Support-Vector-Machine-Based Ranking Significantly Improves the Effectiveness of Similarity Searching Using 2D Fingerprints and Multiple Reference Compounds

Hanna Geppert,[‡] Tamás Horváth,[†,§] Thomas Gärtner,[†] Stefan Wrobel,[†,§] and Jürgen Bajorath*[,‡]

Fraunhofer IAIS, Schloss Birlinghoven, D-53754 Sankt Augustin, Germany, Department of Life Science Informatics, B-IT, LIMES Program Unit Chemical Biology and Medicinal Chemistry, Rheinische Friedrich-Wilhelms-Universität, Dahlmannstr. 2, D-53113 Bonn, Germany, and Institute of Computer Science III, Rheinische Friedrich-Wilhelms-Universität, Römerstr. 164, D-53117 Bonn, Germany

Received December 14, 2007

Similarity searching using molecular fingerprints is computationally efficient and a surprisingly effective virtual screening tool. In this study, we have compared ranking methods for similarity searching using multiple active reference molecules. Different 2D fingerprints were used as search tools and also as descriptors for a support vector machine (SVM) algorithm. In systematic database search calculations, a SVM-based ranking scheme consistently outperformed nearest neighbor and centroid approaches, regardless of the fingerprints that were tested, even if only very small training sets were used for SVM learning. The superiority of SVM-based ranking over conventional fingerprint methods is ascribed to the fact that SVM makes use of information about database molecules, in addition to known active compounds, during the learning phase.

1. INTRODUCTION

Ligand-based virtual screening (LBVS) methods are designed to efficiently process millions of database molecules and select a limited number of candidate compounds that are most likely to possess a desired biological activity. LBVS has become an integral part of the hit identification process in pharmaceutical research, and it is also used to complement high-throughput screening.[1] A long-established strategy for LBVS is similarity searching using molecular fingerprints,[2–4] which transforms compounds into bit string representations of chemical structures and properties and compares them in fingerprint space. Representative state-of-the-art fingerprint designs include hashed connectivity pathways,[5] structural dictionary-based designs,[6] layered-atom environments,[7] and pharmacophore-type fingerprints.[8] Similarities between database and active reference molecules are quantitatively determined by calculating the pairwise overlap of their fingerprint representations. For this purpose, a variety of similarity metrics have been introduced, the most prominent being the Tanimoto coefficient (Tc).[3]

Similarity searching using fingerprints can be applied in situations where only a single active reference structure is available, different from many other similarity methods that require multiple reference compounds such as, for example, clustering or partitioning. However, search performance usually improves when multiple active compounds are available. Accordingly, various approaches have been introduced to utilize multiple reference molecules in fingerprint calculations, including consensus[9] or centroid[10] fingerprints, scaling procedures,[11] and nearest-neighbor methods[10,12] (i.e., data fusion techniques). Several studies have been conducted to compare these different search strategies, and nearest-neighbor as well as centroid calculations were often found to perform best.[10,12] The relative performance of these search strategies is often influenced by differences in structural diversity between compound classes. For example, for structurally homogeneous classes, the 1-NN approach easily detects active compounds, whereas averaging of fingerprints or similarity values often produces better results than 1-NN for moderately diverse classes.

Recently, Wilton et al.[13] have evaluated binary kernel discrimination (BKD) in virtual compound screening using Tripos' Unity 2D fingerprint as a descriptor and a support vector machine[14,15] (SVM) for comparison. Calculations were carried out on three sets of pesticides using differently sized training sets. For mid-sized learning sets of ca. 200 active and inactive compounds, the results for BKD, SVM, and various Unity 2D similarity rankings were rather heterogeneous on the three data sets, and no clear preferences could be observed. When large training sets of 6000 active and up to 60 000 inactive molecules were used on the combined pesticide data sets, the machine learning methods, especially SVM, outperformed Unity 2D similarity rankings.[13]

