Screening for New Antidepressant Leads of Multiple Activities by Support Vector Machines

Zsolt Lepp, Takashi Kinoshita, and Hiroshi Chuman*

Institute of Health Biosciences, The University of Tokushima, Shomachi, Tokushima 770-8505, Japan

Received July 25, 2005

Virtual screening was carried out against 21 biological targets related to depression by support vector machine classification using the same atom-type descriptors. The models were effective as 0.2−0.8 of theoretical enrichments of the external test data sets could be achieved, depending on the target. The set of predicted active molecules had large diversity and contained examples with high dissimilarity to the compounds of training sets. Filtering the database of known antidepressants by all 21 models it was found that on average compounds were classified active for 2.3 targets.

INTRODUCTION

Antidepressants constitute one of the largest therapeutic areas of current drug market. Despite this fact there is solid interest for newer type of drugs as, for example, existing antidepressant treatments exhibit limited efficacy and slow onset of action. The state of research involving new biological targets was reviewed recently.[1] One conclusion is that it might be favorable to develop new molecules with activity against various targets.

However the large number of valid targets makes it difficult to identify all possible mechanism of all candidate compounds. Virtual screening is a promising tool to help get this work done.

The aim of this work was to develop virtual screening method(s) for a number of targets related to depression. The models were expected to be used at the early phase drug discovery namely, in new lead identification. The requirements were that the filters should classify molecules as potentially positive or negative in a high throughput fashion.

The chosen method was using support vector machines with simple atom-type descriptors. There were various reasons for the choice. First of all, there is very little 3D structural information available about the potential target proteins. This fact dictates that the approach should be to examine known active molecules and to extract characteristic structural information that is suitable to evaluate novel compounds. The purpose of the procedure was to identify novel lead candidates by classifying compounds as active or nonactive.

Recently an effective method for such classification problem was reported.[2] Probabilistic neural networks with atom-type descriptors were used. It was concluded that the effectiveness of the method probably lies in the combination of the strong discriminating power of neural networks and the generality of atom-type descriptors. The advantages of characterizing a molecule by simply counting the number of atom-types it is made of are various. These are mainly the accuracy (theoretically 100%), the ease, and the speed of generating these features. Furthermore the method can be used for any kind of compounds with large diversity.

Because of these advantages, the procedure seemed to be very suitable for high throughput virtual screening. With partial intent to further elaborate the referred work,[2] the neural network was substituted for another statistical learning method. A relatively novel method, support vector machine (SVM), was applied. This method was first published in 1995 and soon became popular for solving various classification problems. It is now more and more frequently used for performing chemoinformatics and chemometrics tasks.[3−6]

To select biological targets and data sets of active and nonactive molecules, the Drug Data Report database from MDL (MDDR) was used.[7] The selection of targets was driven by the data sets of MDDR. The database contains a general, “antidepressants” category. Additional targets having common molecules with this category were chosen. It was a prerequisite that enough examples had to be available for the successful training of SVM. Finally altogether 20+1 targets were selected with +1 being the general “antidepressants” category bearing more than 5000 diverse structures.

The generated SVM models were used to predict activities for compounds from the MDL Screening Compounds Directory (SCD)[7] and to perform various statistical analyses.

METHODS

Data Sets. All molecular structures for training and testing SVM models were selected from the MDDR. Table 1 contains the number of compounds selected for training and test sets for each target. For cross-validation sets 20% of the training sets were selected during the training process (5-fold cross-validation). Target **1** is the general antidepressants class. It is a diverse set; molecules with activity to a broad range of biological targets belong here. Of these compounds, 393 do not belong to any other target group than **1** (no more activity other than the antidepressant effect was reported for those). The last column in Table 1 shows the number of shared molecules between **1** and each of the other targets. The existence of these common molecules was one of the reasons for selecting **2**−**20** into the study. The

* Corresponding author phone and fax: +81-88-633-9508; e-mail: hchuman@ph.tokushima-u.ac.jp.

10.1021/ci050301y CCC: $33.50
Published on Web 11/12/2005

Table 1. List of Biological Targets and the Size of Train and Test Sets Used for Creating Screening Models

ID	code[a]	target name	no.[b]	train set[c]	test set[c]	∩08000[d]
1	08000	general antidepressants	5472	2452	3020	na
		Serotonin Receptor Agonists				
2	06235	HT1A	986	489	497	590
3	06237	HT1C	139	68	71	43
4	06246	HT1D	624	310	314	60
		Serotonin Receptor Antagonists				
5	06233	HT3	845	419	426	44
6	06248	HT2A	649	321	328	127
7	06249	HT2B	109	53	56	52
8	06250	HT2C	235	115	120	154
9	06240	HT1A	750	489	261	363
10	06245	5HT reuptake inhibitors	745	371	374	662
		Adenosine Agonists				
11	08450	A1	256	126	130	115
12	08451	A2	289	143	146	126
13	31262	α2 blockers	248	122	126	165
		Dopamine Antagonists				
14	07702	D1	179	87	92	54
15	07701	D2	484	241	243	72
16	08415	norepinephrine uptake inhibitors	223	109	114	173
		Monoamino-Oxidase Inhibitors				
17	08410	MAO A	93	44	49	89
18	08420	MAO B	116	56	60	52
19	11126	dopamine reuptake inhibitors	172	83	89	130
20	06215	CRF antagonists	401	198	203	298
21	42731	substance P antagonists	1623	643	980	148
neg	na	negatives[e]	30 469	10 652	19 817	0
neg	na	negatives[f]	30 469	2498	27 971	0

[a] Activity code in MDDR. [b] Number of entries. [c] Number of entries selected for train and test sets. [d] Number of compounds belonging to that also a member of general antidepressants category (**1**). [e] Train and test sets for **1**. [f] Train ad test sets for **2**−**21**.

collective number of data entries of Table 1 is 14 667, but because of redundancies 10 704 distinct active (positive) compounds were used. The nonactive (negative) examples were chosen among the remaining molecules in MDDR after removing all positives. Two negative train sets were made, containing 10 000 and 2500 entries for **1** and for all the other targets, respectively.

