Assessing Different Classification Methods for Virtual Screening

Dariusz Plewczynski,*[,†,‡] Stéphane A. H. Spieser,[§] and Uwe Koch*[,§]

BioInfoBank Institute, Limanowskiego 24A/16, 60-744 Poznan, Poland, Interdisciplinary Centre for Mathematical and Computational Modeling, University of Warsaw, Warsaw, Poland, and Department of Chemistry, Istituto di Ricerche di Biologia Molecolare P. Angeletti, Merck Research Laboratories, Rome, Via Pontina km 30600, Pomezia 00040, Italy

Received November 28, 2005

How well do different classification methods perform in selecting the ligands of a protein target out of large compound collections not used to train the model? Support vector machines, random forest, artificial neural networks, *k*-nearest-neighbor classification with genetic-algorithm-optimized feature selection, trend vectors, naïve Bayesian classification, and decision tree were used to divide databases into molecules predicted to be active and those predicted to be inactive. Training and predicted activities were treated as binary. The database was generated for the ligands of five different biological targets which have been the object of intense drug discovery efforts: HIV-reverse transcriptase, COX2, dihydrofolate reductase, estrogen receptor, and thrombin. We report significant differences in the performance of the methods independent of the biological target and compound class. Different methods can have different applications; some provide particularly high enrichment, others are strong in retrieving the maximum number of actives. We also show that these methods do surprisingly well in predicting recently published ligands of a target on the basis of initial leads and that a combination of the results of different methods in certain cases can improve results compared to the most consistent method.

INTRODUCTION

High-throughput screening (HTS) is the method of choice for identifying novel ligands of biological targets. Virtual screening is the name given to a range of computational methods that are used to prioritize the biological testing of large chemical data sets. HTS results are not free from errors, and different assay formats for the same target can give different results.[1,2] In this context, virtual screening is employed to analyze HTS results and support prioritization and the identification of false positives and negatives.[3] Virtual screening is also used to prioritize compounds for testing, for example, to select nonproprietary compounds which have to be bought from external vendors.[4] In this article, seven different methods are compared in terms of their capability to divide a database into molecules predicted to be active and those predicted to be inactive. In regard to this classification problem, we were interested in comparing the methods in terms of their capability to enrich actives in the subset of compounds predicted to be active and to what degree they are capable of retrieving as many actives as possible. The methods compared in this article are support vector machines (SVM), artificial neural networks (ANN), naïve Bayesian classification (NB), *k*-nearest-neighbor with genetic-algorithm-optimized feature selection (kNN), random forest (RF), decision tree (DT), and trend vectors (TV), which will be discussed in more detail later.

The choice of the method for virtual screening depends on what is known about the target and its ligands. First, if a single active molecule is known, then similarity searching can be used.[5] Second, pharmacophore mapping can be applied if several actives have been identified to ascertain common patterns of features that may be responsible for the observed activity.[6] Third, in particular for heterogeneous sets of actives, the actives can be used as training data for a machine-learning system.[7−15] Finally, if the 3D structure of the biological target is known, a docking study can be carried out to identify molecules that are complementary to the binding site.[16,17]

In this report, the focus is on the third class of virtual screening methods. Seven different methods are applied to a classification problem regarding the selection of actives for different target proteins. To cover a wide range of possible structure−activity relationships, five protein targets were selected which are unrelated in terms of their function and do not show structural similarity between their ligands. For all five targets, dihydrofolatereductase (DH), cyclooxygenase-2 (COX2), thrombin (TH), HIV reverse transcriptase (RT), and the estrogen receptor (AE), many ligands are known and some are marketed as drugs. For each target protein, four data sets were generated, allowing a study of the performance of each method on data sets containing a different number of actives. Each data set was divided into one set for training and one for testing the model; input activities were provided as binaries. In two data sets, the active compounds in the training collection were chosen as the ones published first, and in the test set, those published later were chosen. This selection allows the use of quantitative structure−activity relationship (QSAR) methods in a

* Corresponding authors. Tel.: +48-61-8653520 (D.P.), +390691093644 (U.K.). E-mail: darman@bioinfo.pl (D.P.), uwe_koch@merck.com (U.K.).
† BioInfoBank Institute.
‡ Interdisciplinary Centre for Mathematical and Computational Modeling, University of Warsaw.
§ Department of Chemistry. Istituto di Ricerche di Biologia Molecolare P. Angeletti.

10.1021/ci050519k CCC: $33.50
Published on Web 04/27/2006

Table 1. Number of Compounds Constituting the Training and Test Sets for Each Protein Target[a]

protein target	COX2	DH	TH	RT	AE
		VA			
train (positive/negative)	75/2106	17/2151	77/2032	80/2353	23/2353
test (positive/negative)	37/8346	11/8390	35/8423	35 /8323	12/9053
		AA			
train (positive/negative)	608/2106	120/2151	797/2032	426/2353	170/2353
test (positive /negative)	296/8346	62/8390	357/8423	235/8323	85/9053

[a] Validated actives (VA) annotated as (MDDR): "launched", "Phase III", "Phase II", and "Preclinical". All actives (AA) also include those annotated as "biologically tested" for the target.

Table 2. Prediction of Recently Published Compounds on the Basis of Earlier Published Compounds[a]

protein target	COX2	DH	TH	RT	AE
		Set I			
train (positive/negative)	36/2106	12/2151	34/2032	34/2353	13/2353
test (positive/negative)	77/8346	16/8390	78/8423	81 /8323	22/9053
		Set II			
train (positive/negative)	77/2106	16/2151	78/2032	81/2353	22/2353
test (positive /negative)	36/8346	12/8390	34/8423	34/8323	13/9053

[a] All compounds were taken from the VA set. Set I: one-third of the compounds published first used to generate a model to predict the later two-thirds of the compounds. Set II: two-thirds of the compounds published first used to generate a model to predict the later one-third of the compounds.

predictive sense to study whether recently developed ligands of a target can be predicted on the basis of earlier compounds. Finally, we studied whether the combination of results from different methods improves the performance of the classification. In each case, the predicted activities were obtained as binaries.

