Drug Discovery Using Support Vector Machines. The Case Studies of Drug-likeness, Agrochemical-likeness, and Enzyme Inhibition Predictions

Vladimir V. Zernov,[†] Konstantin V. Balakin,[‡] Andrey A. Ivaschenko,[‡] Nikolay P. Savchuk,[‡] and Igor V. Pletnev*,[†]

Department of Chemistry, Lomonosov Moscow State University, Leninskie Gory 1, 119899 GSP-3 Moscow, Russia, and Chemical Diversity Labs, Inc., 11575 Sorrento Valley Road, San Diego, California 92121

Received May 8, 2003

Support Vector Machines (SVM) is a powerful classification and regression tool that is becoming increasingly popular in various machine learning applications. We tested the ability of SVM, in comparison with well-known neural network techniques, to predict drug-likeness and agrochemical-likeness for large compound collections. For both kinds of data, SVM outperforms various neural networks using the same set of descriptors. We also used SVM for estimating the activity of Carbonic Anhydrase II (CA II) enzyme inhibitors and found that the prediction quality of our SVM model is better than that reported earlier for conventional QSAR. Model characteristics and data set features were studied in detail.

INTRODUCTION

Computer-aided screening of new drugs relies heavily upon various filters aimed at retaining promising, drug-like, compounds while throwing away those unlikely to be drugs. During stepwise filtering, the complexity and specificity of filters gradually increase: from simple rules similar to Lipinsky's rule[1,2] to sophisticated QSAR models. As a computational technique behind the most sophisticated filters, artificial neural networks (ANN) are becoming the de facto standard. In general, ANN are a relatively easy to use, powerful, and versatile tool, but there are some drawbacks associated with this prediction method. Among them are (i) the "black-box" character of ANN, which may hamper the interpretation of derived models and fine-tuning; (ii) the risk of overfitting (i.e., ability to fit to training data noise rather than to true data structure, resulting in poor generalization); and (iii) a relatively long training time.

Recently, a relatively novel method has become popular in the machine learning community, which seems to be at least as powerful and versatile as ANNs. These are the so-called Support Vector Machines (originally proposed and developed by Vladimir Vapnik[3]), which exist in classification and regression versions. SVM applications are being actively pursued in various areas, from genomics to face recognition.[4−6]

The first brief reports on the application of SVMs to drug design problems are quite promising. Burbidge et al.[7] compared the performance of SVMs, ANNs, and C5.0 decision trees for predicting the inhibition of dihydrofolate reductase by 55 substituted pyrimidines. SVM classifiers demonstrated the best prediction rating and were also much less time-consuming than ANNs. A subsequent paper[8] reported similar results for blood-brain barrier (BBB) permeability predictions (learning set of 172 compounds; overall prediction quality on SVM slightly outperformed ANNs). Another type of SVM application was reported by Warmuth and co-workers,[9] who performed the selection of "actives" in large data sets. The authors demonstrated that on two data sets provided by DuPont Pharmaceuticals to predict the permeability of compounds, selection strategy based on SVM performs efficiently and is much better than random selection.

In the research presented here, we applied SVM to real-life large-scale drug discovery problems, specifically, the creation of drug- and agro-likeness filters[10] for screening large compound collections. One particular objective was to compare the performance of SVM and ANN classifiers on the same data. Additionally, we tested SVM in QSAR-related analysis of Carbonic Anhydrase II inhibition data and compared our results with literature reports on more conventional techniques.

METHODOLOGY

There exist a number of excellent introductions into SVM, both printed[3,11,12] and electronically available.[13] For this reason we will only briefly summarize the main ideas and terms of SVM classification here.

From the practioner's viewpoint, a particularly important feature of SVM is that it explicitly relies on *Statistical Learning Theory*[3] and directly addresses the issue of avoiding overfitting. The key concept here is *Structural Risk Minimization* principle (SRM)[14] proposed by Vapnik and Chervonenkis in the early 1970s.

Suppose we have a set of m training data points $\{(\mathbf{x}_1,y_1)...(\mathbf{x}_m,y_m)\}$ where $\mathbf{x}$ are features (descriptors; $\mathbf{X}$ is called input space) and y_m is a class label, typically, -1 and 1 in binary classification tasks. Suppose also that there exists an

* Corresponding author e-mail: pletnev@analyt.chem.msu.ru.
[†] Lomonosov Moscow State University.
[‡] Chemical Diversity Labs, Inc.

10.1021/ci0340916 CCC: $25.00
Published on Web 10/16/2003

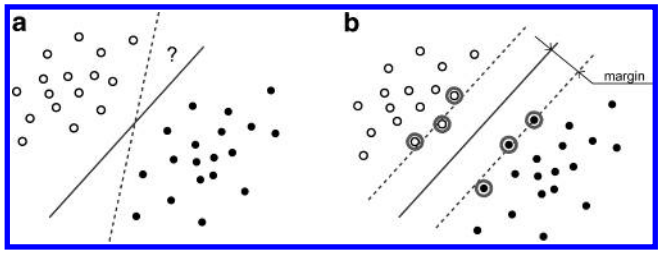


Figure 1. a. Two possible linear discriminant planes. b. Best plane maximizes the margin.

unknown probability distribution P(**x**,y), which describes a relation of features to classes. We attempt to associate the descriptors with classes by introducing prediction, or decision, function f(**x**,*a*), which value changes from −1 to +1, dependent on the class.

