic nitrogen appears to undergo a π–cation interaction with Tyr652, and (3) the pore shape constrains possible conformations of HERG inhibitors. Discriminant models have been constructed to filter out potential HERG inhibitors.[6,7] Keserü[6] showed that his discriminant model predicts 83% of actives and 87% of inactives correctly by hologram QSAR[8] method using 55 compounds for the model building and 13 compounds for the validation. Aronov and Goldman[7] combined a two-dimensional (2D) topological similarity filter with a 3D pharmacophore ensemble procedure to discriminate between 85 actives and 329 inactives. In their study, the 50-fold cross validation resulted in the overall classification accuracy of 82%. We suggest, in this paper, a novel HERG filter that gives a higher degree of classification accuracy.

Literature was surveyed to collect IC_{50} values of as many drugs as possible, as determined by the patch clamp HERG current inhibition assay using the mammalian cell line, HEK, or CHO. This resulted in IC_{50} values of 73 drugs. Many of the collected values were taken from the Fenichel's database[9] and Ref. 6 and the rest from recent literature.[10–17] Since the collected data are from different sources, we did not predict IC_{50} values themselves, but used them only for the purpose of making a boundary defining two classes. Thus, an error introduced should be kept minimal. Also note that no data from Xenopus Oocyte, which would skew the quality of data set, are included in our set. The collected IC_{50} values of these drugs are given in Table 1, which also shows that these drugs cover a diversity of therapeutic categories. Those therapeutic categories

Keywords: HERG; In silico; SVM; Prediction; Discriminant analysis.
* Corresponding author. Tel.: +81 438523975; fax: +81 438523986; e-mail: toby@rd.hitachi.co.jp

0960-894X/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.bmcl.2005.03.080

Table 1. Experimentally measured pIC_{50} values and therapeutic categories of 73 HERG inhibiting compounds used in this study

Compound name	pIC_{50}	Therapeutic category	Compound name	pIC_{50}	Therapeutic category
Astemizole	8.00	Antihistamine	Ebastine	5.52	Antihistamine
Dofetilide	8.00	Antiarrhythmic	Alosetron	5.49	Antidiarrheal
Sertindole	8.00	Antipsychotic	Sildenafil	5.48	PDE5 inhibitor
Ibutilide	8.00	Antiarrhythmic	Imipramine	5.47	Antidepressant
Lidoflazine	7.80	Ca^{2+} blocker	Granisetron	5.43	Antiemetic
Tolterodine	7.77	Muscarinic antagonist	Flecainide	5.41	Antiarrhythmic
E-4031	7.70	Antiarrhythmic	Citalopram	5.40	Antidepressant
Haloperidol	7.52	Antipsychotic	Mefloquine	5.25	Antimalarial
Droperidol	7.49	Antipsychotic	Cocaine	5.14	Narcotic
Cisapride	7.40	Prokinetic	Buprenorphine	5.12	Opioid blocker
Pimozide	7.30	Antipsychotic	Methadone	5.01	Opioid blocker
Ziprasidone	6.92	Antipsychotic	Nitrendipine	5.00	Ca^{2+} blocker
Verapamil	6.85	Ca^{2+} blocker	Amiodarone	5.00	Antiarrhythmic
Risperidon	6.82	Antipsychotic	Amitriptyline	5.00	Antidepressant
Domperidone	6.79	Prokinetic	Carvedilol	4.98	Antiarrhythmic
Loratadine	6.77	Antihistamine	Dolasetron	4.92	Antiemetic
Olanzapine	6.74	Antipsychotic	Diltiazem	4.76	Ca^{2+} blocker
Thioridazine	6.72	Antipsychotic	Sparfloxacin	4.74	Antibiotics
Terfenadine	6.70	Antihistamine	Chlorpheniramine	4.68	Antihistamine
Halofantrine	6.70	Antimalarial	Diphenhydramine	4.57	Antihistamine
Terikalant	6.60	Antiarrhythmic	Cetirizine	4.52	Antihistamine
Quinidine	6.49	Antiarrhythmic	Grepafloxacin	4.30	Antibiotics
Meperidine	6.49	Opioid blocker	Nifedipine	4.30	Ca^{2+} blocker
Clozapine	6.49	Antipsychotic	EDDP	4.30	Opioid blocker
Mizolastine	6.36	Antihistamine	Clarithromycin	4.23	Antibiotics
Mesoridazine	6.26	Antipsychotic	Disopyramide	4.04	Antiarrhythmic
Bepridil	6.26	Antianginal	Epinastine	4.00	Antihistamine
Ondansetron	6.09	Antiemetic	Moxifloxacin	3.89	Antibiotics
Desipramine	5.86	Antidepressant	Gatifloxacin	3.89	Prokinetic
Azimilide	5.85	Antiarrhythmic	Procainamide	3.86	Antiarrhythmic
Mibefradil	5.84	Antiemetic	Nicotine	3.61	Cholinergic
Chlorpromazine	5.83	Antipsychotic	Codeine	3.52	Opioid blocker
Fluoxetine	5.82	Antidepressant	Levofloxacin	3.04	Prokinetic
Prazosine	5.80	α_1-Adrenoreceptor antagonist	Ciprofloxacin	3.02	Antibiotics
Fentanyl	5.74	Opioid blocker	Morphine	3.00	Opioid blocker
Ketoconazole	5.72	Antifungal	Ofloxacin	2.85	Prokinetic
Laam	5.66	Opioid blocker			