SVMs were originally developed for binary classification problems and have become popular in the chemoinformatics field.[16–18] In a typical SVM analysis, training compounds belonging to two different classes (e.g., active versus inactive) are projected into chemical reference space and a separating hyperplane is derived. Then, test compounds are evaluated in this reference space to predict their class labels dependent on which side of the hyperplane they fall. SVMs were adopted for virtual screening by using the signed distance between a molecule and the hyperplane to rank database compounds in order of decreasing priority for

* To whom correspondence should be addressed: Tel: +49-228-2699-306, Fax: +49-228-2699-341, E-mail: bajorath@bit.uni-bonn.de.
† Fraunhofer IAIS.
‡ Department of Life Science Informatics, Rheinische Friedrich-Wilhelms-Universität.
§ Institute of Computer Science III, Rheinische Friedrich-Wilhelms-Universität.

10.1021/ci700461s CCC: $40.75

Published on Web 03/05/2008

biological testing.[19] In the context of hit identification, the advantage of ranking compared to classification methods is that the number of molecules submitted to biological testing can arbitrarily be chosen.

The SVM approach is not dependent on the use of a specific chemical space representation. For example, the reference space can be defined by use of numerical property descriptors[20] that are important elements of chemical space design[21,22] but can be defined also by molecular fingerprints, which makes SVM-based ranking directly comparable to conventional similarity searching.

In this study, we focus on a performance evaluation of the SVM-based ranking approach and conventional ranking methods utilized in fingerprint similarity searching. We did not aim at comparing standard similarity searching with a supervised learning technique such as SVM but, rather, aimed at evaluating the predictive value of fingerprint descriptors using ranking approaches that do or do not assign different weights to individual bit positions. Here, we report the results of systematic search calculations on 10 different sets of pharmaceutically relevant compounds and five different types of fingerprints to compare the different ranking techniques. While SVMs are usually trained with fairly large compound sets, for example, see Wilton et al.,[13] we used data sets for training comprising only five active molecules and between 14 and 14 423 randomly chosen database compounds. These compound sets were chosen in order to mimic practical virtual screening situations where often only a few active compounds are available. In our analysis, SVM-based ranking consistently produced higher recall rates for our compound classes than the nearest neighbor or centroid search strategy, regardless of the tested fingerprint design.

2. RANKING STRATEGIES FOR VIRTUAL SCREENING USING FINGERPRINTS

2.1. Compound Ranking Based on Support Vector Machines. The SVM approach utilizes a set of n active/inactive training molecules. In this study, compounds are represented as binary fingerprints with m bit positions such that each training molecule is assigned a vector $\mathbf{u_i} \in \{0,1\}^m$ ($i = 1, \ldots, n$) that defines a point in m-dimensional Euclidean vector space. In addition, training compounds are labeled with their activity $y_i \in \{-1,+1\}$, where $y_i = -1$ means inactive and $y_i = +1$ means active. The principal idea of the SVM approach is to determine a hyperplane $\{\mathbf{x} \in \mathbf{R}^m: \mathbf{w}\cdot\mathbf{x} + b = 0\}$ such that all active training molecules are located in the positive half-space $\{\mathbf{x} \in \mathbf{R}^m: \mathbf{w}\cdot\mathbf{x} + b \geq 0\}$ and all inactive training molecules in the negative half-space $\{\mathbf{x} \in \mathbf{R}^m: \mathbf{w}\cdot\mathbf{x} + b \leq 0\}$; $\mathbf{w}$ is the normal vector of the hyperplane and b a scalar. If the training data are linearly separable, an infinite number of such hyperplanes exist, and the one that maximizes the distance from the nearest training examples is called the *maximum-margin hyperplane* and determined by solving a convex optimization problem. If training data cannot be linearly separated, a maximum-margin hyperplane can still be deduced by permitting training errors and penalizing misplaced molecules during the optimization procedure.[23] Optimization yields a normal vector $\mathbf{w}$ that is used to rank database compounds according to the value of $g(\mathbf{x}) = \mathbf{w}\cdot\mathbf{x}$.