COMPUTATIONAL METHODS

Data Sets and Descriptors. The set of compounds to which we applied the machine-learning algorithms was extracted from the 2004 edition of the MDDR (MDL Drug Data Report).[18] For each target protein, two sets of ligands were extracted. A larger set contains all ligands of the target according to the annotation in the MDDR. For the second, smaller set, we exclude those compounds which are annotated as "biologically tested" and include only ligands which have gone beyond the first step of drug discovery. These are annotated as "launched", "Phase III", "Phase II", and "Preclinical". From here on, we refer to the larger "all actives" set with the label AA and to the smaller set of validated actives with the label VA. These compounds represent the positives or actives used in training or testing. In each case, the actives were chosen irrespective of eventual differences in their modes of binding. Thus, RT actives contain non-nucleosidic allosteric inhibitors as well as nucleosides, and AE actives are comprised of agonists and antagonists. The negatives or inactives are the compounds which have not been annotated as a ligand of this target. In addition, compounds annotated as "biologically tested" were excluded from the list of inactives. All activities, input and predicted, are treated as binary without considering compound-related scores.

Because the number of negatives is larger than the number of positives, we have tried to rebalance the data set to some extent by selecting a relatively larger number of actives and a smaller number of nonactives in the training set than that in the test set. The compounds were divided into training and test sets by randomly selecting two-thirds of the actives and one-third of the nonactives for training and one-third of the actives and two-thirds of the nonactives for testing (Table 1). To test whether the random selection of compounds for training and testing creates a bias, we repeated the selection 25 times and applied the machine-learning algorithms to each set. We did not detect any significant difference between the results for the data sets (vide infra).

In another experiment, we chose the active compounds for the training and test sets not randomly but according to the date of their first publication. The actives were taken from the set of validated actives. Two sets were created, one in which the oldest third was used for training and the newer two-thirds for the test set and one in which we choose the oldest two-thirds for training and the more recent third for testing (Table 2).

We have utilized the regular atom pair (AP) descriptors[19,20] because of their proven success in classifying compounds, ease of use, and interpretability.

Classification and Model Validation. There are many ways to present the performance of a classifier. We use here precision (P) and recall (R) values. In some cases, we also tabulate the enrichment factor (EF). The definitions are given below:

$$E = \frac{\mathrm{fp} + \mathrm{fn}}{\mathrm{tp} + \mathrm{fp} + \mathrm{tn} + \mathrm{fn}} \times 100\% \quad (1)$$

$$R = \frac{\mathrm{tp}}{\mathrm{tp} + \mathrm{fn}} \times 100\% \quad (2)$$

$$P = \frac{\mathrm{tp}}{\mathrm{tp} + \mathrm{fp}} \times 100\% \quad (3)$$

$$\mathrm{EF} = \frac{\mathrm{tp}/(\mathrm{tp} + \mathrm{fp})}{(\mathrm{tp} + \mathrm{fn})/(\mathrm{tp} + \mathrm{fp} + \mathrm{tn} + \mathrm{fn})} \times 100\% \quad (4)$$

where tp is the number of true positives, fp is the number of false positives, tn is the number of true negatives, and fn is the number of false negatives. The classification error E

provides an overall error measure, whereas recall R measures the percentage of correct predictions, and precision P gives the percentage of observed positives that are correctly predicted (the measure of the reliability of positive instances prediction). The enrichment factor describes to what degree true positives are over-represented in the selected data set compared to the whole set of compounds.

Machine-Learning Methods. We have applied to our data several QSAR methods, including SVM, ANN, NB, kNN, RF, DT, and TV. A mathematical description of the algorithms employed in this study is beyond the scope of this paper. We therefore briefly outline each method. For more details, we refer to the references given in each of the following paragraphs.

The basic concept of the SVM algorithm[21,22] is, first, to project the input data vectors representing descriptors to a high dimensional feature space. This mapping is described by the kernel function, which is, in our case, the simple linear one. The algorithm then constructs the optimal hyperplane separating two sets of positives and negatives. For each test case, descriptor vectors are first mapped into the same feature space, and the class membership of these instances is predicted using the hyperplane. The resulting prediction score measures the distance between the border and a point representing a compound. A higher score indicates a higher confidence of the predictions. In this study, we employed the freely available LIBSVM software[23] with the linear kernel. The penalty parameter for the error term was determined by a 5-fold leave-half-out cross validation.

Another machine-learning algorithm we have used is ANN.[24−27] The network is comprised of three layers (input, hidden, and single-output neuron). This is an ensemble artificial neural network method. We use 100 neural nets. The input parameters are the principal components of the original descriptors. There are three nodes in the hidden layer. Our implementation trains the ANN on 75% of the data and stops training when the error on the remaining 25% is minimized.[28] The 75/25 split is done randomly and repeated $N = 100$ times. The final model is then averaged over N models.[29]

We also used the (kNN[25] method that subdivides a set of input cases (characterized by the vectors of descriptors) into different classes. kNN predicts a classification for test cases on the basis of the majority voting of its k nearest neighbors in the feature space. In our implementation, $k = 5$ is used.[30] We used the Euclidian metric for calculating distances and the same set of descriptors used for other methods. The most discriminatory descriptors are calculated using a genetic algorithm with four generations and 40 chromosomes.[30,31]

The NB method is a simple classification method based on the Bayes rule for conditional probability. NB may be used in high-throughput screening data analysis as a simple classifier between "active" and "inactive" compounds. It is guided by the frequency of occurrence of various molecular descriptors in a training set. NB classification is based on two core assumptions. First, the descriptors in the training samples are equally important. Second, it assumes independence of the descriptors from each other. Therefore, their combined probability is obtained by multiplying the individual probabilities. These two assumptions are often violated, and full independence of the descriptors is rarely observed. However, NB is very robust to violations of these assumptions and tolerant toward noise.[32]

Trend vector analysis in combination with topological descriptors such as AP has proved useful in drug discovery for ranking large collections of chemical compounds in order of predicted biological activity.[33] TV is a samples-based version of partial least squares based on the absence or presence of AP descriptors. The activity of an unknown compound could be predicted from its descriptors by calculating the trend vector. The trend vector is the first moment of the activity in the space of the descriptors. A trend vector represents the one-dimensional array of correlations between the biological activity of interest and a set of properties or "descriptors" of compounds in a training set. The trend vector in the space of the descriptors points in the general direction of highest activity. Our implementation is a sample-based version of partial least squares based on the presence or absence of substructure descriptors.