common in HERG inhibitors include antiarrhythmics, antipsychotics, antihistamines, opioid blockers, and Ca^{2+} blockers. For each of these drugs, 57 2D descriptors, defined as a function of the two-dimensional structure, were computed by MOE.[18] Also, 51 molecular fragment-count descriptors were computed. The used molecular fragments were a subset of the public 166-bit MACCS key set.[19] Discriminant models were constructed and evaluated by the following three steps: (1) selection of important descriptors from the computed 108 (57 + 51) descriptors using support vector machine (SVM), (2) construction of a classifier using SVM and optimization of related parameters, and (3) evaluation of the accuracy of the classifier by the 10-fold cross validation. SVM is a machine learning method that is superior to the other methods in two factors.[20] First, non-linear class boundary can be implemented using linear models made by descriptors. Second, over-fitting is unlikely to occur. Those strengths are achieved as a result of using relatively long training time as a tradeoff.

In order to critically evaluate the predictive ability of our approach, 73 drugs were separated into actives and inactives in two different ways. First separation boundary was set at pIC_{50} (the negative of $\log(IC_{50})$) = 4.4, which defined 58 actives and 15 inactives. Another separation at $pIC_{50} = 6.0$ defined 28 actives and 45 inactives. For both separations, we found that selecting eight descriptors resulted in the most accurate classifier. This preliminary analysis and the following construction of an SVM classifier were done using WEKA[20] package. The chosen descriptors are listed in Table 2. The number in parenthesis is the average 'merit' of the corresponding descriptor, which is related to the contribution of the descriptor to the classification. For the separation at $pIC_{50} = 4.4$, five 2D descriptors and three molecular fragment-count descriptors were selected. $S\log P$ gives insight into the hydrophobicity of molecules. Two PEOE_VSA descriptors indicate the surface area of strongly negatively charged region (PEOE_VSA6) and slightly positively charged region (PEOE_VSA + 1). DIAMETER is a suggestion of the size of the inhibitor molecule. SMR_VSA5 indicates the surface area of a molecular fragment having a value of molar refractivity between 0.440 and 0.485. The number of 'NH_2' fragments was found to be important, which might relate to possible hydrogen bonding sites. The fragment 'ACH_2CH_2A' may relate to the flexibility of a molecule and suggests that the existence of a rather long chain is important. This chain also relates to the local hydropho-

Table 2. The chosen descriptors used to construct the classifier

Descriptor	Separation at $pIC_{50} = 4.4$	Separation at $pIC_{50} = 6.0$
1	*S*log *P* (7.0)	# OAAAO (8.0)
2	PEOE_VSA + 1 (6.3)	# ACH_2AAACH_2A (5.4)
3	PEOE_VSA-6 (5.8)	# Nnot%A%A (4.8)
4	DIAMETER (5.3)	VSA_BASE (4.7)
5	SMR_VSA5 (4.3)	PEOE_VSA0 (4.6)
6	# NH_2 (3.3)	# ACH_2AACH_2A (4.1)
7	# ACH_2CH_2A (2.8)	SMR_VSA0 (3.9)
8	# A$A!A$A (1.2)	# 8-membered or larger ring (1.4)

