Available online at www.sciencedirect.com
ScienceDirect
Journal of Molecular Graphics and Modelling 26 (2008) 1276–1286
Journal of Molecular Graphics and Modelling
www.elsevier.com/locate/JMGM

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

L.Y. Han [a], X.H. Ma [a], H.H. Lin [a], J. Jia [a], F. Zhu [a], Y. Xue [c], Z.R. Li [c], Z.W. Cao [b], Z.L. Ji [d], Y.Z. Chen [a,b,*]

[a] *Bioinformatics and Drug Design Group, Department of Pharmacy, National University of Singapore, Blk S16, Level 8, 3 Science Drive 2, Singapore 117543, Singapore*
[b] *Shanghai Center for Bioinformation Technology, Shanghai 201203, PR China*
[c] *College of Chemistry, Sichuan University, Chengdu 610064, PR China*
[d] *Bioinformatics Research Group, School of Life Sciences, Xiamen University, Xiamen 361005, FuJian Province, PR China*

Received 29 April 2007; received in revised form 5 December 2007; accepted 5 December 2007
Available online 15 December 2007

Abstract

Support vector machines (SVM) and other machine-learning (ML) methods have been explored as ligand-based virtual screening (VS) tools for facilitating lead discovery. While exhibiting good hit selection performance, in screening large compound libraries, these methods tend to produce lower hit-rate than those of the best performing VS tools, partly because their training-sets contain limited spectrum of inactive compounds. We tested whether the performance of SVM can be improved by using training-sets of diverse inactive compounds. In retrospective database screening of active compounds of single mechanism (HIV protease inhibitors, DHFR inhibitors, dopamine antagonists) and multiple mechanisms (CNS active agents) from large libraries of 2.986 million compounds, the yields, hit-rates, and enrichment factors of our SVM models are 52.4–78.0%, 4.7–73.8%, and 214–10,543, respectively, compared to those of 62–95%, 0.65–35%, and 20–1200 by structure-based VS and 55–81%, 0.2–0.7%, and 110–795 by other ligand-based VS tools in screening libraries of $\geq$1 million compounds. The hit-rates are comparable and the enrichment factors are substantially better than the best results of other VS tools. 24.3–87.6% of the predicted hits are outside the known hit families. SVM appears to be potentially useful for facilitating lead discovery in VS of large compound libraries.

Keywords: Computer aided dug design; Drug discovery; High-throughput screening; Lead discovery; Machine learning method; Virtual screening

1. Introduction

Virtual screening (VS) has been extensively explored for facilitating lead discovery [1–4] and for identifying agents of desirable pharmacokinetic and toxicological properties [5,6]. Machine learning (ML) methods have recently been used for developing ligand-based VS (LBVS) tools [7–14] to complement or to be combined with structure-based VS (SBVS) [1,15–26] and other LBVS [2,27–30] tools aimed at improving the coverage, performance and speed of VS tools.

ML methods have been used as part of the efforts to overcome several problems that have impeded progress in more extensive applications of SBVS and LBVS tools [1,31]. These problems include the vastness and sparse nature of chemical space needs to be searched, limited availability of target structures (only 15% of known proteins have known 3D structures), complexity and flexibility of target structures, and difficulties in computing binding affinity and solvation effects. LBVS may in some cases limit the diversity of hits due to the bias of training molecules [15]. Therefore, alternative approaches that enhance screening speed and compound diversity without relying on target structural information are highly desired. ML methods have been explored for developing

* Corresponding author at: Bioinformatics and Drug Design Group, Department of Pharmacy, National University of Singapore, Blk S16, Level 8, 3 Science Drive 2, Singapore 117543, Singapore. Tel.: +65 6874 6877; fax: +65 6774 6756.
E-mail address: phacyz@nus.edu.sg (Y.Z. Chen).

1093-3263/$ – see front matter
doi:10.1016/j.jmgm.2007.12.002

Table 1
Comparison of the reported performance of different virtual screening (VS) methods in screening large libraries of compounds

Type of VS method and size of compound libraries screened	VS method [references]	Compounds screened			Virtual hits selected by VS method		Known hits selected by VS method			
		No of compounds	No of known hits	Percent of known hits	No of compounds selected as virtual hits	Percent of screened compounds selected as virtual hits	No of known hits selected	Yield	Hit-rates	Enrichment factor
Structure-based VS, extremely large libraries (≥1 M)	Docking + pre-screening filter [2,18,19]	1 M–2 M	355–630	~0.03%	1 K–60 K	0.08–3%	340–390	62–95%	0.65–35%	20–1200
Structure-based VS, large libraries	Docking + pre-screening filter [11,20–26]	134 K–400 K	100–1016	0.12–0.76%	375–4.5 K	0.28–3%	5–231	2–30%	0.11–17%	4–66
Ligand-based VS (machine learning), extremely large libraries (≥1 M)	Machine learning–SVM [2,8,11,13]	2.5 M	22–46	0.0009–0.0018%	2.5 K–11 K	0.1–0.45%	18–25	55–81%	0.2–0.7%	110–795
Ligand-based VS (machine learning), large libraries	Machine learning–SVM [2,9]	172 K	118–128	~0.07%	1.7 K	1%	26–70	22–55%	1.5–4.1%	22–55
	Machine learning–SVM [11,12]	98.4 K	259–1146	0.26–1.16%	984	1%	131–710	44–69%	14–72%	44–69
	Machine learning–BKD [12,9,11,13,14]	101 K–103 K	259–1166	0.25–1.2%	5.1 K	5%	65–972	14–94%	1.2–18.9%	3–19
	Machine learning–LMNB [1,11,13]	172 K	118	0.069%	1.7 K	1%	19	16%	1%	15
	Machine learning–CKD [18,12]	98.4 K	259–1211	0.26–1.23%	984	1%	132–960	34–94%	13–98%	53–94
Ligand-based VS (clustering), large libraries	Hierachical k-means [5,28]	344.5 K	91–1556	0.026–0.45%	3750–21285	1.1–6.2%	27–761	23–55%	0.72–5%	7.97–31.2
	NIPALSTREE [5,28]	344.5 K	91–1556	0.026–0.45%	3469–28125	1.0–8.2%	17–625	18–50%	0.49–2.8%	3.51–18.7
	Hierachical k-means + NIPALSTREE disjunction [5,28]	344.5 K	91–1556	0.026–0.45%	7317–43165	2.1–12.3%	30–980	33–72%	0.41–2.9%	4.86–17.6
	Hierachical k-means + NIPALSTREE conjunction [5,28]	344.5 K	91–1556	0.026–0.45%	538–6692	0.16–1.9%	14–406	6–32%	1.1–10.2%	7.77–98
Ligand-based VS (structural signatures), extremely large libraries (≥1 M)	Pharmacophore [3,29,80,81]	1.77 M–3.8 M	55–144	0.0014–0.0081%	20 K–1 M	1.15–26%	6–39	11–70%	0.0039–0.084%	3–10.3
Ligand-based VS (structural signatures), large libraries	Pharmacophore [1,30]	380 K	30	0.0079%	6917	1.82%	23	76.7%	0.33	41.8

such alternative VS tools [7–9] because of their high-CPU speed (100 K data points per hour on 3 GHz PC) [11] and capability for covering highly diverse spectrum of compounds [32].