The decision function parameters *a* are to be found via minimization of the functional of expected error

$$I(a) = \int Q(x,a,y)P(x,y)dxdy$$

where $Q(x,a,y)$ is a loss function. For example, $Q = (y - f(x,a))^2$ corresponds to the common least-squares estimate. The obvious problem is that the integral depends on the unknown true distribution *P* defined for the whole input space, but all we have is a sampling from that distribution, our training set. So for practical purposes, the integral should be replaced with the sum over the training points only, *empirical risk*. However, there could be a number of different functions which all give a good classification. Our selection criterion among these functions is then that we should select the decision function that performs best not only at training set examples but also on previously unseen data, that is, the function with the best *generalization* ability. According to SRM, this may be achieved by minimizing both empirical risk and *confidence interval*. The latter term is proportional to the ratio of model complexity (measured by the so-called *Vapnik-Chervonenkis dimension*) to the number of training data points. Omitting formulas, SRM states that the optimal classifier is given by a tradeoff between reduction of the training error and limiting model complexity, that is limiting the chances of overfitting.

Consider an example of classification in two-dimensional input space (Figure 1). Given the depicted training set, both solid and dashed separation lines (Figure 1a) are acceptable; but which one is better? Intuitively, it is clear that the better-generalizing line is one less sensitive to small perturbations in position of data points, so it is the solid line in Figure 1a. In other words, the decision line must lie in some sense maximally far apart from the training points of different classes. This is exactly what follows from SRM application to the task and what constitutes the essence of SVM: the optimal classifier is the one providing the largest *margin* separating the classes (margin is defined as the sum of shortest distances from decision line to the closest points of both classes, Figure 1b). Geometrically, the optimal line bisects the shortest line between the convex hulls of the two classes. Notably, it appears that a relatively small number of data points which are closest to the line (lie on the margin; so-called *support vectors, SV*) are completely enough to determine the position of optimal separation line (*optimal separation hyperplane, OSH*, for a high-dimension case).

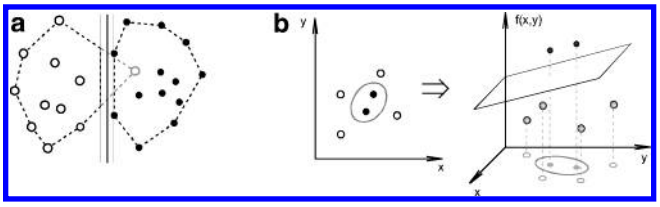


Figure 2. a. Explanation of the soft margin technique: best plane bisects the reduced convex hulls. b. Nonlinear mapping into higher dimensions: kernel application.

Both SVs and OSH can be found by solving a related quadratic programming problem. If the separating hyperplane is $Wx + b = 0$, which implies $y_i(Wx_i+b) \geq 1$, $i = 1..m$, the decision is found by minimization Euclidian norm $1/2||W||^2$:

$$W = \sum_{i=1}^{m} y_i \alpha_i \cdot x_i$$

Only if the corresponding Lagrange multipliers $\alpha_i > 0$, these x_i are support vector *x*. After minimization, the decision function is written as

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{m} y_i \alpha_i \cdot x \cdot x_i + b)$$

Note that only a limited subset of training points, named support vectors, contribute to the expression.

In a linearly inseparable case, where no error-less classification can be achieved by hyperplane, there still exist two ways to proceed with SVM.

The first one is to modify linear SVM formulation to allow misclassification. Mathematically, this is achieved by introducing classification-error (*slack*) variables $\xi_i > 0$ and minimizing the combined quantity

$$||W||^2 + C\sum_{i=1}^{m} \xi_i$$

under the separation constraints as $y_i(Wx_i+b) \geq 1 - \xi_i$, $i = 1..m$. Here the parameter *C* regulates a tradeoff between minimization of training error and maximization of margin. Such an approach known as *soft margin technique* is exemplified geometrically by Figure 2a.

Another way is *nonlinear SVM*, which has achieved a great deal of attention in the past decade. The most popular current approach is "transferring" data points from initial descriptor space to higher-dimensional space, which is derived by adding new degrees of freedom through nonlinear transformations of initial dimensions (Figure 2b). The hope is that problems that are nonlinear in original space may be linear in higher dimensions, so linear solution techniques become applicable.

Importantly, direct transfer of the points from original to higher-dimensional space is not necessary, as all SVM mathematics deals with dot products of variables $(\mathbf{x}_i,\mathbf{x}_j)$ rather than with variable values $\mathbf{x}_i,\mathbf{x}_j$ themselves. All that is necessary is to replace dot products $(\mathbf{x}_i,\mathbf{x}_j)$ with their higher-dimensional analogues, functions $K(\mathbf{x}_i,\mathbf{x}_j)$ expressed over *original* variables **x**. The suitable functions *K* are called *kernels*, and the whole approach is known as *kernel trick*.

Table 1. Descriptors Used for Building Drug- and Agro-likeness Estimation Models

no.	descriptor	description	software used for caclulation	drug-likeness model	agro-likeness model
1	MW	molecular weight	Chemosoft[18]	+	+
2	FA	fractional absorption	Chemosoft	+	+
3	LogD	log of 1-octanol/water partition coefficient at pH 7.4	Chemosoft	+	+
4	LogP	log of 1-octanol/water partition coefficient (neutral form)	Chemosoft	−	+
5	LogSw	log of water solubility (g/mL) at pH 7.4	Chemosoft	+	+
6	H_don	number of hydrogen bond donors	Chemosoft	+	+
7	H_acc	number of hydrogen bond acceptors	Chemosoft	+	+
8	B_rot	number of rotatable bonds	Chemosoft	+	+
9	RG	molecular radius of gyration	Cerius[2] [a]	−	+
10	AP	atomic polarizability	Cerius[2]	−	+
11	DM	dipole moment	Cerius[2]	−	+
12−21	JDD	set of Jurs descriptors	Cerius[2]	−	+

[a] Accelrys, Inc. 2000. URL: http://www.accelrys.com/.

