

Predictive Activity Profiling of Drugs by Topological-Fragment-Spectra-Based Support Vector Machines

Kentaro Kawai, Satoshi Fujishima, and Yoshimasa Takahashi*

Laboratory for Molecular Information Systems, Department of Knowledge-Based Information Engineering,
Toyohashi University of Technology, Hibarigaoka 1-1, Tempaku-cho, Toyohashi 441-8580, Japan

Received December 26, 2007; Revised Manuscript Received April 20, 2008; Accepted April 23, 2008

Aiming at the prediction of pleiotropic effects of drugs, we have investigated the multilabel classification of drugs that have one or more of 100 different kinds of activity labels. Structural feature representation of each drug molecule was based on the topological fragment spectra method, which was proposed in our previous work. Support vector machine (SVM) was used for the classification and the prediction of their activity classes. Multilabel classification was carried out by a set of the SVM classifiers. The collective SVM classifiers were trained with a training set of 59 180 compounds and validated by another set (validation set) of 29 590 compounds. For a test set that consists of 9864 compounds, the classifiers correctly classified 80.8% of the drugs into their own active classes. The SVM classifiers also successfully performed predictions of the activity spectra for multilabel compounds.

INTRODUCTION

In the pharmaceutical industry, much effort has been devoted to developing new drugs. It is true that the development of new drugs can lead us to a better life, but serious adverse effects are often reported.¹ Pharmaceutical companies have to ensure the safety of their drugs in early research phases, because finding critical adverse effects in a late clinical phase leads to enormous financial losses. Although a versatile and powerful tool for estimating side effects is desired, nothing has been established so far. For approaching this problem, computational models obtained by traditional quantitative structure–activity relationships, machine learning methods, and knowledge-based systems have been adopted. Several computational programs are commercially available^{2,3} that might be useful for accelerating the research progress of drug safety. Many works for predicting the toxicity of chemicals have been reported.^{4–6} Most of the efforts in the computational toxicity prediction of drugs were mainly focused against cardiotoxicity,⁷ genotoxicity,^{8,9} hepatotoxicity,¹⁰ and drug-metabolizing enzyme inhibition.¹¹

When a drug shows plural biological actions (referred to as pleiotropic effects^{12,13}), it becomes a task for risk assessment to investigate the overall behavior of drugs, which would be relevant to undesirable effects. According to Davignon,¹³ pleiotropic effects are actions that are not specifically developed. The effects may be related or unrelated to the primary action mechanism of the drug, and those may be undesirable, neutral, or beneficial. A serious side effect of a drug is part of the pleiotropic effects, but it is really undesirable. However, for the beneficial case, identifying some of unanticipated effects would often lead us to the discovery of new drug candidates. Therefore, it is quite useful for us to predict the pleiotropic effects of a drug.

To predict unanticipated effects of a drug, Xie et al.¹⁴ investigated similarity of the binding site on the proteome-wide scale, which is based on the assumption that similar binding sites of proteins are likely to bind similar ligands, and they elucidated that the mechanism of the adverse effect of selective estrogen receptor modulators is based on the inhibition of sacroplasmic reticulum Ca^{2+} ion channel ATPase. In contrast, Bender et al.¹⁵ adopted a ligand-based approach, in which they applied Bayesian models of 70 preclinical safety-pharmacology-related targets to analyze unanticipated effects of the drugs, and they showed that 90% of known adverse reactions could be correctly identified, on average. In addition, there are many approaches^{16–19} available to investigate pharmacology targets of a drug. Poroikov and co-workers^{19–22} reported on a computer program, called PASS, which can be used to predict multiple biological actions of a drug. It was based on a statistical approach using MNA descriptors.²¹ The robustness of the PASS prediction was investigated too.²² It was shown that the prediction accuracy was still retained as much as before when they excluded 60% of the training data.

Machine learning techniques are widely applied in the various disciplines of classification problems, and this is the case in medicinal chemistry. There are several types of machine learning such as decision trees, artificial neural networks (ANN), support vector machines (SVM), and Bayesian models.²³ Recently, a number of SVM application studies were reported.^{24,25} Byvatov et al.²⁵ compared SVM and ANN for drug and nondrug classification abilities, and they concluded that SVM was superior to ANN in overall prediction accuracy. We investigated the usage of SVM and ANN for drug discovery too. In a previous work, we reported that an ANN approach²⁶ combined with the topological fragment spectra²⁷ (TFS) allowed us to successfully classify dopamine antagonists that interact with four different types of dopamine receptors, and it could be applied to the prediction of activity for class-unknown compounds. It was

* Corresponding author phone: +81-532-44-6878; fax: +81-532-44-6873; e-mail: taka@mis.tutkie.tut.ac.jp.

also shown that SVM works for this type of problem^{28–30} much better. However, in those cases, each compound belonged exactly to one single class, and no multilabeled compounds were included for the data sets. Multilabel classification^{31,32} is a more complex problem because each compound may be classified into multiple classes at a time or not classified into any of the classes.

In the present work, we have investigated the multilabel classification of a large number of drugs that have one or more of 100 different kinds of activity labels to explore a comprehensive model which can be used to predict individual biological activities and pleiotropic effects.

DATA SET AND METHODS

Data Set. In this work, we employed 98 634 compounds that belong to one or more of 100 different activity classes. All of the data were taken from the MDDR.³³ The MDDR is a structure database of investigative new drugs. It is an appropriate data source for the present work because the database includes a large number of compounds for which their biological activities are multiply labeled with different activities. The data sets used in the work were prepared as follows.

The MDDR database (release 2001.1) consists of 119 110 compounds, and each compound is labeled by one or more classes of 701 activities. First, we chose 108 591 compounds which are involved in the first 100 largest classes, in descending order of the number of entries of the database. All chemical structures of the compounds were desalted if they were salts, and the compounds that had more than 50 heavy atoms were removed. The purpose of this treatment is to remove large molecules such as natural products and large peptides. Those preparations resulted in a data set of 98 634 compounds.

The whole data set prepared above was divided into two groups, a modeling set and a test set. The modeling set consists of 88 770 compounds (90% of the total data), and the test set consists of 9864 compounds (10% of the total data). Compounds were randomly chosen for each set. Further, the modeling set was divided into two groups again, a training set and a validation set. The training set involves 59 180 compounds, and 29 590 compounds were used for the validation set. Members of these sets were randomly selected too. The test set was used only to examine the prediction ability in regard to class-unknown compounds.

Feature Representation. To describe structural features of a chemical compound, the TFS method²⁷ was used. The TFS of a chemical structure is derived from the enumeration of all possible substructures (or structural fragments) that have a specified number of bonds and a characterization of the fragments as a numerical quantity. We count the occurrence of individual fragments with the same characteristic value, and we sort them according to the characteristic value. Thus, the TFS is a digital spectrum, and it can be described as a multidimensional numerical pattern vector. In the present work, we enumerated possible structure fragments that have five or less connected bonds. All of the fragments were characterized by the sum of atomic mass numbers of the constituent atoms. In this manner, every compound was expressed by a 257-dimensional numerical pattern vector.

