Comparison of Ranking Methods for Virtual Screening in Lead-Discovery Programs

David Wilton* and Peter Willett

Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Sheffield S10 2TN, UK

Kevin Lawson and Graham Mullier

Syngenta, Jealott's Hill International Research Centre, Bracknell, Berkshire RG42 6EY, UK

Received August 20, 2002

This paper discusses the use of several rank-based virtual screening methods for prioritizing compounds in lead-discovery programs, given a training set for which both structural and bioactivity data are available. Structures from the NCI AIDS data set and from the Syngenta corporate database were represented by two types of fragment bit-string and by sets of high-level molecular features. These representations were processed using binary kernel discrimination, similarity searching, substructural analysis, support vector machine, and trend vector analysis, with the effectiveness of the methods being judged by the extent to which active test set molecules were clustered toward the top of the resultant rankings. The binary kernel discrimination approach yielded consistently superior rankings and would appear to have considerable potential for chemical screening applications.

INTRODUCTION

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, with the aim of ensuring that those molecules with the largest a priori probabilities of activity are assayed first in a lead discovery program.[1−3] At the risk of oversimplification, it is possible to identify four main classes of virtual screening methods, these differing in the amount of structural and bioactivity data that is available. If just a single active molecule is available, such as a competitor's compound or a natural product, then similarity searching can be used, in which a database is ranked in decreasing order of similarity to the known active structure.[4] If several actives have been identified, then pharmacophore mapping can be carried out to ascertain common patterns of features that may be responsible for the observed activity, with a 3D substructure search of the database then being carried out to identify further molecules that contain the pharmacophore.[5] If it is not possible to identify a common pharmacophore, as often occurs with heterogeneous sets of actives (e.g., the initial weak leads from a high-throughput screening program), and if a fair number of both active and inactive molecules are available, then these can be used as training data for a machine learning system.[6] Finally, if the 3D structure of the biological target is known, then a docking study can be carried out to identify those database molecules that are complementary to the binding site.[7]

This paper reports a comparison of methods that can be used for the third class of virtual screening methods, which includes such common approaches as substructural analysis,[8,9] genetic algorithms,[10] neural networks,[11,12] and decision trees.[13−15] Two main classes of approach are possible: ranking methods order a database in order of decreasing probability of activity and classification methods divide a database into those molecules that are predicted to be active and those that are predicted to be inactive. We focus here on ranking methods. The next section of the paper describes the various methods that we have chosen to test, and this is then followed by an evaluation of their effectiveness using two large databases that contain both structural and bioactivity data: the first of these is the well-known NCI AIDS file[16] while the second of these is a large file of *in vivo* data from the multinational agrochemicals company Syngenta. The paper concludes with a summary of our major findings.

* Corresponding author e-mail: d.j.wilton@sheffield.ac.uk.

VIRTUAL SCREENING METHODS

Similarity Methods. The simplest way of predicting the likely activities of a set of compounds is by computing their similarities to a training set of known actives and or inactives, i.e., a *k*-nearest neighbor classifier. Four different versions of this basic approach were tested as follows. Assume that the training set contains N_A actives and N_I inactives and that $S(i,j)$ is the similarity between a test set molecule j and a training set molecule i. Then S_A is defined to be the mean similarity to all the actives, i.e.

$$S_A(j) = \frac{1}{N_A} \sum_{i \in Actives} S(i,j)$$

D_I is defined to be the mean dissimilarity to all the inactives (not something that is expected to be a good predictor but included here for purposes of comparison), i.e.

$$D_I(j) = \frac{1}{N_I} \sum_{i \in Inactives} \{1 - S(i,j)\}$$

10.1021/ci025586i CCC: $25.00
Published on Web 12/10/2002

S_{max} is defined to be the similarity to the most similar training set active, i.e.

$$S_{max}(j) = \max\{S(i,j)\} \quad i \in Actives$$

and, finally, S_{A-I} is defined to be the mean similarity to all the actives minus the mean similarity to all the inactives

$$S_{A-I} = \frac{1}{N_A} \sum_{i \in Actives} S(i,j) - \frac{1}{N_I} \sum_{i \in Inactives} S(i,j)$$

In these methods, as with all the others considered here, the test set molecules are ranked in descending order of the calculated scores, i.e., similarity values in the present context, with the expectation that the top-ranked molecules have the greatest probability of activity.