The reported performance of various LBVS and SBVS tools in screening compound libraries of >90,000 compounds is summarised in Table 1. Caution needs to be raised about straightforward comparison of these reported results, which might be misleading because the outcome of VS strongly depends on the datasets used. The dataset-dependence of VS performance can be illustrated by a test shown in a subsequent section of this paper. Therefore, the listed results should be viewed as providing very crude pictures about the reported VS performances. While exhibiting equally good hit selection performance, in screening extremely large (≥1 million) and large (100,000–900,000) libraries, the currently developed ML tools tend to show lower hit-rate (ratio of known hits and the predicted hits) and, in some cases, lower enrichment factor (magnitude of hit-rate improvement over random selection) than the best performing SBVS tools. For instance, in screening extremely large libraries, the reported yield (percentage of known hits predicted), hit-rate and enrichment factor of ML tools are in the range of 55–81%, 0.2–0.7% and 110–795, respectively [8,11,13], compared to those of 62–95%, 0.65–35% and 20–1200 by SBVS tools [18,19]. While in screening libraries of ∼98,000 compounds the reported hit-rates of some ML tools are comparable to those of SBVS tools, their enrichment factors are substantially smaller. A lower hit-rate gives rise to a higher number of false-hits and a lower enrichment factor suggests that there might be bigger room for further optimizing a VS tool. Hence, there is a need for further improving the hit-rate and enrichment factor of ML tools. It is not uncommon for the pharmaceutical industry to screen >1 million compounds per high-throughput screening campaign [33]. Therefore, improvement of hit-rate and enrichment factor is highly desirable for developing practically useful ML tools for LBVS.

Two approaches have been explored for minimizing false hits. One is the selection of top-ranked hits, which has been extensively used in LBVS [8,9,13,14,34,35] and SBVS [18,20–22,36,37]. The other is the elimination of potentially unpromising hits in pre-screening stage by using such filters as Lipinski's rule of five [38,19], and recognition of pharmacophore [21] and specific chemical groups or interaction patterns [18,20,24,39]. In addition to the application of these approaches, the performance of ML tools in screening large libraries may be further improved by using training sets of more diverse spectrum of compounds to develop more optimally performing ML models. These models have been generated by using two-tier supervised classification ML methods [7–9,11–14,40], which require training sets of diverse spectrum of active and inactive compounds. The training inactive compounds in these models have been collected from up to a few hundred known inactive compounds or/and putative inactive compounds from up to a few dozen biological target classes in MDDR database [7–9,11–14,40], which may not always be sufficient to fully represent inactive compounds in the vast chemical space, thereby making it difficult to optimally minimize false hit prediction rate of ML models.

In this work, we examined to what extent hit-rate and enrichment factor of ML tools can be improved by using training-sets of more diverse spectrum of inactive compounds. A widely used and better performing ML method, support vector machines (SVM) [8–12,14], was used to develop SVM models for identifying active compounds of single mechanism (HIV-1 protease inhibitors, dihydrofolate reductase (DHFR) inhibitors, dopamine receptor antagonists) and multiple mechanisms (central nervous system (CNS) active agents). HIV-1 protease inhibitors form an important class of anti-HIV agents some of which have been successfully used clinically [41]. DHFR inhibitors are useful for the treatment of microbial infections [42], cancer [43], and parasitic diseases [44]. Dopamine antagonists have been used as antipsychotic agents [45] and for the treatment of cervical dystonia [46], vertigo [47], and gastrointestinal motility disorders [48]. CNS active agents are composed of a diverse spectrum of CNS acting compounds that produce anxiolytic, antipsychotic, antidepressant, analgesic, anticonvulsant, antimigraine, antiischemic, antiparkinsonian, nootropic, neurologic, epileptic, neuroleptic, neurotropic, neuronal injury inhibiting, narcotics antagonizing, and CNS stimulating effects [49]. Because of their diverse therapeutic applications and structural frameworks, these compounds are highly useful for testing the performance of SVM and other ML tools in LBVS of large compound libraries.

Our SVM models were trained by using known active compounds and putative inactive compounds extracted from compound families that contain no known active compound. Compound families can be generated by clustering distinct compounds of chemical databases into groups of similar structural and physicochemical properties [28]. The developed SVM models were tested in screening libraries of 2.986 million compounds from the PUBCHEM database that are not in the training sets of these SVM models. The yields, hit-rates and enrichment factors derived from these tests were compared with those of SBVS and other LBVS tools applied in the screening of extremely large libraries to determine to which extent the overall performance of SVM models can be enhanced and whether it is comparable to that of the best performing VS tools reported in the literatures. To further evaluate whether our SVM models predict active and inactive compounds rather than membership of certain compound families, distribution of the predicted active and inactive compounds in the compound families were analyzed.

2. Methods

2.1. Collection of active compounds

Table 2 gives the statistics of collected active compounds for the four active compound classes and their structural diversity index (DI) (defined in subsequent section). The structures of a few selected compounds for each class are shown in Fig. 1. For comparison of structural diversity of the compounds in these and those of the other structurally diverse classes, the statistics

Table 2
Diversity index (DI) and number of HIV protease inhibitors, DHFR inhibitors, dopamine antagonists, and CNS active agents used for developing support vector machines ligand-based virtual screening tools

Chemical class	No. of active compounds	DI Value
Blood–brain barrier penetrating agents [87]	276	0.430
FDA approved drugs	1,121	0.495
NCI diversity set	1,804	0.544
P-Glycoprotein substrates [69]	116	0.555
CYP D6 inhibitors	180	0.575
CNS active agents (this work)	16,182	0.578
CYP 2D6 substrates	198	0.588
Human intestine absorbing agents [55]	131	0.596
Estrogen receptor agonists [10]	243	0.618
HIV protease inhibitors (this work)	5,161	0.626
DHFR inhibitors (this work)	755	0.719
Dopamine antagonists (this work)	1,184	0.741

For comparison, relevant data of several other compound classes of highly diverse structures are also included. These compound classes are arranged in descending order of structural diversity.

and DI values of several such classes are also listed in Table 2. A total of 5161 HIV-1 protease inhibitors, with log(IC50) values in the range of −7.85 to −3.30, were selected from the HIV/OI Enzyme Inhibition Database of the National Institute of Allergy and Infectious Diseases of NIH. 76.6% of which are peptide-based inhibitors (66% and 5% are peptidomimetics and symmetry-based inhibitors, respectively) and 23.4% are non-peptide-based inhibitors. The quality of these inhibitors were further validated against literature reports we found from the literature database PUBMED to ensure that they have been described as HIV-1 protease inhibitors with IC50 values in the range of binding potencies considered to be important in various cases.

DHFR inhibitors were collected from a publication [50]. We were able to use our software [51] to generate molecular descriptors of 755 of the 756 collected inhibitors. We collected 1184 distinct dopamine antagonists from three separate sources, which include 1163 from MDDR database, 126 from PUBCHEM database, and 41 from a publication [52]. CNS active agents were retrieved from those compounds in MDDR database annotated as anxiolytic, antipsychotic, antidepressant, analgesic (non-opioid and opioid), anticonvulsant, antimigraine, antiischemic (cerebral), antiparkinsonian, stimulant in central, antagonist to narcotics, centrally acting agent, nootropic agent, neurologic agent, epileptic, and neuronal injury inhibitor/neuroleptic/neurotropic. We were able to use our software [51] to derive molecular descriptors for 16,182 of the collected 16,390 non-redundant CNS active compounds.