Support Vector Machine. SVM³⁴ has become popular as a powerful nonlinear classifier due to the introduction of a kernel trick in the past decade. Basically, SVM is a binary classifier with a maximum margin between the class boundaries. Suppose there is a data set represented by $(\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_n)$, where \mathbf{x}_i is a pattern vector and n is the number of patterns, and those patterns are linearly separable with class labels $y_i \in \{-1, 1\}$, $i = 1, \dots, n$. The discriminant function can be described as the following equation:

$$f(\mathbf{x}_i) = (\mathbf{w}^T \mathbf{x}_i) + b \quad (1)$$

where \mathbf{w} is a weight vector and b is a bias. When the equation is $f(\mathbf{x}_i) = 0$, it shows the discriminant surface. The discriminant surface with the maximum margin can be found by minimizing

$$\|\mathbf{w}\|^2 = \mathbf{w}^T \mathbf{w} = \sum_{i=1}^d w_i^2 \quad (2)$$

with constraints, $y_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1$ ($i = 1, \dots, n$).

This basic concept can be generalized to a linearly inseparable case by introducing Lagrangian multipliers and by using the concept of nonlinear mapping by a kernel function. It maps the input vectors \mathbf{x} into a higher dimensional feature space \mathbf{z} through some nonlinear mapping chosen a priori. In this space, an optimal discriminant hyperplane with a maximum margin is constructed. Finally, the discriminant function of interest can be described as eq 3:

$$f(\mathbf{x}) = \sum_{i=1}^n \alpha_i y_i K(\mathbf{x}, \mathbf{x}_i) + b \quad (0 \leq \alpha_i \leq C) \quad (3)$$

where α_i is a Lagrangian multiplier, y_i is a class label, b is a bias, and C is a regularization parameter. $K(\mathbf{x}, \mathbf{x}_i)$ is a kernel function. In the present work, we used a radial basis function (eq 4) as the kernel function for mapping the data into a higher-dimensional feature space.

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\frac{\|\mathbf{x} - \mathbf{x}'\|^2}{\sigma^2}\right) \quad (4)$$

For this case, it would be important to choose appropriate values for the two parameters, σ and C . Optimal values of the parameters were determined by a simple grid search technique.

The training of SVM is apparently time-consuming because it requires the solving of a very large quadratic programming (QP) optimization problem. The problem also involves a matrix that has a number of elements equal to the square of the number of training examples. Platt et al.³⁵ proposed a sequential minimal optimization (SMO) method, which is a simple and fast training algorithm for SVM and which is easy to implement. SMO breaks this large QP problem into a series of smallest possible QP problems in which only two examples are considered for each. Then, the QP problems can be solved analytically to avoid time-consuming numerical optimization in the inner loop. In addition, it allows us to handle a very large data set for the training without a large matrix. In this work, all of the computations for the SVM training were carried out using an in-house software tool which was implemented according to Platt's SMO algorithm and pseudocode reported in the literature.³⁵

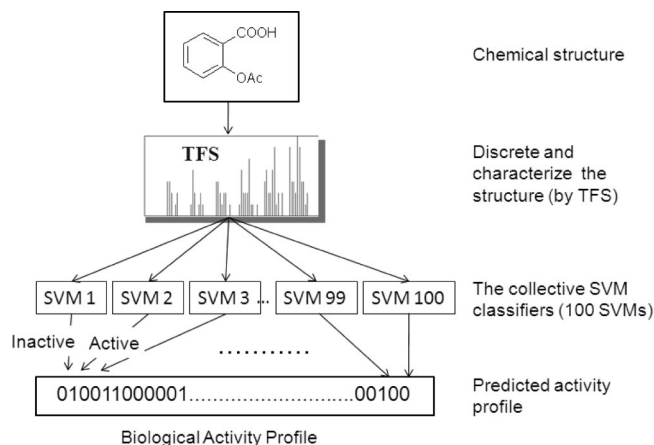


Figure 1. An illustrative schema of the predictive activity profiling of a drug. Every chemical structure is featured by the TFS method, and their biological activities are classified or predicted by the 100 collective SVM classifiers.

Multilabel Classification. As mentioned above, basically, SVM is a binary classifier. Multilabel classification³² is one of the multicategory classification problems. It can be accomplished by means of combinations of independent binary classifiers trained for individual classes of interest. There are two different types of the multicategory classification. One is for single-label classification in which each classifier learns a set of single-label data; then, the collective classifiers allow us to assign a testing sample to a single-label class. The other is for multilabel classification in which each classifier learns a set of multilabel data, and the collective classifiers allow us to assign a testing sample to two or more classes. The most common approach to multilabel classification is to independently train a binary classifier for each class (positive or negative) and then assign an object to all of the classes for which the corresponding classifier gives a “positive”. To perform this, we adopted the “one against the rest” method, which decomposes a multilabel problem into multiple binary classification problems. This approach requires a number of binary classifiers equal to the number of classes to be considered in the work. Thus, we developed 100 SVM classifiers for the present case. Each SVM was trained to separate the samples associated with a class of interest from the others. Figure 1 illustrates the basic concept of our present approach to predictive activity profiling of drugs.

RESULTS AND DISCUSSION

Determination of SVM Parameters. First, using the modeling set (training set and validation set), we tried to determine an appropriate value of σ in the Gaussian kernel function and that of the regularization parameter C for each class. SVM classifiers with different values for σ and C were investigated in a systematic manner with several different values. The appropriate values for the parameters were selected on the basis of sensitivity, which expresses a classification performance of a model. When the number of true positives (TP) and that of false negatives (FN) are obtained, the sensitivity is calculated by eq 5.

$$\text{sensitivity (\%)} = \frac{\text{TP}}{\text{TP} + \text{FN}} \times 100 \quad (5)$$

The sensitivity indicates a ratio of correctly classified positive examples to all of the positive examples.

As an example of systematic analysis for the appropriate values of σ and C in terms of the sensitivity, the result for the class of antineoplastics is shown in Figure 2. To prevent overfitting, appropriate parameter values were determined that maximize the sensitivity against the validation set. For this case, the highest sensitivity was obtained when the values of σ and C were 35 and 30, respectively. The specificities were 95% or more for most of the cases in the present data sets. In the same way, the appropriate values of the parameters were determined for every activity class, and they were used in the following analysis. Those values for individual classes are listed in Tables 1 and 2.