Trend Vectors. The trend vector (TV) is the first moment of activity in descriptor space[17,18] and is analogous to the dipole moment (the first moment of electronic charge in normal 3D space). In general for a set of N compounds, the j-th element ($1 \le j \le K$, where K is the total number of descriptors used for characterizing the molecules in the training set and the test set) of the trend vector is defined by

$$T_j = \frac{1}{N} \sum_{i=1}^{N} (a_i - A) D_{ij}$$

where D_{ij} is the value of descriptor j for compound i, a_i is the activity of compound i, and A is the average activity of all N compounds. A score can then be calculated for any molecule i by calculating the dot product of its vector with the trend vector representing the training set, i.e., for an arbitrary compound i we compute the score

$$\sum_{j=1}^{K} T_j D_{ij}$$

Substructural Analysis. Substructural analysis was first described by Cramer et al.,[8] and many different weighting schemes have been described for this purpose.[9] For each fragment or bit, j, in the binary fingerprints which characterize the training set molecules, a weight is calculated that is a function of the numbers of active and inactive molecules in the training set that have the j-th bit set. A score is then computed for a test set molecule by summing (or otherwise combining) the weights of those bits that are set in its fingerprint.

Many different weighting schemes have been described for substructural analysis, and in this study, we have evaluated four such schemes: these are the R1, R2, AVID, and WT2 weights that performed well in the comparative study of Ormerod et al.[9] Let A_j and I_j be the numbers of active and inactive molecules with bit j set, with T_j being the total number of compounds with bit j set (i.e., A_j+I_j). Similarly, let N_A and N_I be the total numbers of active and inactive molecules, with a total of N_T (i.e., N_A+N_I) molecules in the training set. Then the four weighting schemes tested are shown in Table 1. For the R1 and R2 schemes, the score for each molecule is computed by summing the weights for each bit present, while for the AVID and WT2 weights this sum is then divided by the number of bits set for that particular molecule.[9] Note that TV is a substructural analysis method; indeed, it is identical to the MAS weight that was one of the weights studied by Ormerod et al. However, we have kept it separate here as it is also applicable to nonbinary data, whereas this is not the case for R1, R2, AVID, and WT2.

Table 1. Weighting Schemes for Substructural Analysis[a]

weighting scheme	weight
AVID	$(A_j + 1) \Big/ \left(\frac{T_j N_A}{N_T} + 1\right)$
R1	$\log\left(\frac{A_j/N_A}{T_j/N_T}\right)$
R2	$\log\left(\frac{A_j/N_A}{I_j/N_I}\right)$
WT2	$\frac{A_j - I_j}{N_j}$

[a] In this table, A_j and I_j are the numbers of active and inactive training set compounds with bit j set, T_j is the total number of compounds with bit j set, i.e., $A_j + I_j$, N_A and N_I are the total numbers of training set active and inactive compounds, and N_T is the total number of training set compounds, i.e., $N_A + N_I$.

Bioactivity Profiles. Gillet et al. describe a method, called a *bioactivity profile* (BP), that is similar to substructural analysis and that computes weights for a range of a high-level structural features using a genetic algorithm.[10] Each feature is divided into bins covering a single value or range of values. For example, the feature 'H-bond donors' is divided into bins for molecules containing one H-bond donor, molecules containing two H-bond donors, etc. up to ≥ 20 H-bond donors; while each bin for a continuous feature such as molecular weight covers a range of values. A molecule will be assigned to one bin for each feature depending on the value of that feature in that molecule: the features used here were molecular weight, number of H-bond donors, number of H-bond acceptors, number of aromatic rings, number of rotatable bonds, and the $^2k_\alpha$ shape index.

