Virtual Screening of Abl Inhibitors from Large Compound Libraries by Support Vector Machines

X. H. Liu,[†] X. H. Ma,[†,⊥] C. Y. Tan,[‡] Y. Y. Jiang,[‡] M. L. Go,[§] B. C. Low,[⊥] and Y. Z. Chen*[,†]

Bioinformatics and Drug Design Group, Department of Pharmacy, Centre for Computational Science and Engineering, National University of Singapore, Blk S16, Level 8, 3 Science Drive 2, Singapore 117543, The Key Laboratory of Chemical Biology, Guangdong Province, The Graduate School at Shenzhen, Tsinghua University, Shenzhen, Guangdong, P. R. China, Department of Pharmacy, National University of Singapore, BLK S4, 18 Science Drive 4, Singapore 117543, and Department of Biological Science, National University of Singapore, Blk S2, Level 5, Science Drive 4, Singapore 117543

Received April 14, 2009

Abl promotes cancers by regulating cell morphogenesis, motility, growth, and survival. Successes of several marketed and clinical trial Abl inhibitors against leukemia and other cancers and appearances of reduced efficacies and drug resistances have led to significant interest in and efforts for developing new Abl inhibitors. In silico methods of pharmacophore, fragment, and molecular docking have been used in some of these efforts. It is desirable to explore other in silico methods capable of searching large compound libraries at high yields and reduced false-hit rates. We evaluated support vector machines (SVM) as a virtual screening tool for searching Abl inhibitors from large compound libraries. SVM trained and tested by 708 inhibitors and 65 494 putative noninhibitors correctly identified 84.4 to 92.3% inhibitors and 99.96 to 99.99% noninhibitors in 5-fold cross validation studies. SVM trained by 708 pre-2008 inhibitors and 65 494 putative noninhibitors correctly identified 50.5% of the 91 inhibitors reported since 2008 and predicted as inhibitors 29 072 (0.21%) of 13.56M PubChem, 659 (0.39%) of 168K MDDR, and 330 (5.0%) of 6 638 MDDR compounds similar to the known inhibitors. SVM showed comparable yields and substantially reduced false-hit rates against two similarity based and another machine learning VS methods based on the same training and testing data sets and molecular descriptors. These suggest that SVM is capable of searching Abl inhibitors from large compound libraries at low false-hit rates.

INTRODUCTION

Abl plays key roles in cancers by regulating morphogenesis and motility and by promoting cell growth and survival via Bcr-Abl mediated activation of Src-family kinases and PI3K, Ras, Myc, c-jun, and STAT pathways.[1] Abl inhibitors are effective in the treatment of leukemia and in clinical trials of other cancers.[2−4] In some cases, these inhibitors show negligible activity against common mutations and modest effects in advanced cancer phases, and some patients develop resistance associated with Abl kinase domain mutations.[4] The successes and problems of these inhibitors have raised significant interest in and has led to intensifying efforts for discovering new Abl inhibitors.[4,5] Several in silico methods have been used for facilitating the search and design of Abl inhibitors, which include pharmacophore,[6] QSAR,[7] scaffold assembly,[8] molecular docking,[9,10] and their combinations.[11,12]

These in silico methods have shown impressive capability in the identification of potential Abl inhibitors, but their applications may be affected by such problems as the vastness and sparse nature of chemical space that needs to be searched, the complexity and flexibility of target structures, the difficulties in accurately estimating binding affinity and solvation effects, and the limited diversity of training active compounds.[13−15] Therefore, it is desirable to explore other in silico methods that complement these methods with expanded coverage of chemical space, increased screening speed, and reduced false-hit rates without necessarily relying on the modeling of target structural flexibility, binding affinity, and salvation effects.

A ligand-based virtual screening (VS) method, support vector machines (SVM), has been explored as such a method that produces high yields and low false-hit rates in searching active agents of single and multiple mechanisms from large compound libraries (i.e., with an expanded applicability domain)[16] and in identifying active agents of diverse structures.[16−20] Good VS performance can also be achieved by SVM trained from sparsely distributed active compounds.[21] SVM classifies active compounds based on differentiating physicochemical profiles between active and inactive compounds rather than on structural similarity to active compounds per se, which has the advantage of not relying on the accurate computation of structural flexibility, activity-related features, binding affinity, and solvation effects. Moreover, the fast speed and expanded applicability domain of SVM enables an efficient search of a vast chemical space. Therefore, SVM may be a potentially useful VS tool

* Corresponding author: Telephone: 65-6874-6877. Fax: 65-6774-6756. E-mail: phacyz@nus.edu.sg.
[†] Bioinformatics and Drug Design Group, Department of Pharmacy, National University of Singapore.
[⊥] Department of Biological Science, National University of Singapore.
[‡] Tsinghua University.
[§] Department of Pharmacy, National University of Singapore.

10.1021/ci900135u CCC: $40.75
Published on Web 08/18/2009

to complement other in silico methods for searching ABL inhibitors from large libraries.

In this work, we developed a SVM VS model for identifying Abl inhibitors and evaluated its performance by both a 5-fold cross validation test and a large compound database screening test. In the 5-fold cross validation test, a data set of Abl inhibitors and noninhibitors was randomly divided into five groups of approximately equal size, with four groups used for training a SVM VS tool and one group used for testing it; the test process is repeated for all five possible compositions to derive an average VS performance. In the large database screening test, a SVM VS tool was developed by using Abl inhibitors published before 2008, its yield (percent of known inhibitors identified as virtual-hits) was estimated by using Abl inhibitors reported since 2008 and not included in the training data sets. Virtual hits and false-hit rate in searching large libraries were evaluated by using 13.56 M PubChem, 168K MDDR, and 6 638 MDDR compounds similar in structural and physicochemical properties to the known Abl inhibitors.

PubChem and MDDR contain high percentages of inactive compounds significantly different from the Abl inhibitors, and the easily distinguishable features may make VS enrichments artificially good.[22] Nonetheless, certain percentages of PubChem and MDDR compounds are kinase inhibitors or are similar to known Abl inhibitors. For instance, about 1 500 MDDR and 10 000 PubChem compounds are kinase inhibitors, and 6 638 MDDR compounds are similar to at least one known Abl inhibitor. Therefore, VS performance may be more strictly tested by using these and other compounds that resemble the physicochemical properties of the known Abl inhibitors, so that enrichment is not simply a separation of trivial physicochemical features.[23] To further evaluate whether our SVM VS tool predicts Abl inhibitors and noninhibitors rather than membership of certain compound families, distribution of the predicted active and inactive compounds in the compound families was analyzed.

Moreover, VS performance of SVM was compared to those of two similar VS methods, Tanimoto similarity searching and k nearest-neighbor (kNN), and to an alternative but equally popular machine learning method, probabilistic neural network (PNN), which is based on the same training and testing data sets (same sets of PubChem and MDDR compounds) and molecular descriptors. In a study that compares the performance of SVM to 16 classification and nine regression methods, it has been reported that SVMs show mostly good performances both on classification and regression tasks, but other methods proved to be very competitive.[24] Therefore, it is useful to evaluate the VS performance of SVM in searching large compound libraries by comparison with those of similarity based approaches and other typical machine learning methods.