Results of Classification and Prediction. Individual SVM classifiers were trained with their appropriate parameters determined above. The details of classification results for individual data sets are shown in the Tables 1 and 2. Here, it should be noticed that there are two different types of activity class labels, the labels of molecular targets (or action mechanisms) and those of biological readout effects (or therapeutic treatments). They are listed in separate tables for convenience. Table 1 shows the results for the classes of molecular targets, and Table 2 shows those for the classes of biological readout effects. The specificity is defined as a ratio of correctly classified negative examples to all of the negative examples. The value is calculated using the number of true negatives (TN) and false positives (FP).

$$\text{specificity (\%)} = \frac{\text{TN}}{\text{TN} + \text{FP}} \times 100 \quad (6)$$

It is shown that the value of the specificity is considerably high compared to that of the sensitivity in every class. For the validation set, the specificity ranged from 92.8 to 100, and the total average of 100 classes was 99.2. It is considered that the remarkable specificity is due to there being so many negative examples for each classification. On the other hand, the sensitivity ranged from 60.9 to 99.7, and the total average was 82.7. The first five classes of compounds that have the highest sensitivity values are carbapenem (no. 10), cephalosporin (no. 3), quinolone (no. 14), renin inhibitor (no. 13), and H⁺/K⁺-ATPase inhibitor (no. 32). All five are classes

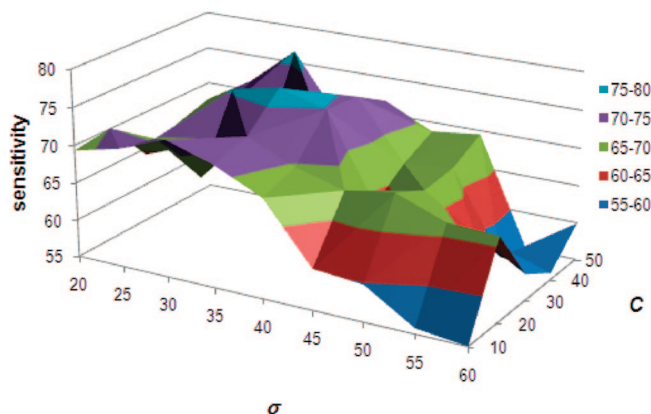


Figure 2. Result of the systematic analysis for the appropriate values of the SVM parameters, σ and C , for the class of antineoplastics.

Table 1. Classification Performances of the SVM Classifiers for Molecular Target Classes (Nos. 1–44)

No.	class label	parameter		training set				validation set				test set			
		σ	C	n^a	sens ^b	spec ^c	ppv ^d	n	sens	spec	ppv	n	sens	spec	ppv
1	lipoxygenase inhibitor	30	30	1634	99.8	99.7	90.6	803	85.3	98.9	68.6	266	78.2	99.0	68.4
2	calcium channel blocker	35	20	980	99.2	100	99.5	490	82.2	99.8	87.4	157	82.8	99.7	82.8
3	cephalosporin	60	20	859	99.9	100	100	430	99.1	100	99.5	143	100	100	98.6
4	PAF antagonist	40	30	858	99.8	100	99.9	442	90.0	99.6	75.5	147	89.8	99.6	77.6
5	NMDA receptor antagonist	25	10	850	100	100	99.4	437	83.5	99.7	82.4	147	77.6	99.7	82.0
6	ACAT inhibitor	45	20	832	99.6	100	99.8	414	90.8	99.9	89.7	140	88.6	99.8	86.1
7	phosphodiesterase IV inhibitor	55	40	819	98.5	99.9	94.5	400	86.3	99.6	72.9	131	89.3	99.4	68.0
8	substance P antagonist	50	50	767	99.7	100	100	377	92.3	99.9	91.1	123	87.8	99.9	89.3
9	gPIIb/IIIa receptor antagonist	40	20	752	99.3	100	97.8	380	87.4	99.6	74.8	128	86.7	99.6	75.0
10	carbapenem	55	10	768	99.9	100	98.8	384	99.7	99.9	95.3	127	99.2	100	96.9
11	angiotensin II blocker	60	10	747	97.2	99.8	83.2	373	92.0	99.5	68.6	127	89.8	99.4	66.3
12	leukotriene antagonist	35	40	740	100	100	99.6	366	83.3	99.8	87.4	113	81.4	99.8	86.0
13	renin inhibitor	60	20	404	99.3	100	95.7	201	95.5	99.9	88.9	80	95.0	100	95.0
14	quinolone	45	50	650	99.7	100	97.0	328	95.7	99.9	90.0	103	99.0	99.8	87.2
15	angiotensin II AT1 antagonist	60	50	619	99.5	99.9	87.3	314	86.0	99.5	66.7	103	79.6	99.5	61.7
16	thrombin inhibitor	50	30	558	99.5	100	98.9	278	90.6	99.5	64.9	97	91.8	99.5	66.4
17	HMG-CoA reductase (β) Inhibitor	50	30	615	99.8	100	99.5	303	93.4	99.9	93.7	109	92.7	100	98.1
18	farnesyl protein transferase inhibitor	45	40	562	99.6	100	96.6	293	86.0	99.7	76.4	89	89.9	99.7	72.1
19	tyrosine-specific protein kinase inhibitor	40	40	584	99.7	100	97.2	276	81.9	99.5	58.4	108	75.9	99.5	61.2
20	muscarinic (M1) agonist	30	30	585	99.3	100	98.8	294	89.8	99.7	76.3	95	85.3	99.7	76.4
21	steroid (5 α) reductase inhibitor	40	20	556	100	100	99.1	270	91.9	99.8	83.5	97	91.8	99.8	84.8
22	aldose reductase inhibitor	45	40	548	99.8	100	99.8	280	90.0	99.7	73.0	88	93.2	99.4	60.3
23	5 HT1A agonist	35	50	529	99.8	99.1	49.3	278	86.3	98.3	32.6	90	82.2	98.2	29.8
24	potassium channel activator	55	50	513	98.4	99.9	90.3	259	82.2	99.7	69.2	88	81.8	99.7	67.9
25	TNF inhibitor	35	30	489	99.2	100	97.8	240	85.8	99.3	49.4	78	84.6	99.2	46.2
26	thromboxane antagonist	35	40	505	100	100	99.8	249	94.8	99.8	83.1	82	89.0	99.9	84.9
27	HIV-1 protease inhibitor	55	40	397	99.5	100	93.6	198	90.9	99.7	67.7	61	95.1	99.7	64.4
28	5 HT3 antagonist	30	10	506	99.2	100	98.6	244	87.3	99.9	83.2	82	86.6	99.8	76.3
29	endothelin antagonist	55	20	423	98.8	100	98.1	211	86.7	99.8	78.9	70	85.7	99.9	87.0
30	phospholipase A2 inhibitor	35	30	431	99.3	100	100	209	80.4	99.8	72.1	64	79.7	99.7	67.1
31	leukotriene synthesis inhibitor	25	50	457	99.6	97.7	24.8	230	89.1	97.2	19.8	75	85.3	97.2	18.8
32	H ⁺ /K ⁺ -ATPase inhibitor	45	30	442	99.6	100	96.7	208	95.2	99.4	53.7	72	87.5	99.4	52.1
33	acetylcholinesterase inhibitor	30	30	449	100	100	99.8	203	80.8	99.9	86.3	68	79.4	99.9	88.5
34	antihistaminic	45	20	404	98.3	100	99.0	209	83.7	99.7	68.1	72	83.3	99.5	57.1
35	cyclooxygenase inhibitor	50	40	412	94.4	100	99.5	195	71.8	99.7	63.6	67	59.7	99.7	58.0
36	elastase inhibitor	40	40	360	99.7	100	100	182	92.3	99.8	70.9	63	87.3	99.8	75.3
37	protease inhibitor	55	50	302	97.4	99.6	54.7	148	77.0	99.2	32.2	40	85.0	99.2	29.3
38	calcium regulator	50	50	270	99.6	99.6	55.5	137	85.4	99.2	34.4	44	86.4	99.1	30.6
39	lipid peroxidation inhibitor	35	40	350	100	100	99.7	183	76.0	99.7	59.7	64	65.6	99.6	51.9
40	cyclooxygenase-2 inhibitor	60	40	371	97.3	100	98.4	177	90.4	99.8	70.2	60	93.3	99.6	61.5
41	leukotriene D4 antagonist	60	50	362	99.2	99.8	79.1	183	91.3	99.5	52.0	58	93.1	99.4	47.8
42	5 HT1D agonist	45	20	355	98.9	99.9	86.5	176	93.8	99.6	60.0	58	93.1	99.6	57.4
43	reverse transcriptase inhibitor	40	20	320	97.2	100	97.5	158	80.4	99.7	63.2	61	72.1	99.8	69.8
44	dopamine (D4) antagonist	55	50	352	99.4	99.7	66.8	169	88.2	99.2	38.2	53	90.6	99.2	39.3
	1–44 average (molecular targets)				99.2	99.9	91.3		87.8	99.6	69.8		86.2	99.5	68.2