From SCD 3 247 299 compounds were chosen and screened by the models of **1**−**21**.

Software Tools. To carry out all SVM calculations was used the software LibSVM.[8] The atom-type descriptors were created by a proprietary program. Various chemoinformatics tasks, such as standardizing molecular structures, calculating molecular fingerprints, comparing the similarities of libraries, and ward clustering, were done by JChem[9] and the statistical software, R.[10] K-means clustering were calculated by Xlstat.[11]

Training and Test Sets. All compounds were neutralized, counterions were removed, and hydrogens were added by JChem before calculating descriptors.

Altogether 39 atom-type descriptors were used and summarized in Table 2. To create test and train sets for a given target all of the active examples were clustered in the descriptor space by k-means clustering. Each of the clusters was roughly divided into two halves (in some cases into 40−60%, for train and test sets, respectively), in a random fashion.

The negative examples were selected as follows. Only compounds that had detailed activities were selected from MDDR. Molecules belonging to targets **1**−**21** were removed, and none of those was placed into the negative sets. The remaining compounds were clustered in the descriptor space. From each cluster were selected about 10 000 and 2500 entries into the train sets of **1** and **2**−**21**, respectively. It was ensured that the same percentages of all the clusters were placed into the given sets. The remaining compounds were put into the test sets.

Model Creation. C-SVM was used with radial base function (RBF) also called the Gaussian function. The positive and negative examples were labeled as 1 and −1, respectively. The "easy.py" script of LibSVM was used to automate parameter selection during training. C-SVM with the Gaussian kernel only needs two parameters: C that is the penalty parameter of the error term and γ that is the kernel parameter for the radial base function type kernel. The script optimizes the two parameters by a systematic grid search. The 20% cross-validation set derived from the train set was used to evaluate the accuracy of prediction during the systematic search. Using the optimal γ and C values a final training was performed but this time with the training set being united with the cross-validation set.

Similarity Comparison of Molecules. Compound libraries were compared using the JChem software. To describe the molecules 1024 bit hashed fingerprints were used. The maximum pattern length was set to 6 with 2 bits for each pattern. Similarity between two compounds was determined by means of calculating the Tanimoto coefficients.

Validation of Models. The SVM models were built using the training sets. The external test sets were used to evaluate the methods using four different performance measures. These measures were recall, precision, enrichment, and the product of the former two.

Table 2. Atom-type Descriptors Used in All the Models

no.	name	description
1−3	C1−C3	sp1−sp3 carbon, respectively
4	Car	sp2 aromatic carbon
5	C21	carboxylic type carbon (R-**C**X1-X2-Y; X1=O,N,S X2=O,N,S Y=any, R=alkyl)
6−9	F,Cl,Br,I	halogens
10	P	phosphorus
11	O2	sp2 oxygen with no adjacent pi orb (R−CO−R, etc.)
12	O21	sp2 oxygen with adjacent pi orb (X1 with C21 type, R−C=C=O, aromatic-C=O)
13	O3	sp3 oxygen in ether alcoholic R−O−R
14	O31	sp3 oxygen in ester or conj ether (X=C−O−R,XC−O−R) (X2 with C21 type)
15	Oar	oxygen in aromatic ring
16−17	Ono,Ono2	oxygen in NO and in NO2 groups, respectively
18	N1	sp1 nitrogen
19	N2	sp2 nitrogen with no adjacent pi RR′C=NH−R″
20	N21	sp2 nitrogen with adjacent pi orb (C=NH, R−C=C=N−R, etc.) (X1 with C21)
21	N3	sp3 nitrogen
22	N31	sp3 nitrogen in ester or con. ether (X=C−O−R) (X2 in case of C21 type)
23	Nar	nitrogen in aromatic ring
24−25	Nno,Nno2	nitrogen NO and in NO2 groups, respectively
26	H	hydrogen
27	Hdnr	hydrogen in H donor group
28	Har	aromatic hydrogen
29	S2	sp2 sulfur (C=S)
30	S3	sp3 sulfur in thioether
31−32	S4,S6	sulfinyl and sulfonyl S, respectively (R2SO,RSO2)
33	Sh	sp3 S in thiol
34	Sar	sulfur in aromatic ring
35	Chan	atoms inside chains
36	Nring	atoms inside rings
37	Nrot	rotatable bonds
38	Hbac	H-bond acceptors
39	Nrng	rings

Recall values for both positive and negative samples defined by

$$\text{recall} = t_p/(t_p+f_n) \qquad (1)$$

for positives, where t_p is the number of true positives and f_n is the number of false negatives. The value describes the ratio of correctly classified members of a data set.

Precision measures the percentage of true pieces among all instances a model has classified as belonging to the given category

$$\text{precision} = t_p/(t_p+f_p) \qquad (2)$$

where f_p is the number of false positives. Because in this work the precision of actives is of interest only the values for those are reported.

Enrichment factor is a frequently used measure for virtual screening. It measures the enrichment of the method if compared to random selection. It can be interpreted as the ratio of true positives in the set of instances classified as active (i.e., the precision) vs the ratio of actives in the original set to be classified.

$$\text{enrichment factor} = \text{precision}/((t_p+f_n)/(t_p+f_p+t_n+f_n)) \qquad (3)$$

Theoretical Enrichment. The resulting enrichment value for a model depends on the theoretical enrichment of a data set. It tells the maximum enrichment that can be achieved. The theoretical enrichment can be obtained if the both recall and precision equals 1 thus, all and solely the actives were found.

$$\text{theoretical enrichment} = (t_p+f_p+t_n+f_n)/(t_p+f_n) \qquad (4)$$

The aforementioned measures are very useful when comparing the effectiveness of various methods or models on the same data set. However, in the current case the same method was used for a number of different targets which means that the data sets were not identical. To compare the wellness of models for different targets the enrichment factor cannot be used, because it depends on the theoretical enrichment.

