

VIRTUAL SCREENING USING BINARY KERNEL DISCRIMINATION

JÉRÔME HERT, PETER WILLETT AND DAVID J. WILTON

Krebs Institute for Biomolecular Research and Department of Information Studies,
University of Sheffield, Western Bank, Sheffield S10 2TN, U.K.

E-Mail: *p.willett@sheffield.ac.uk

Received: 7th June 2004 / Published: 22nd July 2005

ABSTRACT

This paper discusses a range of procedures for virtual screening of chemical databases in which the molecules are represented by 2D fragment bit-strings. Training-sets containing active and inactive molecules from the NCI AIDS dataset and from the Syngenta corporate pesticide file were processed using binary kernel discrimination (BKD), similarity searching, and substructural analysis methods. The effectiveness of these methods was judged by the extent to which active test-set molecules were clustered towards the top of the resultant rankings. The BKD approach was found consistently to yield the best rankings, and its general effectiveness was confirmed in similarity searches of the MDL Drug Data Report database based on multiple reference structures. As well as being effective, BKD is reasonably efficient, and the method would hence appear to be well suited to virtual screening of 2D structure databases.

INTRODUCTION

Increasing use is being made of virtual screening methods to maximize the cost effectiveness of biological screening programmes by prioritizing a test-set, such as a corporate database of previously unassayed molecules, in decreasing order of the *a priori* probability of activity [1, 2].

There are many different types of virtual screening method: in the first part of this paper, we discuss machine learning methods that can be used when heterogeneous sets of both active and inactive molecules are available for use as training data, e.g., after an initial round of high-throughput screening (HTS).

Specifically, we report a comparison of several such methods when they are used in simulated virtual screening experiments with the NCI AIDS database [3] and with pesticide data from the Syngenta corporate file. These experiments, based on 2D fragment bit-strings, demonstrate the effectiveness of the machine learning technique known as *binary kernel discrimination* (BKD) [4].

BKD is studied further in the second part of the paper, which considers the use of similarity searching as a virtual screening mechanism. Similarity searching is normally used to identify those database molecules that are most similar to a single active *reference structure*, using some quantitative definition of inter-molecular structural similarity [5,6]. Following Schuffenhauer *et al.* [7] and Xue *et al.* [8], we consider here similarity-based virtual screening techniques that can be used when not one but several different bioactive reference structures are available, i.e., when the training-set contains only active molecules. A comparison of several search methods that make use of multiple reference structures, this time using the MDL Drug Data Report (MDDR) database [9], further demonstrates the general effectiveness of BKD for virtual screening of 2D chemical structure databases. Many additional studies, including the use of other structure representations and other screening methods, are reported by Wilton *et al.* [10] and by Hert *et al.* [11].

VIRTUAL SCREENING WITH ACTIVE AND INACTIVE TRAINING DATA

Virtual Screening Methods

Similarity methods

The simplest way of predicting the likely activities of a set of molecules is by computing their similarities to a training-set of known actives and inactives, i.e., a *k*-nearest neighbour classifier. Two approaches are reported here; experiments with other, less effective similarity approaches are discussed by Wilton *et al.* [10]. Given a test-set molecule, *j*, to be predicted, S_{max} is defined to be the similarity to the most similar training-set active, i.e.,

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

while *SA-I* is defined to be the mean similarity to all the training-set actives minus the mean similarity to all the training-set inactives, i.e.,

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

The similarities in these approaches are calculated using the Tanimoto Coefficient [5]. If a , b and c are the numbers of bits set in the one fingerprint, set in the other fingerprint, and set in both fingerprints, respectively, then the Tanimoto coefficient is defined to be:

$$\frac{c}{a + b - c}$$

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.

Substructural analysis methods

Substructural analysis was first described by Cramer *et al.* [12], and many different weighting schemes have been described for this purpose [13]. For each fragment or bit, j , in the binary fingerprints that 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. Here, we report the use of the R1 and R2 weights that performed well in the comparative study of Ormerod *et al.* [13]. Let A_j (I_j) be the number of active (inactive) molecules with bit j set, and let T_j be the total number of molecules with bit j set. Similarly, let N_A (N_I) be the total number of active (inactive) molecules, and let N_T be the total number of molecules in the training-set. Then the two weights are given by:

$$\text{R1} = \log \left(\frac{A_j / N_A}{T_j / N_T} \right), \quad \text{R2} = \log \left(\frac{A_j / N_A}{I_j / N_I} \right).$$