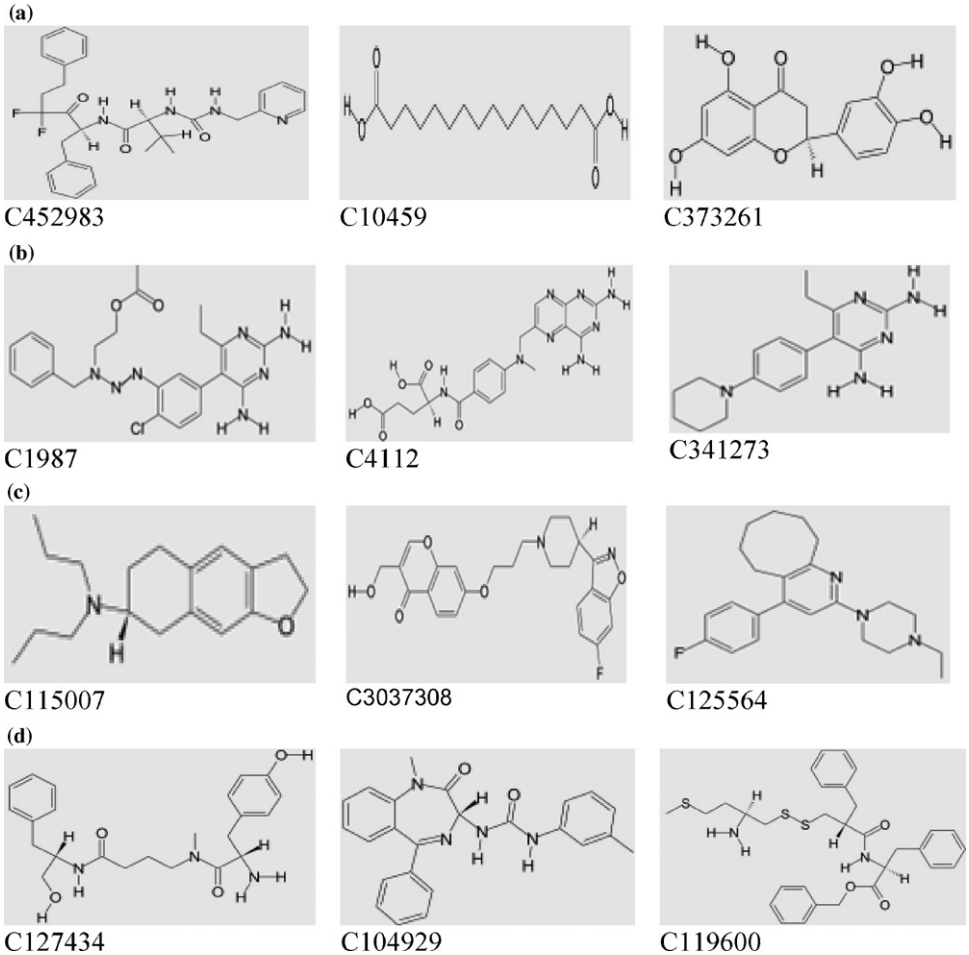

Fig. 1. Structure of the selected HIV protease inhibitors, DHFR inhibitors, dopamine antagonists, and CNS active agents. The PUBCHEM accession number of these compounds is given: (a) HIV-1 protease inhibitors; (b) DHFR inhibitors; (c) Dopamine antagonists; (d) CNS active agents.

2.2. Generation of putative inactive compounds

Apart from the use of known inactive compounds and active compounds of other biological target classes as putative inactive compounds [7–9,11–14,40], a new approach extensively used for generating inactive proteins in ML classification of various classes of proteins [53–55] may be applied for generating putative inactive compounds. An advantage of this approach is its independence on the knowledge of known inactive compounds and active compounds of other biological target classes, which enables more expanded coverage of the ''inactive'' chemical space in cases of limited knowledge of inactive compounds and compounds of other biological classes. A drawback of this approach is the possible inclusion of some undiscovered active compounds in the ''inactive'' class, which may affect the capability of ML methods for identifying novel active compounds. As will be demonstrated, such an adverse effect is expected to be relatively small for many biological target classes.

In applying this approach to proteins, all known proteins are clustered into ~8900 protein families based on the clustering of their amino acid sequences [56], and a set of putative inactive proteins can be tentatively extracted from a few representative proteins in those families without a single known active protein. Undiscovered active proteins of a specific functional class typically cover no more than a few hundred families, which give a maximum possible ''wrong'' family representation rate of <10% even when all of the undiscovered active proteins are misplaced into the inactive class [57]. Importantly, inclusion of the representative of a ''wrong'' family into the inactive class does not preclude other active family members from being classified as active. Statistically, a substantial percentage of active members can be classified by ML methods as active even if its family representative is in the inactive class [57]. Therefore, in principle, a reasonably good ML model can be derived from these putative inactive samples, which has been confirmed by a number of studies [53–55,57].

In a similar manner, known compounds can be grouped into compound families by clustering them in the chemical space defined by their molecular descriptors [28,58]. As ML methods predict compound activities based on their molecular descriptors, in developing ML tools, it makes sense to cluster as well as to represent compounds in terms of molecular descriptors. By using a k-means method [28,58] and molecular descriptors computed from our own software [51], we generated 7990 cluster families from the available compounds in PUBCHEM database, which is consistent with the 12,800 compound-occupying neurons (regions of topologically close structures) for 26.4 million compounds of up to 11 atoms [59], and the 2851 clusters for 171,045 natural products [60]. Analogue groups such as steroids and catecholamines are distributed in a few families. Active compounds in extensively studied target classes such as those of HIV-1 protease inhibitors, DHFR inhibitors, and dopamine antagonists are distributed in 770, 135, and 799 families, respectively. Because of the extensive effort in searching the known compound libraries for identifying active compounds in these target classes, the number of undiscovered ''active'' families in PUBCHEM database is expected to be relatively small, most likely no more than several hundred families. The ratio of the undiscovered ''active'' families (hundreds on less) and the families that contain no known active compound (6000–7000 based on current version of PUBCHEM) for these and possibly many other target classes is expected to be <15%. Therefore, putative inactive compounds can be generated by extracting a few representative compounds of those families that contain no known active compound, with a maximum possible ''wrong'' family representation rate of <15% even when all of the undiscovered active compounds are misplaced into the inactive class.

CNS active agents are distributed in numerous biological target classes such as agonists, antagonists, regulators of G-protein coupled receptors and nuclear receptors, blockers and regulators of ion channels, substrates, inhibitors, activators, and regulators of transporters, and inhibitors and regulators of enzymes involved in the synthesis and metabolism of signalling molecules in the CNS system [49]. Therefore, agents in this multi-target class are expected to cover a significantly larger portion of the chemical space than those of a single target class, leading to a possibly higher ''wrong'' family representation rate because of the likelihood of higher number of undiscovered active families in the limited chemical space covered by the currently available compounds in existing databases. As a result, the quality of the putative non-CNS active compounds generated by the new approach may be affected to some extent. The new approach is expected to become more and more useful for multi-target classes when the coverage of chemical space can be significantly expanded as a result of increasing volume of the chemical databases.

There are 7220, 7855, 7191, 3440 families that contain no known HIV-1 protease inhibitor, DHFR inhibitor, dopamine antagonist, and CNS active agent, respectively. Thus datasets of 41,254 putative non-HIV-1 protease inhibitors, 44,856 putative non-DHFR inhibitors, 42,804 putative non-dopamine antagonists, and 20,465 putative non-CAN active compounds were generated by random selection of 5–6 representative compounds from each of these families, respectively.