$$\text{enrichment factor} = \text{precision} * \text{theoretical enrichment} \qquad (5)$$

The precision is a better measure for the purpose, but it does not tell anything about the absolute number of predicted actives. A good model should not only supply correct actives but the more of the actives found the better. To take both criteria into consideration the product of precision and recall *(*PR*)* of positive examples was used.

$$\text{PR} = \text{precision} * \text{recall} \qquad (6)$$

RESULTS AND DISCUSSION

Support Vector Machine. SVM is a relatively new, promising method for learning separation functions in classification tasks or performing functional estimation in regression and is originated from the statistical learning theory developed by Vapnik and Chervonenkis.[3,11] It is a supervised learning technique applicable to both classification and regression. It offers a possibility to find a solution by making a nonlinear transformation of the original input space

Table 3. Model Parameters and the Accuracy of Predictions for Train Sets

		model params[a]		C.V.[b] [%]	recall [%]	
ID	target name	C	γ	neg&pos	neg	pos
M1	general antidepressants	2	45.25	89.2	99.01	89.56
	Serotonin Receptor Agonists					
M2	HT1A	4	8	92.42	97.29	95.5
M3	HT1C	1024	0.35	99.42	99.84	97.06
M4	HT1D	8192	0.13	96.94	98.52	96.45
	Serotonin Receptor Antagonists					
M5	HT3	45.3	0.71	95.18	97.99	89.5
M6	HT2A	181	0.5	95.33	97.14	87.81
M7	HT2B	256	0.59	98.59	99.89	84.91
M8	HT2C	256	16	97.86	99.16	99.13
M9	HT1A	2	8	94.93	99.16	85.04
M10	serotonin reuptake inhibitors	8	4	93.34	98.21	92.99
	Adenosine Agonists					
M11	A1	1.4	11.31	97.95	99.61	92.86
M12	A2	2048	2	97.24	98.84	97.87
M13	α2 blockers	1	16	96.61	99.68	80.33
	Dopamine Antagonists					
M14	D1	1024	16	98.46	99.8	98.85
M15	D2	8	8	93.88	98.72	95.44
M16	norepinephrine uptake inhibitors	4	8	97.78	99.66	84.55
	Monoamino-Oxidase Inhibitors					
M17	MAO A	2	16	99.18	99.99	81.82
M18	MAO B	32	16	98.71	99.9	92.86
M19	dopamine reuptake inhibitors	1.4	22.63	98.14	99.77	93.98
M20	CRF antagonists	7.4×10^3	0	98.44	99.61	92.93
M21	substance P antagonists	8	4	93.34	97.3	95.96

[a] Model parameters for SVM. [b] Recall values for counter validation sets.

into a high dimensional feature space, where an optimal separating hyperplane can be found. Optimal in this case means that a maximal margin classifier with respect to the training data set can be obtained. An important and unique feature of this approach is that the solution is based only on those data points, which are at the margin. These points are called support vectors.

An ingredient of SVMs and other kernel methods is the use of kernels, which makes it possible to map the data implicitly into a feature space and to train a linear machine in such a space, potentially side-stepping the computational problems inherent in evaluating the feature map.[13]

The method has some advantages over other learning techniques. By changing of the margin the capacity of method can be controlled. There are no local minima to be trapped into. There are usually a few parameters (e.g. only two for the RBF kernel, that was used), and these are nonrandom ones meaning that the final model is stable and reproducible. Further advantage is that the method is relatively simple, and it is not necessary to be an SVM-expert to successfully apply the existing software.

Support vector machine is already a frequently used method in chemical informatics. It was shown in various references[14−20] that it is comparable to other classification procedures. SVM gave similar or sometimes superior predicting capability than PLS[14−16,18] or Neuro Network[14,16,17−20] methods.

SVM is not without disadvantages, either. It is still necessary to select the attributes to be included in the problems and the type of kernel including its parameters. In addition, the existing knowledge about the system can only be added to the model only through this selection. It is also difficult to interpret the results, and the feature selection (i.e., sorting out the most important descriptors) is especially difficult, and although there are various advances in achieving this aim,[6] these methods are still not commonly integrated into available software. Fortunately, the continuous advances in computer hardware and algorithms are easing the problem of high resource-intensity of SVM that was one of the main reasons blocking the widespread application of the method, so far.

Evaluation of Models. SVM classification models for all the 21 targets in Table 1 were prepared as described in the Methods section. The success rates of predictions for SVM models are shown in Tables 3 and 4. Table 3 contains the two optimal parameters (γ and C) for SVM-training and the results for the cross-validation and train sets. The second table is a summary of performance measures of the models based on the test sets.

By examining the results for train and test sets it can be concluded that all the models bear very high predictive capability, especially true for the negative samples. The success rate is higher than 95% for the negative train and test sets in the case of all the targets. It is especially good given that the test set consisted of almost 30 000 (20 000 for **1**) molecules with biological activity. Furthermore some of the molecules in the negative set were not really negatives. It contained active molecules to some targets that are valid antidepressant targets but were not included on the list of targets to model for various reasons (mainly the low number of examples). It can also be expected that some compounds active against other targets have antidepressant activity too but have yet to be determined. It means that the classification capacity of models for negative examples is even higher than the already good values of Tables 3 and 4.