Binary kernel discrimination

The use of binary kernel discrimination (BKD) for chemical applications has been described by Harper *et al.* [4]. For two molecules i and j , characterized by binary fingerprints of length M that differ in d_{ij} positions, they suggest the use of the kernel function K_λ

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

where λ is a smoothing parameter the value of which is to be determined. Training-set molecules are then ranked using the scoring function:

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

with the optimum value of λ being found from analysis of the training-set. The optimum is obtained by computing scores for each training-set molecule using the other training-set molecules 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 molecules is computed. If this is plotted against λ a clear minimum should be observed indicating the optimum λ , i.e. the value that minimizes the summed ranks of the actives in the training-set. 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 machine-learning technique such as this without an assumption.

Results and Discussion

We carried out simulated virtual screening experiments on two datasets, one public and one corporate. The initial experiments used the NCI AIDS file [3], which contains molecules that have been checked for anti-HIV activity, with 1129 confirmed actives or confirmed moderately actives and 34,862 inactives. Training-sets were randomly generated, each containing 200 actives and 200 inactives, with the remaining 35,591 molecules forming the test-set: three such training-sets were generated for the experiments.

 Virtual Screening Using Binary Kernel Discrimination

The Syngenta dataset contained 132,784 molecules that had been tested in various in vivo whole organism screens; of these 7127 were active in at least one screen, with the remaining 125,657 having a response in the screens less than a pre-defined 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 molecules in each case forming the test-sets.

The test-set and training-set molecules were represented by 988-bit Tripos Unity fingerprints [14] and scores were calculated for each of the test-set molecules. The test-set was then ranked, and the effectiveness of the various methods determined by noting the numbers of actives 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; we hence consider the results for only one of the training-sets, as detailed in Table 1.

Table 1. Percentages of the active molecules retrieved in the top 1% and in the top 5% of the rankings for the NCI and Syngenta datasets.

Ranking Method	NCI Dataset		Syngenta Dataset	
	Top-1%	Top-5%	Top-1%	Top-5%
S_{max}	8.18	38.86	5.85	19.05
S_{A-I}	10.44	21.42	6.61	19.10
RI	11.52	21.42	6.81	18.69
$R2$	1.51	16.47	6.93	20.55
BKD	13.67	42.73	9.84	27.14

Inspection of this table leads to a simple, unequivocal conclusion: that the BKD method gives rankings that are far superior to those of the other ranking methods considered here, and thus that this is the method of choice in terms of the effectiveness of virtual screening. The difference between BKD and the other approaches in the table (and the many other approaches discussed by Wilton *et al.* [10]) is particularly marked with the large Syngenta dataset.

A highly effective virtual screening method is of little practical use if it cannot be applied to datasets of realistic size. The computational requirements are in three parts: the analysis of the training-set, which in BKD requires repeated processing of the training-set to identify the optimal value of the parameter λ ; the scoring of each of the molecules in the test-set using the $L_A(j)$ scoring scheme defined previously; and then the ranking of the test-set in descending order of the calculated scores. Even so, the method's computational requirements are not overly large. Using programs written in C and run on a Silicon Graphics R12000 processor, training for the very large Syngenta dataset took about 36 seconds for each value of l that was tested, and the subsequent scoring using the optimal value of l took 3730 CPU seconds.

VIRTUAL SCREENING WITH ACTIVE TRAINING DATA

The increasing use of HTS means that large amounts of active and inactive training data are likely to become available shortly after the commencement of an agrochemical or pharmaceutical lead-discovery programme. Right at the start of a programme, however, the medicinal chemist may have knowledge of just a few active molecules, such as natural products, or patented, competitor molecules. In such cases, an alternative type of virtual screening can be used, based on similarity searching [5, 6]. This involves matching a single bioactive reference structure against each of the database structures to identify those that are most similar (and hence most likely to exhibit the same activity as the reference structure); here, we evaluate three distinct approaches that can be used when multiple reference structures are available, i.e., when the training-set contains only actives.

Virtual Screening Methods

Modal fingerprint method

Shemetulskis *et al.* first described the single fingerprint approach in their work on Stigmata [15]. The method generates a modal fingerprint from an input set of molecules that seeks to capture the common chemical features present in the members of this training-set. A bit j is set to "on" in the modal fingerprint if that bit is found in more than a user-defined threshold percentage of the training-set molecules.