2.3. Molecular descriptors

Molecular descriptors are quantitative representations of structural and physicochemical features of molecules, which have been extensively used in deriving structure-activity relationships [61,62], quantitative structure activity relationships [63,64] and ML prediction models for pharmaceutical agents [64–72]. A total of 199 descriptors derived by using our software [51] were used in this work. These descriptors were selected from more than 1000 descriptors described in the literature by eliminating those descriptors that are obviously redundant or unrelated to the prediction of pharmaceutical agents [69,73]. The resulting 199 descriptors include 18 descriptors in the class of simple molecular properties, 28 descriptors in the class of molecular connectivity and shape, 97 descriptors in the class of

electro-topological state, 31 descriptors in the class of quantum chemical properties, and 25 descriptors in the class of geometrical properties. They were computed from the 3D structure of each compound by using our own designed molecular descriptor-computing program.

2.4. Determination of structural diversity

Structural diversity of the collected active compounds can be measured by using the diversity index (DI) value, which is the average value of the similarity between pairs of compounds in a dataset [74]:

$$\mathrm{DI} = \frac{\sum_{i=1}^{N} \sum_{j=1, i \neq j}^{N} \mathrm{sim}(i, j)}{N(N-1)}$$

where sim(i, j) is a measure of the similarity between compound i and j, and N is the number of compounds in the dataset. The structural diversity of a dataset increases with decreasing DI value. In this work, sim(i, j) is computed by using the Tanimoto coefficient [75].

$$\mathrm{sim}(i, j) = \frac{\sum_{d=1}^{l} x_{di} x_{dj}}{\sum_{d=1}^{l} (x_{di})^2 + \sum_{d=1}^{l} (x_{dj})^2 - \sum_{d=1}^{l} x_{di} x_{dj}}$$

where l is the number of descriptors computed for the molecules in the dataset.

2.5. Support vector machines method

SVM is a supervised ML method based on the structural risk minimization principle for minimizing both training and generalization error [76]. As illustrated in Fig. 2, when used for classification, SVM separates positive (active compounds) and negative (inactive compounds) training samples in a multi-dimensional space by constructing a hyper-plane optimally positioned between the positive and negative samples. A testing sample is then projected onto this multi-dimensional space to determine its class affiliation based on its relative position to the hyper-plane (active if on the active side, inactive if on the inactive side of the hyper-plane).

There are linear and nonlinear SVMs. Linear SVM is applicable for samples separable by linear mapping of their feature vectors. Nonlinear SVM is used for samples unseparable by linear mapping of their feature vectors, which is more useful for classifying compounds of diverse structures [8,9,11,13,40,71,77,78]. In nonlinear SVM, each feature vector $\mathbf{x}_i$ is projected into a higher dimensional feature space by using a kernel function such as Gausian function $K(\mathbf{x}_i, \mathbf{x}_j) = \mathrm{e}^{-||\mathbf{x}_j - \mathbf{x}_i||^2/2\sigma^2}$, where a hyper-plane is constructed by finding a vector $\mathbf{w}$ and a parameter b that minimizes $||\mathbf{w}||^2$ which satisfies the conditions: $\mathbf{w}\cdot\mathbf{x}_i + b \geq +1$, for $y_i = +1$ (active) and $\mathbf{w}\cdot\mathbf{x}_i + b \leq -1$, for $y_i = -1$ (inactive). Based on the derived $\mathbf{w}$ and b, a new compound $\mathbf{x}$ can be classified as active or inactive when sign[$(\mathbf{w}\cdot\mathbf{x})$] + b] is positive or negative.

2.6. Development of support vector machines virtual screening tools

SVM models for identifying HIV protease inhibitors, DHFR inhibitors, dopamine antagonists, and CNS active agents were developed by a procedure widely used for developing SVM protein classification models of optimal performance [53–55]. In the first step, active and inactive compounds were each divided into separate training, testing and independent evaluation sets. Specifically, active and inactive compounds were each clustered into groups based on their distance in the molecular descriptor space by using a hierarchical clustering method [79]. An upper-limit of the largest separation of 20 was used for each cluster. One representative compound was randomly selected from each group to form a training set that is sufficiently diverse and broadly distributed in the descriptor space. One or up to 50% of the remaining compounds in each group were randomly selected to form the testing set. The selected compounds from each group were further checked to ensure that they are distinguished from those of other groups. The remaining compounds were used as the independent evaluation set, which are also of reasonable level of diversity. Moreover, an analysis of the compounds in each cluster shows that the majority of the compounds in a cluster are substantially different. Thus, the testing and independent evaluation sets are expected to have certain level of usefulness for performing their task of fine-tuning the parameter of a SVM model and for evaluating its prediction performance. In the second step, SVM models were trained by using the training set and their parameters were optimized by using the testing set. The SVM model with the best overall performance on both the testing and independent evaluation sets was selected as a VS tool.

2.7. Virtual screening performance measurement

The frequently used measures of VS performance yield hit-rate and enrichment factor [1,2,7–9,15,27]. In terms of true positives TP (true active), true negatives TN (true inactive), false positives FP (false active), and false negatives FN (false inactive). These are given by TP/(TP + FN), TP/(TP + FP) and TP(TP + FP + TN + FN)/(TP + FP) (TP + FN), respectively.

3. Assessment of virtual screening performance

The developed SVM models for identifying HIV protease inhibitors, DHFR inhibitors, dopamine antagonists, and CNS active agents in screening 2.986 million distinct compounds from the PUBCHEM database that is not in the training sets of our developed SVM models. The performance of these SVM models is given in Table 2, which can be compared with the reported performance of other SBVS and LBVS tools listed in Table 1. There are 2351, 225, 37, and 664 known HIV protease inhibitors, DHFR inhibitors, dopamine antagonists, and CNS active agents in the PUBCHEM database not in the training sets of our SVM models. Our SVM models were able to identify 78.0%, 52.4%, 62.2%, and 66.6% of these known hits, which are comparable to the range of 62–95% by the SBVS tools