A decision tree takes as input an object or situation described by a set of properties and outputs a yes or no decision. Decision trees, therefore, represent Boolean functions, even in the case of a larger range of outputs.[34] The version used here is Quinlan's C4.5.[35] A decision tree is an arrangement of tests that prescribes an appropriate test at every step in an analysis. In general, decision trees represent a disjunction of conjunctions of constraints on the attribute values of instances. Each path from the tree root to a leaf corresponds to a conjunction of attribute tests and the tree itself to a disjunction of these conjunctions. More specifically, decision trees classify instances by sorting them down the tree from the root node to some leaf node, which provides the classification of the instance. Each node in the tree specifies a test of some attribute of the instance, and each branch descending from that node corresponds to one of the possible values for this attribute. An instance is classified by starting at the root node of the decision tree, testing the attribute specified by this node, and then moving down the tree branch corresponding to the value of the attribute. This process is then repeated at the node on this branch and so on until a leaf node is reached.

Random forests grow many classification trees,[36,37] 100 in this application. Each tree is constructed on a bootstrap sample of the data set. To classify a new object from an input vector, this algorithm applies the input vector to each tree of the forest. Each tree gives a classification, and the tree "votes" for that class. The forest chooses the classification having the most votes (over all of the trees in the forest). The forest error rate depends on the correlation between any two trees in the forest (increasing the correlation increases the forest error rate) and the strength of each individual tree in the forest (a tree with a low error rate is a strong classifier, and increasing the strength of the individual trees decreases the forest error rate). The random forest can handle thousands of input variables without variable selection and gives estimates of what variables are important in the classification. Random forest generates an internal unbiased estimate of the generalization error as the forest building progresses and has an effective method for estimating missing data and maintains accuracy when a large proportion of the data is missing. Random forest does not overfit and is fast.

Table 3. Recall (*R*) and Precision (*P*) Values for the Models Generated for Each Target (Training) Considering the VA Set[a]

	COX2		DH		TH		RT		AE	
protein target	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*
SVM(avg)	82.68	99.49	100	99.8	98.36	100	94.19	98.96	91.30	96.49
SVM(stdev)	11.94	0.08	0.0	1.0	1.48	0.0	9.31	1.18	9.65	3.9
RF	100	100	100	100	100	100	100	100	99	100
ANN	39	97	71	92	17	100	1	100	5	100
kNN	59	92	53	90	87	92	58	92	73	76
TV	53	98	41	100	79	97	18	93	*b*	*b*
NB	55	82	100	22	17	93	30	83	*b*	*b*
DT	64	98	100	100	94	99	50	93	77	85

[a] For SVM, the average values SVM(avg) and standard deviations SVM(stdev) obtained from calculations using 25 different, randomly selected training sets are shown. [b] No model was produced.

Table 4. Recall (*R*) and Precision (*P*) Values for the Models Generated for Each Target (Training) Considering the AA Set for Each Target[a]

	COX2		DH		TH		RT		AE	
protein target	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*
SVM(avg)	97.29	99.76	99.67	99.08	99.39	99.68	94.61	99.54	96.33	97.50
SVM(stdev)	1.72	0.20	0.58	0.76	0.25	0.22	1.16	0.31	2.74	1.49
RF	100	100	100	100	100	100	100	100	100	100
ANN	85	98	96	96	97	98	75	98	74	98
kNN	94	92	97	94	97	97	87	88	95	91
TV	73	100	82	98	86	98	55	98	37	100
NB	78	85	97	60	58	95	45	84	24	72
DT	91	99	98	95	96	96	56	99	93	96

[a] For SVM, the average values SVM(avg) and standard deviations SVM(stdev) obtained from calculations using 25 different, randomly selected training sets are shown.

RESULTS AND DISCUSSION

We were particularly interested in the performance of different classification methods in the context of virtual screening using the same data set and molecular descriptors. Models for each method were generated for different training sets of compounds and then applied to a separate test set. We were interested in studying whether there are methods which consistently perform better on most targets and test sets; whether different methods have different strengths, one better in reducing the number of false positives and the other in reducing the number of false negatives; and, finally, whether a combination of methods improves results. We use the enrichment factor (EF) referring to the group of compounds classified as active, the recall value (*R*) to describe the percentage of actives found by the algorithm, and the precision value (*P*) for the percentage of nonactives classified as actives.

Training. Both SVM and RF perform particularly well in the training procedure on all targets. Both methods show recall and precision values either close or equal to 100%. As regards training, both methods perform equally well on the data set with a small number of actives (Table 3) and on the data set with a larger number of actives (Table 4). Tables 3 and 4 include, as well, average and standard deviations for 25 randomly selected data sets for each target in the case of SVM.

The other methods also predict, in most cases, a small number of false positives and negatives if applied to the training set. The results are better for the more balanced data set with a larger number of actives. For some of the methods, such as TV, NB, and DT, we observe target-dependent variations.

Retrieval of Compounds against a Large, Diverse Collection. The models obtained by training the various methods were applied to large, diverse collections of compounds which are not part of the training data set. Two test sets were used. For the VA set, the number of actives for each target is less than 1%. For the larger AA set, the number of actives is less than 5% for each target.