METHODS

Compound Collections and Construction of Training and Testing Data Sets. A total of 708 Abl inhibitors, with IC50 < 50 μM, were collected from the literatures[11,25−27] and the BindingDB database.[28] The inhibitor selection criterion of IC50 < 50 μM was used because it covers most of the reported HTS and VS hits.[29,30] The structures of representative Abl inhibitors are shown in Figure 1. As few noninhibitors have been reported, putative noninhibitors were generated by using our method for generating putative inactive compounds.[16,21] This method requires no knowledge of known inactive and active compounds of other target classes, which enables more expanded coverage of the "noninhibitor" chemical space. Although the yet to be discovered inhibitors are likely distributed in some of these noninhibitor families, a substantial percentage of these inhibitors are expected to be identified as inhibitors rather than noninhibitors, even though representatives of their families are putatively assigned as noninhibitors.[16] The 13.56M PubChem and 168K MDDR compounds were grouped into 8 423 compound families by clustering them in the chemical space defined by their molecular descriptors.[31,32] The number of generated families is consistent with the 12 800 compound-occupying neurons (regions of topologically close structures) for 26.4M compounds of up to 11 atoms[33] and the 2,851 clusters for 171,045 natural products.[34]

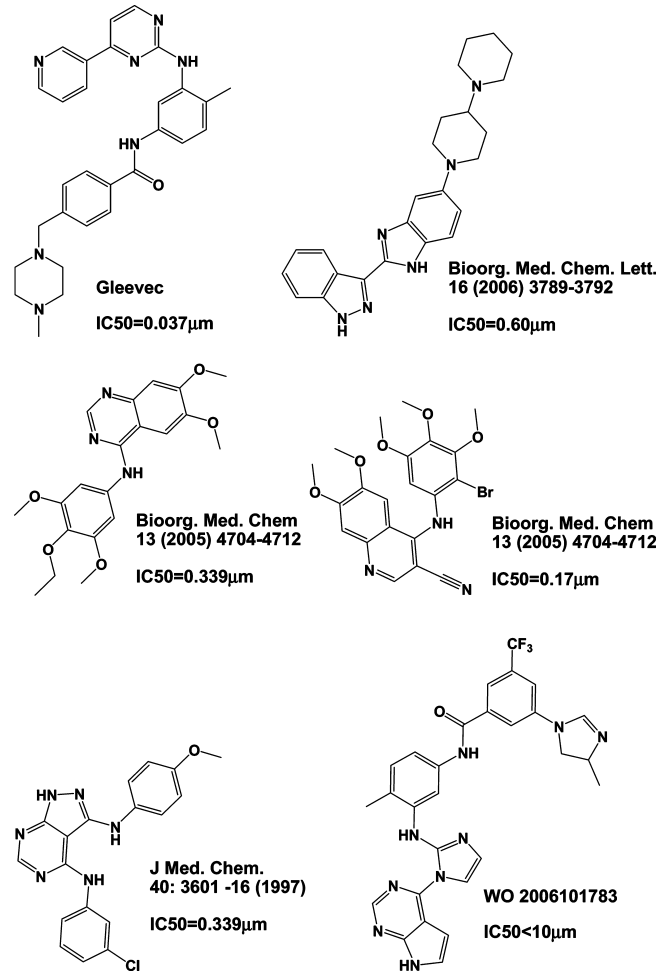

Figure 1. The structures of representative Abl inhibitors.

The collected Abl inhibitors are distributed in 221 families. Because of the extensive efforts in searching kinase inhibitors from known compound libraries, the number of undiscovered Abl inhibitor families in PubChem and MDDR databases is expected to be relatively small, most likely no more than several hundred families. The ratio of the discovered and undiscovered inhibitor families (hundreds) and the families that contain no known inhibitor of each kinase (8 423 based on the current versions of PubChem and MDDR) is expected

Table 1. Molecular Descriptors Used in This Work

descriptor class	number of descriptors in class	descriptors
simple molecular properties[78]	18	number of C,N,O,P,S, number of total atoms, number of rings, number of bonds, number of non-H bonds, molecular weight, number of rotatable bonds, number of H-bond donors, number of H-bond acceptors, number of five-member aromatic rings, number of six-member aromatic rings, number of N heterocyclic rings, number of O heterocyclic rings, number of S heterocyclic rings.
chemical properties[79]	3	Sanderson electronegativity, molecular polarizability, aLogp
molecular connectivity and shape[78,80]	35	Schultz molecular topological index, Gutman molecular topological index, Wiener index, Harary index, gravitational topological index, molecular path count of length 1−6, total path count, Balaban Index J, 0−2th valence connectivity index, 0−2th order delta chi index, Pogliani index, 0−2th solvation connectivity index, 1−3th order Kier shape index, 1−3th order kappa alpha shape index, Kier molecular flexibility index, topological radius, graph-theoretical shape coefficient, eccentricity, centralization, Logp from connectivity.
electro-topological state[78,81]	42	sum of estate of atom type sCH3, dCH2, ssCH2, dsCH, aaCH, sssCH, dssC, aasC, aaaC, sssC, sNH3, sNH2, ssNH2, dNH, ssNH, aaNH, dsN, aaN, sssN, ddsN, aOH, sOH, ssO, sSH; sum of estate of all heavy atoms, all C atoms, all hetero atoms, sum of estate of H-bond acceptors, sum of H estate of atom type HsOH, HdNH, HsSH, HsNH2, HssNH, HaaNH, HtCH, HdCH2, HdsCH, HaaCH, HCsats, HCsatu, Havin, sum of H estate of H-bond donors

to be <15%. Therefore, a putative noninhibitor training data set can be generated by extracting a few representative compounds from each of those families that contain no known inhibitor, with a maximum possible "wrong" classification rate of <15%, even when all of the undiscovered inhibitors are misplaced into the noninhibitor class. The noise level generated by up to 15% wrong negative family representation is expected to be substantially smaller than that of the maximum 50% false-negative noise level tolerated by SVM.[19] Based on earlier studies[16,21] and this work, it is expected that a substantial percentage of the undiscovered inhibitors in the putative noninhibitor families can be classified as inhibitor despite that their family representatives are placed into the noninhibitor training sets.

It is noted that, in the database screening test, 50% of families that contain Abl inhibitors reported since 2008 are not covered by the Abl inhibitor training data set (inhibitors reported before 2008), and the representative compounds of these families were deliberately placed into the inactive training sets as these inhibitors are not supposed to be known in our study. As shown in earlier studies[16,21] and in this work, a substantial percentage of the inhibitors in these misplaced inhibitor-containing noninhibitor families were predicted as inhibitors by our SVM VS tool. Moreover, a small percentage of the compounds in these putative noninhibitor data sets are expected to be unreported and undiscovered inhibitors, and their presence in these data sets is not expected to significantly affect the estimated false-hit rate of SVM.

Molecular Descriptors. Molecular descriptors are quantitative representations of structural and physicochemical features of molecules, which have been extensively used in deriving structure−activity relationships,[35,36] quantitative structure activity relationships,[37,38] and VS tools.[39−46] All of the 98 1D and 2D descriptors available from our software[47] were used in this work so as to optimally represent the chemical space covered by the 13.56M PubChem and 168K MDDR compounds. These descriptors and the relevant references are given in Table 1, which include 18 descriptors in the class of simple molecular properties, three descriptors in the class of chemical properties, 42 descriptors in the class of electro-topological state, and 35 descriptors in the class of molecular connectivity and shape. Descriptors in the first three classes are nonredundant. Some descriptors in the fourth class have some degree of overlap in describing the topological features in spite of their differences in mathematical expression. These descriptors include the Schultz molecular topological index, the Gutman molecular topological index, the Wiener index, the Harary index, and the gravitational topological index. The partial overlap in the topological descriptors is not expected to be a serious problem for SVM classification because SVM is less penalized by descriptor redundancy.[48,49]