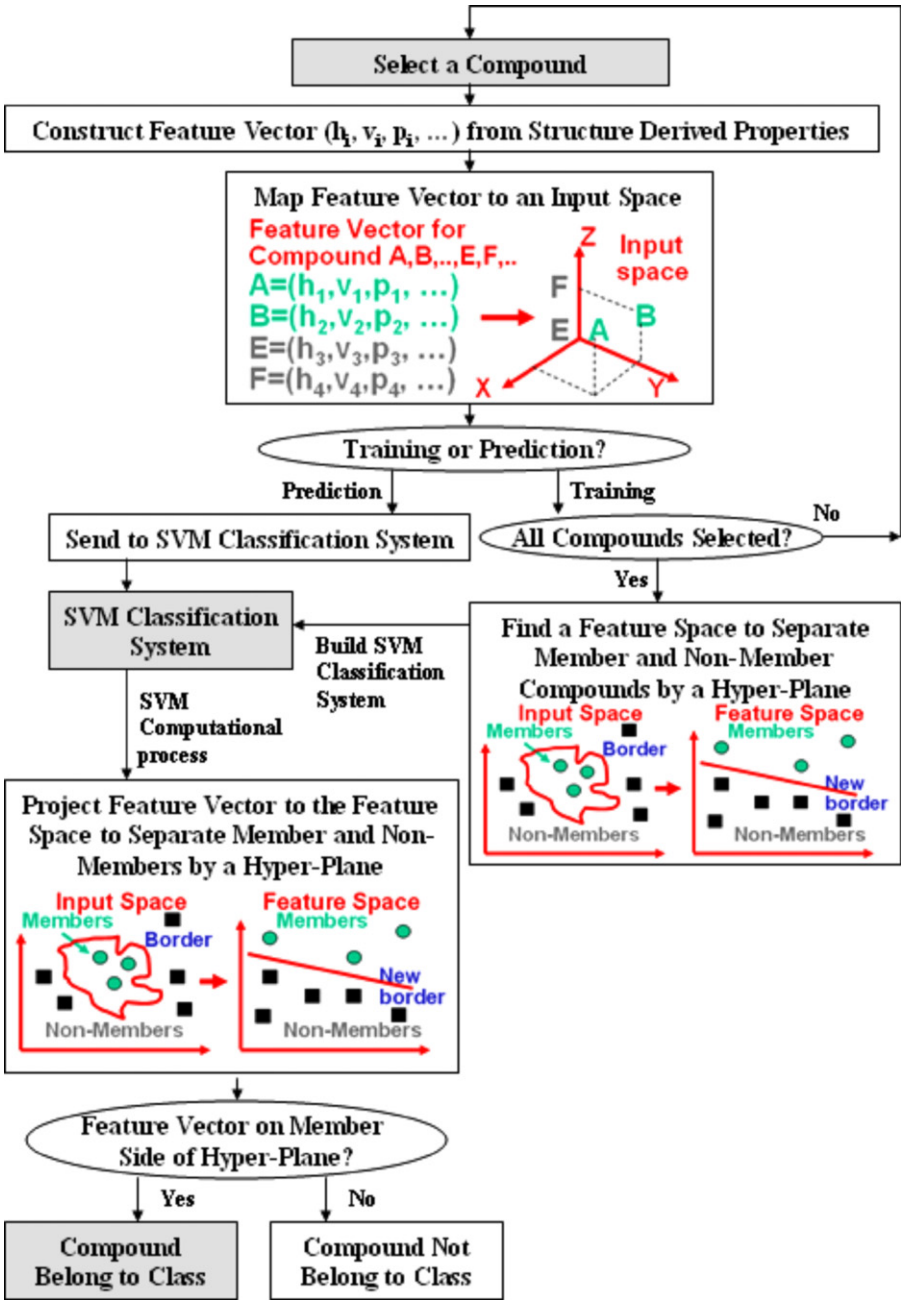

Fig. 2. Schematic diagram illustrating the process of the prediction of active compounds of a compound class from their structural-derived properties (molecular descriptors) by using a machine learning method—support vector machines. A, B, E, F and (h_j, p_j, v_j, etc.) represents such structural and physicochemical properties as hydrophobicity, volume, polarizability, etc.

[18,19] and 55–81% by other LBVS [8,11,13] tools in screening libraries of ≥1 million compounds, and they are also comparable to the percentages in screening libraries of 98,400–344,500 compounds by other SBVS [15–17,20–26] and LBVS tools [9,11–14,28,30]. These results suggest that our developed SVM models are equally effective in selecting potential hits in VS of large libraries.

In addition to the exhibition of equally effective hit selection performance, our SVM models appear to show relatively lower "false" hit identification rate. Without the use of top-ranked cut-off or additional filter, our SVM models identified a total of 8157, 160, 299, and 9502 virtual hits for the four compound classes, respectively, which are comparable to and in some cases smaller than those identified by SBVS [15–26] and other LBVS [8,9,11–13,28,73,77,78] tools even though a substantially larger number of compounds (2.983 M vs. 98.4 K to 2.5 M) were screened. As a result, smaller percentages of screened compounds were selected as virtual hits, which are in the range of 0.0054–0.32% as compared to those of 0.08–3% by SBVS tools [15–26], 0.1–5% by other reported ML models [9,11–14], 0.16–82% by clustering methods [28], and 1.15–26% by pharmacophore models [29,30,80,81]. By using Lipinski's rule of five [38] as a filter, the numbers of identified virtual hits are further reduced to 333, 115, 209, and 8035 for

Table 3
Performance of support vector machines virtual screening tools developed in this work for identifying HIV protease inhibitors, DHFR inhibitors, dopamine antagonists, and CNS active agents in screening 2.986 million compounds

Screening task	Compounds screened				Virtual hits selected by SVM					Known hits selected by SVM			
	No of compounds	No of known hits not in training sets of SVM-LBVS tool	Percent of known hits	No of families covered by known hits	No of selected virtual hits	Percent of selected virtual hits not in the families covered by known hits	Percent of screened compounds selected as virtual hits	No of selected virtual hits passed rule-of-five	Percent of selected virtual hits passed rule-of-five and not in the families covered by known hits	No of known hits selected	Yield	Hit-rates	Enrichment factor
HIV protease inhibitors	2.986 M	2351	0.076%	496	8157	24.3%	0.27%	333	42.6%	1833	78.0%	22.5%	296
DHFR inhibitors	2.986 M	225	0.007%	60	160	71.3%	0.0054%	115	64.4%	118	52.4%	73.8%	10,543
Dopamine antagonists	2.986 M	37	0.0012%	29	299	87.6%	0.01%	209	82.8%	23	62.2%	7.7%	6,417
CNS active agents	2.986 M	664	0.022%	519	9502	85.7%	0.32%	8035	84.1%	442	66.6%	4.7%	214

the four compound classes, respectively, suggesting that introduction of such filters or combination with other VS methods may enable further reduction of the number of falsely predicted hits.

The hit-rates of our SVM models are 22.5%, 73.8%, 7.7%, and 4.7% for the four classes of compounds, respectively, which are comparable to those of 0.65–35% by SBVS tools [18,19] and substantially improved against those of 0.2–0.7% by other reported SVM models [8,11,13] in screening extremely large libraries. These hit-rates are also greater than the majority of the hit-rates in screening large libraries of 98,400–344,500 compounds by SBVS [15–17,20–26] and other LBVS tools [9,11–14,28–30,80,81]. The enrichment factors of our SVM models are 296, 10,543, 6417, and 214 for the four classes of compounds, respectively, which are substantially improved against those of 20–1200 by SBVS tools [18,19] and 110–795 by other reported SVM models [8,11,13] in screening extremely large libraries. Therefore, our method is useful in improving the hit-rate and enrichment factor of SVM while maintaining the equally high hit identification rate as other SBVS and LBVS tools.

To further evaluate whether our SVM models predict active compounds rather than membership of certain compound families, Compound family distribution of the predicted active and inactive compounds for the four compound classes were analyzed. As shown in Table 3, 24.3%, 71.3%, 87.6%, 85.7% of the predicted HIV protease inhibitors, DHFR inhibitors, dopamine antagonists, and CNS active agents belong to the families that contain no known active compound. For those families that contain at least one known active compound, >70% of the compounds (>90% in majority cases) in each of these families were predicted as inactive compounds by our SVM models. These results suggest that our SVM models predict active compounds rather than membership to certain compound families. Some of the predicted active compounds not in the family of known active compounds may serve as potential ''novel'' active compounds. Therefore, as in the case shown by an earlier study [82], SVM methods have certain capacity for predicting novel active compounds.