2.2. Conventional Ranking Strategies for Similarity Searching. The SVM-based ranking method was compared to two popular conventional ranking techniques for fingerprint searching using multiple reference molecules, the centroid[10] and nearest-neighbor[12] approach in combination with Tanimoto similarity.[3] For a set R of n active reference molecules, $\mathbf{u_i} = (u_{i1}, u_{i2}, \ldots, u_{im})$ for all $i = 1, \ldots, n$, and a database compound $\mathbf{x} = (x_1, x_2, \ldots, x_m)$, the centroid approach determines an "average" fingerprint $\mathbf{u}_{\bar{R}}$ of all active reference molecules with

$$\mathbf{u}_{\bar{R}} = (u_{\bar{R}1}, u_{\bar{R}2}, ..., u_{\bar{R}m})$$

and

$$u_{\bar{R}j} = \frac{1}{n}\sum_{i=1}^{n} u_{ij}$$

for all $j = 1, \ldots, m$.

Then, $\mathbf{x}$ is compared to the "centroid" $\mathbf{u}_{\bar{R}}$ using, for example, the Tc [i.e., $\mathrm{Tc}(\mathbf{u}_{\bar{R}},\mathbf{x})$]. By contrast, the k nearest-neighbor method (k-NN) separately calculates the Tanimoto similarity of $\mathbf{x}$ to each individual reference compound $\mathbf{u_i}$, yielding n different similarity values: $s_i = \mathrm{Tc}(\mathbf{u_i},\mathbf{x})$. The similarity scores s_i are sorted, and the k ($1 \leq k \leq n$) highest values, corresponding to the k nearest-neighbors of $\mathbf{x}$ in fingerprint space, are selected. The average of these k values represents the final similarity score for $\mathbf{x}$. Thus, the centroid approach "merges" chemical information provided by several active reference molecules before comparison to a database compound, whereas the k-NN method conducts an individual similarity search for each active reference molecule and then "fuses" the resulting similarity values.

3. FINGERPRINT DESIGNS

In order to make the comparison of different ranking methods independent of the characteristics of a specific fingerprint design, we included five different fingerprints in our analysis: MACCS,[6,24] Daylight,[5] Molprint2D,[7,25] TGD,[26,27] and TGT.[27] MACCS represents an ensemble of 166 structural fragments that are assigned to 166 bit positions monitoring the presence or absence of the fragments in a molecule. The Daylight fingerprint determines connectivity pathways in molecules and maps them to overlapping bit segments using a hash function. We used a Daylight fingerprint version that consists of 2048 bit positions and monitors pathways of length 0−7. Molprint2D derives atom environments from the connectivity table of a molecule. Since the total number of possible atom environments can, in principle, become exceedingly large, the environments are not assigned to predefined fingerprint bit positions but described as a set of strings. To represent Molprint2D as a bit string, we determined all different atom environments present in our activity classes and the screening database, enumerated them, and assigned them to unique fingerprint positions. This resulted in a fingerprint representation with 84 560 bit positions. However, for a typical test compound, only 15–25 of these bits were set on. TGD and TGT are two- and three-point pharmacophore-type fingerprints with 420 and 1704 bit positions, respectively, that are determined from the 2D molecular graph representation and implemented in the Molecular Operating Environment (MOE).[27]

Table 1. Compound Activity Classes

class code	biological activity	number of compounds
ANG	angiotensin-II antagonists	27
ETA	endothelin antagonists	22
GPA	glycoprotein IIb/IIIa receptor antagonists	25
HIV	HIV protease inhibitors	24
IL1	IL-1β converting enzyme–inhibitors	23
INO	inosine monophosphate dehydrogenase inhibitors	35
SQE	squalene epoxidase inhibitors	25
SSI	squalene synthetase inhibitors	29
THR	thrombin inhibitors	23
ULD	upregulators of LDL receptor	21