Each bin for each feature has a weight associated with it, and the score for a specific molecule is the sum of the weights for each bin to which it has been assigned, thus enabling the training set molecules to be ranked in order of their bioactivity scores. A measure of the extent to which the training set actives occur at the top of this ranking is used as the fitness function of a genetic algorithm that seeks to optimize the bin-weights so as maximize the degree of separation of the actives and inactives when the summed scores are ranked. The resulting weights can then be applied to test set data. Although such high-level descriptors are very simple, they proved sufficient to discriminate between drug and (presumed) nondrug compounds in the World Drug Index and SPRESI databases, respectively, and were subsequently applied successfully to screening data at Glaxo-SmithKline.[10]

Binary Kernel Discrimination. The final ranking approach considered here are the binary kernel discriminators (BKD) that have recently been described for chemical applications by Harper et al.[19,20] For two compounds i and j, characterized by binary fingerprints of length M, that differ

in d_{ij} positions, the kernel function K_λ, suggested by Aitchison and Aitken,[21] is

$$K_\lambda(i,j) = \lambda^{M-d_{ij}}(1-\lambda)^{d_{ij}} \quad (1)$$

where λ is a smoothing parameter to be determined. This function can then be used to estimate the probability that acompound is active. Following Harper et al., we have used the scoring function

$$L_A(j) = \frac{\sum_{i \in Active} K_\lambda(i,j)}{\sum_{i \in Inactive} K_\lambda(i,j)} \quad (2)$$

to rank the molecules in the test set, using the optimum value of λ found for the training set. This is obtained by computing scores for each training set compound using the other training set compounds for a number of different values of λ in the range 0.50 to 0.99. For each value of λ the sum of the ranks of the active compounds is computed. If this is plotted against λ a clear minimum should be observed indicating the optimum λ, i.e., that which minimizes the summed ranks of the actives in the training set. If a clear minimum cannot be observed then a modified form of eq 1 may be used, as described by Harper[20]

$$K_\lambda(i,j) = [\lambda^{M-d_{ij}}(1-\lambda)^{d_{ij}}]^{k/M} \quad (3)$$

where k is an integer less than M. In principle, this modified kernel function should also include a constant multiplier derived from the values of λ, M, and k, but as the same λ is used in the numerator and denominator of (2) this term would cancel out when the scores are computed. It is assumed that the optimal value in the training set is also optimal for the test set. This is clearly a strong assumption, but the results we have obtained suggest that it does not result in poor predictive performance and it is difficult to use a technique such as this without such an assumption. All of the optimum values of λ obtained in our experiments described below were found in the range 0.52 to 0.70.

EXPERIMENTAL DETAILS AND RESULTS

NCI Data Set. Our initial experiments used the NCI AIDS file[16] which contains compounds that have been checked for anti-HIV activity. We used a total of 1129 confirmed actives or confirmed moderately actives and 34 862 inactives. Sets of 200 actives and 200 inactives were selected at random from this file to provide the training set data with the remaining 35 591 compounds forming the test set: three such training sets were generated for the experiments. The test set and training set molecules were represented in one of three ways: by 988-bit Tripos UNITY fingerprints;[22] by 1024-bit Barnard Chemical Information (BCI) fingerprints;[23] and by the set of six high-level structural features noted previously as having been used by Gillet et al. in their work on bioactivity profiles. The effectiveness of the various methods was determined by plotting cumulative recall curves[24] over the top 5% of the ranked test set and by noting the numbers of actives retrieved in the top 1% and the top 5% of the ranking. The results obtained with the three different training sets were all very similar, in that though there were variations in the precise values obtained with the different test sets there was very little difference in the relative performance of the various methods, and we hence consider only one of them, as summarized in Table 2. In this table, the best result in each column has been italicized and starred, and any result within 10% of this best result has been italicized.