The modal fingerprint is then used as a query and compared to the fingerprints of the molecules in the test-set. Shemetulskis *et al.* used two metrics to rank the molecules of the database: the modal percent and the Tanimoto coefficient, and we have used the latter approach here.

Data fusion method

Data fusion is the name given to a range of techniques that combine inputs from different sensors, with the expectation that using multiple information sources enables more effective decisions to be made than if just a single sensor is employed. The approach has been made in many different fields; when applied to chemoinformatics applications (where it is sometimes referred to as consensus scoring) the fusion is effected by combining the results of several database searches using different descriptors or scoring functions [16]. In conventional applications of data fusion, a single active reference structure is searched against a database in several different ways; in the present context, we have several different reference structures that are all searched against a database in exactly the same way (specifically using 2D fingerprints with the Tanimoto coefficient). We have fused the similarity scores, $S(i,j)$ ($1 \leq i \leq n$, the number of active reference structures) for some molecule j using the S_{max} fusion rule discussed in the first section of the paper, i.e., the test-set molecules were ranked on the basis of :

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

Substructural analysis method

A weighting scheme for substructural analysis normally requires access to training-set data for both actives and inactives. In the present context, however, we do not have access to all the necessary information as the training-set consists of just active molecules. However, if we restrict our attention to those weighting schemes that do not make explicit use of information about the inactives and also make the assumption that the overall characteristics of the training-set are mirrored by those of the entire database that is to be searched, then we can use the R1 weight. The equation for the weight is as given previously; here, however, T_j is the total number of molecules in the database with the bit j set and N_T is the total number of molecules in the database (rather than the total numbers of molecules in the training-set, as in conventional substructural analysis).

Binary kernel discrimination

An analogous approximation can be used to enable BKD to be used when only actives are available for training purposes. The approach we have taken is to make the assumption that the overall characteristics of the inactives are approximated with a high degree of accuracy by the overall characteristics of the entire database that is to be searched. If this assumption is accepted then a training-set can be generated by taking the set of reference structures and adding to it molecules randomly selected from the database, with the expectation that most, if not all, of these added molecules are inactive. Since actives are inherently very rare, both this expectation and that underlying the R1 approximation are not unreasonable.

Results and Discussion

We have evaluated the various approaches above by means of simulated virtual screening searches on the MDDR database [9]. After removal of the duplicates and molecules that could not be processed using local software, a total of 102,535 molecules was available for searching and these were represented by Unity fingerprints [14]. These molecules were searched using the eleven sets of active molecules from MDDR that are listed in Table 2. A rough guide to the diversity of each of the chosen sets of bioactives is provided by matching each compound with every other in its activity class, calculating similarities using the Unity fingerprint and Tanimoto coefficient and computing the mean of these intra-set similarities. The resulting similarity scores are listed in the second column of Table 2, where it will be seen that the renin inhibitors are the most homogeneous and the cyclooxygenase inhibitors are the most heterogeneous.

For each of the 11 activity classes, ten active molecules were selected for use as the training-set. The selections were done at random, subject to the constraint that no pair-wise similarity in a group exceeded 0.80 (using Unity fingerprints and the Tanimoto coefficient). Each searching method was repeated ten times using different training-sets, and in each search, a note was made of the percentage of the active molecules (i.e., those in the same class as those in the training-set) that occurred in the top 5% of the ranking resulting from that search. The results presented below are the mean and standard deviations for these recall values, averaged over each set of ten searches (very similar results to those listed here were obtained if the top-1%, rather than the top-5%, of the rankings were evaluated).

Virtual Screening Using Binary Kernel Discrimination

Table 2. Mean percentage of active molecules retrieved by multiple-reference methods over the top 5% of the ranked test-set.

Activity Class	Self-Similarity	Modal	Data Fusion	Substructural Analysis	BKD
5HT3 antagonist	0.35	30.31	49.03	29.27	52.32
5HT1A agonist	0.34	21.85	37.15	30.13	38.19
5HT Reuptake inhibitor	0.34	39.63	49.68	33.12	45.82
D2 antagonist	0.34	27.12	37.40	27.51	38.65
Renin inhibitor	0.57	88.77	88.62	52.94	93.34
Angiotensin II AT1 antagonist	0.40	73.63	80.44	43.40	84.47
Thrombin inhibitor	0.41	49.43	58.58	35.64	63.06
Substance P antagonist	0.39	36.80	47.14	36.52	58.39
HIV protease inhibitor	0.44	53.53	61.62	34.05	68.45
Cyclooxygenase inhibitor	0.26	10.96	26.52	19.20	33.15
Protein kinase C inhibitor	0.32	35.60	48.01	35.58	49.37
Average over all classes	0.38	42.51	53.11	34.31	56.84

A large number of searches was carried out to identify the best parameter settings for the various methods discussed above (and also several other, less effective methods that are described by Hert *et al.* [11]). The resulting settings were then used in the main experiments, the results of which are detailed in the body of Table 2. Inspection of this table shows that the fusion of the similarity scores and BKD are the clear methods of choice, consistently out-performing modal fingerprints and substructural analysis.