4. COMPOUND SETS AND CALCULATION PROTOCOLS

Alternative ranking methods were compared on 10 different activity classes listed in Table 1. Activity classes were selected to contain similar numbers of compounds such that the numbers of potential database hits was also comparable. These classes consisted of between 21 and 35 compounds and were originally assembled from the MDL Drug Data report[28] as described elsewhere.[29,30] As a source database, a "2D-unique" version of ZINC[31] (termed 2D-ZINC) was generated by removing duplicate molecules producing identical 2D molecular graphs. 2D-ZINC contained about 1.44 million molecules that were all considered potential false positives in our virtual screening trials. For each activity class, 100 different sets of five active compounds were randomly selected as reference or training molecules. The remaining 16–30 compounds were added as potential hits to 2D-ZINC. Unlike the *k*-NN and centroid approach, the SVM-based ranking method required training sets that included not only active but also inactive molecules. Since no confirmed inactive molecules were available, we used random subsets of increasing size (0.001%, 0.01%, 0.1%, and 1%) from 2D-ZINC as negative training examples. For each combination of a fingerprint and activity class, the 1-NN, 5-NN, centroid, and SVM-based ranking methods were applied, and the recall of active compounds was monitored among the 100 and 1000 top-scoring database molecules and averaged over the 100 trials corresponding to the different subsets of active compounds used in training. Perl scripts were written to facilitate the 1-NN, 5-NN, and centroid searches. SVM-based ranking was carried out using a publicly available SVM implementation, SVM^{light},[32,33] its standard parameter settings, and the linear kernel, that is, the inner product in the Euclidean space of the fingerprints. We did not attempt to further optimize SVM parameters or test alternative kernel functions because the naïve SVM application using SVM^{light} could be readily repeated with different fingerprint descriptors and compared with the results of conventional similarity searching.

5. VIRTUAL SCREENING TRIALS

The focal point of our study has been the comparison of state-of-the-art ranking methods for fingerprint searching with multiple reference molecules and SVM-based ranking using fingerprints as feature vectors. Results of our systematic similarity search trials are summarized in Table 2. Recovery rates are reported for each activity class, fingerprint, and ranking approach for selection sets of 100 compounds. In addition, Figure 1 graphically represents virtual screening results averaged over all activity classes for database selection sets of 100 and 1000 compounds. The conventional finger-

Table 2. Performance of Different Ranking Methods[a]