For the AA set of compounds, the highest enrichment factors are observed for TV and RF for each target (Figure 1a). SVM is placed third in four of the five targets, followed by the other methods. Significant differences in the enrichment factors are observed for different targets. Most methods perform particularly well on DH and AE. Although the study of structure−activity relationships is not within the scope of this paper, we speculate that ligands of DH and AE contain a subset of compounds with particular features, such as steroids, which render them more easily distinguishable from the bulk of nonactives. Because, in this study, the sample size is not kept fixed but allowed to vary according to the classification, high enrichment factors can be achieved with a high number of false negatives if the sample size is small (Table 5). For example, TV, yielding high enrichment factors, predicts in many cases a smaller number of compounds to be active than do other methods (Table 5). For this reason, recall values were also calculated, which give the percentage of all actives which have been retrieved (Figure 1b). SVM and kNN retrieve, on average, more than 90% of the active compounds, and RF and DT retrieve more than 80%. TV and ANN perform less well on this measure. The target dependence of the recall values is small for all methods except TV and NB. It is noteworthy that the performance in terms of the enrichment factor and recall for the same method can be quite different. This is most obvious for TV, which achieves high enrichment factors but often does not perform very well in terms of recall. Accordingly, the number of compounds to be acquired on the basis of a selection with TV is the smallest, yielding a high enrichment but a relatively poor recall.

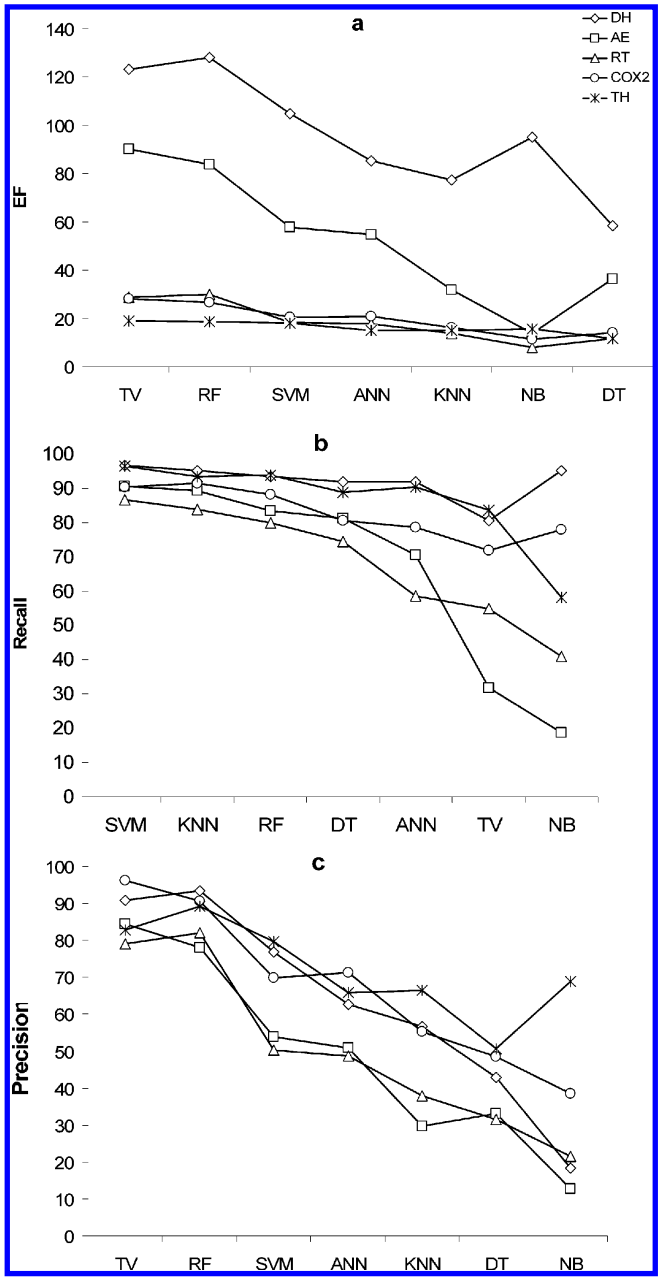

Figure 1. Enrichment factor (a), recall (b), and precision (c) for the AA set, different methods and targets.

To complete the analysis, we give the percentage of true actives in the set of compounds predicted to be active by each method (precision, Figure 1c). Here again, TV and RF perform the best, giving less than 20% false positives (precision 80%). Also, SVM (average precision 66%) and ANN perform well, whereas kNN, DT, and NB each predict more than 50% false positives.

This analysis shows that different methods have different strengths. If the priority is on acquiring a small number of compounds with as high an enrichment factor as possible and a high number of false negatives is not considered to be a problem, TV is the method of choice. If the size of the set of samples is not of major concern but one wants to retrieve actives as completely as possible, SVM, kNN, or RF should be chosen. SVM and RF offer the additional advantage of high precision, that is, a small number of false positives in the data set predicted to be active.

Table 5. Size of the Sample of Compounds Classified to Be Active by Each Method as a Percentage of the Whole Test Set for Each Target

	COX2	DH	TH	RT	AE
		VA Set			
observed actives	0.4	0.1	0.4	0.4	0.1
SVM	0.8	0.2	0.7	1.1	0.3
RF	0.4	0.02	0.2	0.3	0.4
kNN	0.8	0.06	0.8	0.7	0.4
TV	0.2	0.02	0.5	0.2	*a*
ANN	0.01	0.01	0.04	0.01	0.0
NB	1.4	2.3	0.3	0.6	*a*
DT	0.7	0.4	1.4	0.9	0.5
		AA Set			
observed actives	3.5	0.7	4.2	2.8	0.9
SVM	4.5	0.9	5.4	4.9	1.6
RF	3.4	0.7	5.1	2.7	1.0
kNN	5.8	1.2	6.3	6.3	2.9
TV	2.6	0.7	4.5	1.9	0.4
ANN	3.8	1.1	6.2	3.4	1.3
NB	7.3	3.9	3.8	5.5	1.4
DT	5.9	1.6	8.1	6.8	2.3

[a] Not applicable because no model was produced.