Table 2. Numbers of NCI Actives Retrieved in the Top 1% and in the Top 5% of the Rankings[a]

	top 1% of ranking			top 5% of ranking		
method	Unity	BCI	features	Unity	BCI	features
S_A	26	52	16	78	119	48
D_I	9	7	5	27	27	70
S_{max}	76	132	24	361	*343*(*)	114
S_{A-I}	97	83	*72*(*)	199	166	*167*(*)
TV	112	69	62	181	183	160
R1	107	110		199	209	
R2	14	49		153	138	
AVID	58	19		177	122	
WT2	50	8		170	112	
BP			30			138
BKD	*127*(*)	*138*(*)		*397*(*)	317	

[a] The best result in each column is italicized and starred, and any result within 10% of this best result is italicized. For BKD Unity, the value of λ used was 0.59 with the modified kernel function (3) and $k = 100$. For BKD BCI, the value of λ used was 0.66 with the unmodified kernel function (1).

Inspection of Table 2 demonstrates that the nonbinary structural features perform consistently poorly, in some cases retrieving a number of actives that is little different from the number that might be expected using random selection. The two bit-string representations are much to be preferred, with Unity bit-strings being superior in some cases, and the BCI bit-strings in others. S_{max} is clearly the method of choice from among the four similarity methods. This has the advantage of being extremely simple, based as it is just on the single most similar member of the training set; it does, however, mean that it might not be particularly effective in the absence of well-marked series in the test set and the training set. The D_I method does very poorly, but this is hardly surprising as it was included to provide a baseline of performance for the similarity methods, rather than as a realistic tool for ranking a test set. R1 is the best of the substructural analysis methods but is significantly outperformed by BKD: this yields the largest number of actives with the BCI fingerprints at the 1% level and with the Unity fingerprints at 5%.

Figure 1 shows the cumulative recall plots for several of the methods to illustrate diagrammatically the range of performance levels that were achieved. This figure also includes the plots for the best possible performance (i.e., where all of the actives are clustered right at the top of the ranking) and for completely random performance (i.e., where the actives are distributed equally throughout the entire ranking).

Similar qualitative results were obtained using the other two training sets, BKD with Unity fingerprints always yielding the largest number of actives at the 5% level.

Syngenta Data Set. The Syngenta data set contained 132784 molecules that had been tested in various *in vivo* whole organism screens; of these 7127 were active in at least

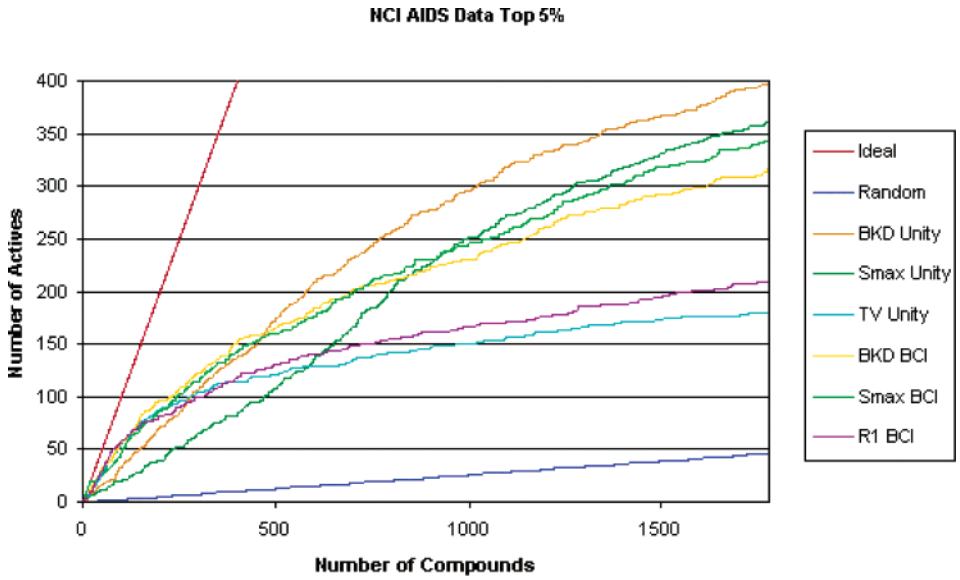

Figure 1. Cumulative recall plot for ranking of the NCI data set.