	similarity searching			SVM-based ranking			
	1-NN	5-NN	centroid	0.001%	0.01%	0.1%	1%
			(a) Daylight				
ANG	13.9	32.1	21.6	43.8	52.3	62.7	69.5
ETA	1.4	3.6	1.1	6.9	9.4	11.0	16.0
GPA	2.3	8.1	5.1	9.3	14.8	24.5	36.7
HIV	18.5	26.6	12.3	30.8	47.4	58.9	66.4
IL1	5.6	11.0	6.8	17.6	32.7	44.7	53.5
INO	57.5	75.6	68.5	87.4	89.7	90.6	90.4
SQE	6.3	9.5	4.8	14.6	26.6	36.3	44.5
SSI	15.3	10.9	5.8	14.6	17.3	20.6	27.2
THR	1.3	2.8	0.6	2.9	7.0	13.2	19.6
ULD	8.4	11.3	5.3	12.4	16.8	21.6	29.4
average	13.0	19.1	13.2	24.0	31.4	38.4	45.3
			(b) MACCS				
ANG	3.8	24.7	23.4	19.4	33.6	42.1	44.6
ETA	0.0	0.9	0.7	1.8	2.7	4.9	6.2
GPA	4.7	10.8	12.0	12.3	25.0	32.2	35.9
HIV	4.7	11.4	11.2	15.5	26.7	34.5	38.9
IL1	0.0	4.8	3.6	4.0	10.5	15.1	19.1
INO	19.8	20.4	18.0	17.9	25.3	32.2	37.8
SQE	3.9	10.7	9.0	12.3	23.2	29.2	34.2
SSI	9.1	3.9	3.1	8.2	11.5	15.8	18.5
THR	1.6	3.4	4.1	7.6	13.1	20.6	23.9
ULD	0.0	1.6	1.8	0.6	2.3	3.9	7.3
average	4.8	9.3	8.7	9.9	17.4	23.1	26.6
			(c) Molprint2D				
ANG	56.2	57.4	45.8	65.6	66.9	66.6	69.7
ETA	7.4	7.1	1.8	16.9	21.5	23.4	27.3
GPA	14.8	29.7	17.5	48.8	56.6	60.1	65.9
HIV	34.9	49.6	26.6	72.0	75.8	78.3	81.8
IL1	34.6	41.2	25.4	56.8	58.2	63.1	67.0
INO	57.7	87.8	86.8	90.7	91.3	91.4	91.3
SQE	19.5	21.3	12.2	34.4	40.7	46.6	51.1
SSI	18.1	15.2	7.3	22.8	23.8	25.3	28.2
THR	9.7	6.2	1.2	17.1	22.4	27.5	31.3
ULD	11.9	13.1	5.3	22.3	27.1	29.1	35.6
average	26.5	32.9	23.0	44.7	48.4	51.1	54.9
			(d) TGD				
ANG	22.4	28.6	27.3	39.8	52.5	56.6	59.1
ETA	4.2	7.1	5.7	11.7	13.1	16.1	15.6
GPA	17.6	30.7	25.3	50.1	62.3	69.4	74.2
HIV	15.8	13.2	8.7	36.6	43.1	47.6	51.5
IL1	14.0	20.9	18.4	45.2	46.9	48.4	49.7
INO	18.6	22.5	22.4	8.5	20.0	27.4	36.7
SQE	3.6	0.6	0.4	0.7	2.1	3.7	4.9
SSI	3.9	8.7	5.1	22.1	22.8	25.0	24.9
THR	5.2	1.8	0.9	18.9	26.4	29.2	31.0
ULD	6.5	13.6	9.8	16.8	25.7	31.1	34.8
average	11.2	14.8	12.4	25.0	31.5	35.5	38.2
			(e) TGT				
ANG	9.3	17.2	15.5	21.9	28.6	37.1	43.7
ETA	6.9	6.2	3.4	6.1	7.4	8.1	9.3
GPA	14.2	12.1	9.4	19.2	29.6	39.5	50.1
HIV	24.6	41.7	31.1	47.0	54.8	59.3	63.3
IL1	25.0	24.5	19.1	39.9	40.1	44.2	50.2
INO	24.9	31.9	30.9	25.7	50.8	70.5	80.0
SQE	5.3	5.8	3.1	1.8	3.9	6.3	10.0
SSI	1.3	2.0	1.6	7.3	9.5	15.1	21.0
THR	13.6	10.6	6.4	17.6	25.6	31.7	39.4
ULD	8.6	8.1	5.9	4.8	8.3	12.6	18.4
average	13.4	16.0	12.6	19.1	25.8	32.4	38.5

[a] Recovery rates (in percent) are reported for selection sets of 100 compounds when averaging over 100 different trials for each combination of a fingerprint and ranking technique. In each case, only five active reference compounds were used. In addition, for SVM-based ranking, different percentages of 2D-ZINC were used as negative training examples: 0.001%, 0.01%, 0.1%, and 1% corresponding to 14, 144, 1442, and 14 423 compounds, respectively. Activity classes are abbreviated according to Table 1.