^a The number of compounds. ^b Sensitivity. ^c Specificity. ^d Positive predictive value.

that have any labels involved with action mechanisms. For these five classes, the present SVM classifiers classified 95% or more of the compounds of individual classes into their own correct classes. For all 44 classes labeled by action mechanisms, an average value of the sensitivities was 87.8% (see Table 1).

On the other hand, the worst five classes for sensitivity were neuronal injury inhibitor (no. 53), antiparkinsonian (no. 79), immunosuppressant (no. 69), agent for restenosis (no. 89), and anti-inflammatory, topical (no. 96). These are all therapeutic treatment classes. The sensitivities for these five classes were 69.5, 69.4, 67.4, 62.7, and 60.9, respectively. The average value of the sensitivities for 56 therapeutic treatment classes was 78.6% (see Table 2). The difference between these average values of sensitivity was 9.2. The results suggest that the activity classes labeled in terms of action mechanism tend to give us better classification compared to the classes labeled in terms of

therapeutic treatment. Generally, it is considered that there are several mechanisms for a particular affection. For example, we have ACE inhibitors, Ca channel blockers, and renin inhibitors as antihypertensives for hypertension disease. It is expected that such a class with a label of affection or disease name would involve structurally diverse compounds. Therefore, it is reasonable that the activity classes labeled by biological readout effects tend to give lower sensitivities compared to the classes labeled by action mechanism or target enzyme.

The computational experiment with the modeling set (training set and validation set) showed that the TFS-based SVM may successfully classify many kinds of drugs into their own activity classes. Then, we examined the prediction ability of the SVM classifiers using the test set, which is regarded as a set of class-unknown compounds. The prediction results were also given in Tables 1 and 2. The total average of 100 classes is 80.8 for the sensitivity and

Table 2. Classification Performances of the SVM Classifiers for Therapeutic Action Classes (Nos. 45–100)

No.	class label	parameter		training set				validation set				test set			
		σ	C	n^a	sens ^b	spec ^c	ppv ^d	n	sens	spec	ppv	n	sens	spec	ppv
45	antineoplastic	35	30	6240	98.8	99.4	94.9	3121	76.7	96.4	71.3	1041	75.3	96.3	70.4
46	antihypertensive	40	30	5703	98.2	99.3	93.4	2851	81.9	97.8	79.7	949	82.5	97.7	79.3
47	antiallergic/asthmatic	25	20	4910	99.7	100	99.5	2452	78.2	97.8	76.6	818	75.6	97.6	74.2
48	cognition disorders, agent for	35	50	3653	99.2	99.7	95.8	1832	77.8	97.4	66.3	611	75.6	97.6	67.8
49	antiarthritic	35	40	3475	99.7	99.8	96.9	1733	77.7	97.4	65.2	578	75.3	97.2	62.8
50	hypolipidemic	35	20	2961	98.9	100	99.9	1475	83.8	99.3	86.1	494	79.4	99.2	83.9
51	anxiolytic	30	30	2994	99.6	100	99.7	1499	73.4	99.0	79.5	497	70.8	98.8	76.5
52	antiinflammatory	25	50	2727	99.8	99.5	90.9	1364	73.5	98.3	67.2	444	69.6	98.0	62.2
53	neuronal injury inhibitor	30	20	2638	97.9	100	99.6	1332	69.5	99.1	78.8	443	67.0	99.0	76.3
54	platelet antiaggregatory	45	40	2471	97.7	99.4	86.7	1245	80.6	98.0	63.7	408	78.7	97.8	60.8
55	antidepressant	25	50	2415	99.7	100	99.8	1186	73.3	98.9	74.0	402	70.9	98.7	69.3
56	antipsychotic	55	50	2389	95.0	98.7	76.0	1174	79.5	97.6	57.7	387	77.3	97.3	54.0
57	antiviral (AIDS)	25	50	1623	99.6	100	99.8	813	79.7	99.4	80.1	270	78.1	99.4	78.1
58	antiviral	40	30	1672	98.3	99.9	96.1	830	77.3	99.0	69.1	280	73.9	99.1	69.9
59	analgesic, nonopioid	25	50	1728	99.7	100	99.4	858	71.6	99.0	67.8	299	68.9	99.0	68.7
60	antibacterial	55	40	1567	95.7	99.9	95.2	782	85.5	99.5	81.6	261	82.4	99.5	81.4
61	antidiabetic	25	30	1574	99.8	99.3	80.5	788	84.8	97.3	46.4	263	83.3	97.5	48.0
62	antianginal	30	40	1572	99.3	99.5	84.4	790	75.3	97.8	48.2	265	76.6	97.8	49.3
63	antifungal	45	40	1056	98.5	99.6	81.9	527	85.4	99.1	63.1	167	80.8	99.1	60.8
64	anticonvulsant	35	20	1464	97.9	100	99.4	762	73.1	99.2	71.8	261	68.2	99.1	68.2
65	cardiotonic	25	50	1386	99.5	100	99.9	688	70.3	99.4	72.7	223	70.9	99.3	69.9
66	bronchodilator	45	50	1318	99.5	93.3	25.4	648	88.6	92.8	21.5	212	88.2	92.4	20.4
67	antiarrhythmic	30	20	1291	98.9	100	99.6	652	79.0	99.2	70.1	226	80.5	99.4	76.2
68	antipsoriatic	30	10	1143	95.5	100	97.8	567	71.6	99.7	80.2	195	70.3	99.6	78.3
69	immunosuppressant	50	40	703	92.9	100	97.6	347	67.4	99.2	50.9	121	62.8	99.2	48.7
70	anticoagulant	50	30	952	97.8	99.8	89.7	475	80.8	99.5	73.6	155	81.9	99.5	70.6
71	antiulcerative	40	50	1097	99.0	99.9	93.1	541	74.7	99.2	64.7	174	77.6	99.2	63.1
72	antisecretory, gastric	25	30	983	99.9	100	99.7	489	83.0	99.8	89.4	161	82.0	99.8	86.8
73	treatment for osteoporosis	55	50	886	97.5	99.4	70.5	445	80.0	98.2	40.4	145	81.4	98.4	42.4
74	antiischemic, cerebral	30	20	927	99.5	100	99.4	465	72.7	99.5	68.3	151	70.9	99.5	68.2
75	prostate disorders, agent for	40	20	869	99.3	100	99.8	432	84.0	99.8	85.0	148	81.1	99.7	82.8
76	antimigraine	45	50	837	99.3	99.9	94.4	411	81.5	99.2	60.5	137	87.6	99.3	63.8
77	antiacne	30	30	753	99.6	100	99.3	377	84.4	99.5	69.4	127	80.3	99.5	67.5
78	antibiotic	30	30	476	99.0	100	98.9	227	74.4	99.4	50.6	75	76.0	99.5	51.8
79	antiparkinsonian	35	20	709	98.2	100	99.7	356	69.4	99.6	67.3	115	70.4	99.7	73.6
80	septic shock, treatment for	60	50	614	92.8	99.4	61.3	302	74.2	98.3	31.0	99	79.8	98.2	30.7
81	antiobesity	30	50	621	99.8	100	99.4	310	79.7	99.6	69.8	101	79.2	99.7	72.7
82	antidiabetic, symptomatic	30	40	662	99.7	100	100	326	83.7	99.7	74.6	109	81.7	99.6	69.5
83	antiglaucoma	35	30	650	99.9	100	99.2	332	85.2	99.5	65.8	118	80.5	99.4	62.9
84	antiemetic	35	10	660	98.6	100	96.0	316	84.2	99.6	71.7	104	81.7	99.6	68.5
85	antineoplastic antibiotic	45	30	530	99.4	99.8	83.8	259	86.5	99.6	64.7	90	81.1	99.7	68.2
86	analgesic, opioid	35	50	636	99.5	100	99.8	317	83.3	99.8	81.2	106	79.2	99.7	73.7
87	immunomodulator	35	40	596	99.7	99.9	94.7	317	71.0	99.0	42.6	97	63.9	98.9	35.6
88	vasodilator	25	50	530	99.3	100	99.8	276	74.6	99.5	58.4	83	79.5	99.5	58.4
89	restenosis, agent for	45	20	408	92.7	100	99.2	201	62.7	99.8	64.6	63	66.7	99.7	60.0
90	irritable bowel syndrome, agent for	60	50	453	99.3	93.1	10.1	227	85.5	93.1	8.7	75	81.3	93.5	8.7
91	urinary incontinence, agent for	50	40	441	98.2	100	97.3	225	82.2	99.5	53.5	74	77.0	99.3	46.0
92	antibiotic, macrolide	50	10	74	100	100	100	37	91.9	100	85.0	12	75.0	100	100
93	stimulant, peristaltic	60	40	398	98.0	99.8	76.8	200	87.0	99.5	56.1	62	79.0	99.4	44.5
94	diagnostic agent	35	50	247	99.6	99.9	87.2	124	83.1	99.7	53.9	41	85.4	99.6	50.0
95	antiangiogenic	30	50	325	99.4	100	100	162	70.4	99.5	44.5	60	65.0	99.4	41.1
96	antiinflammatory, topical	35	20	368	97.8	100	100	169	60.9	99.8	68.2	63	58.7	99.8	71.2
97	diuretic	60	40	329	97.6	99.3	44.8	165	84.2	98.9	29.2	52	78.8	98.8	25.5
98	pulmonary emphysema, agent for	45	40	345	98.8	100	100	168	89.9	99.9	81.2	62	83.9	99.9	83.9
99	immunostimulant	50	40	218	98.2	100	99.5	102	81.4	99.8	53.9	35	85.7	99.7	53.6
100	antimetabolite	55	30	338	93.2	100	90.8	167	82.0	99.7	59.6	49	73.5	99.6	50.7
45–100 average (therapeutic actions)					98.4	99.6	90.6		78.6	98.8	63.9		76.6	98.8	62.5
1–100 average					98.7	99.7	90.9		82.7	99.2	66.5		80.8	99.1	65.0