The results for the VA set in both training and testing must be interpreted with care because of the statistical effect small variations in the prediction can have. The results for DH and AE were not included in this analysis because of the small number of validated actives for this target (28 and 34, respectively).

As before, large enrichments are observed for each target (Figure 2a). On all three measures, the relative performance of the different methods is similar to what we observed for the first data set. The best performing methods are TV, ANN, and RF, with enrichments factors larger than 100, followed by SVM and kNN, with EF values above 80. Also, in the case of the recall values, the same relative ranking of methods is observed as that for the data set with a larger number of actives. It is remarkable that SVM and kNN are capable of retrieving, on average, more than 60% of the actives from data sets with less than 0.5% active molecules, and RF and DT are close to 50%. In terms of precision, RF performs the best, with on average less than 35% false positives, followed by TV and SVM.

Finally, we tried to retrieve recently published active compounds on the basis of models generated from older compounds. This reflects a situation common in drug discovery, where some initial leads are known and one wants to retrieve active analogues. In this case, the VA set was ordered according to the year of the compounds' first publication. Two data sets were created with, in one case, the oldest third and, in the other, the oldest two-thirds included in the training data set. Because SVM had the highest recall values for each target using the VA set and RF had the highest precision score for three out of four targets, we decided to focus on the performance of these two methods. The models developed for this data set were applied to a test set containing as actives the more recently published compounds.

Again, the numbers of actives in both, the data sets for training and testing, are very small, calling for a cautious interpretation of the data. For both methods, SVM and RF, high enrichment factors are observed for each target (Table 6). For RF, enrichment is larger than for SVM in most cases.

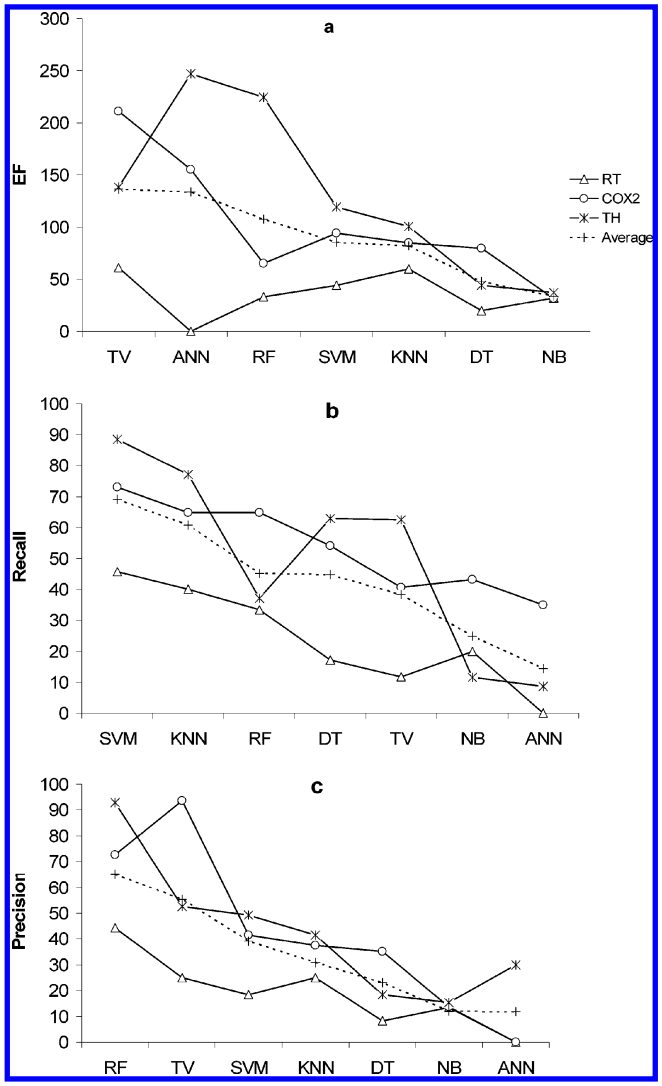

Figure 2. Enrichment factor (a), recall (b), and precision (c) for the VA set, different methods and targets.

Table 6. Percentage of Compounds Predicted to Be Active (PA), Enrichment Factor (EF), Recall and Precision (%) for SVM, and RF Applied to the Prediction of the Most Recently Published Third or Two-Thirds of Ligands Using a Model Based on Ligands Published Earlier

	COX2		DH		TH		RT	
protein target	1/3	2/3	1/3	2/3	1/3	2/3	1/3	2/3
PA(SVM)	0.9	0.6	0.2	0.2	0.5	0.4	1.0	1.2
PA(RF)	0.3	0.4	0.01	0.05	0.4	0.08	0.3	0.3
EF SVM	68	122	233	432	122	74	47	42
EF RF	100	125	525	1050	96	139	607	129
recall SVM	64	74	50	88	33	26	46	50
recall RF	27	49	6	50	8	18	20	40
precision SVM	66	45	44	41	66	43	33	15
precision RF	96	91	100	100	86	82	46	46

However, as the recall values show, this is mainly due to the fact that RF predicts a smaller number of compounds to be active. Remarkably, with the exception of thrombin, SVM is capable of identifying more, or close to, 50% of the actives, whereas on this measure, RF does not perform as well (Table 6). RF, however, identifies only a small number of false positives.

This example confirms that useful models can be generated on the basis of a very small number of actives. This is relevant for drug discovery where, at the beginning of the search for ligands of a new target, only very limited information is available. Moreover, at least for these targets, with the possible exception of TH, compounds discovered early are sufficiently similar to later compounds, often patented by competitors, to allow the generation of models capable of recognizing the latter.