Table 4. Accuracy Measures for Predictions of Test Sets

		recall (%)					
ID	target name	pos[a]	neg[b]	E[c]	TE[d]	P[e]	P*R[f] (%)
M1	general antidepressants	58.7	96.4	5.4	7.6	0.716	42.0
		Serotonin Receptor Agonists					
M2	HT1A	79.7	95.2	13.2	59.2	0.223	17.8
M3	HT1C	80.3	99.6	144.7	408.6	0.354	28.4
M4	HT1D	88.9	97.9	29.3	93.2	0.314	27.9
		Serotonin Receptor Antagonists					
M5	HT3	84.5	96.9	19.6	68.9	0.284	24.0
M6	HT2A	80.8	96.2	17.3	89.2	0.194	15.7
M7	HT2B	73.2	99.1	67.3	500.5	0.134	9.8
M8	HT2C	55.8	99.4	63.2	242.2	0.261	14.6
M9	HT1A	67.4	98.9	39.0	111.9	0.349	23.5
M10	serotonin reuptake inhibits	70.9	97.5	21.0	78.4	0.268	19.0
		Adenosine Agonists					
M11	A1	67.7	99.5	83.7	223.6	0.374	25.3
M12	A2	71.9	97.9	29.4	199.2	0.148	10.6
M13	α2 blockers	43.7	99.6	68.6	230.7	0.297	13.0
		Dopamine Antagonists					
M14	D1	57.6	99.8	132.4	315.6	0.419	24.2
M15	D2	70.8	97.5	23.4	120.1	0.195	13.8
M16	norepinephrine uptake i.	80.8	99.2	63.3	254.9	0.248	20.1
		Monoamino-Oxidase Inhibitors					
M17	MAO A	57.1	100.0	473.3	591.7	0.800	45.7
M18	MAO B	58.3	99.9	260.3	483.4	0.538	31.4
M19	dopamine reuptake inhibs	52.8	99.8	131.1	315.3	0.416	22.0
M20	CRF antagonists	88.2	99.2	60.0	143.6	0.418	36.9
M21	substance P antagonists	82.0	96.6	13.5	30.5	0.443	36.4

[a] Positive examples. [b] Negative examples. [c] Enrichment. [d] Theoretical enrichment. [e] Precision. [f] Precision multiplied by recall.

This highlights the wellness of the method for virtual screening, as the low number of false negatives is a very important criterion for this kind of task.

Another remarkable result is the very good prediction capability (59% recall for actives and 96% for negative ones) of model for target **1**. It is surprising because this data set contains a very diverse collection of compounds including actives against probably all the possible depression related targets. The precision of the model (0.72) is exceptionally good, meaning that the most diverse set of all targets bore one of the best precisions. Only the MAO A model has a better value but that is a relatively small set of quite homologous compounds, thus the good generalization capability of the model is understandable. It seems, that SVM with only atom-type descriptors is quite capable of separating the chemical space of antidepressants from the chemical space of other drugs. It should be noted that the negative set contained a number of the central nervous system (CNS) active molecules, thus this good separation probably cannot be explained with simple ADME-related differences between the antidepressants, which are mainly CNS drugs, and other drugs.

Comparing Tables 3 and 4, some discrepancies can be observed between recall of positives in train and test sets in the case of some targets. It means that in some cases the model is less general, and it is more specific to the training set than it should be in the ideal case. There are two possible reasons for it. One is that the data set is very diverse. It is true probably only for the case of **1**. It contains active structures against a large number of diverse targets, and even the large training set is not enough for a perfect model. The other reason (e.g. **M14**) the number of descriptors is big for the number of examples, resulting in overtraining. Obviously the solution would be to select an optimum set of descriptors. However, as it was mentioned, the relative difficulty of feature selection is one of the problematic areas of using SVM, and it was not feasible to carry out this task for all the 21 targets. On the other hand, the lack of good prediction power (thus, low recall of positives in test set) for some data sets were always accompanied by good precision (i.e., not only less true but also less false positives are predicted). It makes the use of these models feasible for virtual screening.

As it was mentioned the PR measure was applied to compare the effectiveness of models for various targets as it gives a more balanced view than recall or precision, alone. For example the precision of three models (**M14, M19, M20**) were uniformly 0.42. But looking at the PR values (24.16%, 21.96%, 36.85%, respectively) it can be concluded that the model for CRF (corticotropine releasing factor) antagonists is the most successful, with quite a margin. Comparing the enrichment values would give the wrong hint: the best model would come out as the worst (ca. 60 vs 130 for the other two) because of the different theoretical enrichments of these three targets. And while all the models equally realize a 0.42 fraction of their theoretical enrichments, the CRF model does it with identifying the largest ratio of positives of the three, thus its PR value is the biggest of the three.

Similarity Comparison of the Molecules of Train Set of Model 1. Model **1** was built from a large data set with a negative set consisting of a 10 000 member representative subset of MDDR. To determine the statistical significance of the model similarity comparisons were carried out between the active and nonactive examples of the data set. Pairwise similarity comparisons were carried out including all the compounds (thus ca. 13 000 × 13 000 comparisons were done). To define the similarity between two members regular