^a The number of compounds. ^b Sensitivity. ^c Specificity. ^d Positive predictive value.

99.1 for the specificity. It is shown that both the sensitivity and the specificity are still high, even for the test set. For more details, the sensitivities of the classes carbapenem, cephalosporin, quinolone, renin inhibitor, and H⁺/K⁺-ATPase inhibitor, which had the top-five highest sensitivities in the training, are 99.2, 100, 99.0, 95.0, and 87.5 for the test set, respectively. The SVM classifiers still gave

us good performances in those predictions. On the other hand, those of the five classes (neuronal injury inhibitor, antiparkinsonian, immunosuppressant, restenosis, and anti-inflammatory) that resulted in the lowest sensitivities at the training are 67.0, 70.4, 62.8, 66.7, and 58.7, respectively. These results are also comparable to the results for the training.

Alternatively, we were concerned with the appearance of overprediction (especially for false-positive prediction) to each activity class because there are so fewer positive examples compared to the negatives. To contend with the issue, another statistical index was employed, which is referred to as the positive predictive value (PPV).³⁶ The PPV is defined as follows:

$$\text{PPV (\%)} = \frac{\text{TP}}{\text{TP} + \text{FP}} \times 100 \quad (7)$$

where TP is the number of true positives and FP is the number of false positives, again. It is clear that a smaller PPV shows more overprediction. The PPVs calculated for each class are listed in the Tables 1 and 2, too. The total average of 100 classes is 90.9 for the training set, 66.5 for the validation set, and 65.0 for the test set. This shows that the SVM classifiers do not give us serious overpredictions.

In addition, it might be noticed that the data sets used in the work are highly unbalanced because the number of positive instances is much smaller than that of negative instances in every case. Nevertheless, the present approach worked very well. In another work,²⁹ we investigated TFS-based machine learning using unbalanced data sets. The results suggested that training by TFS-based ANN strongly depends on the size of the data set for each class. ANN gave us better predictions, even for the test set, when a class had more instances in the training, but it gave poor results for a class with a small number of instances. And, it was shown that TFS-based SVM worked much better and more successfully compared to ANN, even in the case of an unbalanced data set. The results of the present work again show us the usage of the TFS-based SVM approach in such a case.

Predictive Activity Profiling of Multilabel Compounds.

Next, our interest was whether the SVM classifiers could identify multilabel compounds. A simple statistical analysis of the present data set showed us that more than half of the compounds (52 684 compounds) have multiple activity labels. The largest number of class labels for a single compound is eight in the present data set. The 100 classifiers obtained above can be applied to predictive profiling of biological activities, in which the individual classifiers give us a “positive” or “negative” to the corresponding activity to assess multilabel biological function. Table 3 summarizes the results described above from the viewpoint of multilabel classification.

Here, we defined a new index to assess performance for multilabel classification:

$$\text{ML-sensitivity (\%)} = \frac{m_t}{n} \times 100 \quad (8)$$

In eq 8, m_t is the number of compounds in which all of the class labels are correctly identified, and n is the number of compounds in the data set. The average ML-sensitivity for the training set was 98.1, as shown in Table 3. That for the validation set is 75.2. It is still good.