With some minor exceptions, the performance of all of the methods tends to increase as the self-similarity of the active molecules increases. The correlation with intra-class similarity is not unexpected; what is of importance is that good screening performance is obtained even with quite diverse activity classes (such as the protein kinase C inhibitors and the D2 antagonists). The worst results are obtained with the most diverse set of actives, i.e., the cyclooxygenase inhibitors; even here, however, the data fusion and BKD runs represent 5.3-fold and 6.6-fold enrichments, respectively, over a random ranking of the dataset (with average enrichment factors for these two approaches over all classes being 10.6 and 11.3, respectively).

In the final set of experiments, we sought to quantify the benefit that can be achieved using multiple reference structures, rather than single reference structures as in conventional similarity searching. This was done by using every single active molecule in turn in each of the 11 chosen activity classes as the reference structure, and recording the minimum, mean and maximum performance, as detailed in Table 3.

Table 3. Percentage of active molecules retrieved by single similarity searches over the top 5% of the ranked test-set.

Activity Class	Mean	Max	Min
5HT3 antagonist	21.15	40.97	1.89
5HT1A agonist	18.43	39.29	2.45
5HT Reuptake inhibitor	24.02	42.69	1.43
D2 antagonist	17.35	35.58	0.26
Renin inhibitor	80.54	93.21	2.95
Angiotensin II AT1 antagonist	48.04	81.67	3.64
Thrombin inhibitor	33.51	63.56	0.63
Substance P antagonist	26.87	57.69	0.57
HIV protease inhibitor	37.60	63.65	1.89
Cyclooxygenase inhibitor	9.39	21.09	0.32
Protein kinase C inhibitor	19.42	46.05	0.68
Average over all classes	30.57	53.22	1.52

The mean values correspond to the performance that might be expected using an average, individual reference structure and are clearly much lower than the figures reported in Table 2 for the BKD and data fusion methods (30.57% as against 56.84% and 53.11%, respectively). Thus, the use of ten actives, rather than just one, results in an increase of over two-thirds in the numbers of actives retrieved. Perhaps the most interesting figures in Table 3 are those listed under "Max". These represent the best single similarity searches possible from the many hundreds of individual bioactive molecules (this number ranges from 349 for the 5HT re-uptake inhibitors up to 1236 for the substance P antagonists). If we consider the average over all activity classes, it will be seen that this upper-bound is only fractionally better than the data fusion result in Table 2 and is actually worse than the BKD figure. Thus, on average, picking any ten active reference structures and combining them using S_{max} data fusion or BKD will enable searches to be carried out that are comparable to even the best possible conventional similarity search using a single active reference structure.

Choosing between S_{max} data fusion and BKD is difficult; the latter is more effective here, but is much more time-consuming; however, the first set of experiments show that it is far more effective than data fusion when large amounts of training-data are available, rather than the very small sets of ten actives used here. We hence conclude that BKD will, in general, result in better rankings of a test-set than will the far simpler data fusion approach.

CONCLUSIONS

Virtual screening provides an ideal domain for the application of machine learning techniques, many of which are designed specifically for binary categorization problems (the categories here being active and inactive). At the same time, the characteristics of chemical datasets are very different from those common in much machine-learning research: very extensive use is made of binary, rather than real-valued, object representations (i.e., fragment bit-strings); the data are very numerous (chemical datasets are typically of size 10^5 - 10^6); and the two categories are markedly different in size (since actives are far, far less common than inactives). It is thus of some interest to test new machine-learning methods in this domain, and this has been the principal driver for our studies of the use of kernel discrimination methods for virtual screening. It is very important not to over-emphasize the advantages of some new computational approach. Even so, the results presented here do suggest that BKD provides an attractive focus for future research: it is effective in operation, in both exact and approximate forms; it has been applied successfully to both pharmaceutical and agrochemical datasets; and it can be used with very large files of 2D fingerprints.