Table 3. Numbers of Syngenta Actives Retrieved in the Top 1% and in the Top 5% of the Rankings[a]

	top 1% of ranking			top 5% of ranking		
method	Unity	BCI	features	Unity	BCI	features
S_A	119	186	62	498	597	324
D_I	65	66	47	485	401	397
S_{max}	375	412	*238*(*)	1222	1278	*755*(*)
S_{A-I}	424	372	144	1225	1256	587
TV	379	303	65	1078	1108	431
R1	437	407		1199	1279	
R2	445	187		1318	1146	
AVID	422	237		1296	1223	
WT2	424	233		1308	1222	
BP			86			512
BKD	*631*(*)	*426*(*)		*1741*(*)	*1493*(*)	

[a] The best result in each column is italicized and starred, and any result within 10% of this best result is italicized. For BKD Unity, the value of λ used was 0.57 with the modified kernel function (3) and $k = 100$. For BKD BCI, the value of λ used was 0.53 with the unmodified kernel function (1).

one screen, with the remaining 125 657 having a response in the screens less than a predefined threshold value. As before, three different training sets were randomly generated, each containing 713 actives (i.e., 10% of the total actives) and 713 inactives with the remaining 131 358 compounds in each case forming the test sets. As with the NCI data set there was little variation from one training set to another so we have again reported only one set of results, as shown in Table 3.

Inspection of Table 3 reveals both similarities and differences between it and Table 2. The nonbinary structural features perform just as badly, when compared with the two types of bit-string, as previously, and it is again the case that neither of these is consistently superior to the other. That said, some of the individual differences here are proportionally quite large (especially with the 1% threshold), and where this is the case then it is normally Unity that is the better of the two. Broadly speaking, the bit-string results can be divided into three groups: S_A and D_I both do poorly; BKD does very well, and all the rest are broadly the same (especially at the 5% threshold). The BKD/Unity combination is especially noteworthy here: it retrieved 29.5% more actives than the next-best combination (R2/Unity) at the 1% threshold and 14.2% more actives than the next-best combination (BKD/BCI) at the 5% threshold (if this other BKD combination is excluded then the margin of difference rises to 24.3%).

The cumulative recall curves for some of the best-performing methods are plotted in Figure 2, again with the best-possible and completely random lines for comparison.

Again, similar qualitative results were obtained using the other two training sets, BKD with Unity fingerprints always yielding the largest number of actives at the 5% level.

Other Experiments. We carried out many other experiments in addition to those reported above, the majority of them involving the smaller NCI data set. The most substantive were those involving a support vector machine (SVM),[25] a classification approach that has been suggested as being particularly appropriate for chemical applications[26] and that involved use of the popular SVM^{light} software.[27] Unlike the ranking methods, the SVM score for a test set compound merely classifies it as either likely active (if the score is positive) or likely inactive (if the score is negative). Each of the ranking methods can be converted to a classification method by applying a threshold to the ranked scores, with everything above (or below) the threshold being classified as active (or inactive); the application of such a threshold then allows a direct comparison of the performance of the SVM with the ranking methods. In an extensive comparative study using the NCI data set, we found that the SVM results were never better than the other methods considered here and were normally so markedly inferior that we did not consider it worth applying the procedure to the much larger Syngenta data set. For example Table 4 shows a comparison of the results for an SVM and BKD Unity using the same NCI data set as that for Table 2 and Figure 1. In this case the SVM classified 6286 compounds as active; therefore, for comparison, we regard the top 6286 BKD ranked compounds as classified active. Both methods correctly classify just over half the actives, but of those compounds predicted active most are actually inactive. Similar experiments with the other training sets and other descriptors yielded equally poor classifications.