It should be noted that correct classification (or correct prediction) in the table refers to the instance where all of the class labels of a tested compound were correctly classified but allowing for overpredictions. For the validation set, for example, Table 3 shows that there are six compounds that have multilabels for seven different activities, and only two of them were correctly identified for all of the activity labels they had. The other four were not. Details of the results for these six compounds are shown in Table 4. The SVM classifiers correctly identified all of the seven activities for compounds 1 and 2, but overprediction of an activity resulted for compound 2. The classifiers also correctly identified six of their seven activities for compounds 3 and 4 as well. In both cases, one of the observed activities was missing in prediction and an alternative activity was overpredicted. Four of the seven activities for compound 5 were correctly predicted; three of them were missing in prediction, and another kind of activity was overpredicted. For compound 6, nothing was correctly predicted, and no alternative activities were overpredicted. It may be considered that compound 6 was regarded as noise because of there being few similar compounds in the TFS data space.

Then, we assessed performance on the test set, as well. It was shown that the classifiers had an ML-sensitivity of 73.1 on average (Table 3). Figure 3 shows a correlation plot between the number of activity labels annotated in MDDR and the number of activity labels predicted by the SVMs for the test set. The horizontal axis shows the number of activities annotated in MDDR, and the vertical axis shows the average value of number of activities predicted. For example, it is shown that the single-labeled compounds in MDDR were predicted in 1.7 activity classes for each, on average. The plot of Figure 3 shows that there is no serious overprediction for the test set. From these results, it is considered that the collective SVM classifiers can be used to profile multiple biological activities of the drugs.

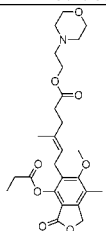
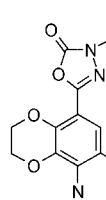
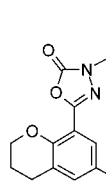
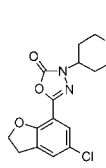
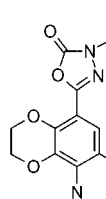
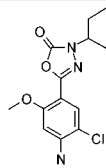
Consideration on Structure Diversity and Scaffold Dependency. In chemical application of supervised learning approaches, the quality of a classifier obtained highly depends on the data set of chemical compounds for the training. It is

Table 3. Details of Classification and Prediction Results for the Multilabel Compounds

number of labels	training set	validation set	test set
1	27200/27549 (98.7%) ^a	10979/13810 (79.5%)	3555/4591 (77.4%)
2	21644/22081 (98.0%)	8314/11302 (75.4%)	2729/3703 (73.7%)
3	6636/6862 (96.7%)	2188/3449 (63.4%)	683/1153 (59.2%)
4	1917/2014 (95.2%)	576/983 (58.6%)	193/319 (60.5%)
5	497/522 (95.2%)	149/240 (62.1%)	40/76 (52.6%)
6	134/137 (97.8%)	41/70 (58.6%)	12/20 (60.0%)
7	14/14 (100%)	2/6 (33.3%)	1/2 (50.0%)
8	0/1 (0%)		
total	58042/59180 (98.1%)	22249/29590 (75.2%)	7213/9864 (73.1%)

^a The numerator is the number of compounds for which every MDDR annotation label was correctly identified (or predicted), and the denominator is the number of compounds in the data set. The figure in the parentheses is the value of the ML-sensitivity.

Table 4. Prediction Results for Six Multilabel Compounds That Have Seven Different Activities for Each

No.	Structure	Observed activity	Predicted activity
1		Antineoplastic Antiarthritic Antiinflammatory Antiviral Antipsoriatic Immunosuppressant Immunomodulator	Antineoplastic Antiarthritic Antiinflammatory Antiviral Antipsoriatic Immunosuppressant Immunomodulator
2		Antihypertensive Cognition Disorders, Agent for Anxiolytic Antidepressant Antiarrhythmic Irritable Bowel Syndrome, Agent for Urinary Incontinence, Agent for	Antihypertensive Cognition Disorders, Agent for Anxiolytic Antidepressant Antiarrhythmic Irritable Bowel Syndrome, Agent for Urinary Incontinence, Agent for <u>Stimulant, Peristaltic^{a)}</u>
3		<u>Antihypertensive</u> Cognition Disorders, Agent for Anxiolytic Antidepressant Antiarrhythmic Irritable Bowel Syndrome, Agent for Urinary Incontinence, Agent for	Cognition Disorders, Agent for Anxiolytic Antidepressant Antiarrhythmic Irritable Bowel Syndrome, Agent for Urinary Incontinence, Agent for <u>Stimulant, Peristaltic</u>
4		<u>Antihypertensive</u> Cognition Disorders, Agent for Anxiolytic Antidepressant Antiarrhythmic Irritable Bowel Syndrome, Agent for Urinary Incontinence, Agent for	Cognition Disorders, Agent for Anxiolytic Antidepressant Antiarrhythmic Irritable Bowel Syndrome, Agent for Urinary Incontinence, Agent for <u>Antianginal</u>
5		<u>Antihypertensive</u> <u>Anxiolytic</u> <u>Antidepressant</u> Cognition Disorders, Agent for Antiarrhythmic Irritable Bowel Syndrome, Agent for Urinary Incontinence, Agent for	Cognition Disorders, Agent for Antiarrhythmic Irritable Bowel Syndrome, Agent for Urinary Incontinence, Agent for <u>Stimulant, Peristaltic</u>
6		<u>Antihypertensive</u> <u>Cognition Disorders, Agent for</u> <u>Anxiolytic</u> <u>Antidepressant</u> <u>Antiarrhythmic</u> <u>Irritable Bowel Syndrome, Agent for</u> <u>Urinary Incontinence, Agent for</u>	No activity was predicted.

^a Missing activities and unanticipated activities were underlined.

considered that, if the whole data set includes a large number of very similar compounds that share particular molecular scaffolds, it makes it easy to get good predictions.

We investigated the molecular scaffolds embedded in the present data sets by means of nonterminal vertex graph (NTG)³⁷ analysis. A NTG is defined as a vertex graph which does not have any terminal vertex nor any isolated vertex. It is quite similar to a Murko scaffold³⁸ but is a more general graph expression. The NTG was described fully elsewhere. Here, four different types of expressions (simple graph (SG), edge-weighted graph (EG), vertex-weighted graph (VG), and chemical graph (CG)) defined in our work were examined against the present data sets. The results of the NTG analyses are shown in Table 5. For the whole data set (98 634 compounds), 38 492 different NTGs in the CG expression were found. A simple ratio of the total number of compounds

to the NTGs is 2.5. In other words, there are less than three compounds that share a NTG/CG on average. For the test set, the ratio is 1.3. It is concluded that the present data set maintains structural diversity.