SVM Method. The process of training and using a SVM VS model for screening compounds based on their molecular descriptors is schematically illustrated in Figure 2. SVM is based on the structural risk minimization principle of the statistical learning theory,[50,51] which consistently shows outstanding classification performance, is less penalized by sample redundancy, and has lower risk for overfitting.[48,49] In linearly separable cases, SVM constructs a hyperplane to separate active and inactive classes of compounds with a maximum margin. A compound is represented by a vector $\mathbf{x}_i$ composed of its molecular descriptors. The hyperplane is constructed by finding another vector $\mathbf{w}$ and a parameter b that minimizes$||\mathbf{w}||^2$ and satisfies the following conditions:

$$\mathbf{w}\mathbf{x}_i + b \geq +1, \quad \text{for } y_i = +1 \quad \text{class 1(active)} \quad (1)$$

$$\mathbf{w}\mathbf{x}_i + b \leq -1, \quad \text{for } y_i = -1 \quad \text{class 2(inactive)} \quad (2)$$

where y_i is the class index, $\mathbf{w}$ is a vector normal to the hyperplane, $|b|/||\mathbf{w}||$ is the perpendicular distance from the hyperplane to the origin and $||\mathbf{w}||^2$ is the Euclidean norm of $\mathbf{w}$. Based on $\mathbf{w}$ and b, a given vector $\mathbf{x}$ can be classified by $f(\mathbf{x}) = \text{sign}[\mathbf{w}\mathbf{x} + b]$. A positive or negative $f(\mathbf{x})$ value indicates that the vector $\mathbf{x}$ belongs to the active or inactive class respectively.

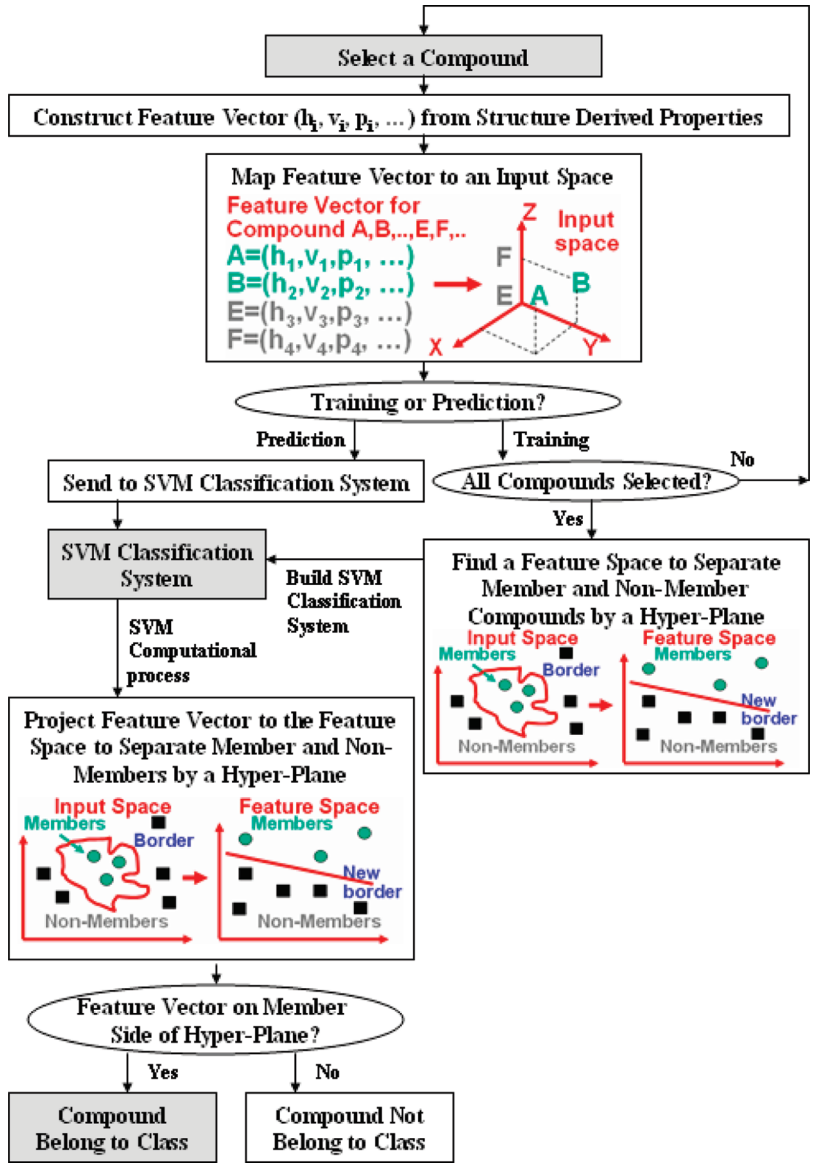

Figure 2. Schematic diagram illustrating the process of training a prediction model and using it for predicting active compounds of a compound class from their structurally derived properties (molecular descriptors) by using SVM. The symbols h_j, p_j, vj,..., represent structural and physicochemical properties as hydrophobicity, volume, polarizability, etc. Compounds A, B,..., E,F,..., are represented using one feature vector (h_j, p_j, vj,...).

In nonlinearly separable cases, which frequently occur in classifying compounds of diverse structures,[17−20,45,52−54] SVM maps the input vectors into a higher dimensional feature space by using a kernel function $K(\mathbf{x}_i, \mathbf{x}_j)$. We used RBF kernel $K(\mathbf{x}_i, \mathbf{x}_j) = e^{-\|\mathbf{x}_j - \mathbf{x}_i\|^2/2\sigma^2}$, which has been extensively used and has consistently shown better performance than other kernel functions.[55−57] Linear SVM can then applied to this feature space based on the following decision function: $f(\mathbf{x}) = \text{sign}(\sum_{i=1}^{l} \alpha_i^0 y_i K(\mathbf{x}, \mathbf{x}_i) + b)$, where the coefficients α_i^0 and b are determined by maximizing the following Langrangian expression: $\sum_{i=1}^{l} \alpha_i - 1/2 \sum_{i=1}^{l} \sum_{j=1} \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$ under the conditions $\alpha_i \geq 0$ and $\sum_{i=1}^{l} \alpha_i y_i = 0$. A positive or negative $f(\mathbf{x})$ value indicates that the vector $\mathbf{x}$ is an inhibitor or noninhibitor, respectively.

In developing our SVM VS tool, a hard margin $c = 1 \times 10^5$ was used, and the σ values were found to be 0.06. Its performance indicators can be derived from the numbers of true positives TP (true inhibitors), true negatives TN (true noninhibitors), false positives FP (false inhibitors), and false negatives FN (false noninhibitors). In 5-fold cross validation studies, the inhibitor and noninhibitor prediction accuracies are given by sensitivity SE = TP/(TP + FN)100 and specificity SP = TN/(TN + FP)100, respectively. Prediction accuracies have also been frequently measured by overall prediction accuracy (Q) and Matthews correlation coefficient (C):[58]

$$Q = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{1a}$$

$$C = \frac{(\text{TP} \times \text{TN}) - (\text{FN} \times \text{FP})}{\sqrt{(\text{TP} + \text{FN})(\text{TP} + \text{FP})(\text{TN} + \text{FN})(\text{TN} + \text{FP})}} \tag{2b}$$

In the large database screening tests, the yield and false-hit rate are given by TP/(TP + FN) and FP/(TP + FP), respectively.