Table 7. Recall and Precision Dependent on the Number of Methods Having Predicted the Compounds To Be Active

	COX2		DH		TH		RT		AE	
consensus	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*
					AA					
SVM	91	70	97	77	97	80	87	50	91	54
1	92	27	100	15	99	39	93	17	94	16
2	88	58	98	54	96	65	89	44	92	50
3	86	77	97	76	97	78	83	73	87	72
4	81	88	97	86	92	83	78	85	82	77
5	75	94	92	91	88	88	63	91	77	80
6	71	99	87	100	80	91	47	95	34	85
7	65	100	74	100	53	95	20	100	8	100
					VA					
SVM	73	42	73	48	89	49	46	18	42	17
1	84	17	100	6	91	16	71	12	67	11
2	70	47	100	31	83	48	40	29	58	30
3	68	78	81	70	71	68	20	23	25	25
4	54	87	18	100	69	79	14	25	17	50
5	49	86	9	100	39	93	6	40	*a*	*a*

[a] Models were generated by only four methods for AE.

Retrieval of Compounds Using More Than One Method. We were interested in studying whether we can reduce the number of false positives and false negatives by combining the results of two or more methods. For this purpose, we have calculated recall and precision values on the basis of compounds which have been retrieved by at least one method, at least two methods, and so on. Again, this consensus approach was applied to both AA and VA data sets (Table 7).

Considering all of the compounds retrieved by at least one of the seven methods for each target, almost all actives are found for the AA set. However, because recall values obtained with SVM alone are already close to or above 90%, the improvement compared to SVM is small. In addition, this small improvement in recall comes at the price of a significant reduction in precision compared to SVM. If, on the other hand, only compounds retrieved by all methods are considered, recall drops to around 20% for RT and AE but is still substantial for the other targets (Table 7). For this consensus set, the number of false positives is very small for each target (precision > 95% for all targets). At intermediate consensus numbers, an improvement of precision is observed when going from compounds found at least once to those found twice and three times. At the same time, recall does not decrease significantly (Figure 3). Thus, whereas the consensus approach does not offer a significant advantage in terms of recall compared to SVM, precision improves significantly for the consensus results compared to SVM or RF applied alone.

A more substantial improvement in recall is observed when the consensus approach is applied to the VA set (Figure 4). When all of the compounds identified by at least one method as active are considered, we find for three targets a more than 25% improvement in recall and for the other two an improvement of around 10% compared to the results from

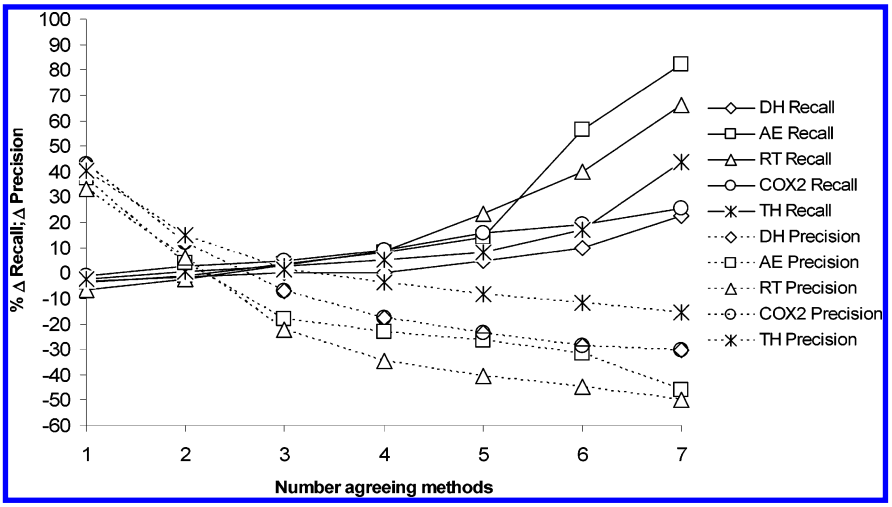

Figure 3. Recall and precision obtained from the consensus approach are subtracted from the values for SVM alone [Δ recall = *R*(SVM) − *R*(consensus); Δ precision = *P*(SVM) − *P*(consensus)]. Each value is calculated on the basis of compounds predicted to be active by at least one, two, three, etc. methods based on the AA set.

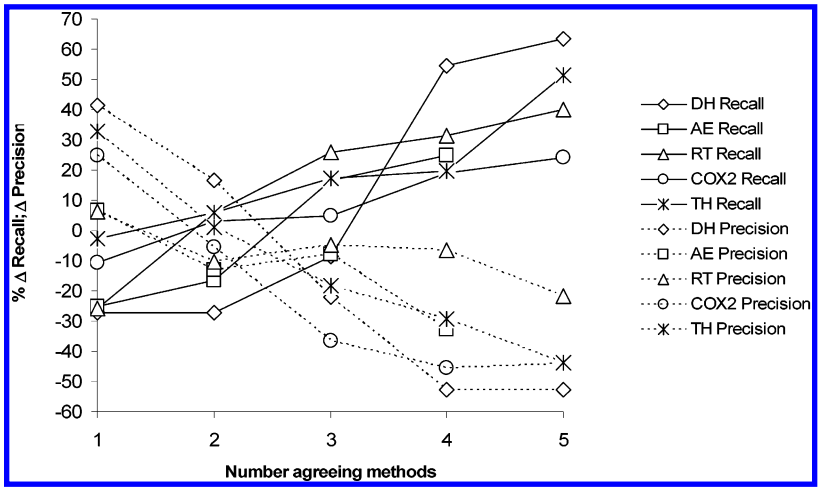

Figure 4. Recall and precision obtained from the consensus approach are subtracted from the values for SVM alone [Δ recall = *R*(SVM) − *R*(consensus); Δ precision = *P*(SVM) − *P*(consensus)]. Each value is calculated on the basis of compounds predicted to be active by at least one, two, three, etc. methods based on the VA set.

SVM only. The corresponding reduction in precision is between 6% and 41%. These data suggest two possible applications of the consensus approach. In cases in which only a small number of actives is known, selecting all actives identified by at least one methods helps to improve recall. The consensus approach is particularly successful in identifying false positives and thus improving precision.