Furthermore, we examined how many scaffolds (NTG/CGs) appeared in every data set (in the test set, too). It was found that the data sets share 3389 unique NTG/CGs, and 5586 compounds in the test set (9864 compounds) have any of the NTG/CGs. Conversely, 4278 compounds in the test set have none of them. To assess the scaffold dependency in the predictions, the sensitivity, specificity, and PPV were calculated for these two groups of the test set separately. The results are summarized in Table 6. The sensitivity and the PPV for the group of NTG-shared compounds are 86.6 and 69.8, respectively. On the other hand, those for the NTG-unshared group are 73.0 and 58.7, respectively. It is clear

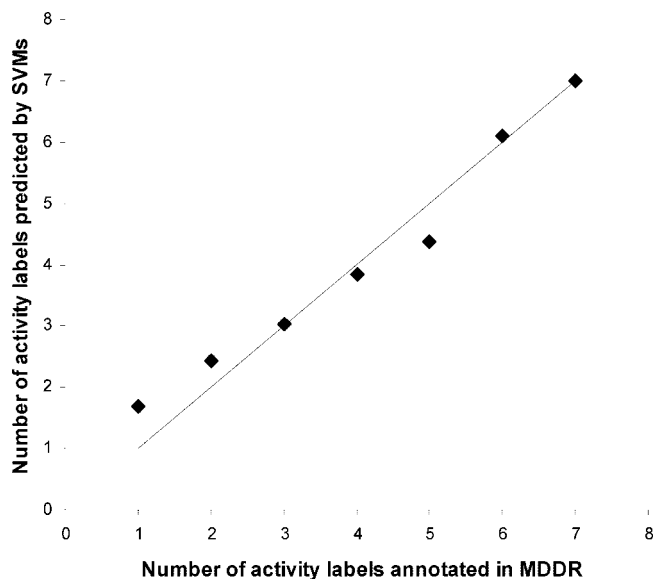


Figure 3. Correlation plot between the number of activities annotated in the MDDR and that obtained by the multilabel prediction for the test set. The horizontal axis shows the number of activities annotated in MDDR, and the vertical axis shows the average value of the number of activity labels predicted by the SVMs. The test set includes from single-label compounds to eight-label compounds (see Table 3). It is shown that there is no serious overprediction in the case.

Table 5. The Results of NTG Analysis against the Individual Data Sets Used in the Work^a

	the whole set	training set	validation set	test set
compounds	98634	59180	29590	9864
NTG/SG	14855	11454	8004	4018
NTG/EG	24142	18199	12345	5797
NTG/VG	34780	25326	16318	7162
NTG/CG	38492	27751	17665	7598

^a The table shows the number of unique NTGs in each graph expression. SG, EG, VG, and CG are simple graph, edge-weighted graph, vertex-weighted graph, and chemical graph (edge and vertex weighted graph), respectively.

Table 6. Results of Multilabel Prediction in the NTG-Shared Compounds and the NTG-Unshared Compounds

	test set		
	all	NTG-shared	NTG-unshared
compounds	9864	5586	4278
sensitivity (%)	80.8	86.6	73.0
specificity (%)	99.1	99.2	99.0
PPV (%)	65.0	69.8	58.7

that the prediction results for the NTG-unshared compounds were less than those for the NTG-shared compounds. But, we think that the results are still good for a large number of activity classes.

Alternatively, Figure 4 shows a comparison of the results of multilabel prediction for the NTG-shared compounds and NTG-unshared compounds by means of a correlation plot similar to that in Figure 3. It is clear that there is no serious overprediction even for the group of NTG-unshared compounds. It seems that the predictions for the NTG-unshared compounds are, rather, providing slight underpredictions, and the difference between the two groups becomes large as the number of activity labels increases. The results show that

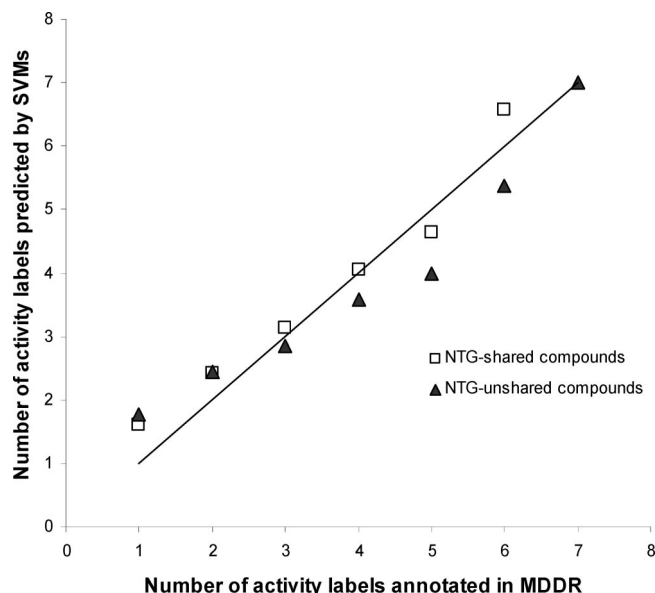


Figure 4. Correlation plot between the number of activities annotated in the MDDR and that obtained by the multilabel prediction for the test set. The figure shows the comparison of the NTG-shared compounds and the NTG-unshared compounds at the multilabel prediction of Figure 3. It is clear that there is no serious overprediction even for the NTG-unshared compounds.

there apparently exists a shared scaffold dependency in the classification and prediction of the drugs. However, it may also be concluded that the present approach to predictive activity profiling of drugs is useful even for NTG-unshared compounds because both the sensitivity and PPV are still good, and they are acceptable in use.

CONCLUSIONS

In the present work, we have developed the collective SVM classifiers for 100 different types of activity to drug molecules. Structural feature representation of each drug molecule was based on the TFS. The TFS-based SVM classifiers developed using 59 180 compounds for the training and 29 590 compounds for the validation could correctly predict 80% of the biological activities for the test set of 9864 compounds. The SVM classifiers also performed well for predictive activity profiling of the multilabel compounds. We believe that the present model can be used to predict biological activities of unknown compounds or to alert to adverse effects of drug candidates.

ACKNOWLEDGMENT

We thank Mr. Kazusa Noto for his assistance in NTG analysis. We also thank the reviewers for their helpful comments to improve the manuscript. This research was partially supported by the Ministry of Education, Culture, Sports, Science and Technology of Japan, Grant-in-Aid for Scientific Research (A) 14208032.

REFERENCES AND NOTES

- (1) FDA's Drug Safety Initiative. <http://www.fda.gov/cder/drugSafety.htm> (accessed Feb 19, 2008).
- (2) *ADMET Descriptors in Discovery Studio*; Accelrys: San Diego, CA.
- (3) *DEREK for Windows*; Lhasa limited: Leeds, West Yorkshire, U.K.
- (4) Mohan, C. G.; Gandhi, T.; Garg, D.; Shinde, R. Computer-Assisted Methods in Chemical Toxicity Prediction. *Mini Rev. Med. Chem.* **2007**, *7*, 499–507.