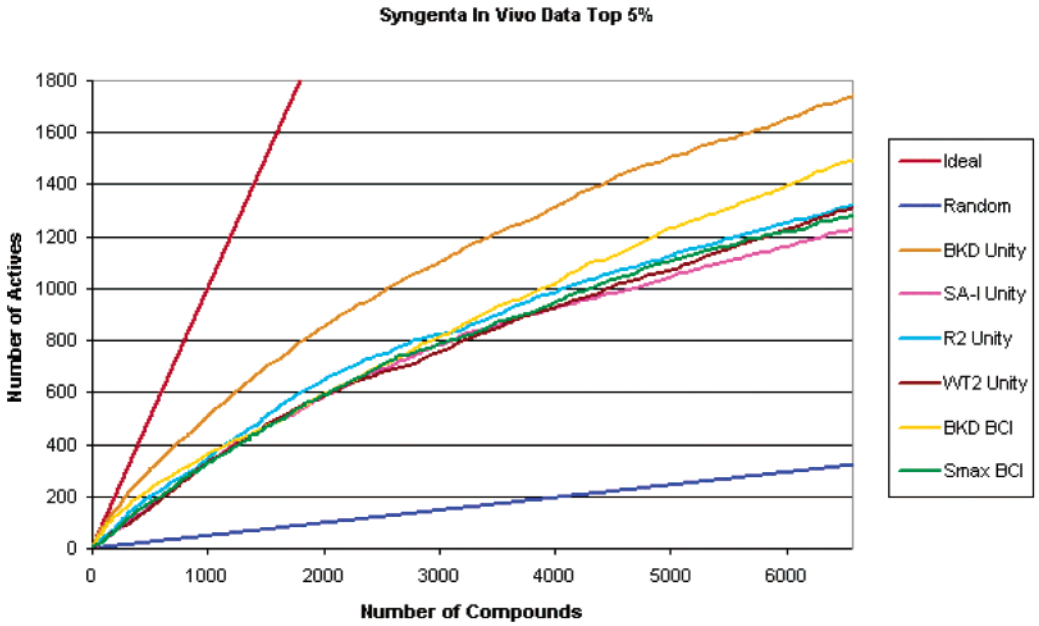

Figure 2. Cumulative recall plot for ranking of the Syngenta data set.

Table 4. Classification of the NCI Data Set with a SVM and BKD Unity, Classifying 6286 Compounds as Active[a]

	SVM	BKD
number of actives classified active	512	552
number of actives classified inactive	417	377
number of inactives classified active	5774	5734
number of inactives classified inactive	28888	28928

[a] The SVM was implemented with the SVM^{light} software[27] using a polynomial kernel function and all other parameters as default.

Another set of experiments involved using the cosine coefficient and the Euclidean distance[4] as an alternative to the Tanimoto coefficient for the four similarity methods. Neither of these provided any obvious improvement over the Tanimoto coefficient with any of the representations, and they were sometimes noticeably less effective. We also looked at the down-weighting of fragments in substructural analysis that occurred in nearly-equal numbers of actives and inactives. Specifically, for each of these weighting schemes there will be a weight, normally 0, that corresponds to a bit being found in the same number of actives as inactives: we refer to this as the *neutral* value. We set an integer value N. If the number of actives containing a bit differs from the number of inactives containing that bit by less than N we set that bit's weight to the neutral value. With $N = 10$, this change had little effect on the performance of TV, AVID, R1, and WT2 weights but improved the results for R2 to the extent that they were then almost as good as those for R1 and TV. Experiments were also carried out on one of the NCI training sets using the PASS weighting scheme described recently by Anzali et al.:[28] this gave results comparable to, but no better than, the best of the schemes in Table 1.

Finally, experiments were also carried out with the Syngenta data where the training sets were selected so as to be diverse, rather than representative. Specifically, three sets were generated so that no pair of actives had a Tanimoto/Unity similarity more than 0.85, and three sets so that no pair of actives and no pair of inactives had such a threshold similarity. Perhaps unsurprisingly, the results from these diverse training sets were consistently lower than for the representative training sets; however, the relative performance of the various ranking methods was broadly similar.