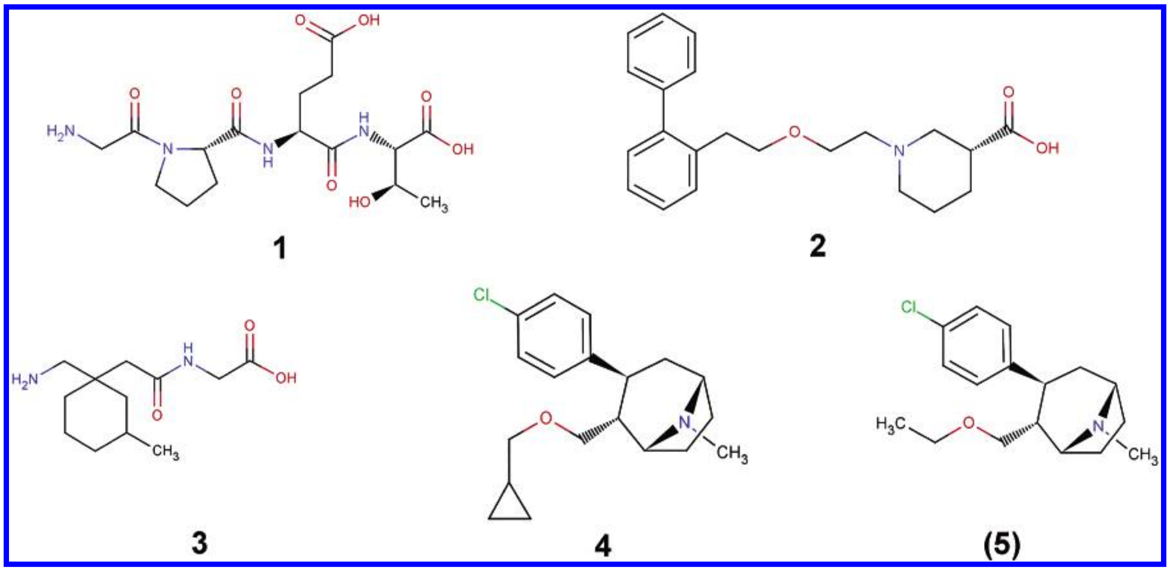

Figure 1. Some characteristic false negative structures from the cluster of Ward clusters of **1** in which the *smallest* number of molecules were successfully classified active by **M1**. **5** was correctly predicted as active.

Table 5. Summary of Ward Clustering of Data Set **1** and the Success Rate of Prediction by **M1** Grouped by Clusters

no. of clusters	**62**
No. of Compounds in Each Cluster	
minimum	45
maximum	165
median	80
Successfully Predicted (%)	
minimum	44
maximum	93
median	70
average	71

Table 6. Average Dissimilarities between the Active (Positive) and Nonactive (Negative) Molecules of the Train Set **1**

	mean	std dev	median
positive−positive	0.119	0.051	0.116
negative−negative	0.140	0.071	0.131
positive−negative	0.133	0.061	0.126

Euclidean distances were used that are the root sum-of-squares of differences of feature vectors. The feature vector of a molecule consisted of the same 39 atom-type descriptors that were used for building SVM models. Before applying the similarity calculations all the descriptors were scaled between 0 and 1, the same way as for SVM training. The dissimilarity distributions are shown for three cases: positive−positive, negative−negative, and positive−negative comparisons. All the dissimilarities were scaled between 0 and 1 for clarity. The results are summarized in Table 6 and are visualized as histograms in Figure 3. There is a remarkable similarity between the three histograms. It is clear that using these atom-type descriptors it is impossible to distinguish the antidepressants from other drugs using simple similarity comparisons. Almost all the members of the data sets are located between 0 and 0.3. Because the negative data set is a diverse and representative subset of druglike compounds, this low diversity shows that these descriptors compress the chemical space and can effectively be used for a large variety of (druglike) compounds.

Multiple-Target Classification of Antidepressants. Various analyses were done to examine the success of models in predicting antidepressants. The aims were to find out which type of structures can be predicted correctly, whether the classification capacity can be improved by using a consensus scoring of multiple target predictions, and how the models could help to select the best compounds for further evaluation.

The bases of analyses were the data set of antidepressants (**1**) and the 30 000-member negative set, for which activities were predicted by all 21 models.

First it was examined for which kind of structures the general model (**M1**) was unsuccessful to predict activities on data set **1**. Altogether 3871 true positives (t_p) and 1500 false negatives (f_n) were predicted. The similarity of t_p and f_n were compared by JChem using molecular fingerprint descriptors. It was found that the average similarity of molecules of f_n to their nearest neighbor in set t_p is 0.82. This relatively large similarity suggests that for most f_n molecules it is possible to find a t_p with very similar core structure. Other statistics were carried out on this data set. The compounds of **1** were clustered using a different clustering method than the one described in the Methods section (k-means clustering in the atom-type descriptor space). It was ward clustering using fingerprint descriptors by JChem. The number of optimal clusters was determined by the Kelly method,[21] and it was 62. The ratios of successfully predicted t_ps in each clusters were determined and are summarized in Table 5. The median and average success ratio for the 62 clusters (which contain the compounds of train set, too) is ca. 70%. The prediction for even the worst case was 44% that indicates that there was no cluster for which the model was entirely unsuccessful.

Figures 1 and 2 show some characteristic structures from the clusters with the worst and best success rate, respectively. Compound **5** of Figure 1 was successfully classified while **4** was not. Thus, the common core of the two molecules was found. This also shows the limits of using atom-type descriptors: because topological information is coded into the descriptors only in a limited way, this kind of small differences cannot be well treated. However, it is also true that the most important application of the method is finding new lead structures from a diverse set of molecules, for which

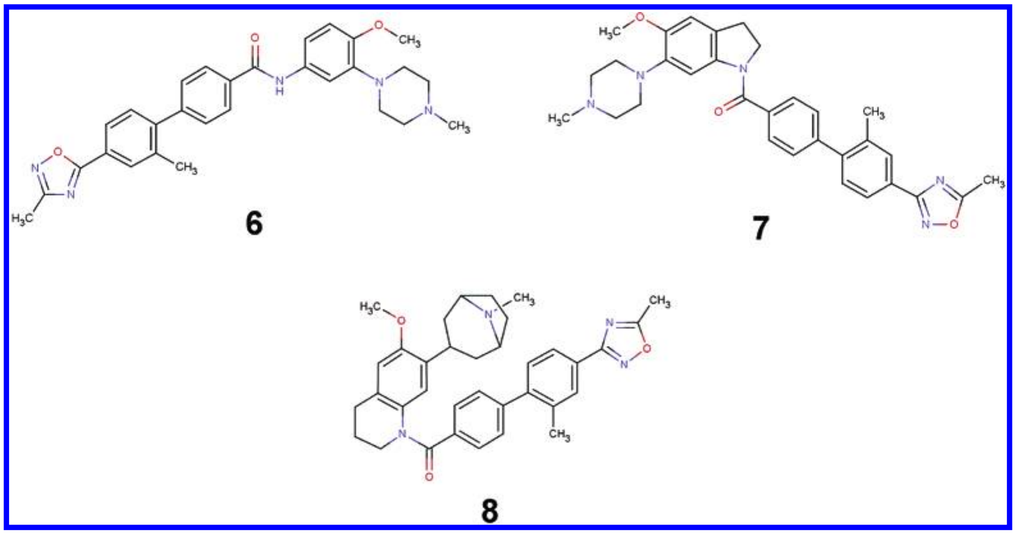

Figure 2. Some characteristic true positive structures from the cluster of Ward clusters of **1** in which the *largest* number of molecules were successfully classified active by **M1**.