Tanimoto Similarity Searching Method. Compounds similar to at least one Abl inhibitor in a training data set can be identified by using the Tanimoto coefficient sim(i, j):[59]

$$\mathrm{sim}(i,j) = \frac{\sum_{d=1}^{l} \mathbf{x}_{di}\mathbf{x}_{dj}}{\sum_{d=1}^{l} (\mathbf{x}_{di})^2 + \sum_{d=1}^{l} (\mathbf{x}_{dj})^2 - \sum_{d=1}^{l} \mathbf{x}_{di}\mathbf{x}_{dj}} \quad (3)$$

where l is the number of molecular descriptors. A compound i is considered to be similar to a known active j in the active data set if the corresponding sim(i, j) value is greater than a cutoff value. In this work, the similarity search was conducted for MDDR compounds. Therefore, in computing sim(i, j), the molecular descriptor vectors $\mathbf{x}_i$'s were scaled with respect to all of the MDDR compounds. The cutoff values for similarity compounds are typically in the range of 0.8−0.9.[23,60] A stricter cutoff value of 0.9 was used in this study.

K-Nearest-Neighbor Method. kNN measures the Euclidean distance $D = \sqrt{|\mathbf{x} - \mathbf{x}_i|^2}$ between a compound $\mathbf{x}$ and each individual inhibitor or noninhibitor $\mathbf{x}_i$ in the training set.[61] A total of k number of vectors nearest to the vector $\mathbf{x}$ are used to determine the decision function $f(\mathbf{x})$:

$$\hat{f}(\mathbf{x}) \leftarrow \arg\max_{v \in \mathbf{V}} \sum_{i=1}^{k} \delta(v, f(\mathbf{x}_i)) \quad (4)$$

where $\delta(a, b) = 1$, if $a = b$ and $\delta(a, b) = 0$, if $a \neq b$, arg max is the maximum of the function, $\mathbf{V}$ is a finite set of vectors {$\mathbf{v}$1,..., $\mathbf{v}$s}, and $\hat{f}(\mathbf{x})$ is an estimate of $f(\mathbf{x})$. Here estimate refers to the class of the majority compound group (i.e., inhibitors or noninhibitors) of the k nearest neighbors. The parameter $k = 1$ was found to give the best performance of this work.

Probabilistic Neural Network Method. PNN is a form of neural network that classifies objects based on Bayes' optimal decision rule[62] $h_i c_i f_i(\mathbf{x}) > h_j c_j f_j(\mathbf{x})$, where h_i and h_j are the prior probabilities, c_i and c_j are the costs of misclassification, and $f_i(\mathbf{x})$ and $f_j(\mathbf{x})$ are the probability density function for class i and j, respectively. A compound $\mathbf{x}$ is classified into class i if the product of all the three terms is greater for class i than for any other class j (not equal to i). In most applications, the prior probabilities and costs of misclassifications are treated as being equal. The probability density function for each class for a univariate case can be estimated by using the Parzen's nonparametric estimator:[63]

$$g(\mathbf{x}) = \frac{1}{n\sigma} \sum_{i=1}^{n} W\left(\frac{\mathbf{x} - \mathbf{x}_i}{\sigma}\right) \quad (5)$$

where n is the sample size, σ is a scaling parameter which defines the width of the bell curve that surrounds each sample point, $W(d)$ is a weight function which has its largest value at $d = 0$ and $(\mathbf{x} - \mathbf{x}_i)$ is the distance between the unknown vector and a vector in the training set. The Parzen's nonparametric estimator was later expanded by Cacoullos[64] for the multivariate case:

$$g(\mathbf{x}_1, ..., \mathbf{x}_p) = \frac{1}{n\sigma_1, ..., \sigma_p} \sum_{i=1}^{n} W\left(\frac{\mathbf{x}_1 - \mathbf{x}_{1,i}}{\sigma_1}, ..., \frac{\mathbf{x}_p - \mathbf{x}_{pi}}{\sigma_p}\right) \quad (6)$$

The Gaussian function is frequently used as the weight function because it is well behaved, easily calculated, and satisfies the conditions required by Parzen's estimator. Thus, the probability density function for the multivariate case becomes

$$g(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} \exp\left(-\sum_{j=1}^{p} \left(\frac{\mathbf{x}_j - \mathbf{x}_{ij}}{\sigma_j}\right)^2\right) \quad (7)$$

The network architectures of PNN are determined by the number of compounds and descriptors in the training set. There are four layers in a PNN. The input layer provides input values to all neurons in the pattern layer and has as many neurons as the number of descriptors in the training set. The number of pattern neurons is determined by the total number of compounds in the training set. Each pattern neuron computes a distance measure between the input and the training case represented by that neuron and then subjects the distance measure to the Parzen's nonparameteric estimator. The summation layer has a neuron for each class and the neurons sum all the pattern neurons' output, corresponding to members of that summation neuron's class to obtain the estimated probability density function for that class. The single neuron in the output layer then estimates the class of the unknown compound $\boldsymbol{x}$ by comparing all the probability density functions from the summation neurons and by choosing the class with the highest probability density function. The σ value of our developed PNN model is 0.003.

RESULTS AND DISCUSSION

Performance of SVM Identification of Abl Inhibitors Based on 5-Fold Cross Validation Test. The 5-fold cross validation test results of SVM in identifying Abl inhibitors

Table 2. Performance of Support Vector Machines for Identifying Abl Inhibitors and Noninhibitors Evaluated by 5-Fold Cross Validation Study

	Abl inhibitors				Abl noninhibitors					
crossvalidation	Number of training/ testing inhibitors	TP	FN	SE(%)	Number of training/ testing noninhibitors	TN	FP	SP(%)	Q (%)	C
1	566/142	131	11	92.25%	52395/13099	13098	1	99.99%	99.91%	0.915
2	566/142	125	17	88.03%	52395/13099	13094	5	99.96%	99.83%	0.845
3	566/142	128	14	90.14%	52395/13099	13097	2	99.98%	99.88%	0.886
4	567/141	119	22	84.40%	52395/13099	13094	5	99.96%	99.80%	0.808
5	567/141	128	13	90.78%	52396/13098	13093	5	99.96%	99.86%	0.872
average				**89.12%**				**99.97%**	**99.86%**	**0.865**
SD				0.0304				0.000149	0.000434	0.0407
SE				0.0136				0.00007	0.00019	0.0182

and putative noninhibitors are given in Table 2. The accuracies for predicting inhibitors and noninhibitors are 84.4 to 92.3% and 99.96 to 99.99%, respectively. The overall prediction accuracy Q and Matthews correlation coefficient C are 99.79 to 99.90% and 0.808 to 0.915, respectively. The inhibitor accuracies of our SVM are comparable to or slightly better than the reported accuracies of 58.3 to 67.3% for protein kinase C inhibitors by SVM-RBF and CKD methods,[65] 83% for Lck inhibitors by SVM method,[66] and 74 to 87% for inhibitors of any of the eight kinases (three Ser/Thr and five Tyr kinases) by SVM, ANN, GA/kNN, and RP methods.[67] The noninhibitor accuracies are comparable to the value of 99.9% for Lck inhibitors[66] and substantially better than the typical values of 77 to 96% of other studies.[65,67] Caution needs to be raised about straightforward comparison of these results, which might be misleading because the outcome of VS strongly depends on the data sets and molecular descriptors used. Based on these rough comparisons, SVM appears to show good capability in identifying Abl inhibitors at low false-hit rates. Similar prediction accuracies were also found from two additional 5-fold cross validation studies conducted by using training testing sets separately generated from different random number seed parameters.