Finally, the effect of the combination of two methods on recall and precision was evaluated. Because we tend to be more interested in avoiding false negatives, we combined the results for any pair of methods such that compounds identified by any of the two methods as active were considered to be active. Thus, only compounds predicted to be not active by both methods were counted as nonactives. Most combinations did not show any significant improvement in recall compared to the results obtained from SVM only. In Table 8 only the results for those combinations which lead to an improved recall are listed. For comparison, we also included results of a combination of SVM with a random selection of a number of compounds corresponding to the total number of actives.

With respect to the larger AA set, the improvement of recall is modest for the combinations because SVM alone yields excellent recall values. Not surprisingly, precision is reduced for each combination with respect to SVM alone. Combination of SVM with a random selection does not change recall but reduces precision significantly. The best performing pairs are combinations of SVM with kNN, DT,

Table 8. Recall (*R*) and Precision (*P*) for All Combinations of Methods Improving the Recall Compared to the Application of SVM Alone or SVM/Random

	COX2		DH		TH		RT		AE	
protein target	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*	*R*	*P*
AA Set										
SVM	90	70	97	77	97	80	87	50	91	54
SVM/random	90	37	97	39	97	41	87	27	91	25
SVM/kNN	96	49	98	51	99	61	92	30	93	28
SVM/DT	92	44	100	41	98	51	90	28	93	32
SVM/NB	93	39	100	19	98	69	88	29	91	31
kNN/DT	94	38	98	36	96	44	90	24	91	22
VA Set										
SVM	73	42	73	47	89	49	46	18	42	17
SVM/random	73	21	73	22	89	25	46	10	42	10
SVM/DT	73	28	91	27	91	22	49	12	67	14
SVM/NB	73	17	100	6	89	37	54	15	42	17
SVM/kNN	73	27	73	40	91	34	57	18	58	14
kNN/DT	70	27	91	29	80	18	43	13	67	12

and NB as well as the combination of kNN and DT. For the VA set, the best combinations of two methods demonstrate a more substantial improvement of recall (Table 8). This reduction in the number of false negatives is, in each case, accompanied by a reduction of precision with respect to SVM alone. The combinations with the largest improvement of recall are the same as those for the larger data set. Again, a combination of methods seems to be particularly useful for data sets poor in active compounds.

CONCLUSION

We have applied seven different QSAR methods to identify ligands of five different biological targets. For each target, many ligands are known, and when a portion of these is used to create models which are used to identify the ligands not used for training, we observe a significant enrichment of true actives in the selected sample set for each method. For data sets rich in actives, such as the AA set discussed above, one would choose TV if the purpose is to select a small subset with the highest enrichment of true actives. RF, SVM, and ANN also perform well in terms of enrichment. If the aim is to retrieve as many true actives as possible, SVM is the method of choice for four out of five targets, closely followed by kNN and RF. Because for the AA data set SVM achieves recall values close to or above 90%, combination with the results from other methods does not offer a significant improvement. In terms of precision, reducing the number of false positives, TV and RF are particularly successful, followed by SVM and kNN. Because SVM and RF perform very well in both recall and precision, they offer an economic choice for selecting a subset of compounds for screening. This is in agreement with similar studies regarding biological classification problems.[22]

For the VA set, the percentage of actives is less than 0.5% for each target and, thus, closer to the number of ligands known at the beginning of a drug discovery project. SVM achieves recall values close to 70% and still more than 40% in its least successful target (RT). kNN, RF, and DT perform also well in terms of recall. Again, for RF, the number of false positives is particularly small (average precision = 65%). In this case, a more substantial improvement is observed when the results of several methods are combined, suggesting that the consensus approach is more relevant for unbalanced data sets.

These results show that, even in the absence of many known ligands of a new target, methods such as SVM and RF can generate models, allowing the retrieval of actives out of large compound collections. To further validate these two methods, we have shown that models generated using early compounds can predict the more recently synthesized ligands of the corresponding targets. This is also an indication that, at least for these five targets, the newer compounds have maintained a significant similarity to the initial ligands in the AP descriptor space.

ACKNOWLEDGMENT

This work was partially supported by the EC BioSapiens (LHSG-CT-2003-503265) and SEPDSA(SP22-CT-2004-003831) 6FP projects. We thank our colleagues Dr. Robert Sheridan, Dr. Subhas Chakravoy, Dr. Bradley Feuston, Dr. Vladimir Maiorov, Dr. Andy Liaw, and Dr. Eugene Fluder for the implementation of the methods described in this article. We also thank the reviewers for valuable comments. This work was supported in part by a grant from the MIUR.