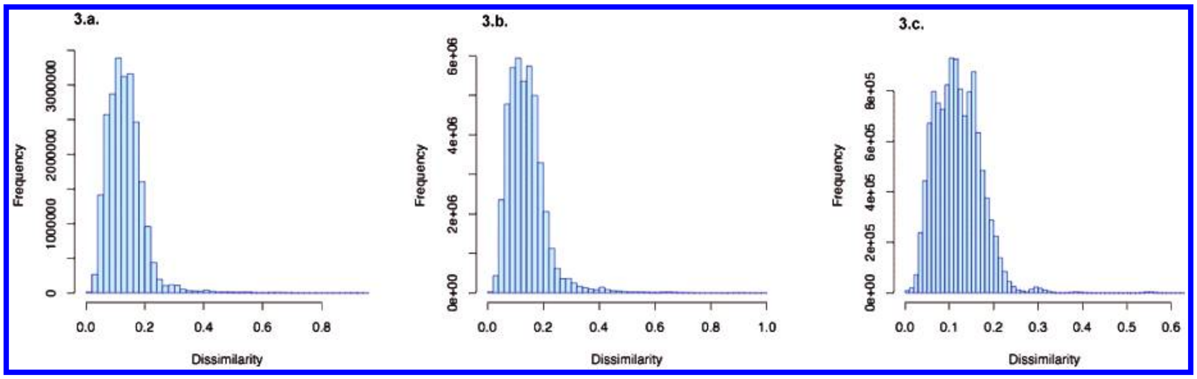

Figure 3. Histograms of dissimilarity distribution of the compounds of the train set **1** using atom-type descriptors. Dissimilarities between a. any one of actives vs any one of nonactives; b. between any two of negatives; and c. between any two of positives. (Note the different scales of y axes.)

Table 7. Summary of Consensus Scoring Using Multiple Models on Data Set **1** and Negatives

model	recall (%)		
	1	negatives	PR[c] for **1** (%)
M1	72.32	97.34	60.34
any[a]	86.57	86.48	46.31
M1 AND any[b]	55.57	98.17	46.96

[a,b] A compound was considered to be classified active, if any of the models (*a*) or **M1** plus at least one of the other models (*b*) predicted so. [c] Precision multiplied by recall.

it seems to be very successful. Optimizing the leads would be a following task, and for the smaller set of leads other methods or more specific descriptors could be used.

It is interesting that the average molecular weight of t_ps is 390 vs 377 of that of f_ns. If it is not an artifact, it might mean that "more elaborated" structures with larger molecular weight are easier to predict, because those are more typical to a given target.

Because not only the aforementioned general antidepressant model but also 20 other models were made, it is of interest to study if the inclusion of these models could improve the classification performance. The full set of **1** and the negatives were used for the evaluation. The results are in Table 7. Only three extreme cases were examined: (1) only **M1** were used to classify compounds, (2) all models were used, and the compounds classified positive by any of the models were considered active, and (3) only those hits were judged active which were predicted as so by **M1** and at least one of the other models.

The data reveal that the most optimal solution is still using **M1** only (although it should be noted that this comparison also included the train set of **M1**, and if the test set was used, only the results were worse (e.g. PR = 42 instead of 60). On the other hand, in the second case the recall of actives has drastically increased (together with false positives), while in the third case the number of false positives has decreased (so as the actives). It is up to the task to choose the approach that suits the requirements the best.

Having information from 21 targets can also be helpful to rank the compounds. The simplest way is counting how many times a molecule was a hit for a model. Some of the molecules are known to be active against more than one target.

As it is not likely that all the molecules in the database were biologically tested against all the possible targets, virtual screening could reveal if a compound might be active against other than the reported proteins in the database. Furthermore, it is also expected that an antidepressant active against one target will more likely be a false positive hit during virtual screening for another structurally similar protein, than a compound of no such activity. Thus, a molecule with many predicted hits (even if it is false positive in some cases) is

Table 8. Screening against Multiple Targets

	no. of hits[b]		% of all compds[c]		
hit/compd[a]	1	neg[e]	1	neg	av weight[d] 1
11	1	0	0.02	0.00	297.5
10	7	1	0.13	0.00	316.9
8	39	4	0.71	0.01	375.3
7	120	18	2.19	0.06	363.3
6	215	53	3.93	0.17	376.2
5	340	122	6.21	0.40	377.5
4	536	186	9.80	0.61	375.0
3	707	369	12.92	1.21	385.7
2	1388	740	25.37	2.43	392.5
1	1385	2627	25.31	8.62	403.4
0	734	26349	13.41	86.48	364.5
summary	5472	30469	100.00	100.00	366.2[e]

[a] Number of models that predicted a given molecule active. [b,c] The number (*b*) of molecules classified active by the given number of models, and the same as percent of the full set (*c*). [d] Average molecular weight of compounds belonging to the given group. [e] Average.

Table 9. Summary of Hits from SCD

hit/compd	no. of compds[a]	av mwgt[b]	with **M1** = 1[c] no.	%	most dsim[d]	av sim[e]
11	9	286.4	9	100.0	0.36	0.64
10	1	405.5	1	100.0	0.44	0.67
9	58	347.3	58	100.0	0.43	0.63
8	314	351.6	310	98.7	0.42	0.58
7	1019	336.4	919	90.2	0.42	0.55
6	2560	338.6	2278	89.0	0.42	0.48
5	4931	338.2	3854	78.2	0.41	0.27
4	9266	350.1	6005	64.8	0.41	0.27
3	18160	360.4	9661	53.2	0.40	0.38
2	49357	371.1	19244	39.0	0.41	0.31
1	231000	403.0	24827	10.7	0.41	0.30
sum[f] or av[g]	316675[f]	352.1[g]	67166[f]	65.1[g]	0.41[g]	0.46[g]

[a] Number of compounds. [b] Average molweight. [c] Only those molecules which were predicted active by **M1**, too. [d] Similarity of that compound to its nearest neighbor in data set **1** for which this value is the smallest (i.e., most dissimilar to its nearest neighbor). [e] Average similarity. [f] Sum of rows. [g] Average (mean) of rows.