4. Comparative analysis of virtual screening performance of our method

The performance of our method can be more appropriately evaluated by using it to develop VS tools and test them based on the same dataset construction and testing procedures as those used in other VS methods. In this work, we specifically developed additional VS prediction models by using the same dataset construction method and same data source of a standard similarity-based method, the data fusion method [13], the performance of both methods were then compared by using the same data source. The data fusion method is based on Taminoto-based similarity searching using multiple reference compounds, which have shown good performances for a number of active compound groups by using only a small number of training active compounds [13], and thus is a good

reference method for evaluating the performance of out method.

We developed three separate HIV protease inhibitor VS tools by using our method and datasets of similar sizes and from the same sources as that used by the reported studies of the data fusion method [13,83]. Our training and testing datasets were generated from 1054 HIV protease inhibitors extracted from the MDDR database. Based on the training set generation procedure of the data fusion method [83], three sets of 60, 80 and 100 inhibitors were selected from this full set of 1054 inhibitors as the active compound training sets, from which the inactive compound training sets were generated by using our method. Using the same testing method of the data fusion method, the performance of the three developed SVM VS tools were evaluated by using the remaining 994, 974 and 954 HIV protease inhibitors, respectively, which showed that 59.5%, 62.2% and 67.3% of these remaining inhibitors were correctly identified. The performance of these SVM VS tools is similar to and in some cases slightly improved against that of 55.2–58.0% of the data fusion method that used a similar number of training HIV protease inhibitors [13]. This suggests that, by using the equally small active compounds as training data, our SVM model is capable of performing at the same level and in some cases slightly improved level than that of the data fusion method.

5. Discussions

The performance of SVM and other ML methods critically depends on the diversity of compounds in a training dataset and the appropriate description of the compounds. The datasets used in developing ML models described in Table 1 and in this work are not expected to be fully representative of all of the active and inactive compounds. Known inactive compounds, particularly those structurally similar to an active compound, may serve to further refine ML models at higher ''structural resolutions'' than those achievable by using only the putative inactive compounds generated from this work. Mining of known active compounds and inactive compounds from the literature [84] and other sources [85,86] is a key to developing more optimally performing ML models for VS.

Examination of incorrectly predicted compounds by ML models consistently suggests that the currently used molecular descriptors are insufficient to adequately represent some of the compounds that contain complex structural or chemical configurations [69,71,87]. Examples of these agents are those with large rigid structure combined with a short flexible hydrophilic tail, compounds that contain multi-rings with various heteroatoms such as nitrogen, oxygen, sulphur, fluorine and chlorine. Due to the limited coverage of the number of bond links in a heteroatom loop, the currently available topological descriptors are not yet capable of describing the special features of a complex multi-ring structure that contains multiple heteroatoms. It appears that none of the currently available descriptors are capable of fully representing molecules containing a long flexible chain. Therefore, it might be helpful to explore different combination of descriptors and to select more optimal set of descriptors by using more refined feature selection algorithms and parameters [69,73]. However, indiscriminate use of many existing topological descriptors, some of which are overlapping and redundant to each other, may introduce noise as well as extending the coverage of some the aspects of these special features. Thus, it may be necessary to introduce new descriptors for more appropriately representing these and other special features.

6. Concluding remarks

By using training sets of more diverse spectrum of inactive compounds, the hit-rates and enrichment factors of SVM models can be substantially improved to the level comparable to and in some cases higher than those of the best performing SBVS and LBVS tools reported in the literatures. Because of their high computing speed and capability for covering highly diverse spectrum compounds, SVM and other ML methods 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 lead discovery [19,23,81].

References

[1] B.K. Shoichet, Virtual screening of chemical libraries, Nature 432 (2004) 862–865.
[2] T. Lengauer, C. Lemmen, M. Rarey, M. Zimmermann, Novel technologies for virtual screening, Drug Discov. Today 9 (2004) 27–34.
[3] J.W. Davies, M. Glick, J.L. Jenkins, Streamlining lead discovery by aligning in silico and high-throughput screening, Curr. Opin. Chem. Biol. 10 (2006) 343–351.
[4] P. Willett, Similarity-based virtual screening using 2D fingerprints, Drug Discov. Today 11 (2006) 1046–1053.
[5] H. van de Waterbeemd, E. Gifford, ADMET in silico modelling: towards prediction paradise? Nat. Rev. Drug Discov. 2 (2003) 192–204.
[6] W.B. Matthew, S.B.H. Trotter, Support vector machines for ADME property classification, QSAR & Combinatorial Sci. 22 (2003) 533–548.
[7] G. Harper, J. Bradshaw, J.C. Gittins, D.V. Green, A.R. Leach, Prediction of biological activity for high-throughput screening using binary kernel discrimination, J. Chem. Inf. Comput. Sci. 41 (2001) 1295–1300.
[8] R.N. Jorissen, M.K. Gilson, Virtual screening of molecular databases using a support vector machine, J. Chem. Inf. Model. 45 (2005) 549–561.
[9] M. Glick, J.L. Jenkins, J.H. Nettles, H. Hitchings, J.W. Davies, 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. 46 (2006) 193–200.
[10] H. Li, C.Y. Ung, C.W. Yap, Y. Xue, Z.R. Li, Y.Z. Chen, Prediction of estrogen receptor agonists and characterization of associated molecular descriptors by statistical learning methods, J. Mol. Graph Model. 25 (2006) 313–323.
[11] Z. Lepp, T. Kinoshita, H. Chuman, Screening for new antidepressant leads of multiple activities by support vector machines, J. Chem. Inf. Model. 46 (2006) 158–167.
[12] B. Chen, R.F. Harrison, G. Papadatos, P. Willett, D.J. Wood, X.Q. Lewell, P. Greenidge, N. Stiefl, Evaluation of machine-learning methods for ligand-based virtual screening, J. Comput. Aid. Mol. Des. (2007).
[13] J. Hert, P. Willett, D.J. Wilton, P. Acklin, K. Azzaoui, E. Jacoby, A. Schuffenhauer, 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. 46 (2006) 462–470.
[14] L. Franke, E. Byvatov, O. Werz, D. Steinhilber, P. Schneider, G. Schneider, Extraction and visualization of potential pharmacophore points using