Thus far, we have focused on the effectiveness of the various methods; however, a highly effective virtual screening method is of little practical use if it cannot be applied to data sets of realistic size. Assume that the training set and the test set contain M and N molecules, respectively, and that these are characterized by a total of K features. The computational requirements are in three parts: the analysis of the training set; the scoring of each of the molecules in the test set; and then the ranking of the test set in descending order of the calculated scores. This last component has an time complexity of O(Nlog$N$) for all of the methods and is not considered further. The analysis of the training set and then the scoring of the test set for TV and for the substructural analysis methods is of complexity O(MK) and O(NK), respectively. The complexities for BKD are O(M^2K) for analysis and O(MNK) for scoring, but there is an additional factor that needs to be considered as the training set here must be processed repeatedly to identify the optimal value for the parameter λ. For the similarity methods, there is no training set analysis and the scoring is of complexity O(MNK); however, the complexity can be reduced to O(NK) if the cosine coefficient is used and if the similarities are calculated using the centroid approach.[29] In practice, BKD is the most time-consuming: not only must λ be optimized but each calculation, whether during training or scoring, involves exponentiations that are not required by the other methods. Even so, its requirements are not overly large. Using programs written in C and run on a Silicon Graphics R12000 processor, training for the Syngenta data set took about 36 s for each value of λ that is tested during the training phase; the subsequent scoring using the optimal value of λ took 3730 CPU seconds.

CONCLUSIONS

In this paper, we have reported a comparison of several virtual screening methods that can be used given a training

set for which both structural and bioactivity data are available. Our experiments suggest that the application of the binary kernel discrimination method of Harper et al. to sets of molecules characterized by Unity fingerprints provides an effective way of prioritizing compounds for biological testing.

Although a large number of experiments have been carried out, there are many other approaches that could have been considered. Thus, the three representations are all inherently 2D in character: not only are there other, widely used 2D representations (such as Daylight fingerprints, molecular holograms, or topological indices) but there is also an increasing number of 3D representations that are sufficiently rapid for large-scale virtual screening applications (such as four-point pharmacophores and topomeric shape descriptors). Again, considering methods rather than representations, there is much interest in alternative approaches to chemical machine learning, such as hierarchic decision trees and neural networks. Accordingly, our experiments can in no sense be regarded as comprehensive; however, they do provide sufficient evidence for us to consider further investigations of the binary kernel discriminant approach for chemical screening applications. Areas for development that we intend to investigate include the following: its use with other types of representation; the structural diversity of the top-ranked structures that result from its use; the optimization of the λ parameter in the kernel; its combination with other ranking approaches using data fusion; and its application in a predictive mode to real lead-discovery problems.

ACKNOWLEDGMENT

We thank Syngenta for funding, the referees for comments on an earlier version of this manuscript, and Barnard Chemical Information Ltd., the Royal Society, Tripos Inc. and the Wolfson Foundation for hardware, laboratory, and software support. The Krebs Institute for Biomolecular Research is a Biomolecular Sciences Centre of the Biotechnology and Biological Sciences Research Council.