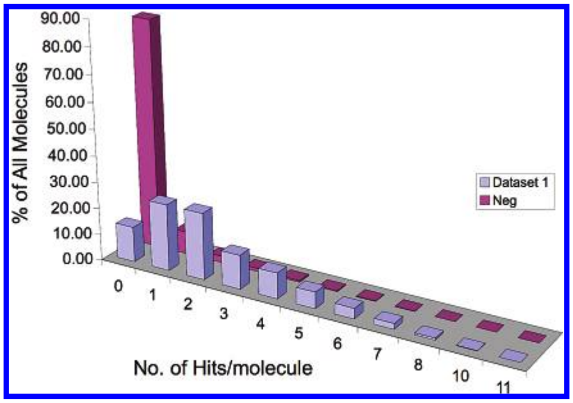

Figure 4. Comparing molecules of **1** and the negative set by the number of molecules that were hits by a given number models (see also Table 8).

presumably a better candidate to follow up with additional analyses. Another advantage of this compound is that it might be a "smallest common denominator"-like structure for many targets, thus during the later phases it can later be tailored to exhibit the required activity(ies).

Table 8 and Figure 4 show the number of compounds of data sets **1** and neg, which were positive hits by a given number of models. The most "successful" molecule was predicted to be active in up to 11 cases. Obviously most (86%) of the negative set was not classified positive by any of the models. However, some of those were predicted positive for more than one target. It can be expected that the large negative set contains some active molecules, which were not yet identified as an antidepressant; especially the ones with many positive results are very likely to belong into this group.

Figure 2 shows the distribution of how many times a compound was positive vs the number of compounds of the aforementioned two data sets. The tendency of the two curves is clearly different. The one of negatives exponentially decreases with most data points belonging to the category of zero hit. On the other hand the graph of antidepressants peaks at above 2 (with the average being 2.31 hit/compound vs 0.23 of the former set) and declines less steeply. As there is no correlation between any two of the models (based on the results for **1**, Table S1 in the Supporting Information), the large umber of multihit per compounds cannot be interpreted as an artifact of the method. It might rather mean that due to the similarity of some of the targets, most antidepressants develop activity through multiple ways.

Table 8 summarizes the average molecular weight of **1** for each group. Although the numbers of samples are statistically insignificant for compounds of 8−11 hits, for the other data points a small decrement in average molecular weights can be observed. The same tendency was noticed for the actives of SCD database (discussed later), thus it can be suggested that that a larger (more "elaborated" or optimized) structure is necessary for a selective hit and the molecules with many hits have smaller common structure that might function better as a lead compound.

Virtual Screening of SCD. Some attempts were made to use the developed models in some screening tasks. One of these was the use of MDL Screening Compounds Database (SCD). Altogether about 3.2 million molecules were selected and classified from this database.

One of the advantages of using SVM and atom-type descriptors is the speed of prediction. The calculation speed for SCD and for the 21 targets *altogether* using one standard 3GHz PC CPU was roughly 100 000 data points per hour. Thus, the average speed for one target was ca. 2 million cases/h. This speed makes even the HTS evaluation of large virtual libraries possible.

The results are summarized in Table 9. It can be concluded that the reverse proportion of average molecular weight with the number of hits/molecules is even more obvious (and statistically significant) than in the case of set **1**. The maximum hit was 11 and could be achieved for 9 molecules. If a stricter criterion was used—**M1** should be positive, the number of hits decreased to roughly one-fifth, with a ratio decreasing with the number of hits/compounds i.e., the compounds that were hits for more targets were more possibly to be predicted active by **M1**, too. A very important advantage of using only atom-type descriptors is that the method can be applied to a diverse set of compounds. As it is shown in Table 9, where the similarity of those molecules of SCD that were predicted active by **M1**, are compared to reference set **1** using molecular fingerprints. Two values are calculated for each group: 1 the similarity of that SCD compound to its nearest neighbor, which is the most dissimilar to any of the members of reference set; 2 the average similarities of the given group of hits to all the molecules of **1**. The first statistics reveal the most dissimilar structure that can be found by the method. For example among the 9 molecules with 11 hits the most dissimilar one to set **1** had a similarity value of 0.36 to its nearest neighbor in **1**. Thus, if we are to screen SCD by means of molecular fingerprints and using the over 5000 members of **1** as seed compounds, we should set the similarity criterion to 0.35 in order to find that given structure. This is significantly beyond the generally used value of 0.85.[22] Usually a more dissimilar lead compound is more desirable, so using SVM with atom-type descriptors has an edge over methods based on a similarity search by fingerprints in this regard, and because of its speed, it is more suitable for high throughput screening than most 2D or 3D methods which are usually used for obtaining diverse lead structures.[23−26]

CONCLUSIONS

Screening models were developed for 21 biological targets related to depression, by using computational learning method (support vector machines) to classify between active and nonactive chemical structures characterized by atom-type descriptors, alone. Altogether about 10 000 positive and 30 000 negative examples were used from MDDR which was used for making the models, with one of the data sets (**1**) being a collection of about 5500 diverse molecules of antidepressant activity, regardless of the mode of effect.

Despite the identical method with the same descriptors for all the cases, the models showed satisfactory prediction ability by realizing 0.2−0.8 fraction of the theoretical enrichment for a given data set of external test sets. Recall values for positive examples were between 45 and 90% range with recall values above 95% in all cases for the negative sets. In the case of **1** the ratios were 59% and 96% with a precision value of 0.72, respectively, which is a remarkably good result given the size and diversity of the data set. It suggests that separation of the chemical space of antidepressants, in general, from other drugs might be possible. It should be noted that the success rates could be further improved by selecting optimal sets among descriptors specifically to a given target (called feature selection).