support vector machines: application to ligand-based virtual screening for COX-2 inhibitors, J. Med. Chem. 48 (2005) 6997–7004.
[15] S. Ghosh, A. Nie, J. An, Z. Huang, Structure-based virtual screening of chemical libraries for drug discovery, Curr. Opin. Chem. Biol. 10 (2006) 194–202.
[16] B.K. Shoichet, S.L. McGovern, B. Wei, J.J. Irwin, Lead discovery using molecular docking, Curr. Opin. Chem. Biol. 6 (2002) 439–446.
[17] J.M. Jansen, E.J. Martin, Target-biased scoring approaches and expert systems in structure-based virtual screening, Curr. Opin. Chem. Biol. 8 (2004) 359–364.
[18] J.C. Mozziconacci, E. Arnoult, P. Bernard, Q.T. Do, C. Marot, L. Morin-Allory, Optimization and validation of a docking-scoring protocol: application to virtual screening for COX-2 inhibitors, J. Med. Chem. 48 (2005) 1055–1068.
[19] D. Vidal, M. Thormann, M. Pons, A novel search engine for virtual screening of very large databases, J. Chem. Inf. Model. 46 (2006) 836–843.
[20] M.D. Cummings, R.L. DesJarlais, A.C. Gibbs, V. Mohan, E.P. Jaeger, Comparison of automated docking programs as virtual screening tools, J. Med. Chem. 48 (2005) 962–976.
[21] A. Evers, T. Klabunde, Structure-based drug discovery using GPCR homology modeling: successful virtual screening for antagonists of the alpha1A adrenergic receptor, J. Med. Chem. 48 (2005) 1088–1097.
[22] D.M. Lorber, B.K. Shoichet, Hierarchical docking of databases of multiple ligand conformations, Curr. Top Med. Chem. 5 (2005) 739–749.
[23] N. Stiefl, A. Zaliani, A knowledge-based weighting approach to ligand-based virtual screening, J. Chem. Inf. Model. 46 (2006) 587–596.
[24] E. Vangrevelinghe, K. Zimmermann, J. Schoepfer, R. Portmann, D. Fabbro, P. Furet, Discovery of a potent and selective protein kinase CK2 inhibitor by high-throughput docking, J. Med. Chem. 46 (2003) 2656–2662.
[25] T.N. Doman, S.L. McGovern, B.J. Witherbee, T.P. Kasten, R. Kurumbail, W.C. Stallings, D.T. Connolly, B.K. Shoichet, Molecular docking and high-throughput screening for novel inhibitors of protein tyrosine phosphatase-1B, J. Med. Chem. 45 (2002) 2213–2221.
[26] I.J. Enyedy, Y. Ling, K. Nacro, Y. Tomita, X. Wu, Y. Cao, R. Guo, B. Li, X. Zhu, Y. Huang, Y.Q. Long, P.P. Roller, D. Yang, S. Wang, Discovery of small-molecule inhibitors of Bcl-2 through structure-based computer screening, J. Med. Chem. 44 (2001) 4313–4324.
[27] T.I. Oprea, H. Matter, Integrating virtual screening in lead discovery, Curr. Opin. Chem. Biol. 8 (2004) 349–358.
[28] A. Bocker, G. Schneider, A. Teckentrup, NIPALSTREE: a new hierarchical clustering approach for large compound libraries and its application to virtual screening, J. Chem. Inf. Model. 46 (2006) 2220–2229.
[29] D. Schuster, E.M. Maurer, C. Laggner, L.G. Nashev, T. Wilckens, T. Langer, A. Odermatt, The discovery of new 11beta-hydroxysteroid dehydrogenase type 1 inhibitors by common feature pharmacophore modeling and virtual screening, J. Med. Chem. 49 (2006) 3454–3466.
[30] T. Steindl, C. Laggner, T. Langer, Human rhinovirus 3C protease: generation of pharmacophore models for peptidic and nonpeptidic inhibitors and their application in virtual screening, J. Chem. Inf. Model. 45 (2005) 716–724.
[31] H. Li, C.W. Yap, C.Y. Ung, Y. Xue, Z.R. Li, L.Y. Han, H.H. Lin, Y.Z. Chen, Machine learning approaches for predicting compounds that interact with therapeutic and ADMET related proteins, J. Pharm. Sci. 96 (2007) 2838–2860.
[32] H. Li, C.W. Yap, Y. Xue, Z.R. Li, C.Y. Ung, L.Y. Han, Y.Z. Chen, Statistical learning approach for predicting specific pharmacodynamic, pharmacokinetic or toxicological properties of pharmaceutical agents, Drug Dev. Res. 66 (2006) 245–259.
[33] C. Lipinski, A. Hopkins, Navigating chemical space for biology and medicine, Nature 432 (2004) 855–861.
[34] D.J. Wilton, R.F. Harrison, P. Willett, J. Delaney, K. Lawson, G. Mullier, Virtual screening using binary kernel discrimination: analysis of pesticide data, J. Chem. Inf. Model. 46 (2006) 471–477.
[35] B. Chen, R.F. Harrison, K. Pasupa, P. Willett, D.J. Wilton, D.J. Wood, X.Q. Lewell, Virtual screening using binary kernel discrimination: effect of noisy training data and the optimization of performance, J. Chem. Inf. Model. 46 (2006) 478–486.
[36] J.C. Alvarez, High-throughput docking as a source of novel drug leads, Curr. Opin. Chem. Biol. 8 (2004) 365–370.
[37] M. Schapira, B.M. Raaka, S. Das, L. Fan, M. Totrov, Z. Zhou, S.R. Wilson, R. Abagyan, H.H. Samuels, Discovery of diverse thyroid hormone receptor antagonists by high-throughput docking, Proc. Natl. Acad. Sci. U.S.A. 100 (2003) 7354–7359.
[38] C.A. Lipinski, F. Lombardo, B.W. Dominy, P.J. Feeney, Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings, Adv. Drug Deliv. Rev. 46 (2001) 3–26.
[39] E. Perola, Minimizing false positives in kinase virtual screens, Proteins 64 (2006) 422–435.
[40] J. Cui, L.Y. Han, H.H. Lin, H.L. Zhang, Z.Q. Tang, C.J. Zheng, Z.W. Cao, Y.Z. Chen, Prediction of MHC-binding peptides of flexible lengths from sequence-derived structural and physicochemical properties, Mol. Immunol. 44 (2007) 866–877.
[41] A. Spaltenstein, W.M. Kazmierski, J.F. Miller, V. Samano, Discovery of next generation inhibitors of HIV protease, Curr. Top Med. Chem. 5 (2005) 1589–1607.
[42] R.L. Then, Antimicrobial dihydrofolate reductase inhibitors—achievements and future options: review, J. Chemother. 16 (2004) 3–12.
[43] J.J. McGuire, Anticancer antifolates: current status and future directions, Curr. Pharm. Des. 9 (2003) 2593–2613.
[44] G.E. Linares, E.L. Ravaschino, J.B. Rodriguez, Progresses in the field of drug design to combat tropical protozoan parasitic diseases, Curr. Med. Chem. 13 (2006) 335–360.
[45] A. Serretti, D. De Ronchi, C. Lorenzi, D. Berardi, New antipsychotics and schizophrenia: a review on efficacy and side effects, Curr. Med. Chem. 1s1 (2004) 343–358.
[46] C.H. Adler, R. Kumar, Pharmacological and surgical options for the treatment of cervical dystonia, Neurology 55 (2000) S9–S14.
[47] T.C. Hain, M. Uddin, Pharmacological treatment of vertigo, CNS Drugs 17 (2003) 85–100.
[48] P. Demol, H.J. Ruoff, T.R. Weihrauch, Rational pharmacotherapy of gastrointestinal motility disorders, Eur. J. Pediatr. 148 (1989) 489–495.
[49] H.P. Rang, M.M.D., J.M. Ritter Pharmacology, vol. Churchill Livingstone, 2001.
[50] J.J. Sutherland, L.A. O'Brien, D.F. Weaver, Spline-fitting with a genetic algorithm: a method for developing classification structure–activity relationships, J. Chem. Inf. Comput. Sci. 43 (2003) 1906–1915.
[51] Y. Xue, C.W. Yap, L.Z. Sun, Z.W. Cao, J.F. Wang, Y.Z. Chen, Prediction of P-glycoprotein substrates by a support vector machine approach, J. Chem. Inf. Comput. Sci. 44 (2004) 1497–1505.
[52] J. Bostrom, M. Bohm, K. Gundertofte, G. Klebe, A 3D QSAR study on a set of dopamine D4 receptor antagonists, J. Chem. Inf. Comput. Sci. 43 (2003) 1020–1027.
[53] C.Z. Cai, L.Y. Han, Z.L. Ji, X. Chen, Y.Z. Chen, SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence, Nucleic Acids Res. 31 (2003) 3692–3697.
[54] L.Y. Han, C.Z. Cai, Z.L. Ji, Z.W. Cao, J. Cui, Y.Z. Chen, Predicting functional family of novel enzymes irrespective of sequence similarity: a statistical learning approach, Nucleic Acids Res. 32 (2004) 6437–6444.
[55] H.H. Lin, L.Y. Han, C.Z. Cai, Z.L. Ji, Y.Z. Chen, Prediction of transporter family from protein sequence by support vector machine approach, Proteins 62 (2006) 218–231.
[56] JK, W. JOELib/JOELib2. ed. 2005.
[57] L.Y. Han, C.J. Zheng, B. Xie, J. Jia, X.H. Ma, F. Zhu, H.H. Lin, X. Chen, Y.Z. Chen, Support vector machine approach for predicting druggable proteins: recent progress in its exploration and investigation of its usefulness. Drug Discov. Today (in press).
[58] T.I. Oprea, J. Gottfries, Chemography: the art of navigating in chemical space, J. Comb. Chem. 3 (2001) 157–166.
[59] Reymond, T.F.A.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 (published on Web 30/01/2007).