Descriptors staring from '#' are the count of substructures defined by the character string which follows to the '#'. Abbreviations used in the character strings are defined as follows: A, a non-hydrogen atom; $, a bond belonging to a ring(s); !, a bond belonging to a chain; %, an aromatic bond; not%, a non-aromatic bond.

bicity produced by CH_2 groups. Finally, the fragment 'A$A!A$A' indicates a substructure consisting of two rings connected by a bond. Note that each of individual descriptors does not necessarily divide the actives and inactives clearly. Since SVM can include non-linearity in the classifier by nature, the combination of the selected descriptors defines a non-linear boundary between actives and inactives. For the separation at $pIC_{50} = 6.0$, three 2D descriptors and five molecular fragment-count descriptors are selected. High contributors are fragment-count descriptors such as 'OAAAO' fragment-count. The existence of large fragments 'ACH_2AAACH_2A' and 'ACH_2AACH_2A' is noteworthy. Since the size of the channel pore is known to be larger than 12 Å,[4,5] flexible porefillers may be important to achieve the high potency toward HERG. The importance of a basic atom[3] is rediscovered as indicated by VSA_BASE descriptor. Comparing the selected descriptors for the two separations, two of the selected descriptors for the separation at $pIC_{50} = 4.4$ (*S*log *P* and DIAMETER) are global properties of a molecule, while all of the selected descriptors for the separation of $pIC_{50} = 6.0$ are related to a partial structure of a molecule. This fact implies that for the modest HERG inhibition, that is, $pIC_{50} = 4.4$ or higher, a molecule has to satisfy global conditions such as *S*log *P* and DIAMETER. On the other hand, in order to have strong inhibiting potency, that is, $pIC_{50} = 6.0$ or higher, optimization through modification in terms of structural fragments appears to be more important. It seems that the descriptors chosen for the $pIC_{50} = 6.0$ separation are special cases of those chosen for the $pIC_{50} = 4.4$. For example, fragments 'ACH_2AAACH_2A' and 'ACH_2AACH_2A' are special cases contributing to *S*log *P*. Also, fragment 'Nnot%A%A' is a special case of 'A$A!A$A' fragment. Another important point to note is that many of the selected 2D descriptors are 'VSA' descriptors such as PEOE_VSA and SMR_VSA.[21] Those descriptors can, in principle, distinguish between small differences in a local region of two globally similar molecules. The use of such 'local' information as given by the VSA descriptors and molecular fragment-count descriptors seems necessary to construct a robust and accurate in silico model from topological information.

Mapping of the selected descriptors for the separation at $pIC_{50} = 6.0$ onto molecular structure suggests the presence of fragments contributing to the potency toward HERG. Two fragment patterns are shown in Figure 1. Most of the molecules that strongly bind to HERG have either of the fragments shown as pattern 1. These fragments are directly related to 'ACH_2AAACH_2A' and 'ACH_2AACH_2A' descriptors where one of the non-hydrogen atoms is specifically a nitrogen atom. Also, many molecules that strongly bind to HERG have one of the fragments shown as pattern 2. These fragments are commonly characterized as a nitrogen atom connected to an aromatic ring. This structure is described by the 'Nnot%A%A' descriptor. The fragment patterns have correspondence with pharmacophore models derived earlier.[2,3,7] These authors commonly define 'positively ionizable atom' as a pharmacophore. This can be a non-aromatic nitrogen atom in pattern 1 of Figure 1. Also, 'hydrophobic region', and 'ring connected through a nitrogen atom' are picked up as components of pharmacophore. Those components are also seen in Figure 1.

The results of the classification were described. SVM classifiers were constructed for the two test sets. Prior to classifier construction, all the descriptor values are standardized. The complexity parameter, which was passed to the SVM algorithm,[22] was set to 2.0. As the kernel of the classifier, we chose to use a radial basis function. The kernel parameters and the classification accuracy are given in Table 3. First, manually optimized exponent parameters for the radial basis function were 0.062 and 0.040 for the separation at $pIC_{50} = 4.4$ and $pIC_{50} = 6.0$, respectively. For the separation at $pIC_{50} = 4.4$, 56 of 58 active drugs and 13 of 15 inactive drugs were classified correctly, giving the overall classification accuracy of 95%. For the separation at $pIC_{50} = 6.0$, 24 of 28 actives and 42 of 45 inactives were correctly classified, resulting in the overall classification accuracy of 90%. These results indicate significantly better values for classification accuracy than those reported.[6,7] The results reported by Keserü[6] defined the