We can regard a kernel function as a new type of similarity measure and it is hence of interest to consider the three components of a chemical similarity measure [5], *viz* the structure representation, the similarity coefficient and the weighting scheme that are used. We intend to investigate all of these aspects in the context of kernel functions. For example, previous work has shown that the Hamming Distance is far less effective for fingerprint-based similarity searching than coefficients such as the Tanimoto Coefficient and the Cosine Coefficient, and it would therefore be of interest to consider the use of these, and other, coefficients with other types of kernel function. Again, one could employ weighted fingerprints, with the bits being assigned weights derived from substructural analysis of training-set molecules, or look at alternative types of representation. We intend to investigate these developments of the basic BKD approach in the future.

ACKNOWLEDGEMENTS

We thank the following: the Novartis Institutes for Biomedical Research for funding J.H.; Syngenta for funding D.J.W.; John Delaney, Kevin Lawson and Graham Mullier (Syngenta) and Pierre Acklin, Kamal Azzaoui Edgar Jacoby and Ansgar Schuffenhauer (Novartis) for helpful comments on this work; MDL Information Systems Inc. for the provision of the MDDR database; and Barnard Chemical Information Ltd., Daylight Chemical Information Systems Inc., the Royal Society, Tripos Inc. and the Wolfson Foundation for software and laboratory support. The Krebs Institute for Biomolecular Research is a designated biomolecular sciences centre of the Biotechnology and Biological Sciences Research Council.

REFERENCES

- [1] Böhm, H.-J., Schneider, G. (Eds) (2000) *Virtual Screening for Bioactive Molecules*. Wiley-VCH, Weinheim.
 - [2] Klebe, G. (Ed.) (2000) *Virtual Screening: an Alternative or Complement to High Throughput Screening*. Kluwer, Dordrecht.
 - [3] 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>
 - [4] Harper, G., Bradshaw, J., Gittins, J.C., Green, D.V.S., Leach, A.R. (2001) Prediction of biological activity for high-throughput screening using binary kernel discrimination. *J. Chem. Inf. Comput. Sci.* **41**:1295-1300.
 - [5] Willett, P., Barnard, J.M., Downs, G.M. (1998) Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **38**:983-996.
 - [6] Sheridan, R.P., Kearsley, S.K. (2002) Why do we need so many chemical similarity search methods? *Drug Discov. Today* **7**:903-911.
 - [7] Xue, L., Stahura, F.L., Godden, J.W., Bajorath, J. (2001) Fingerprint scaling increases the probability of identifying molecules with similar activity in virtual screening calculations. *J. Chem. Inf. Comput. Sci.* **41**:746-753.
 - [8] Schuffenhauer, A., Floersheim, P., Acklin, P., Jacoby, E. (2003) Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci.* **43**:391-405.
 - [9] The MDL Drug Data Report database is available from MDL Information Systems Inc. at <http://www.mdli.com>
 - [10] Wilton, D.J., Willett, P., Lawson, K., Mullier, G. (2003) Comparison of ranking methods for virtual screening in lead-discovery programs. *J. Chem. Inf. Comput. Sci.* **43**:469-474.
-

Virtual Screening Using Binary Kernel Discrimination

- [11] Hert, J., Willett, P., Wilton, D.J., Acklin, P., Azzaoui, K., Jacoby, E., Schuffenhauer, A. (2004) Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **44**:1177-1185.
- [12] Cramer, R.D., Redl, G., Berkoff, C.E. (1974) Substructural analysis. A novel approach to the problem of drug design. *J. Med. Chem.* **17**:533-535.
- [13] Ormerod, A., Willett, P., Bawden, D. (1989) Comparison of fragment weighting schemes for substructural analysis. *Quant. Struct.-Activ. Relat.* **8**:115-129.
- [14] The Unity software is available from Tripos Inc. at <http://www.tripos.com>
- [15] Shemetulskis, N.E., Weininger, D., Blankley, C.J., Yang, J.J., Humblet, C. (1996) Stigmata: an algorithm to determine structural commonalities in diverse datasets. *J. Chem. Inf. Comput. Sci.* **36**:862-871.
- [16] Ginn, C.M.R., Willett, P., Bradshaw, J. (2000) Combination of molecular similarity measures using data fusion. *Perspect. Drug Discov. Design* **20**: 1-16.
-