REFERENCES AND NOTES

(1) Caron, P. R.; Mullican, M. D.; Mashal, R. D.; Wilson, K. P.; Su, M. S.; Murcko, M. A. Chemogenomic Approaches to Drug Discovery. *Curr. Opin. Chem. Biol.* **2001**, *5*, 464−470.
(2) Sills, M. A.; Weiss, D.; Pham, Q.; Schweitzer, R.; Wu, X.; Wu, J. J. Comparison of Assay Technologies for a Tyrosine Kinase Assay Generates Different Results in High Throughput Screening. *J. Biomol. Screening* **2002**, *7*, 191−214.
(3) Oprea, T. I.; Bologa, C. G.; Edwards, B. S.; Prossnitz, E. R.; Sklar, L. A. Post-High-Throughput Screening Analysis: An Empirical Compound Prioritization Scheme. *J. Biomol. Screening* **2005**, *10*, 419−426.
(4) Li, J.; Chen, J.; Gui, C.; Zhang, L.; Qin, Y.; Xu Q, Zhang, J.; Liu, H.; Shen, X.; Jiang, H. Discovering Novel Chemical Inhibitors of Human Cyclophilin A: Virtual Screening, Synthesis, and Bioassay. *Bioorg. Med. Chem.* **2005**, *14*, 2209−2224.
(5) Willett, P. Similarity-Based Approaches to Virtual Screening. *Biochem. Soc. Trans.* **2003**, *31*, 603−606.
(6) Guner, O. F. The Impact of Pharmacophore Modeling in Drug Design. *IDrugs* **2005**, *8*, 567−572.
(7) Evers, A.; Hessler, G.; Matter, H.; Klabunde, T. Virtual Screening of Biogenic Amine-Binding G-Protein Coupled Receptors: Comparative Evaluation of Protein- and Ligand-Based Virtual Screening Protocols. *J. Med. Chem.* **2005**, *48*, 5448−5465.
(8) Baurin, N.; Mozziconacci, J. C.; Arnoult, E.; Chavatte, P.; Marot, C.; Morin-Allory, L. 2D QSAR Consensus Prediction for High-Throughput Virtual Screening. An Application to COX-2 Inhibition Modeling and Screening of the NCI Database. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 276−285.
(9) Stahura, F. L.; Bajorath, J. Virtual Screening Methods that Complement HTS. *Comb. Chem. High Throughput Screening* **2004**, *7*, 259−269.
(10) Wilton, D.; Willett, P.; Lawson, K.; Mullier, G. Comparison of Ranking Methods for Virtual Screening in Lead-Discovery Programs. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 469−474.
(11) Jorissen, R. N.; Gilson, M. K. Virtual Screening of Molecular Databases Using a Support Vector Machine. *J. Chem. Inf. Model.* **2005**, *45*, 549−561.
(12) Briem, H.; Gunther, J. Classifying "Kinase Inhibitor-Likeness" by Using Machine-Learning Methods. *ChemBioChem* **2005**, *6*, 558−566.
(13) Stahura, F. L.; Bajorath, J. New Methodologies for Ligand-Based Virtual Screening. *Curr. Pharm. Des.* **2005**, *11*, 1189−1202.
(14) Wang, Y. H.; Li, Y.; Yang, S. L.; Yang, L. Classification of Substrates and Inhibitors of P-Glycoprotein Using Unsupervised Machine Learning Approach. *J. Chem. Inf. Model.* **2005**, *45*, 750−757.
(15) Oprea, T. I.; Matter, H. Integrating Virtual Screening in Lead Discovery. *Curr. Opin. Chem. Biol.* **2004**, *8*, 349−358.
(16) Halperin, I.; Ma, B.; Wolfson, H.; Nussinov, R. Principles of Docking: An Overview of Search Algorithms and a Guide to Scoring Functions. *Proteins* **2002**, *47*, 409−443.
(17) Ferrara, P.; Gohlke, H.; Price, D. J.; Klebe, G.; Brooks, C. L., III. Assessing Scoring Functions for Protein−Ligand Interactions. *J. Med. Chem.* **2004**, *47*, 3032−3047.
(18) MDL Information Systems Inc., San Leandro, CA. MACCS Drug Data Report (MDDR). http://www.mdli.com (accessed Mar 2006).
(19) Sheridan, R. P. The Centroid Approximation for Mixtures: Calculating Similarity and Deriving Structure−Activity Relationships. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1456−1469.
(20) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom Pairs as Molecular Features in Structure−Activity Studies: Definition and Applications. *J. Chem. Inf. Comput. Sci.* **1985**, *25*, 64−73.
(21) Vapnik, V. *Statistical Learning Theory*; Wiley: New York, 1998.
(22) Plewczynski, D.; Tkacz, A.; Godzik, A.; Rychlewski, L. A Support Vector Machine Approach to the Identification of Phosphorylation Sites. *Cell. Mol. Biol. Lett.* **2005**, *10*, 73−89.
(23) Chang, C. C.; Lin, C. J. Training nu-Support Vector Classifiers: Theory and Algorithms. *Neural Computation* **2001**, *13*, 2119−2147.
(24) Guha, R.; Jurs, P. C. Interpreting Computational Neural Network QSAR Models: A Measure of Descriptor Importance. *J. Chem. Inf. Model.* **2005**, *45*, 800−806.
(25) Kauffman, G. W.; Jurs, P. C. QSAR and k-Nearest Neighbor Classification Analysis of Selective Cyclooxygenase-2 Inhibitors Using Topologically Based Numerical Descriptors. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1553−1560.
(26) Agrafiotis, D. K.; Cedeno, W.; Lobanov, V. S. On the Use of Neural Network Ensembles in QSAR and QSPR. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 903−911.

(27) Schneider, G.; Wrede, P. Artificial Neural Networks for Computer-Based Molecular Design. *Prog. Biophys. Mol. Biol.* **1998**, *70*, 175−222.
(28) Feuston, B. P. Personal communication, 2004.
(29) Sheridan, R. P.; Feuston, B. P.; Maiorov, V. N.; Kearsley, S. K. Similarity to Molecules in the Training Set Is a Good Discriminator for Prediction Accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1912−1928.
(30) Chakravorty, S. J. Personal communication, 2004.
(31) Raymer, M. L.; Sanschagrin, P. C.; Punch, W. F.; Venkataraman, S.; Goodman, E. D.; Kuhn, L. A. Predicting Conserved Water-Mediated and Polar Ligand Interactions in Proteins Using a K-Nearest-Neighbors Genetic Algorithm. *J. Mol. Biol.* **1997**, *265*, 445−464.
(32) Labute, P. Binary QSAR: A New Method for the Determination of Quantitative Structure Activity Relationships. *Pac. Symp. Biocomput.* **1999**, 444−455.
(33) Sheridan, R. P.; Nachbar, R. B.; Bush, B. L. Extending the Trend Vector: The Trend Matrix and Sample-Based Partial Least Squares. *J. Comput.-Aided Mol. Des.* **1994**, *8*, 323−340.
(34) Rusinko, A., III; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of a Large Structure/Biological Activity Data Set Using Recursive Partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017−1026.
(35) Quinlan, J. R. *C4.5: Programs for Machine Learning*; Morgan Kaufmann Publishers Inc.: San Francisco, CA, 1993.
(36) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random Forest: A Classification and Regression Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.
(37) Breiman, L. Random Forests. *Macine Learning* **2001**, *45*, 5−32.

CI050519K