Decision function in this case is written as

$$f(x) = \mathrm{sgn}\left(\sum_{i=1}^{m} y_i \alpha_i \cdot K(x, x_i) + b\right)$$

The most common kinds of kernels are

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i, \mathbf{x}_j + 1)^d - \text{Polynomial}$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = \exp(-r|x_i - \mathbf{x}_j|^2) - \text{Radial Basis Function (RBF)}$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = \mathrm{sigmoid}(\eta(\mathbf{x}_i \mathbf{x}_j) + a) - \text{Two-layer perceptron}$$

Finally, let us list the main SVM advantages below:[15]

1. We can build any complex classifier, and the solution is guaranteed to be the global optimum (no danger of getting stuck at local minima). This is a consequence of the quadratic programming approach and of the restriction of possible decision space.

2. There are few parameters to elucidate. Besides the main parameter *C*, only one additional parameter is needed to determine polynomial or RBF kernels, which typically (as can be judged from the literature) demonstrate high classification power.

3. The final results are stable, reproducible, and largely independent of the optimization algorithm. The absence of a random constituent in SVM scheme guarantees that two users which apply the same SVM model with the same parameters to the same data will receive identical results (which is often not true with ANNs).

DRUG-LIKENESS ESTIMATION

A drug-likeness model was profiled and validated by using available databases of drugs and pharmaceutical leads (15 000 molecules from the Ensemble database,[16] a licensed database of known pharmaceutical agents compiled from the patent and scientific literature) and 15 000 nondrugs. The active (drug-like) molecules were reported pharmaceutically active agents (compounds at the stages of (pre)clinical trials, launched drugs, or compounds with proven pharmaceutical activity). Care was taken to avoid over-representation of any single class of compounds with the main chemical classes of known drugs having a similar distribution.

Table 2. Subsets Used in Creating and Testing Drug-likeness Models

	train set	validation set	test set
total	15 000	7499	7500
drugs	7465	3755	3751

Inactive (nondrug-like) compounds were representatively selected from the Sigma-Aldrich catalog,[17] based on an assumption that the collection of compounds without a defined type of activity will show minimal drug-likeness. The nondrugs set was filtered to remove compounds with reactive functionalities and other unwanted substructures.

A list of calculated descriptors used for building drug-likeness models is presented in Table 1.

These descriptors were calculated using the ChemoSoft software.[18] The key feature of this descriptor set which distinguishes it from similar ones[19,20] is the presence of ADME-specific (ADME − absorption, distribution, metabolism, excretion) descriptors $\log D_{74}$, $\log S_w$, and FA (calculated by the SLIPPER program[21] integrated into ChemoSoft). We believe that ADME-related properties are highly significant in the context of pharmacokinetic characteristics of the drug candidates and should contribute significantly to the predictive power of the developed model.

The performance of the model can be enhanced by the use of additional 2D/3D descriptors that contain information about specific functional groups within the molecule. However, increasing the number of descriptors is impractical for very large compound sets. The results shown below demonstrate good predictive power of the model built using this minimal set of descriptors.

The whole set of all 30 000 compounds was divided into three parts (see Table 2): training set (for building drug likeness model by neural networks and SVM), validation set (for checking model quality while training neural networks, useful for avoidance of overfitting), and the test set (for checking prediction quality of the best models). The validation set was also used to check SVM models instead of leave-one-out cross-validation, as the latter is too slow for large data sets.

Before creating models, each descriptor was evaluated according to its capacity to separate drugs from nondrugs, as illustrated in Figure 3.

From these diagrams (Figure 3) one can easily observe the threshold value of the descriptor which provides the best