- (5) van de Waterbeemd, H.; Gifford, E. ADMET in Silico Modelling: Towards Prediction Paradise. *Nat. Rev. Drug Discovery* **2003**, *2*, 192–204.
- (6) Lewis, D. F. V. Computer-Assisted methods in the evaluation of chemical toxicity. In *Reviews in Computational Chemistry*; Lipkowitz, K. B., Boyd, D. B., Eds.; John Wiley & Sons: Hoboken, NJ, 2007; Vol. 3, pp 173–222.
- (7) Yoshida, K.; Niwa, T. Quantitative Structure-Activity Relationship Studies on Inhibition of HERG Potassium Channels. *J. Chem. Inf. Model.* **2006**, *46*, 1371–1378.
- (8) Snyder, R. D.; Smith, M. D. Computational Prediction of Genotoxicity: Room for Improvement. *Drug Discovery Today* **2005**, *10*, 1119–1124.
- (9) Dearden, J. C. In Silico Prediction of Drug Toxicity. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 119–127.
- (10) Cheng, A.; Dixon, S. L. In Silico Models for the Prediction of Dose-dependent Human Hepatotoxicity. *J. Comput.-Aided Mol. Des.* **2003**, *17*, 811–823.
- (11) Ekins, S.; Bravi, G.; Binkley, S.; Gillespie, J. S.; Ring, B. J.; Wikel, J. H.; Wrighton, S. A. Three- and Four-dimensional Quantitative Structure Activity Relationship Analyses of Cytochrome P-450 3A4 Inhibitors. *J. Pharmacol. Exp. Ther.* **1999**, *290*, 429–438.
- (12) Farmer, J. A. Pleiotropic Effects of Statins. *Curr. Atheroscler. Rep.* **2000**, *2*, 208–217.
- (13) Davignon, J. Beneficial Cardiovascular Pleiotropic Effects of Statins. *Circulation* **2004**, *109*, 39–43.
- (14) Xie, L.; Wang, J.; Bourne, P. E. In Silico Elucidation of the Molecular Mechanism Defining the Adverse Effect of Selective Estrogen Receptor Modulators. *PLoS Comput. Biol.* **2007**, *3*, e217.
- (15) Bender, A.; Scheiber, J.; Glick, M.; Davies, J. W.; Azzaoui, K.; Hamon, J.; Urban, L.; Whitebread, S.; Jenkins, J. L. Analysis of Pharmacology Data and the Prediction of Adverse Drug Reactions and Off-Target Effects from Chemical Structure. *ChemMedChem* **2007**, *2*, 861–873.
- (16) Chen, Y. Z.; Zhi, D. G. Ligand-Protein Inverse Docking and Its Potential Use in the Computer Search of Protein Targets of a Small Molecule. *Proteins* **2001**, *43*, 217–226.
- (17) Nettles, J. H.; Jenkins, J. L.; Bender, A.; Deng, Z.; Davies, J. W.; Glick, M. Bridging Chemical and Biological Space: “Target Fishing” Using 2D and 3D Molecular Descriptors. *J. Med. Chem.* **2006**, *49*, 6802–6810.
- (18) Zhou, Y.; Zhou, B.; Chen, K.; Yan, S. F.; King, F. J.; Jiang, S.; Winzler, E. A. Large-Scale Annotation of Small-Molecule Libraries Using Public Databases. *J. Chem. Inf. Model.* **2007**, *47*, 1386–1394.
- (19) Poroikov, V. V.; Filimonov, D. A.; Ihlenfeldt, W. D.; Glorizova, T. A.; Lagunin, A. A.; Borodina, Y. V.; Stepanchikova, A. V.; Nicklaus, M. C. PASS Biological Activity Spectrum Predictions in the Enhanced Open NCI Database Browser. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 228–236.
- (20) Anzail, S.; Barnickel, G.; Cezanne, B.; Krug, M.; Filimonov, D.; Poroikov, V. Discriminating between Drugs and Nondrugs by Prediction of Activity Spectra for Substances (PASS). *J. Med. Chem.* **2001**, *44*, 2432–2437.
- (21) Filimonov, D.; Poroikov, V.; Borodina, Y.; Glorizova, T. Chemical Similarity Assessment through Multilevel Neighborhoods of Atoms: Definition and Comparison with the Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 666–670.
- (22) Poroikov, V. V.; Filimonov, D. A.; Borodina, Y. V.; Lagunin, A. A.; Kos, A. Robustness of Biological Activity Spectra Predicting by Computer Program PASS for Noncongeneric Sets of Chemical Compounds. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1349–1355.
- (23) Nidhi; Glick, M.; Davies, J. W.; Jenkins, J. L. Prediction of Biological Targets for Compounds Using Multiple-Category Bayesian Models Trained on Chemogenomics Databases. *J. Chem. Inf. Model.* **2006**, *46*, 1124–1133.
- (24) Lepp, Z.; Kinoshita, T.; Chuman, H. Screening for New Antidepressant Leads of Multiple Activities by Support Vector Machines. *J. Chem. Inf. Model.* **2006**, *46*, 158–167.
- (25) 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*, 1882–1889.
- (26) Fujishima, S.; Takahashi, Y. Classification of Dopamine Antagonists Using TFS-Based Artificial Neural Network. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1006–1009.
- (27) Takahashi, Y.; Ohoka, H.; Ishiyama, Y. Structural Similarity Analysis Based on Topological Fragment Spectra. In *Advances in Molecular Similarity*; Carbo, R., Mezey, P., Eds.; JAI Press: Greenwich, CT, 1998; Vol. 2, pp 93–104.
- (28) Takahashi, Y.; Nishikoori, K.; Fujishima, S. Classification of Pharmacological Activity of Drugs Using Support Vector Machine. In *Active Mining, Second International Workshop, Lecture Notes in Computer Science*; Tsumoto, S., Yamaguchi, T., Numao, M., Motoda, H., Eds.; Springer: Berlin, 2005; Vol. 3430, pp 303–311.
- (29) Takahashi, Y.; Fujishima, S.; Nishikoori, K.; Kato, H.; Okada, T. Identification of Dopamine D1 Receptor Agonists and Antagonists Under Existing Noise Compounds by TFS-based ANN and SVM. *J. Comput. Chem. Jpn.* **2005**, *4*, 43–48.
- (30) Fujishima, S.; Takahashi, Y.; Nishikoori, K.; Kato, H.; Okada, T. Extended Study of the Classification of Dopamine Receptor Agonists and Antagonists using a TFS-based Support Vector Machine. *New Gener. Comput.* **2007**, *25*, 203–212.
- (31) Hristozov, D.; Gasteiger, J.; Costa, F. B. D. Multilabeled Classification Approach To Find a Plant Source for Terpenoids. *J. Chem. Inf. Model.* **2008**, *48*, 56–67.
- (32) Tsoumakas, G.; Katakis, I. Multi-Label Classification: An Overview. *Int. J. Data Warehousing Mining* **2007**, *3*, 1–13.
- (33) MDL Drug Data Report, version 2001.1; MDL Information Systems, Inc.: Santa Clara, CA, 2001.
- (34) Vapnik, V. N. *The Nature of Statistical Learning Theory*; Springer: New York, 1995.
- (35) Platt, J. C. *Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines*; Technical Report MSR-TR-98-14, Microsoft Research: Redmond, WA, 1998.
- (36) Duffy, S. W.; Chen, H. H.; Tabar, L.; Fagerberg, G.; Paci, E. Sojourn Time, Sensitivity and Positive Predictive Value of Mammography Screening for Breast Cancer in Women Aged 40–49. *Int. J. Epidemiol.* **1996**, *25*, 1139–1145.
- (37) Takahashi, Y. Chemical data mining based on non-terminal vertex graph. In *Systems, Man and Cybernetics*, 2004 IEEE International Conference; IEEE: Piscataway, NJ, 2004; Vol. 5, pp 4583–4587.
- (38) Bemis, G. W.; Murcko, M. A. The Properties of Known Drugs. 1. Molecular Frameworks. *J. Med. Chem.* **1996**, *39*, 2887–2893.

CI7004753