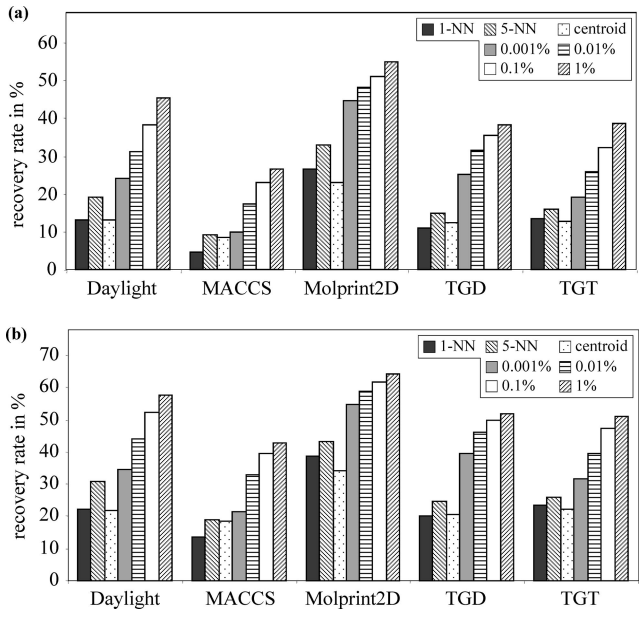

Figure 1. Average search performance for different ranking approaches. Recovery rates were determined in selection sets of (a) 100 and (b) 1000 compounds and averaged over 100 trials and 10 activity classes. "0.001%", "0.01%", "0.1%", and "1%" give the percentages of 2D-ZINC molecules used for SVM learning.

print search calculations mirror trends that are generally observed;[29,30] that is, search performance is compound-class-dependent and varies among different fingerprint designs. In the calculations reported here, on average, Molprint2D produced the highest recovery rates, followed by Daylight, TGT, TGD, and MACCS. In our calculations, an interesting feature of Molprint2D was that 90% of its 84 560 bit positions were set on as individual bits in less than 0.01% of 2D-ZINC (i.e., only 144 compounds). Thus, atom environments in active compounds represented by these bits are highly discriminatory, which is likely to explain the superior performance of Molprint2D over other fingerprints observed here. Independent of the considered fingerprint design, recovery rates for classes ETA, SQE, and THR were consistently lower than for ANG, HIV, and INO, which could at least in part be attributed to the presence of different levels of intraclass structural diversity.[29,30] Among the conventional fingerprint search techniques, the 5-NN approach produced the overall highest recovery rates. For selection sets of 100 compounds, the 1-NN and centroid techniques achieved average recovery rates between 5 and 27% and 5-NN did so between 9 and 33%. Standard deviations of recovery rates were only small, that is, between 0% and 5% in most cases. The compound classes studied here did not contain analog series, which suggests an explanation for the finding that 5-NN calculations were superior to those of 1-NN. For these activity classes, averaging of similarity values was also more effective than calculation of average fingerprints.

The search calculations described above provided the basis for a detailed comparison with SVM-based compound ranking. This comparison revealed two major trends, irrespective of the investigated fingerprint design and activity class. First, the SVM approach produced consistently higher recovery rates than conventional fingerprint search strategies, and second, SVM recall rates increased with increasing numbers of 2D-ZINC compounds used for training, as visualized in Figure 1. In addition, the results in Table 2 show that the ratio between the performance of similarity searching and SVM remained fairly constant. This means that classes with lowest or highest recovery rates in similarity searching displayed the same relative performance in SVM analysis. These results are very likely due to class-dependent differences in the intrinsic ability of fingerprint descriptors to distinguish activity classes from database compounds and are thus not determined by alternative search strategies.

SVM-based ranking using 0.001% of 2D-ZINC as "inactive" training examples (i.e., only 14 molecules) obtained average recovery rates of 10–45%, thus already higher than conventional fingerprint searching. On average, standard deviations of recovery rates were approximately 10%. In only a single instance (for TGD and activity class INO), SVM-"0.001%" produced a considerably lower recovery rate (8.5%) than the 1-NN, 5-NN, and centroid techniques (18.6–22.5%). However, five active and 14 database molecules represent a much smaller compound set than typically used for SVM training (see, for example, Wilton et al.[13]), and the high SVM performance level using this very small training set was not expected. In fact, Table 2 shows that differences in recovery rates of up to 30% were observed in favor of SVM-"0.001%" (e.g., for classes GPA, HIV using Molprint2D or IL1, THR using TGD).