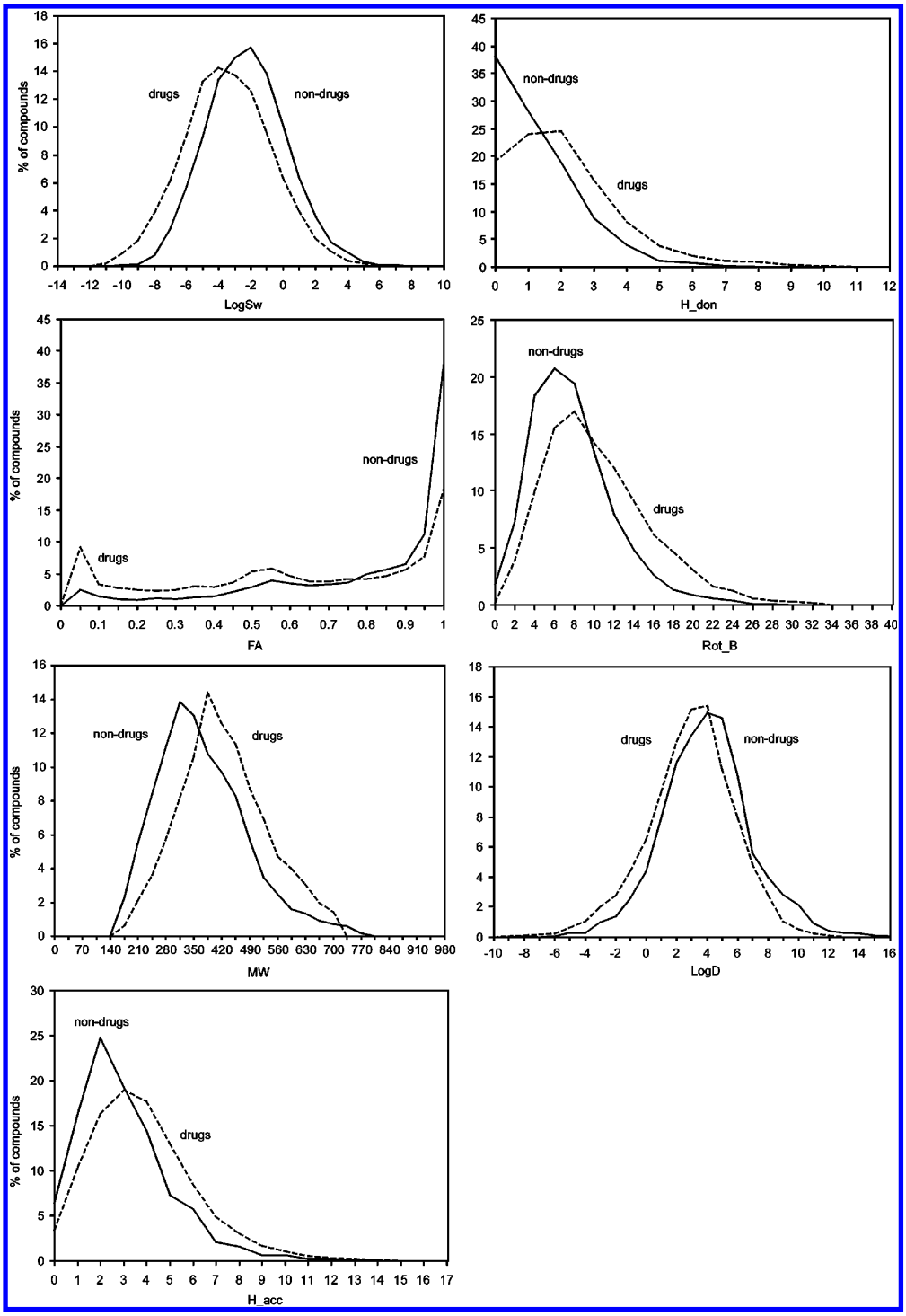


Figure 3. Distribution diagrams of descriptor values for examining the quality of separation between drugs and nondrugs.

separation. The proximity of these thresholds to optimal is confirmed by one-descriptor SVM classifiers that are presented in Table 3.

All seven descriptors examined above were used to build drug-likeness models. SVM classifiers were based on linear or nonlinear (Radial Basis Functions, RBF) kernel. Before modeling each descriptor was scaled to [−1;1] range (by training set; scaled values for other subsets were derived using train set scaling factors). The LibSVM[22] and SVM-light[23] programs were used.

Besides SVM, several ANN models were built with NeuroSolution software[24] (feed-forward networks that consist of input neurons, one hidden layer, and two output neurons). The networks were trained with the molecular descriptors as input values and the drug-likeness scores as output. The final score was calculated by subtracting the "activity" from the "inactivity" value. The back-propagation networks were trained following the momentum-learning rule implemented in the NeuroSolution program. For all the modeling procedures, the training was performed over 1000 iterations.

The results of drug-likeness prediction for the best SVM and ANNs on test set are presented in Table 4.

They are close to typical published[10] quality of 70−80% prediction accuracy. Interestingly, most of the studied

Table 3. Accuracy of Separation between Drugs and Nondrugs for the Best One-Descriptor SVM Models and for Thresholds Taken from Histograms

descriptor	kind of kernel	for train set	for test set	by histogram for train set	threshold (drug is..)
FA	RBF kernel $c=1$ $g=1$	63.02	61.29	62.97	≤ 0.75
	linear kernel $c=1$	62.12	60.79		
LogD	RBF kernel $c=1$ $g=1$	56.56	56.25	56.58	≤ 4.00
	linear kernel $c=1$	56.33	55.16		
LogSw	RBF kernel $c=1$ $g=1$	59.04	58.41	58.92	≥ -3.00
	linear kernel $c=1$	58.81	58.11		
a_acc	RBF kernel $c=1$ $g=1$	59.01	58.87	58.95	≥ 3
	linear kernel $c=1$	58.95	58.79		
a_don	RBF kernel $c=1$ $g=1$	61.59	60.33	61.59	≥ 2
	linear kernel $c=1$	61.59	60.33		
B_rotN	RBF kernel $c=1$ $g=1$	60.63	61.33	60.59	≥ 10
	linear kernel $c=1$	60.32	60.41		
MW	RBF kernel $c=1$ $g=1$	64.00	62.65	61.66	≥ 350
	linear kernel $c=1$				

Table 4. Prediction Quality for the Best Drug-likeness Models, Test Set

model	accuracy (%) overall	accuracy (%) drugs	accuracy (%) nondrugs
SVM, RBF kernel $c=2$ $g=10$[a] (*train. time* $\sim$ 5 min[b])	75.15	72.19	78.10
SVM, linear kernel $c=1$[a] (*train. time* $\sim$ 2 min)	68.68	66.12	71.25
multilayer perceptron, 1 hidden layer (*train. time* $\sim$ 39 min)	72.52	69.63	75.41
modular feedforward network, 2 hidden layers (*train. time* $\sim$ 110 min)	70.92	78.33	63.51
generalized feedforward network, 1 hidden layer (*train. time* $\sim$ 37 min)	69.85	77.53	63.38
Lipinsky rule of 5		62.60	28.60

[a] While searching for the best SVM models, the whole parameter space was scanned (parameter c for linear kernel, 5−10 training cycles in total, and parameters c and g for RBF kernel, 50−70 training cycles in total). In each run a training/validation cycle was performed. Models that have shown the best *predictive* power (i.e. performance on validation set) were selected and applied to *test* set to check for "actual" performance, which is reported in the table. [b] Athlon 1500 MHz, training time for SVM is reported for one training cycle.

compounds meet the requirements of all four conditions of Lipinsky's rule, and nondrugs meet this rule better than drugs.

The distribution of predicted scores for the test set allows the determination of some interesting aspects of the models we created (Figure 4).

The best RBF model has a distinctly nonnormal distribution of predicted scores, especially for drugs, compared to the more balanced linear SVM and neural-network distribution curves. Obviously, this distribution discovers some heterogeneity of compounds of the data set that is presented in initial feature space, but this heterogeneity is nevertheless reflected only in the nonlinear SVM model.

The typical time cost of establishing good SVM decisions is comparable to that necessary to train ANNs, as finding the best SVM model requires scanning the space of parameters, giving rise to 50−70 RBF SVM models. Commonly, it takes 2−3 h on the Athlon 1500 MHz machine.