[60] M.A. Koch, A. Schuffenhauer, M. Scheck, S. Wetzel, M. Casaulta, A. Odermatt, P. Ertl, H. Waldmann, Charting biologically relevant chemical space: a structural classification of natural products (SCONP), Proc. Natl. Acad. Sci. U.S.A. 102 (2005) 17272–17277.
[61] H. Fang, W. Tong, L.M. Shi, R. Blair, R. Perkins, W. Branham, B.S. Hass, Q. Xie, S.L. Dial, C.L. Moland, D.M. Sheehan, Structure–activity relationships for a large diverse set of natural, synthetic, and environmental estrogens, Chem. Res. Toxicol. 14 (2001) 280–294.
[62] W. Tong, Q. Xie, H. Hong, L. Shi, H. Fang, R. Perkins, Assessment of prediction confidence and domain extrapolation of two structure–activity relationship models for predicting estrogen receptor binding activity, Environ. Health Perspect. 112 (2004) 1249–1254.
[63] M.N. Jacobs, In silico tools to aid risk assessment of endocrine disrupting chemicals, Toxicology 205 (2004) 43–53.
[64] J.Y. Hu, T. Aizawa, Quantitative structure–activity relationships for estrogen receptor binding affinity of phenolic chemicals, Water Res. 37 (2003) 1213–1222.
[65] E. Byvatov, U. Fechner, J. Sadowski, G. Schneider, Comparison of support vector machine and artificial neural network systems for drug/nondrug classification, J. Chem. Inf. Comput. Sci. 43 (2003) 1882–1889.
[66] S. Doniger, T. Hofman, J. Yeh, Predicting CNS permeability of drug molecules: comparison of neural network and support vector machine algorithms, J. Computat. Biol. 9 (2002) 849–864.
[67] L. He, P.C. Jurs, L.L. Custer, S.K. Durham, G.M. Pearl, Predicting the genotoxicity of polycyclic aromatic compounds from molecular structure with different classifiers, Chem. Res. Toxicol. 16 (2003) 1567–1580.
[68] R.D. Snyder, G.S. Pearl, G. Mandakas, W.N. Choy, F. Goodsaid, I.Y. Rosenblum, Assessment of the sensitivity of the computational programs DEREK, TOPKAT, and MCASE in the prediction of the genotoxicity of pharmaceutical molecules, Environ. Mol. Mutagen. 43 (2004) 143–158.
[69] Y. Xue, Z.R. Li, C.W. Yap, L.Z. Sun, X. Chen, Y.Z. Chen, Effect of molecular descriptor feature selection in support vector machine classification of pharmacokinetic and toxicological properties of chemical agents, J. Chem. Inf. Comput. Sci. 44 (2004) 1630–1638.
[70] C.W. Yap, C.Z. Cai, Y. Xue, Y.Z. Chen, Prediction of torsade-causing potential of drugs by support vector machine approach, Toxicol. Sci. 79 (2004) 170–177.
[71] C.W. Yap, Y.Z. Chen, Quantitative structure–pharmacokinetic relationships for drug distribution properties by using general regression neural network, J. Pharm. Sci. 94 (2005) 153–168.
[72] V.V. Zernov, K.V. Balakin, A.A. Ivaschenko, N.P. Savchuk, I.V. Pletnev, Drug discovery using support vector machines. The case studies of drug-likeness, agrochemical-likeness, and enzyme inhibition predictions, J. Chem. Inf. Comput. Sci. 43 (2003) 2048–2056.
[73] H. Li, C.W. Yap, C.Y. Ung, Y. Xue, Z.W. Cao, Y.Z. Chen, Effect of selection of molecular descriptors on the prediction of blood–brain barrier penetrating and nonpenetrating agents by statistical learning methods, J. Chem. Inf. Model. 45 (2005) 1376–1384.
[74] J.J. Perez, Managing molecular diversity, Chem. Soc. Rev. 34 (2005) 143–152.
[75] P. Willett, J.M. Barnard, G.M. Downs, Chemical similarity searching, J. Chem. Inf. Comput. Sci. 38 (1998) 983–996.
[76] T. Potter, H. Matter, Random or rational design? Evaluation of diverse compound subsets from chemical structure databases, J. Med. Chem. 41 (1998) 478–488.
[77] C.W. Yap, Y.Z. Chen, Prediction of cytochrome P450 3A4, 2D6, and 2C9 inhibitors and substrates by using support vector machines, J. Chem. Inf. Model. 45 (2005) 982–992.
[78] I.I. Grover, I.I. Singh, I.I. Bakshi, Quantitative structure–property relationships in pharmaceutical research—Part 2, Pharm. Sci. Technol. Today 3 (2000) 50–57.
[79] G.R.C. Rücker, Counts of all walks as atomic and molecular descriptors, J. Chem. Inf. Comput. Sci. 33 (1993) 683–695.
[80] B. Pirard, J. Brendel, S. Peukert, The discovery of Kv1.5 blockers as a case study for the application of virtual screening approaches, J. Chem. Inf. Model. 45 (2005) 477–485.
[81] M. Rella, C.A. Rushworth, J.L. Guy, A.J. Turner, T. Langer, R.M. Jackson, Structure-based pharmacophore design and virtual screening for novel angiotensin converting enzyme 2 inhibitors, J. Chem. Inf. Model. 46 (2006) 708–716.
[82] C.Y. Ung, H. Li, C.W. Yap, Y.Z. Chen, In silico prediction of pregnane X receptor activators by machine learning approaches, Mol. Pharmacol. 71 (2007) 158–168.
[83] M. Whittle, V.J. Gillet, P. Willett, J. Loesel, Analysis of data fusion methods in virtual screening: similarity and group fusion, J. Chem. Inf. Model. 46 (2006) 2206–2219.
[84] PubMed. In National Library of Medicine.
[85] MICROMEDEX MICROMEDEX. In MICROMEDEX, Edition expires 12/2003, Greenwood Village, Colorado.
[86] Bethesda AHFS drug information, vol. American Society of Health-System Pharmacists Inc., 2001.
[87] H. Li, C. Ung, C. Yap, Y. Xue, Z. Li, Z. Cao, Y. Chen, Prediction of genotoxicity of chemical compounds by statistical learning methods, Chem. Res. Toxicol. 18 (2005) 1071–1080.