REFERENCES AND NOTES

(1) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual Screening − an Overview. *Drug Discov. Today* **1998**, *3*, 160−178.
(2) Bohm, H.-J.; Schneider, G. Eds. *Virtual Screening for Bioactive Molecules*; Wiley-VCH: Weinheim, 2000.
(3) Klebe, G. Ed. *Virtual Screening: an Alternative or Complement to High Throughput Screening*; Kluwer: Dordrecht, 2000.
(4) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(5) Guner, O. Ed. *Pharmacophore Perception, Development and Use in Drug Design*; International University Line: La Jolla, CA, 2000.
(6) Gedeck, P.; Willett, P. Visual and Computational Analysis of Structure−Activity Relationships in High-Throughput Screening Data. *Current Opin. Chem. Biol.* **2001**, *5*, 389−395.
(7) 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.
(8) Cramer, R. D.; Redl, G.; Berkoff, C. E. Substructural Analysis. A Novel Approach to the Problem of Drug Design. *J. Med. Chem.* **1974**, *17*, 533−535.
(9) Ormerod, A.; Willett, P.; Bawden, D. Comparison of Fragment Weighting Schemes for Substructural Analysis. *Quant. Struct.-Activ. Relat.* **1989**, *8*, 115−129.
(10) Gillet, V. J.; Willett, P.; Bradshaw, J. Identification of Biological Activity Profiles Using Substructural Analysis and Genetic Algorithms. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 165−179.
(11) Ajay W.; Walters, W. P.; Murcko, M. A. Can we Learn to Distinguish between 'Drug-Like' and 'Nondrug-Like' Molecules? *J. Med. Chem.* **1998**, *41*, 3314−3324.
(12) Sadowski, J.; Kubinyi, H. A Scoring Scheme for Discriminating between Drugs and Nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.
(13) Chen, X.; Rusinko, A.; Young, S. S. Recursive Partitioning Analysis of a Large Structure−Activity Data Set Using Three-Dimensional Descriptors. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 1054−1062.
(14) Roberts, G.; Myatt, G. J.; Johnson, W. P.; Cross, K. P.; Blower, P. E. Leadscope: Software for Exploring Large Sets of Screening Data. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1302−1314.
(15) Tamura, S. Y.; Bacha, P. A.; Gruver, H. S.; Nutt, R. F. Data Analysis of High-Throughput Screening Results: Application of Multidomain Clustering to the NCI Anti-HIV Data Set. *J. Med. Chem.* **2002**, *45*, 3082−3093.
(16) The NCI AIDS database is available at URL http://dtp.nci.nih.gov/. The details of the NCI assay are at URL http://dtp.nci.nih.gov/docs/aids/anti-hiv-screening.html.
(17) 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.
(18) Zaharevitz, D. W. Trend Vector Calculations Using 3D Distance Keys as Molecular Descriptors. At http://dtp.nci.nih.gov/docs/3d_database/stats/trendv.html
(19) Harper, G.; Bradshaw, J.; Gittins, J. C.; Green, D. V. S.; Leach, A. R.Prediction of Biological Activity for High-Throughput Screening Using Binary Kernel Discrimination. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1295−1300.
(20) Harper, G. *The Selection of Compounds for Screening in Pharmaceutical Research*; Ph.D. Thesis, University of Oxford, 1999.
(21) Aitchison, J.; Aitken, C. G. G. Multivariate Binary Discrimination by the Kernel Methodol. *Biometrika* **1976**, *63*, 413−420.
(22) Tripos Inc. is at http://www.tripos.com.
(23) Barnard Chemical Information Ltd.is at http://www.bci.gb.com/.
(24) Edgar, S. J.; Holliday, J. D.; Willett, P. Effectiveness of Retrieval in Similarity Searches of Chemical Databases: a Review of Performance Measures. *J. Mol. Graph. Model.* **2000**, *18*, 343−357.
(25) Burges C. J. C. A Tutorial on Support Vector Machines for Pattern Recognition. *Data Mining Knowledge Discov.* **1998**, *2*, 121−167.
(26) Burbridge, R.; Trotter, M.; Buxton, B.; Holden, S. Drug Design by Machine Learning: Support Vector Machines for Pharmaceutical Data Analysis. *Comput. Chem.* **2001**, *26*, 5−14.
(27) The SVM^{light} system is at http://svmlight.joachims.org/.
(28) Anzali, S.; Barnickel, G.; Cezanne, B.; Krug, M.; Filimonov, D.; Poroikov, V. Discriminating between Drugs and Nondrugs by Prediction of Activity Spectra for Substances (PASS). *J. Chem. Inf. Comput. Sci.* **2001**, *44*, 2432−2437.
(29) Holliday, J. D.; Ranade, S. S.; Willett, P. A Fast Algorithm for Selecting Sets of Dissimilar Structures from large Chemical Databases. *Quant. Struct.-Activ. Relat.* **1995**, *14*, 501−506.

CI025586I