AGROCHEMICAL-LIKENESS ESTIMATION

For this study, we selected 500 diverse commercial agrochemicals[25] and 11 000 diverse nonagrochemical molecules from the Sigma-Aldrich catalog of organic compounds.[17] As in the case of nondrug-like compounds, an assumption was made that the compounds without a defined type of activity from the chemical catalog possess minimal agrochemical activity. The structures were characterized with an extended set of descriptors (21 in total, see Table 1) most of which (Jurs descriptors) mainly combine shape and electronic information about molecules.

Table 5. Subsets Used in Creating and Testing Agro-likeness Models

	train set	validation set	test set
total	5830	2902	2915
agrochemicals	208	116	112

For building models, the same approach as in the drug-likeness experiment described above was used, except that a new method was applied in ANNs. It combines feedforward network learning with descriptor selection by a genetic algorithm that removes insignificant and interfering descriptors during learning. This method produces the lowest error.

The whole set was divided into training, validation, and test set. The validation set was used as an alternative to the cross-validation procedure during selection of the model with the most stable predictive results. Prior to SVM calculations, descriptor values were scaled. The size of each of the data subsets is presented in Table 5.

The disparity between the number of agro- and nonagrochemicals also was taken into account while deriving the models.

None of the linear SVM models showed acceptable classification results, so only the RBF kernel model is described below (any linear kernel shows a high number of classification errors). As was the case for drug-likeness estimation, the RBF SVM model discriminates compounds of the test set better than the best found neural network model. Classification results of the SVM and ANNs (best models) on compounds of the test set are shown in Table 6.

Table 6. Summary of Prediction Quality for the Best Agro-likeness Models on Test Set

model	agrochemicals, errors (from 112)	nonagrochemicals, errors (from 2786)
SVM, RBF kernel c=5000 g=1 (*train. time* ~ 3 min[a])	17	821
genetic algorithm + generalized feedforward network, 1 hidden layer (*train. time* ~ 300 min)	22	1109
generalized feedforward network, 1 hidden layer (*train. time* ~ 25 min)	33	1021
modular feedforward network, 1 hidden layer (*train. time* ~ 69 min)	41	1081

[a] Athlon 1500 MHz, training time for SVM is reported for one training cycle.

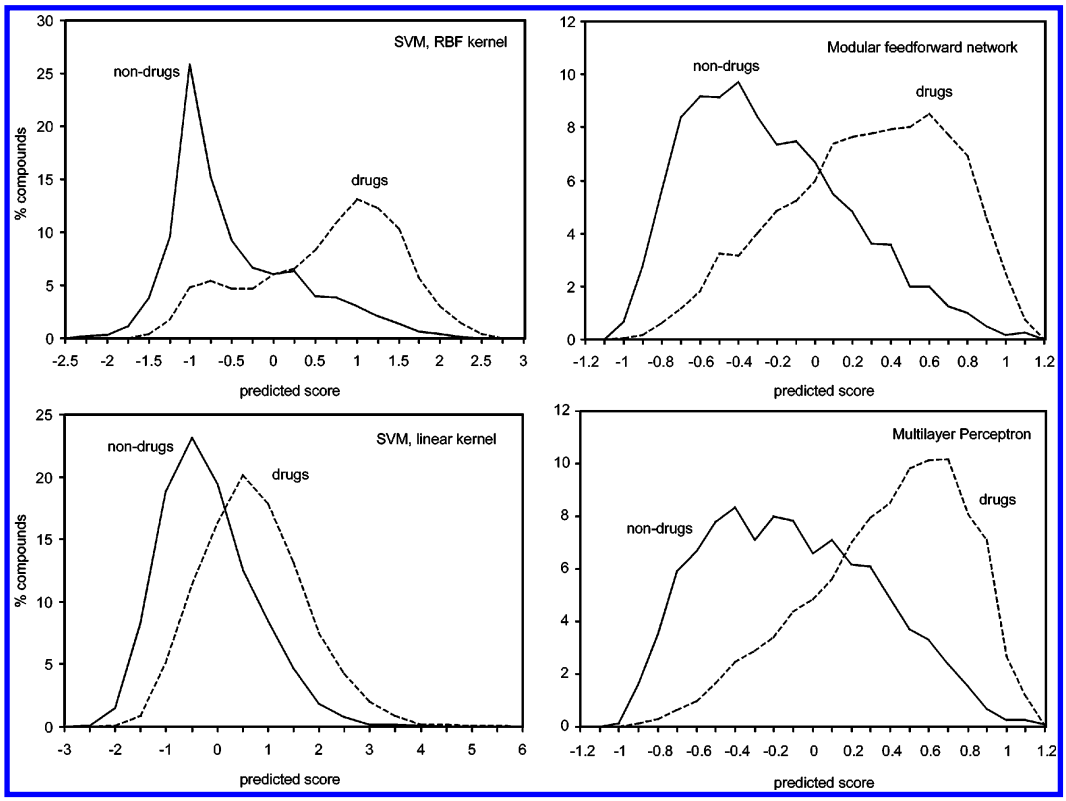


Figure 4. Distribution of drug-likeness predicted scores for test set.

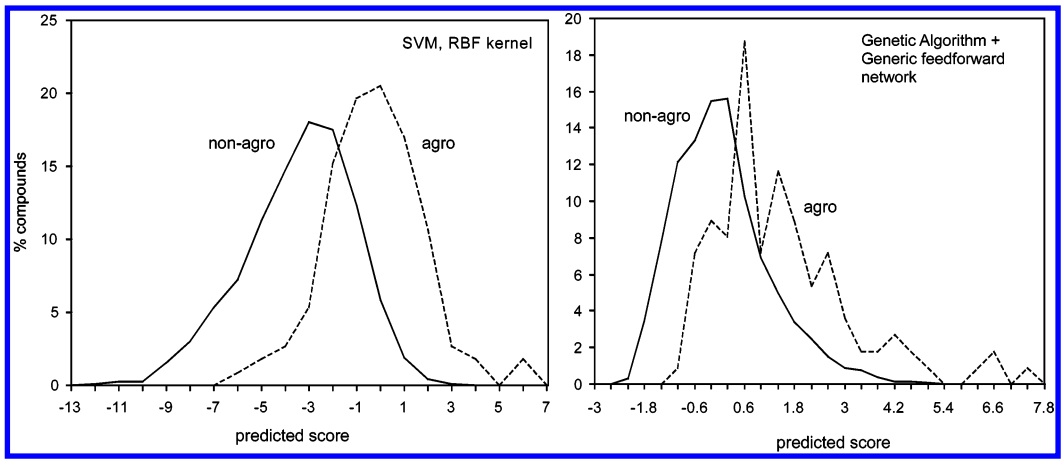


Figure 5. Distribution of agro-likeness predicted scores for test set.