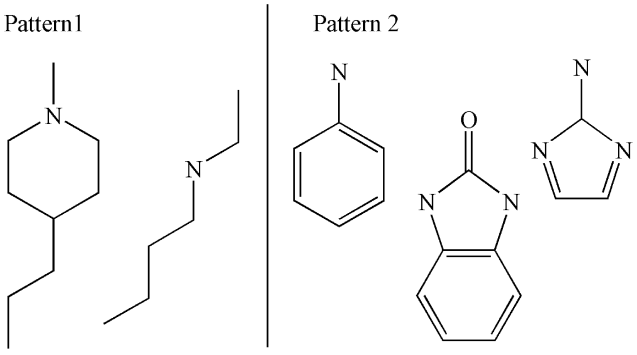

Figure 1. Contributing molecular fragments for strong HERG inhibition.

Table 3. The chosen exponent parameter for the SVM radial basis function kernel and the result of the discriminant analysis compared to prior work

	Separation at $pIC_{50} = 4.4$	Separation at $pIC_{50} = 6.0$
Exponent	0.040	0.062
Active	56/58 (97%)	24/28 (86%)
Inactive	13/15 (87%)	42/45 (93%)
Overall	69/73 (95%)	66/73 (90%)
Active	60/85 (71%)[a]	5/6 (83%)[a]
Inactive	280/329 (85%)[a]	6/7 (87%)[a]
Overall	340/414 (82%)[a]	11/13 (85%)[a]

[a] Ref. 6.

active/inactive boundary at $IC_{50} = 1\ \mu M$ ($pIC_{50} = 6.0$) and Aronov and Goldman[7] set the active/inactive boundary at $IC_{50} = 40\ \mu M$ ($pIC_{50} = 4.4$). Therefore, the comparisons were made between corresponding data in the active/inactive boundary. A further test was performed using an external data set. The test showed that classifier of predicting whether or not a given molecule had $pIC_{50} = 6.0$ could also predict known cardiovascular adverse effects with an accuracy of about 70%. The test set consists of 827 drugs, which are diverse in therapeutic use and are included in Drugdex database.[23] 58 drugs out of 827 (7%) were extracted as candidates having $pIC_{50} = 6.0$ or higher. Because all the pIC_{50} values for those drugs were unavailable, the result of the prediction was evaluated by correlation to the reported adverse effects which may be related to HERG. Drugs are grouped into four by the occurrence of at least one word listed under each group in the section of 'cardiovascular adverse effects'; group 1: 'sudden death', '*torsade de pointes*', or 'cardiac arrest', group 2: '(cardiac) arrhythmia', 'QT prolongation', 'dysrhythmia', 'extrasystole', or 'bradycardia', group 3: 'tachycardia' or 'palpitation' except rare cases, group 4: others. We found that 39 out of 58 drugs (67%) belonged to either group 1 or group 2. Adding the group 3 drugs covered 45 out of 58 drugs (78%). This coverage is only slightly worse than the corresponding number obtained in the cross validation (86%, 24/28) as shown in Table 3. It is not unnatural to consider that these 45 drugs have an interaction with HERG. Conversely, to see whether drugs with predicted $pIC_{50} = 6.0$ or lower have lesser degrees of cardiovascular adverse effects, one appearing at every 15th drug was picked up to form a sample of 51 drugs from a list of 769 drugs with predicted $pIC_{50} = 6.0$ or lower that were classified according to therapeutic categories. The occurrence of those words indicating cardiovascular adverse effects was investigated. As a result, 38 out of 51 drugs (75%) belonged to group 3 or 4. Group 4 alone covered 31 out of 51 drugs (61%).

The success of our approach can be attributed to (1) the set of descriptors used and (2) the use of SVM for choosing descriptors and constructing the classifier. Because, allowing for certain differences in selection of descriptors, many of the descriptors used in this study and in prior work[6,7] are related to substructures of a molecule, we feel that the achieved high accuracy is due rather to the use of SVM. This is supported by other SVM-based studies. Xue et al.[24] showed that binding affinity to human serum albumin is better predicted by SVM than by a linear model based on heuristics. Further, Burbidge et al.[25] compared various machine learning techniques on the problem of classifying inhibition of dehydrofolate reductase. They concluded that SVM outperformed all the other tested techniques including neural network and decision tree. In fact, while neural network, naive Bayes, and decision tree methods were tested for classifying HERG inhibitors in this study (data not shown), these methods yielded less accurate results than SVM.