Virtual Screening Performance of SVM in Searching Abl Inhibitors from Large Compound Libraries. SVM VS tool for searching Abl inhibitors from large was developed by using Abl kinases reported before 2008, as described in the Methods Section. The VS performance of SVM in identifying Abl inhibitors reported since 2008 and searching MDDR and PubChem databases is summarized in Table 3. The yield in searching Abl inhibitors reported since 2008 is 50.5%, which is comparable to the reported 50 to 94% yields of various VS tools.[68] Strictly speaking, direct comparison of the reported performances of these VS tools is inappropriate because of the differences in the type, composition, and diversity of compounds screened and in the molecular descriptors, VS tools, and their parameters used. The comparison cannot go beyond the statistics of accuracies. A more appropriate comparison based on the same training and testing data sets and molecular descriptors were conducted, which are described in a following section.

Virtual-hit and false-hit rates of SVM in screening compounds that resemble the structural and physicochemical properties of the known Abl inhibitors were evaluated by using 6 638 MDDR compounds similar to an Abl inhibitor in the training data set. Similarity was defined by the Tanimoto similarity coefficient ≥ 0.9 between a MDDR compound and its closest inhibitor.[21] SVM identified 330 virtual-hits from these 6 638 MDDR similarity compounds (virtual-hit rate 4.97%), which suggests that SVM has some level of capability in distinguishing Abl inhibitors from noninhibitor similarity compounds. Significantly lower virtual-hit rates and, thus, false-hit rates were found in screening large libraries of 168K MDDR and 13.56 M PubChem compounds. The number of virtual-hits and virtual-hit rates in screening 168K MDDR compounds are 659 and 0.39%, respectively. The number of virtual-hits and virtual-hit rates in screening 13.56 M PubChem compounds are 29 072 and 0.21%, respectively.

Table 3. Virtual Screening Performance of Support Vector Machines for Identifying Abl Inhibitors from Large Compound Libraries

method	inhibitors in training set		inhibitors in testing set			virtual screening performance			
	number of inhibitors	number of chemical families covered by inhibitors	number of inhibitors	number of chemical families covered by inhibitors	percent of inhibitors in chemical families covered by inhibitors in training set	yield	number and percent of identified true inhibitors outside training chemical families	number and percent of 13.56 M PubChem compounds identified as inhibitors	number and percent of the 168K MDDR compounds identified as inhibitors	number and percent of the 6 638 MDDR compounds similar to the known inhibitors identified as inhibitors
SVM	708	221	91	38	50	50.5	9 (19.6)	29 072 (0.21)	659 (0.39)	330 (5.0)
Tarnimoto similarity						70.3	26 (56.5)	NA	6,638 (3.95)	6,638 (100)
K nearest-neighbor						58.2	10 (21.7)	79,043 (0.58)	1,662 (0.99)	550 (8.3)
probabilistic neural network						58.2	10 (21.7)	83,293 (0.61)	1,686 (1.00)	546 (8.2)

Table 4. MDDR Classes That Contain Higher Percentage (≥6%) of Virtual-Hits Identified by SVMs in Screening 168K MDDR Compounds for Abl Inhibitors

kinase	number of SVM identified virtual hits	MDDR classes that contain higher percentage (>6%) of virtual hits	number of virtual hits in class	percentage of class members selected as virtual hits
ABL	659	antineoplastic	310	1.4
		signal transduction inhibitor	116	5.7
		tyrosine-specific protein kinase inhibitor	105	8.9
		antiarthritic	98	0.9
		antiangiogenic	40	2.5

Substantial percentages of the MDDR virtual-hits belong to the classes of antineoplastic, signal transduction inhibitors, tyrosine-specific protein kinase inhibitors, antiarthritic, and antiangiogenic (Table 4, details in Evaluation of SVM Identified MDDR Virtual-Hits section). As some of these virtual-hits may be true Abl inhibitors, the false-hit rate of our SVM is at most equal to and likely less than the virtual-hit rate. Hence, the false-hit rate is <3.95% in screening 6 638 MDDR similarity compounds, <0.39% in screening 168K MDDR compounds, and <0.21% in screening 13.56 M PubChem compounds, which are comparable and in some cases better than the reported false-hit rates of 0.0054 to 8.3% of SVM,[21,69] 0.08 to 3% of structure-based methods, 0.1 to 5% by other machine learning methods, 0.16 to 8.2% by clustering methods, and 1.15 to 26% by pharmacophore models.[68]

To facilitate the selection of true Abl inhibitors from the SVM identified virtual-hits, one may explore a consensus approach that selects potentially promising virtual-hits based on the consensus scores of multiple VS methods that include molecular docking, similarity methods, and pharmacophore models as well as SVM.[68] Our preliminary study showed that 20% of the 659 SVM virtual-hits from the MDDR database were selected by molecular docking, which include 128 compounds that belong to the tyrosine-specific protein kinase inhibitor class. This suggests that a consensus approach is potentially useful for enriching true-hit selection rates.

Evaluation of SVM Identified MDDR Virtual-Hits. SVM identified MDDR virtual-hits were evaluated based on the known biological or therapeutic target classes specified in MDDR. Table 4 gives the MDDR classes that contain a higher percentage (≥6%) of SVM virtual-hits and the percentage values. We found that 310 or 47% of the 659 virtual-hits belong to the antineoplastic class, which represent 1.4% of the 21 557 MDDR compounds in the class. In particular, 105 or 16% of the virtual-hits belong to the tyrosine-specific protein kinase inhibitor class, which represent 8.9% of the 1 181 MDDR compounds in the class. Moreover, 18% and 6% of the virtual-hits belong to the signal transduction inhibitor and antiangiogenic classes, representing 5.7% and 2.5% of the 2 037 and 1 629 members in the two classes, respectively. Therefore, many of the SVM virtual-hits are antineoplastic compounds that inhibit tyrosine kinases and possibly other kinases involved in signal transduction, angiogenesis, and other cancer-related pathways. While some of these kinase inhibitors might be true Abl inhibitors, the majority of them are expected to arise from false selection of inhibitors of other kinases.

A total of 98 SVM virtual-hits belong to the antiarthritic class. An Abl inhibitor Gleevec has been reported to be effective in treatment of arthritis, which is probably due to its inhibition of other related kinases such as c-kit and PDGFR.[70] Moreover, several other kinases have been implicated in arthritis. EGFR-like receptor stimulates synovial cells, and its elevated activities may be involved in the pathogenesis of rheumatoid arthritis.[69] VEGF has been related to such autoimmune diseases as systemic lupus erythematosus, rheumatoid arthritis, and multiple sclerosis.[71] FGFR may partially mediate osteoarthritis.[72] PDGF-like factors stimulate the proliferative and invasive phenotype of rheumatoid arthritis synovial connective tissue cells.[73] Lck inhibition leads to immunosuppression and has been explored for the treatment of rheumatoid arthritis and asthma.[74] Therefore, some of the SVM virtual-hits in the antiarthritic class may be inhibitors of these kinases or their kinase-likes capable of producing antiarthritic activities.