In addition to classification results, the distribution diagrams of predicted scores for the two best models are shown in Figure 5.

The distribution of predicted scores for SVM RBF models not only provides better separation but also has a more regular shape (close to normal). The last fact allows us to suppose that the SVM separation surface is near optimal for the current descriptor space and, accordingly, will recognize outliers better.

STUDY OF CARBONIC ANHYDRASE II (CA II) INHIBITORS

CA II is a well-characterized zinc-dependent enzyme. Its own spatial structure and the structures of its complexes with inhibitors have been determined by X-ray crystallography. Several hundred diverse CA II inhibitors were synthesized in the last 50 years because of its applicability in different areas: diuretics, antiepileptics, modulators of cancer che-

Table 7. Comparison of Binary QSAR, SVM, and "FN Descriptor" Models of CA II Inhibition Activity, Number of Classification Errors

model	training set errors			training set errors, cross-validation			test set errors			overall errors, from 337
	active, from 225	inactive, from 55	overall, from 280	active	inactive	overall	active, from 52	inactive, from 5	overall, from 57	
FN descriptor[a]	5	18	23				0	1	1	24
binary QSAR	10 (pred. quality 96%)	9 (84%)	19 (93%)	96%[b]	82%[b]	93%[b]	3	1	4	23
SVM linear kernel c=1	5	18	23	5 (98%)	18 (67%)	23 (92%)	0	1	1	24
SVM RBF kernel c=10 g=1	3	6	9	4 (98%)	11 (80%)	15 (95%)	2	1	3	12

[a] "FN descriptor" model means that if FN = 0 then compound predicted as inactive; otherwise, as active. [b] In source work[26] only these values are presented.

motherapy, agents for the treatment of glaucoma, etc. Moreover, structure−activity relationships have been extensively studied for this molecule. It is known that common pharmacophore groups are the terminal sulfonamide group connected to an aromatic ring or heterocyclic portion with several hydrogen bonding groups.

The main reason for using this set to test SVM is the presence of a relatively large collection of CA II inhibitors in Gao's paper[26] from his series[26−28] where the author develops the **binary QSAR** approach (compounds are considered concerning concrete biological activity as actives or inactives).

Results of deriving and testing binary QSAR and SVM classifiers on a set of 337 diverse CA II inhibitors are shown in Table 7. Both classifiers were built in space of six descriptors: 1. HI1, first-order atomic valence connectivity index; 2. HI2, second-order atomic valence connectivity index; 3. K1, Kier first shape index; 4. HBA, number of hydrogen bond acceptors; 5. LogP, calculated octanol/water partition coefficient; 6. FN, number of $R-SO_2NH_2$ fragments. The results of binary QSAR and descriptor values for SVM were taken from Appendix I of Gao's source paper,[26] where they are fully described. The same training and test sets we used for deriving SVM classification models, so the "starting points" of two methods were identical and the results can be directly compared. In contrast to experiments on both drug- and agro-likeness, for building these SVM models the initial descriptor values were used without scaling or other preprocessing.

It is evident from Table 7 that an advantage of the best SVM classifier in common prediction results was found in this case as well.

Further it is reasonable to ask the following questions: (1) What are the advantages and disadvantages of SVM models? (2) What is the inherent "classifying power" of the given set of descriptors? (3) How can one correctly separate compounds into active and inactive?

As shown in Table 7, the overall number of errors for the RBF SVM model is half as much as for the others. Cross-validation and counting of errors on the test set, however, demonstrated that actual RBF SVM model quality surpasses binary QSAR model less than 2-fold. Principal Component Analysis (PCA) on all six descriptors showed that (1) FN descriptor (number of unsubstituted sulfanilamide groups) almost fully describes the presence or absence of activity. (2) It is possible to select two (or at maximum, three) principal components (PCs) that are necessary for reproduction of initial set of descriptor variance with adjusted exactness.

These points are illustrated by Figure 6. Both the scree plot and eigenvalue table demonstrate the contribution of PCs to total variance. Loadings plot shows descriptor similarity (the smaller is the angle between two vectors, the more similar are the corresponding descriptors). Scores plot visualizes distribution of compounds in the space of two first PCs and SVM separation surfaces. Note that this picture is essentially two-dimensional (two first PCs) and incomplete; however, it represents the main features of class-separating surfaces well.