A key finding of this analysis has been the consistently high recall of active compounds using SVM-based ranking on fingerprints. A fundamental difference to conventional ranking methods is that SVM is a supervised machine learning technique that uses learning sets to assign different weights to individual fingerprint bit positions. By contrast, when applying the 1-NN and 5-NN method, all fingerprint bit positions equally contribute to the similarity value. The centroid approach implicitly assigns weights to bit positions through generation of an average bit string. However, it only uses known active compounds to derive these weights. As a supervised learning technique, the SVM approach also adds information about inactive (or randomly selected) database compounds to the learning step. Therefore, we ascribe the superiority of the SVM-based ranking technique in part to this information gain. This idea is consistent with the finding that the use of increasing numbers of database molecules for SVM training systematically improved search performance. Adding 10 times more database compounds to the training step (i.e., from 0.001% of 2D-ZINC to 0.01%, 0.1%, and 1%), on average about 5% more active molecules were recovered each time. The use of 0.01% of 2D-ZINC (144 compounds) doubled recovery rates relative to 1-NN, 5-NN, and centroid calculations in many cases.

Figure 1 also shows that increasing the size of database selection sets from 100 to 1000 compounds had only little influence on the relative search performance. For each search technique and fingerprint, about 10–15% more active molecules were recovered in selection sets of 1000 database compounds. Thus, although absolute recovery rates further increased, as one should expect, differences in relative search performance between the alternative methods remained constant.

6. CONCLUDING REMARKS

We systematically compared state-of-the-art (1-NN, 5-NN, and centroid) strategies for fingerprint searching using

multiple reference molecules and a SVM-based ranking scheme with fingerprint bit patterns used as descriptors. The SVM-based ranking method was found to outperform the 1-NN, 5-NN, and centroid approaches, even when only small training sets were used for SVM learning. To achieve high recovery rates, extensive SVM parameter optimization was not required and the application of the linear kernel function was sufficient. However, we expect that using a different kernel function, in particular, the Tanimoto kernel,[34] might further enhance SVM search performance.

The observed improvements in recovery rates by SVM ranking were likely due to the information gain associated with the addition of randomly chosen database "decoys" to the learning step. Under these conditions, SVM calculations were also found to be robust because it was not necessary to use confirmed inactive compounds for learning, which makes SVM-based ranking attractive for practical applications if only a few active compounds are available. In contrast to SVM, conventional fingerprint searching does not include information about inactive or database compounds. Taken together, our results indicate that support-vector-machine-based ranking using fingerprint descriptors and only a few active molecules for learning is capable of producing significant recall of diverse active compounds, which should make this approach a promising addition to the current repertoire of virtual screening tools.

ACKNOWLEDGMENT

We wish to thank Daylight Chemical Information Systems for making the Daylight fingerprint available to us and thank Andreas Bender for Molprint2D.