In summary, a highly accurate discrimination model for HERG inhibition was constructed using SVM. Using these results, an in silico screening program can be constructed.

Acknowledgements

This work was supported by a grant from NEDO project of the Ministry of Economy, Trade, and Industry of Japan.

References and notes

1. De Ponti, F.; Poluzzi, E.; Montanaro, N. *Eur. J. Clin. Pharmacol.* **2001**, *57*, 185.
2. Ekins, S.; Crumb, W. J.; Sarazan, R. D.; Wikel, J. H.; Wrighton, S. A. *J. Pharmacol. Exp. Ther.* **2002**, *301*, 427.
3. Cavalli, A.; Poluzzi, E.; De Ponti, F.; Recanatini, M. *J. Med. Chem.* **2002**, *45*, 3844.
4. Pearlstein, R.; 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.
5. Jiang, Y.; Lee, A.; Chen, J.; Cadene, M.; Chait, B. T.; MacKinnon, R. *Nature* **2002**, *417*, 523.
6. Keserü, G. M. *Bioorg. Med. Chem. Lett.* **2003**, *13*, 2773.
7. Aronov, A. M.; Goldman, B. B. *Bioorg. Med. Chem.* **2004**, *12*, 2307.
8. Heritage, T. W.; Lowis, D. R.. In *Rational Drug Design*; Parrill, A. L., Reddy, M. R., Eds.; ACS Symposium Series; American Chemical Society: Washington, 2000; Vol. 719, p 212.
9. Fenichel. R. R. http://www.fenichel.net/pages/Professional/subpages/QT/Tables/pbydrug.htm.
10. Ridley, J. M.; Dooley, P. C.; Milnes, J. T.; Witchel, H. J.; Hancox, J. C. *J. Mol. Cell Cardiol.* **2004**, *36*, 701.
11. Kang, J.; Chen, X. L.; Wang, H.; Ji, J.; Reynolds, W.; Lim, S.; Hendrix, J.; Rampe, D. *J. Pharmacol. Exp. Ther.* **2004**, *308*, 935.
12. Kongsamut, S.; Kang, J.; Chen, X. L.; Roehr, J.; Rampe, D. *Eur. J. Pharmacol.* **2002**, *450*, 37.
13. Drolet, B.; Rousseau, G.; Daleau, P.; Cardinal, R.; Turgeon, J. *Circulation* **2000**, *102*, 1883.
14. Katchman, A. N.; Mcgroary, K. A.; Kilborn, M. J.; Kornick, C. A.; Manfredi, P. L.; Woosley, R. L.; Ebert, S. N. *J. Pharmacol. Exp. Ther.* **2002**, *303*, 688.
15. Ko, C. M.; Ducic, I.; Fan, J.; Shuba, Y. M.; Morad, M. *J. Pharmacol. Exp. Ther.* **1997**, *281*, 233.
16. Kuryshev, Y. A.; Brown, A. M.; Wang, L.; Benedict, C. R.; Rampe, D. *J. Pharmacol. Exp. Ther.* **2000**, *295*, 614.
17. Ridley, J. M.; Milnes, J. T.; Benest, A. V.; Masters, J. D.; Witchel, H. J.; Hancox, J. C. *Biochem. Biophys. Res. Commun.* **2003**, *306*, 388.

18. MOE: Molecular operating environment; Chemical computing group; http://www.chemcomp.com/.
19. Maccs II; Molecular Design Ltd.; http://www.mdl.com/.
20. Witten, I. H.; Frank, E. *Data Mining: Practical machine learning tools with Java implementations*; Morgan Kaufmann: San Francisco, 2000.
21. Labute, P. *J. Mol. Graphics Modell.* **2000**, *18*, 464.
22. (a) Platt, J. In *Advances in Kernel Methods Support Vector Learning*; Platt Scholkopf, B., Burges, C., Smola, A., Eds.; MIT; (b) Keerthi, S. S.; Shevade, S. K.; Bhattacharyya, C.; Murthy, K. R. K. *Neural Comput.* **2001**, *13*, 637.
23. Thomson healthcare. http://www.thomsonhc.com/.
24. Xue, C. X.; Zhang, R. S.; Liu, H. X.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1693.
25. Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. *Comput. Chem.* **2001**, *26*, 5.