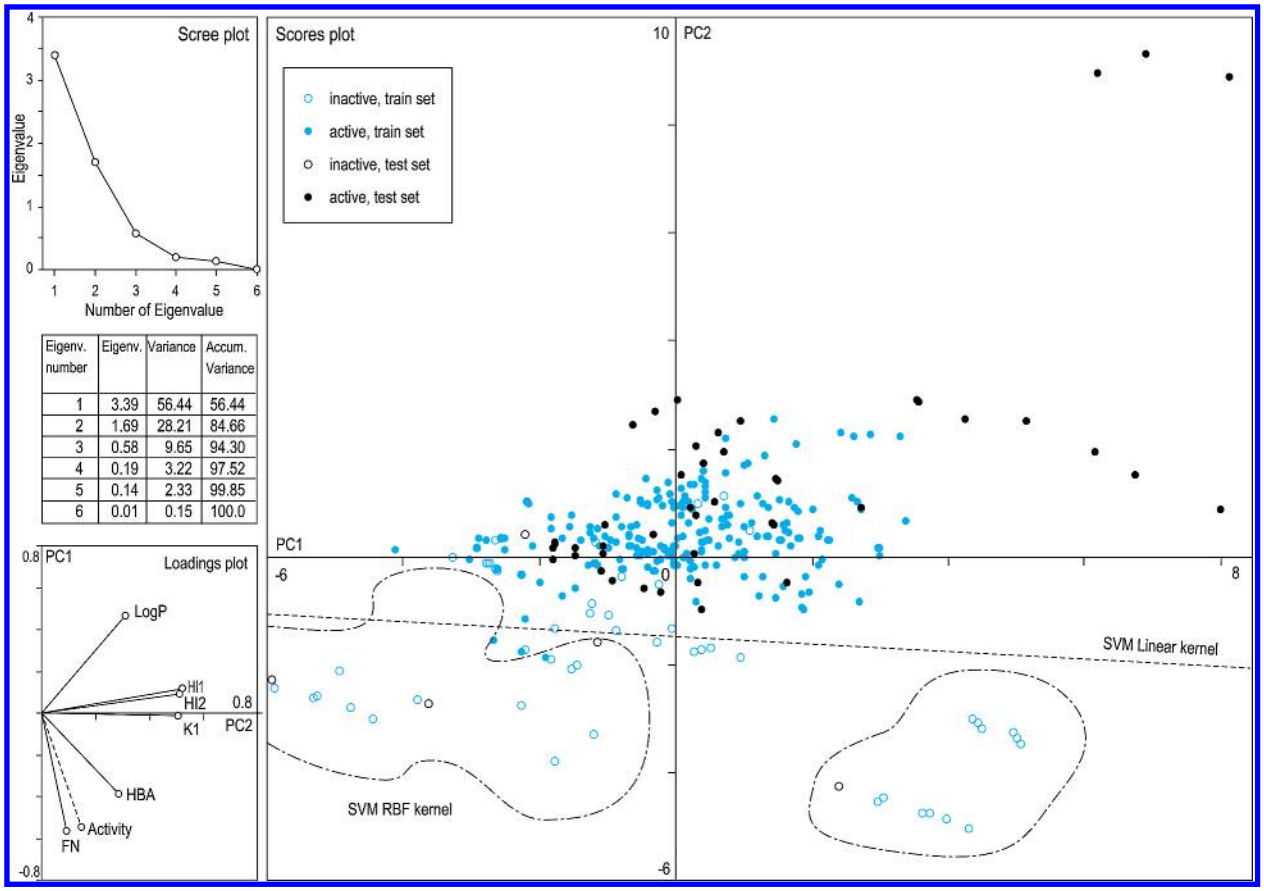
Visual representation of all compounds and separation surface in space of two first PCs allows us to understand how compounds of both sets (train and test) are distributed in descriptor space. Moreover, it allows us to discover a significant feature of SVM RBF separation surface—its locality. The decision surface "envelops" points of inactive compounds (more clearly it is visible in 3D space of three first PCs). In essence, a linear SVM approach that actually takes into account one FN descriptor (linear SVM and FN descriptor models results coincide) derives a rough but more adequate separation surface.

The overall large number of prediction errors for the best models suggests that the data may be poorly preconditioned, which results in a bad representation of data set inner structure. The most probable reason is inadequate division of compounds into active and inactive ones. To illustrate this point, distribution diagram for measured continuous value of biological activity (Log1/C) is shown (Figure 7).

Apparently, the border between active and inactive compounds (Log1/C = 6.0) which was used in the source work[26] is displaced relative to the "statistically justified" range (2−5 values of Log 1/C). If the border is shifted to a smaller value, the number of errors for FN descriptor model decreases, as shown in Table 8.

The minimum number of errors, 5, is found for a border value of 4.5. Further decreasing (or shifting) of the border does not change the number of errors. Note that it is thus possible to estimate the quantitative contribution of a functional group to the measured value of activity, Log1/C. These 5 incorrectly classified compounds have no unsubstituted RSO_2NH_2 fragments but one-substituted RSO_2NHR_1. Thus these errors are due to a failure to take into account the presence of substituted sulfonamide groups, which can appreciably influence value of activity.

As a result of this detailed analysis, it was possible not only to define the main characteristics of derived models but also to reveal a peculiarity relating to the usage of descriptors and definition of the border between active and inactive samples. This peculiarity is critical for deriving



Eigenv. number	Eigenv.	Variance	Accum. Variance
1	3.39	56.44	56.44
2	1.69	28.21	84.66
3	0.58	9.65	94.30
4	0.19	3.22	97.52
5	0.14	2.33	99.85
6	0.01	0.15	100.0

Figure 6. Application of PCA to set of descriptors used for derivation of CA II inhibition activity models. Visualization of SVM separation surfaces in space of two first principal components.

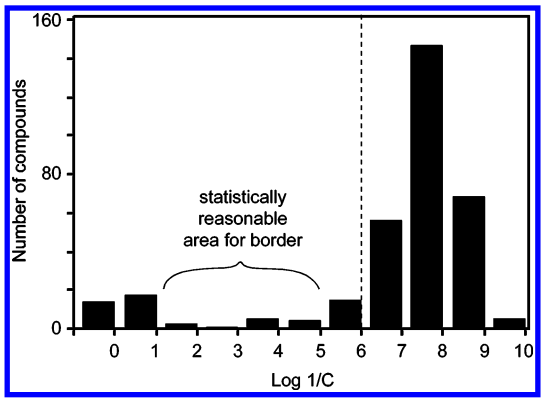


Figure 7. Distribution diagram of measured CA II inhibition activity[26] for all 337 compounds. Dashed line marks the border (taken from source work[26]) used for separating active and inactive compounds. However, the distribution shows that shifting the border to smaller values is statistically more reasonable.