Comparison of Virtual Screening Performance of SVM with Those of Other Virtual Screening Methods. To evaluate the level of performance of SVM and whether the performance is due to the SVM classification models or to the molecular descriptors used, SVM results were compared with those of three other VS methods based on the same molecular descriptors, training data set of Abl inhibitors reported before 2008, and the testing data set of Abl inhibitors reported since 2008, 168K MDDR, and 13.56M PubChem compounds. The three other VS methods include two similarity-based methods, Tanimoto-based similarity searching, and kNN methods and an alternative machine learning method PNN. As shown in Table 3, the yield and maximum possible false-hit rate of the Tanimoto-based similarity searching and the kNN and PNN methods are 70.33% and 3.95%, 58.24% and 0.99%, and 58.24 and 1%, respectively. Compared to these results, the yield of SVM is smaller than but still comparable to the similarity-based VS method, and the false-hit rate of SVM is significantly reduced by 10.1-, 2.5-, and 2.6-fold, respectively. These suggest that SVM performance is due primarily to the SVM classification models rather than the molecular descriptors used, and SVM is capable of achieving comparable yield at substantially reduced false-hit rate as compared to both the similarity-based approach and the alternative machine learning method. Our results are consistent with the report that SVM shows mostly good performances on both classification and regression tasks, but other classification and regression methods proved to be very competitive.[24]

Does SVM Select Abl Inhibitors or Membership of Compound Families? To further evaluate whether our SVM VS tools identify Abl inhibitors rather than membership of certain compound families, the compound family distribution of the identified ABL inhibitors and noninhibitors were analyzed. A total of 19.6% of the identified inhibitors belong to the families that contain no known Abl inhibitors. For those families that contain at least one known Abl inhibitor, >70% of the compounds (>90% in majority cases) in each of these families were predicted as noninhibitor by SVM. These results suggest that our SVM VS tool identifies Abl inhibitors rather than membership to certain compound families. Some of the identified inhibitors not in the family of known inhibitors may serve as potential "novel" Abl inhibitors. Therefore, as in the case shown by earlier studies,[16] SVM has a certain capacity for identifying novel active compounds from sparse- as well as regular-sized active data sets.

CONCLUDING REMARKS

SVM shows substantial capability in identifying Abl inhibitors at comparable yield and in many cases substantially lower false-hit rate than those of typical VS tools reported in literature and evaluated in this work. It is capable of searching large compound libraries at sizes comparable to the 13.56M PubChem and 168K MDDR compounds at low false-hit rates without the need to define an applicability domain, i.e., it has a broad applicability domain that covers the whole chemical space defined by the current versions of PubChem and MDDR databases. The performance of SVM is substantially improved against several other VS methods based on the same data sets and molecular descriptors, suggesting that the VS performance of SVM is primarily due to SVM classification models rather than that of the molecular descriptors used. Because of their high-computing speed and generalization capability for covering highly diverse spectrum compounds, SVM can be potentially explored to develop useful VS tools to complement other VS methods or to be used as part of integrated VS tools in facilitating the discovery of Abl inhibitors and other active compounds.[75−77]

ACKNOWLEDGMENT

This work was supported in part by grants from Singapore Academic Research Fund R-148-000-083-112, National Natural Science Foundation of China Grant 30772651, Ministry of Science and Technology, and 863 High-Tech Program Grant 2006AA020400.

REFERENCES AND NOTES

(1) Hazlehurst, L. A.; Bewry, N. N.; Nair, R. R.; Pinilla-Ibarz, J. Signaling networks associated with BCR-ABL-dependent transformation. *Cancer Control* **2009**, *16* (2), 100–7.

(2) Weisberg, E.; Manley, P. W.; Cowan-Jacob, S. W.; Hochhaus, A.; Griffin, J. D. Second generation inhibitors of BCR-ABL for the treatment of imatinib-resistant chronic myeloid leukaemia. *Nat. Rev. Cancer* **2007**, *7* (5), 345–56.

(3) Gill, A. L.; Verdonk, M.; Boyle, R. G.; Taylor, R. A comparison of physicochemical property profiles of marketed oral drugs and orally bioavailable anti-cancer protein kinase inhibitors in clinical development. *Curr. Top. Med. Chem.* **2007**, *7* (14), 1408–22.

(4) Quintas-Cardama, A.; Kantarjian, H.; Cortes, J. Flying under the radar: the new wave of BCR-ABL inhibitors. *Nat. Rev. Drug Discov.* **2007**, *6* (10), 834–48.

(5) Cao, J.; Fine, R.; Gritzen, C.; Hood, J.; Kang, X.; Klebansky, B.; Lohse, D.; Mak, C. C.; McPherson, A.; Noronha, G.; Palanki, M. S.; Pathak, V. P.; Renick, J.; Soll, R.; Zeng, B.; Zhu, H. The design and preliminary structure-activity relationship studies of benzotriazines as potent inhibitors of Abl and Abl-T315I enzymes. *Bioorg. Med. Chem. Lett.* **2007**, *17* (21), 5812–8.

(6) Manetti, F.; Falchi, F.; Crespan, E.; Schenone, S.; Maga, G.; Botta, M. N-(thiazol-2-yl)-2-thiophene carboxamide derivatives as Abl inhibitors identified by a pharmacophore-based database screening of commercially available compounds. *Bioorg. Med. Chem. Lett.* **2008**, *18* (15), 4328–31.

(7) Falchi, F.; Manetti, F.; Carraro, F.; Naldini, A.; Maga, G.; Crespan, E.; Schenone, S.; Bruno, O.; Brullo, C.; Botta, M., 3D QSAR Models Built on Structure-Based Alignments of Abl Tyrosine Kinase Inhibitors. *ChemMedChem* **2009**.

(8) Aronov, A. M.; Bemis, G. W. A minimalist approach to fragment-based ligand design using common rings and linkers: application to kinase inhibitors. *Proteins* **2004**, *57* (1), 36–50.

(9) Peng, H.; Huang, N.; Qi, J.; Xie, P.; Xu, C.; Wang, J.; Yang, C. Identification of novel inhibitors of BCR-ABL tyrosine kinase via virtual screening. *Bioorg. Med. Chem. Lett.* **2003**, *13* (21), 3693–9.

(10) Schenone, S.; Brullo, C.; Bruno, O.; Bondavalli, F.; Mosti, L.; Maga, G.; Crespan, E.; Carraro, F.; Manetti, F.; Tintori, C.; Botta, M. Synthesis, biological evaluation and docking studies of 4-amino substituted 1H-pyrazolo[3,4-d]pyrimidines. *Eur. J. Med. Chem.* **2008**, *43* (12), 2665–76.

(11) Thaimattam, R.; Daga, P. R.; Banerjee, R.; Iqbal, J. 3D-QSAR studies on c-Src kinase inhibitors and docking analyses of a potent dual kinase inhibitor of c-Src and c-Abl kinases. *Bioorg. Med. Chem.* **2005**, *13* (15), 4704–12.

(12) Manetti, F.; Locatelli, G. A.; Maga, G.; Schenone, S.; Modugno, M.; Forli, S.; Corelli, F.; Botta, M. A combination of docking/dynamics simulations and pharmacophoric modeling to discover new dual c-Src/Abl kinase inhibitors. *J. Med. Chem.* **2006**, *49* (11), 3278–86.

(13) Shoichet, B. K. Virtual screening of chemical libraries. *Nature* **2004**, *432* (7019), 862–5.

(14) Ghosh, S.; Nie, A.; An, J.; Huang, Z. Structure-based virtual screening of chemical libraries for drug discovery. *Curr. Opin. Chem. Biol* **2006**, *10* (3), 194–202.

(15) Li, H.; Yap, C. W.; Ung, C. Y.; Xue, Y.; Li, Z. R.; Han, L. Y.; Lin, H. H.; Chen, Y. Z. Machine learning approaches for predicting compounds that interact with therapeutic and ADMET related proteins. *J. Pharm. Sci.* **2007**, *96* (11), 2838–60.

(16) Han, L. Y.; Ma, X. H.; Lin, H. H.; Jia, J.; Zhu, F.; Xue, Y.; Li, Z. R.; Cao, Z. W.; Ji, Z. L.; Chen, Y. Z. A support vector machines approach for virtual screening of active compounds of single and multiple mechanisms from large libraries at an improved hit-rate and enrichment factor. *J. Mol. Graphics Modell.* **2008**, *26* (8), 1276–86.