REFERENCES AND NOTES

(1) Bajorath, J. Integration of virtual and high-throughput screening. *Nat. Rev. Drug Discovery* **2002**, *1*, 882–894.
(2) Eckert, H.; Bajorath, J. Molecular similarity analysis in virtual screening: foundations, limitations and novel approaches. *Drug Discovery Today* **2007**, *12*, 225–233.
(3) Willett, P. Searching techniques for databases of two- and three-dimensional chemical structures. *J. Med. Chem.* **2005**, *48*, 4183–4199.
(4) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046–1053.
(5) James, C. A.; Weininger, D. *Daylight Theory Manual*; Daylight Chemical Information Systems Inc.: Irvine, CA, 2007.
(6) McGregor, M. J.; Pallai, P. V. Clustering of large databases of compounds: using MDL "Keys" as structural descriptors. *J. Chem. Inf. Comput. Sci.* **1997**, *37*, 443–448.
(7) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular similarity searching using atom environments, information-based feature selection, and a naïve Bayesian classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170–178.
(8) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. New 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem.* **1999**, *42*, 3251–3264.
(9) Shemetulskis, N. E.; Weininger, D.; Blankley, C. J.; Yang, J. J.; Humblet, C. Stigmata: an algorithm to determine structural commonalities in diverse datasets. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 862–871.
(10) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity metrics for ligands reflecting the similarity of the target protein. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391–405.
(11) Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorath, J. Profile scaling increases the similarity search performance of molecular fingerprints containing numerical descriptors and structural keys. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1218–1225.
(12) Hert, J.; Willet, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177–1185.
(13) Wilton, D. J.; Harrison, R. F.; Willett, P.; Delaney, J.; Lawson, K.; Mullier, G. Virtual screening using binary kernel discrimination: analysis of pesticide data. *J. Chem. Inf. Model.* **2006**, *46*, 471–477.
(14) Cristianini, N.; Shawe-Taylor, J. *An introduction to Support Vector Machines and other kernel-based learning methods*; Cambridge University Press: Cambridge, UK, 2000.
(15) Schölkopf, B.; Smola, A. *Learning with Kernels*; MIT Press: Cambridge, MA, 2002.
(16) Burbidge, R.; Trotter, M.; Holden, S.; Buxton, B. Drug design by machine learning: support vector machines for pharmaceutical data analysis. *Comput. Chem.* **2001**, *26*, 5–14.
(17) Warmuth, M. K.; Liao, J.; Rätsch, G.; Mathieson, M.; Putta, S.; Lemmen, C. Active learning with support vector machines in the drug discovery process. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 667–673.
(18) Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1882–1889.
(19) Jorissen, R. N.; Gilson, M. K. Virtual screening of molecular databases using a support vector machine. *J. Chem. Inf. Model.* **2005**, *45*, 549–561.
(20) Livingstone, D. J. The characterization of chemical structures using molecular properties. A survey. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 195–209.
(21) Agrafiotis, D. K.; Lobanov, V. S.; Salemme, F. R. Combinatorial informatics in the post-genomics era. *Nat. Rev. Drug Discovery* **2002**, *1*, 337–346.
(22) Bajorath, J. Selected concepts and investigations in compound classification, molecular descriptor analysis, and virtual screening. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 233–245.
(23) Müller, K.-R.; Rätsch, G.; Mika, S.; Tsuda, K.; Schölkopf, B. An introduction to kernel-based learning algorithms. *IEEE Neural Networks* **2001**, *12*, 181–201.
(24) *MACCS structural keys*; MDL Information Systems Inc.: San Leandro, CA, 2002.
(25) MOLPRINT 2D. http://www.molprint.com (accessed June, 2006).
(26) Sheridan, R. P.; Miller, M. D.; Underwood, D. J.; Kearsley, S. K. Chemical similarity using geometric atom pair descriptors. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 128–136.
(27) *MOE (Molecular Operating Environment)*; Chemical Computing Group Inc.: Montreal, Quebec, Canada, 2007.
(28) *MDL Drug Data Report (MDDR)*; MDL Information Systems Inc.: San Leandro, CA, 2005.
(29) Tovar, A.; Eckert, H.; Bajorath, J. Comparison of 2D fingerprint methods for multiple-template similarity searching on compound activity classes of increasing structural diversity. *ChemMedChem* **2007**, *2*, 225–233.
(30) Eckert, H.; Bajorath, J. Design and evaluation of a novel class-directed 2D fingerprint to search for structurally diverse active compounds. *J. Chem. Inf. Model.* **2006**, *46*, 2515–2526.
(31) Irwin, J. J.; Shoichet, B. K. ZINC - a free database of commercially available compounds for virtual screening. *J. Chem. Inf. Model.* **2005**, *45*, 177–182.
(32) SVM light, version 4.00. http://svmlight.joachims.org/ (accessed Mar, 2002).
(33) 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: Cambridge, MA, 1999.
(34) Ralaivola, L.; Swamidass, S. J.; Saigo, H.; Baldi, P. Graph kernels for chemical informatics. *Neural Networks* **2005**, *18*, 1093–1110.

CI700461S