Table 8. Dependence of Error Amount for "FN Descriptor" Model on the Border Value of Measured Inhibition Activity that Separates Active from Inactive Compounds

separation border for Log1/C (if more or equal is active)	train set overall (active/inactive)	test set overall (active/inactive)
7	51	13
6	23 (5/18)	1 (0/1)
5	8 (5/3)	0
4.5	5 (5/0)	0
4	5 (5/0)	0

adequate classification models (subsequently, Gao derived[28] a new binary model of CA II inhibition activity with genetic selection of descriptors, but these did not lead to a significant improvement of classification). The construction of such models is a subject of ongoing investigation in our laboratory.

CONCLUSION

In summary, we tested Support Vector Machines as a classification tool in several real-life drug-discovery problems and found it typically outperforming other approaches, in particular, artificial neural networks. However, SVM is definitely not a panacea which always produces the best models; the key to the latter often lies in selection of proper descriptors. A particular case of CA II inhibition study provides interesting insights into this relationship, by discovering that the best SVM decision is "local".

ACKNOWLEDGMENT

Two of us (I. Pletnev and V. Zernov) gratefully acknowledge the support from the Russian Foundation for Basic Research (No. 01-07-90383) and "Integraciya" program (No. F0036/883) which enables partial development and usage of COSMOS and DateX software for chemical and statistical calculations.

REFERENCES AND NOTES

(1) Lipinsky, C.; Lombardo, F.; Dominy B.; Feeney, P. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Adv. Drug Delivery Rev.* **1997**, *23*, 3−25.

(2) Oprea, T. Property distribution of drug-related chemical databases. *J. Comput.-Aided Mol. Design* **2000**, *14*, 251−264.

(3) Vapnik, V. *Statistical Learning Theory*; Wiley: New York, 1998.

(4) Guyon, L.; Weston, J.; Barnhill, S.; Vapnik, V. Gene selection for Cancer Classification using Support Vector Machines. *Machine Learning* **2002**, *46*, 389−422.
(5) Cai, Y.; Liu, X.; Xu, X.; Chou, K. Prediction of protein structural classes by support vector machines. *Comput. Chem.* **2002**, *26*, 293−296.
(6) Fernandes, R.; Viennet, E. Face identification using Support Vector Machines. *Proc. Europ. Symp. Artificial Neural Networks (ESANN99)* **1999**, 195−200.
(7) Burbidge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26*, 5−14.
(8) Trotter, M.; Buxton, B.; Holden, S. Support Vector Machines in combinatorial chemistry. Submitted to a special edition of *Measurement & Control*, Sep 2001. URL: http://www.cs.ucl.ac.uk/staff/m.trotter/private/mc_paper-mt-bb-sh.doc.
(9) Warmuth, M.; Ratsch, G.; Mathieson, M.; Liao, J.; Lemmen C. Active learning in the Drug Discovery process. Submitted to *J. Chem. Inf. Comput. Sci.*, December 2002.
(10) Sadowski, J. Optimization of the drug-likeness of chemical libraries. *Perspect. Drug Discovery Des.* **2000**, *20*, 17−28.
(11) *Advances in Kernel Methods − Support Vector Learning*; Schölkopf, B., Burges, C., Smola, A., Eds.; MIT Press: Cambridge, MA, 1999.
(12) Cristianini, N.; Shawe-Taylor, J. *An Introduction to Support Vector Machines*; Cambridge University Press: Cambridge, UK, 2000.
(13) URL: http://www.kernel-machines.org/.
(14) Vapnik, V.; Chervonenkis, A. About structural risk minimization principle. *Automation Remote Control* **1974**, Nos. 8, 9.
(15) Bennet, K.; Campbell, C. Support Vector Machines: Hype or Halleujah *SIGKDD Expl.* **2000**, *2*, 1−13.
(16) Ensemble database of pharmaceutical compounds, Prous Science, 2003. URL: http://www.prous.com/.
(17) Sigma-Aldrich Catalog of Rare Chemicals, July 2001.
(18) Chemical Diversity Labs, Inc. 2003. URL: http://www.chemosoft.com/.
(19) Ajay, A.; Walters, W. P.; Murcko, M. A. Can We Learn to Distinguish between "Drug-Like" and "Nondrug-Like" Molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.
(20) Ajay, A.; Bemis, G. W.; Murcko, M. A. Designing Libraries with CNS Activity. *J. Med. Chem.* **1999**, *42*, 4942−4951.
(21) Raevsky, O.; Trepalin, S.; Trepalina, H.; Gerasimenko, V.; Raevskaya, O. SLIPPER-2001 − Software for Predicting Molecular Properties on the Basis of Physicochemical Descriptors and Structural Similarity. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 540−549.
(22) Chang, C.; Lin, C.-J. LIBSVM: a library for support vector machines, 2001. URL: http://www.csie.ntu.edu.tw/∼cjlin/libsvm.
(23) Joachims, T. Making Large-Scale SVM Learning Practical. In *Advances in Kernel Methods − Support Vector Learning*; Schölkopf, B., Burges, C., Smola, A., Eds.; MIT Press: 1999. URL: http://svmlight.joachims.org/.
(24) NeuroDimension Inc., URL: http://www.nd.com.
(25) Bayer's Catalogue of Agrochemical Information, January 2000.
(26) Gao, H.; Bajorath, J. Comparison of binary and 2D QSAR analyses using inhibitors of human carbonic anhydrase II as a test case. *Mol. Diversity* **1999**, *4*, 115−130.
(27) Gao, H.; Williams, C.; Labute, P.; Bajorath, J. Binary-QSAR analysis of estrogen receptor ligands. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 164−168.
(28) Gao, H.; Lajiness, M.; Van Drie, J. Enhancement of binary QSAR analysis by a GA-based variable selection method. *J. Mol. Graphics Modell.* **2002**, *20*, 259−268.

CI0340916