(17) Jorissen, R. N.; Gilson, M. K. Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model.* **2005**, *45* (3), 549–61.

(18) Lepp, Z.; Kinoshita, T.; Chuman, H. Screening for new antidepressant leads of multiple activities by support vector machines. *J. Chem. Inf. Model.* **2006**, *46* (1), 158–67.

(19) Glick, M.; Jenkins, J. L.; Nettles, J. H.; Hitchings, H.; Davies, J. W. Enrichment of high-throughput screening data with increasing levels of noise using support vector machines, recursive partitioning, and laplacian-modified naive bayesian classifiers. *J. Chem. Inf. Model.* **2006**, *46* (1), 193–200.

(20) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. New methods for ligand-based virtual screening: use of data fusion and machine learning to enhance the effectiveness of similarity searching. *J. Chem. Inf. Model.* **2006**, *46* (2), 462–70.

(21) Ma, X. H.; Wang, R.; Yang, S. Y.; Li, Z. R.; Xue, Y.; Wei, Y. C.; Low, B. C.; Chen, Y. Z. Evaluation of virtual screening performance of support vector machines trained by sparsely distributed active compounds. *J. Chem. Inf. Model.* **2008**, *48* (6), 1227–37.

(22) Verdonk, M. L.; Berdini, V.; Hartshorn, M. J.; Mooij, W. T.; Murray, C. W.; Taylor, R. D.; Watson, P. Virtual screening using protein-ligand docking: avoiding artificial enrichment. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (3), 793–806.

(23) Huang, N.; Shoichet, B. K.; Irwin, J. J. Benchmarking sets for molecular docking. *J. Med. Chem.* **2006**, *49* (23), 6789–801.

(24) Mayer, D.; Leisch, F.; Hornik, K. The support vector machine under test. *Neurocomputing* **2003**, *55* (1−2), 169–186.

(25) McBride, C. M.; Renhowe, P. A.; Gesner, T. G.; Jansen, J. M.; Lin, J.; Ma, S.; Zhou, Y.; Shafer, C. M. 3-Benzimidazol-2-yl-1H-indazoles as potent c-ABL inhibitors. *Bioorg. Med. Chem. Lett.* **2006**, *16* (14), 3789–92.

(26) Traxler, P.; Bold, G.; Frei, J.; Lang, M.; Lydon, N.; Mett, H.; Buchdunger, E.; Meyer, T.; Mueller, M.; Furet, P. Use of a pharmacophore model for the design of EGF-R tyrosine kinase inhibitors: 4-(phenylamino)pyrazolo[3,4-d]pyrimidines. *J. Med. Chem.* **1997**, *40* (22), 3601–16.

(27) Wang, Y.; Shakespeare, W. C.; Huang, W. S.; Sundaramoorthi, R.; Lentini, S.; Das, S.; Liu, S.; Banda, G.; Wen, D.; Zhu, X.; Xu, Q.; Keats, J.; Wang, F.; Wardwell, S.; Ning, Y.; Snodgrass, J. T.; Broudy, M. I.; Russian, K.; Dalgarno, D.; Clackson, T.; Sawyer, T. K. Novel N9-arenethenyl purines as potent dual Src/Abl tyrosine kinase inhibitors. *Bioorg. Med. Chem. Lett.* **2008**, *18* (17), 4907–12.

(28) Liu, T.; Lin, Y.; Wen, X.; Jorissen, R. N.; Gilson, M. K. BindingDB: a web-accessible database of experimentally determined protein-ligand binding affinities. *Nucleic Acids Res.* **2007**, *35* (Database issue), D198–201.

(29) Keseru, G. M.; Makara, G. M. The influence of lead discovery strategies on the properties of drug candidates. *Nat. Rev. Drug Discov.* **2009**, *8* (3), 203–12.

(30) Keseru, G. M.; Makara, G. M. Hit discovery and hit-to-lead approaches. *Drug Discov. Today* **2006**, *11* (15−16), 741–8.

(31) Bocker, A.; Schneider, G.; Teckentrup, A. NIPALSTREE: a new hierarchical clustering approach for large compound libraries and its application to virtual screening. *J. Chem. Inf. Model.* **2006**, *46* (6), 2220–9.

(32) Oprea, T. I.; Gottfries, J. Chemography: the art of navigating in chemical space. *J. Comb. Chem.* **2001**, *3* (2), 157–66.

(33) Fink, T.; Reymond, J. L. Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J. Chem. Inf. Model.* **2007**, *47* (2), 342–53.

(34) Koch, M. A.; Schuffenhauer, A.; Scheck, M.; Wetzel, S.; Casaulta, M.; Odermatt, A.; Ertl, P.; Waldmann, H. Charting biologically relevant chemical space: a structural classification of natural products (SCONP). *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (48), 17272–7.

(35) Fang, H.; Tong, W.; Shi, L. M.; Blair, R.; Perkins, R.; Branham, W.; Hass, B. S.; Xie, Q.; Dial, S. L.; Moland, C. L.; Sheehan, D. M. Structure-activity relationships for a large diverse set of natural, synthetic, and environmental estrogens. *Chem. Res. Toxicol.* **2001**, *14*, 280–294.

(36) Tong, W.; Xie, Q.; Hong, H.; Shi, L.; Fang, H.; Perkins, R. Assessment of prediction confidence and domain extrapolation of two structure-activity relationship models for predicting estrogen receptor binding activity. *Environ. Health Perspect.* **2004**, *112* (12), 1249–1254.

(37) Jacobs, M. N. In silico tools to aid risk assessment of endocrine disrupting chemicals. *Toxicology* **2004**, *205* (1−2), 43–53.

(38) Hu, J. Y.; Aizawa, T. Quantitative structure-activity relationships for estrogen receptor binding affinity of phenolic chemicals. *Water Res.* **2003**, *37* (6), 1213–1222.

(39) Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (6), 1882–1889.

(40) Doniger, S.; Hofman, T.; Yeh, J. Predicting CNS Permeability of Drug Molecules:Comparison of Neural Network and Support Vector Machine Algorithms. *J. Comput. Biol.* **2002**, *9* (6), 849–864.

(41) He, L.; Jurs, P. C.; Custer, L. L.; Durham, S. K.; Pearl, G. M. Predicting the Genotoxicity of Polycyclic Aromatic Compounds from Molecular Structure with Different Classifiers. *Chem. Res. Toxicol.* **2003**, *16* (12), 1567–1580.

(42) Snyder, R. D.; Pearl, G. S.; Mandakas, G.; Choy, W. N.; Goodsaid, F.; Rosenblum, I. Y. Assessment of the sensitivity of the computational programs DEREK, TOPKAT, and MCASE in the prediction of the genotoxicity of pharmaceutical molecules. *Environ. Mol. Mutagen.* **2004**, *43* (3), 143–158.

(43) Xue, Y.; Li, Z. R.; Yap, C. W.; Sun, L. Z.; Chen, X.; Chen, Y. Z. Effect of Molecular Descriptor Feature Selection in Support Vector Machine Classification of Pharmacokinetic and Toxicological Properties of Chemical Agents. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (5), 1630–1638.

(44) Yap, C. W.; Cai, C. Z.; Xue, Y.; Chen, Y. Z. Prediction of torsade-causing potential of drugs by support vector machine approach. *Toxicol. Sci.* **2004**, *79* (1), 170–177.

(45) Yap, C. W.; Chen, Y. Z. Quantitative Structure-Pharmacokinetic Relationships for drug distribution properties by using general regression neural network. *J. Pharm. Sci.* **2005**, *94* (1), 153–68.