A number of molecules were found to be active by more than just one of the models. Averaging the number of models for which a compound was hit over all members of **1** the value of 2.31 was obtained. In opposite, the same value for the negative set was 0.23. A tendency could be observed that the average molecular weight of those molecules that are classified active by a larger number of models is smaller than those with less predicted activity. Thus, using a number of antidepressant targets for virtual screening, it is possible to find less complex lead molecules with structures common to many of these targets.

Because of its speed and the high diversity of predicted actives the method seems adequate for finding new lead molecules by virtual screening of large compound libraries.

This study is one of the growing evidences about the usefulness of support vector machines in virtual screening. Using simple atom-type descriptors, such models could be made that cover large chemical spaces, suggesting the possibility of making more general filters to a given therapeutic area than before.

ACKNOWLEDGMENT

This work was supported by the 21st Century COE Program, Human Nutritional Science on Stress Control, Tokushima, Japan.

Supporting Information Available: Cross correlations matrix for the 21 models based on data set **1** (Table S1), an sdf file containing all the structures of data set **1** with included predicted categories by all models (abbreviations for target names in Table S1), and the full train and test sets of **1** (Table S2 and S3, respectively). This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES AND NOTES

(1) Adell, A.; Castro, E.; Celada, P.; Bortolozzi, A.; Pazos, A.; Artigas, F. Strategies for producing faster acting antidepressants. *Drug Discovery Today* **2005**, *10* (8), 578−585.

(2) Niwa, T. Prediction of Biological Targets Using Probabilistic Neural Networks and Atom-Type Descriptors. *J. Med. Chem.* **2004**, *47* (10), 2645−2650.

(3) Vapnik, V. N. *Statistical Learning Theory*; John Wiley & Sons: 1998.

(4) Chen, N.; Lu, W.; Li, J. Y. G. *Support Vector Machine In Chemistry*; World Scientific Pub Co. Inc.: 2004.

(5) Jorissen, R. N.; Gilson, M. K.; Virtual Screening of Molecular Databases Using a Support Vector Machine. *J. Chem. Inf. Model* **2005**, *45* (3), 549−561.

(6) Gasteiger, J. *Handbook of Chemoinformatics: From Data to Knowledge*; Wiley: 2003.

(7) *MDL Drug Data Report 2004 and Screening Compounds Directory 2005*; Elsevier MDL: Hayward, CA.

(8) Chang, C.-C.; Lin, C.-J. *LIBSVM: a library for support vector machines*; 2001. Software available at http://www.csie.ntu.edu.tw/∼cjlin/lib.

(9) JChem 3.0; ChemAxon Ltd., Budapest, Hungary.

(10) XLSTAT-Pro 7.5; Addinsoft, Brooklyn, NY.

(11) R: A Language and Environment for Statistical Computing. http://www. R-project.org.

(12) Kecman, V. *Learning and Soft Computing*; The MIT Press: 2001.

(13) Kearns, M. S.; Solla, S. A.; Cohn, D. A. *Advances in Neural Information Processing Systems*, 10&11; MIT Press: 1998.

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

(15) Liu, H. X.; Zhang, R. S.; Yao, X. J.; Liu, M. C.; Hu, Z. D.; Fan, B. T. Prediction of the Isoelectric Point of an Amino Acid Based on GA-PLS and SVMs. *J. Chem. Inf. Comput. Sci.* **2004**, *1*, 161−167.

(16) Sorich, M. J.; Miners, J. O.; McKinnon, R. A.; Winkler, D. A.; Burden, F. R.; Smith, P. A. Comparison of Linear and Nonlinear Classification Algorithms for the Prediction of Drug and Chemical Metabolism by Human UDP-Glucuronosyltransferase Isoforms. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2019−2024.

(17) 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*, 1855−1859.

(18) Yao, J.; Panaye, A.; Doucet, J. P.; Zhang, R. S.; Chen, H. F.; Liu, M. C.; Hu, Z. D.; Fan, B. T. Comparative Study of QSAR/QSPR Correlations Using Support Vector Machines, Radial Basis Function Neural Networks, and Multiple Linear Regression. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1257−1266.

(19) Zernov, V. V.; Balakin, K. V.; Ivaschenko, A. A.; Savchuk, N. P.; Pletnev, I. V. Drug discovery Using Support Vector Machines. The Case Studies of Drug-likeness, Agrochemical-likeness, and Enzyme Inhibition Predictions. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 2048−2056.

(20) Merkwirth, C.; Mauser, H.; Schulz-Gasch, T.; Roche, O.; Stahl, M.; Lengauer, T. Ensemble Methods for Classification in Cheminformatics. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1971−1978.

(21) Kelley, L. A.; Gardner, S. P.; Sutcliffe, M. J. An automated approach for clustering an ensemble of NMR-derived protein structures into conformationally related subfamilies. *Protein Eng.* **1996**, *9*, 1063−1065.

(22) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002**, *45* (19), 4350−4358.

(23) Thimm, M.; Goede, A.; Hougardy, S.; Preissner, R. Comparison of 2*D* Similarity and 3*D* Superposition. Application to Searching a Conformational Drug Database. *J. Chem. Inf. Comput. Sci.* **2004**, *44* (5), 1816−1822.

(24) Willett, P. Searching Techniques for Databases of Two- and Three-Dimensional Chemical Structures. *J. Med. Chem.* **2005**, *48* (13), 4183–4199.

(25) Klebe, G. *Virtual Screening: An Alternative or Complement to High Throughput Screening*? Kluwer Academic Publishers: 2000.

(26) Bohm, H.-J.; Schneider, G. *Virtual Screening for Bioactive Molecules*; Wiley-VCH: 2000.

CI050301Y