(46) Zernov, V. V.; Balakin, K. V.; Ivaschenko, A. A.; Savchuk, N. P.; Pletnev, I. V. Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions. *J. Chem. Inf. Comput. Sci.* **2003**, *43* (6), 2048–2056.

(47) Xue, Y.; Yap, C. W.; Sun, L. Z.; Cao, Z. W.; Wang, J. F.; Chen, Y. Z. Prediction of P-glycoprotein substrates by a support vector machine approach. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (4), 1497–505.

(48) Pochet, N.; De Smet, F.; Suykens, J. A.; De Moor, B. L. Systematic benchmarking of microarray data classification: assessing the role of non-linearity and dimensionality reduction. *Bioinformatics* **2004**, *20*, 3185–3195.

(49) Li, F.; Yang, Y. Analysis of recursive gene selection approaches from microarray data. *Bioinformatics* **2005**, *21*, 3741–3747.

(50) Vapnik, V. N. *The Nature of Statistical Learning Theory*; Springer: New York, 1995.

(51) Burges, C. J. C. A tutorial on support vector machines for pattern recognition. *Data Min. and Knowledge Discov.* **1998**, *2* (2), 127–167.

(52) Cui, J.; Lin, L. Y. H., H.H.; Zhang, H. L.; Tang, Z. Q.; Zheng, C. J.; Cao, Z. W.; Chen, Y. Z. Prediction of MHC-Binding Peptides of Flexible Lengths from Sequence-Derived Structural and Physicochemical Properties. *Mol. Immunol.* **2007**, *44*, 866–877.

(53) Yap, C. W.; Chen, Y. Z. Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines. *J. Chem. Inf. Model.* **2005**, *45* (4), 982–92.

(54) Grover, I. I.; Singh, I. I.; Bakshi, I. I. Quantitative structure-property relationships in pharmaceutical research - Part 2. *Pharm. Sci. Technol. Today* **2000**, *3* (2), 50–57.

(55) Trotter, M. W. B.; Buxton, B. F.; Holden, S. B. Support vector machines in combinatorial chemistry. *Meas. Control* **2001**, *34* (8), 235–239.

(56) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26* (1), 5–14.

(57) Czerminski, R.; Yasri, A.; Hartsough, D. Use of support vector machine in pattern classification: Application to QSAR studies. *Quant. Struct. Act. Relat.* **2001**, *20* (3), 227–240.

(58) Matthews, B. W. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta* **1975**, *405* (2), 442–451.

(59) Willett, P. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983–996.

(60) Bostrom, J.; Hogner, A.; Schmitt, S. Do structurally similar ligands bind in a similar fashion. *J. Med. Chem.* **2006**, *49* (23), 6716–25.

(61) Johnson, R. A.; Wichern, D. W. *Applied Multivariate Statistical Analysis*; Prentice Hall: Englewood Cliffs, NJ, 1982.

(62) Specht, D. F. Probabilistic neural networks. *Neural Networks* **1990**, *3* (1), 109–118.

(63) Parzen, E. On estimation of a probability density function and mode. *Ann. Math. Stat.* **1962**, *33*, 1065–1076.

(64) Cacoullos, T. Estimation of a multivariate density. *Ann. I. Stat. Math.* **1966**, *18*, 179–189.

(65) Chen, B.; Harrison, R. F.; Papadatos, G.; Willett, P.; Wood, D. J.; Lewell, X. Q.; Greenidge, P.; Stiefl, N. Evaluation of machine-learning methods for ligand-based virtual screening. *J. Comput. Aided Mol. Des.* **2007**, *21* (1−3), 53–62.

(66) Liew, C. Y.; Ma, X. H.; Liu, X.; Yap, C. W. SVM Model for Virtual Screening of Lck Inhibitors. *J. Chem. Inf. Model.* **2009**, *49* (4), 877−85.

(67) Briem, H.; Gunther, J. Classifying "kinase inhibitor-likeness" by using machine-learning methods. *Chembiochem* **2005**, *6* (3), 558–66.

(68) Ma, X. H.; Jia, J.; Zhu, F.; Xue, Y.; Li, Z. R.; Chen, Y. Z. Comparative analysis of machine learning methods in ligand-based virtual screening of large compound libraries. *Comb. Chem. High Throughput Screen.* **2009**, *12* (4), 344–57.

(69) Yamane, S.; Ishida, S.; Hanamoto, Y.; Kumagai, K.; Masuda, R.; Tanaka, K.; Shiobara, N.; Yamane, N.; Mori, T.; Juji, T.; Fukui, N.; Itoh, T.; Ochi, T.; Suzuki, R. Proinflammatory role of amphiregulin, an epidermal growth factor family member whose expression is augmented in rheumatoid arthritis patients. *J. Inflamm. (London)* **2008**, *5*, 5.

(70) Paniagua, R. T.; Sharpe, O.; Ho, P. P.; Chan, S. M.; Chang, A.; Higgins, J. P.; Tomooka, B. H.; Thomas, F. M.; Song, J. J.; Goodman, S. B.; Lee, D. M.; Genovese, M. C.; Utz, P. J.; Steinman, L.; Robinson, W. H. Selective tyrosine kinase inhibition by imatinib mesylate for the treatment of autoimmune arthritis. *J. Clin. Invest.* **2006**, *116* (10), 2633–42.

(71) Carvalho, J. F.; Blank, M.; Shoenfeld, Y. Vascular endothelial growth factor (VEGF) in autoimmune diseases. *J. Clin. Immunol.* **2007**, *27* (3), 246–56.

(72) Daouti, S.; Latario, B.; Nagulapalli, S.; Buxton, F.; Uziel-Fusi, S.; Chirn, G. W.; Bodian, D.; Song, C.; Labow, M.; Lotz, M.; Quintavalla, J.; Kumar, C. Development of comprehensive functional genomic screens to identify novel mediators of osteoarthritis. *Osteoarthritis Cartilage* **2005**, *13* (6), 508–18.

(73) Remmers, E. F.; Sano, H.; Wilder, R. L. Platelet-derived growth factors and heparin-binding (fibroblast) growth factors in the synovial tissue

pathology of rheumatoid arthritis. *Semin. Arthritis. Rheum.* **1991**, *21* (3), 191–9.

(74) Meyn, M. A., 3rd; Smithgall, T. E. Small molecule inhibitors of Lck: the search for specificity within a kinase family. *Mini Rev. Med. Chem.* **2008**, *8* (6), 628–37.

(75) Vidal, D.; Thormann, M.; Pons, M. A novel search engine for virtual screening of very large databases. *J. Chem. Inf. Model.* **2006**, *46* (2), 836–43.

(76) Stiefl, N.; Zaliani, A. A knowledge-based weighting approach to ligand-based virtual screening. *J. Chem. Inf. Model.* **2006**, *46* (2), 587–96.

(77) Rella, M.; Rushworth, C. A.; Guy, J. L.; Turner, A. J.; Langer, T.; Jackson, R. M. Structure-based pharmacophore design and virtual screening for novel angiotensin converting enzyme 2 inhibitors. *J. Chem. Inf. Model.* **2006**, *46* (2), 708–16.

(78) Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000.

(79) Miller, K. J. Additive Methods in Molecular Polarizability. *J. Am. Chem. Soc.* **1990**, *112*, 8533–8542.

(80) Schultz, H. P. Topological Organic Chemistry. 1. Graph Theory and Topological Indices of Alkanes. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 227–228.

(81) Hall, L. H.; Kier, L. B. Electrotopological State Indices for Atom Types: A Novel Combination of Electronic, Topological and Valence State Information. *J. Chem. Inf. Comput. Sci.* **1995**, *35*, 1039–1045.

CI900135U