Prediction of Biological Activity for High-Throughput Screening Using Binary Kernel Discrimination

Gavin Harper,*,[#] John Bradshaw,[#,†] John C. Gittins,[§] Darren V. S. Green,[#] and Andrew R. Leach[#]

GlaxoSmithKline Research and Development, Medicines Research Centre, Gunnels Wood Road, Stevenage SG1 2NY, U.K., and Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, U.K.

Received August 11, 2000

High-throughput screening has made a significant impact on drug discovery, but there is an acknowledged need for quantitative methods to analyze screening results and predict the activity of further compounds. In this paper we introduce one such method, binary kernel discrimination, and investigate its performance on two datasets; the first is a set of 1650 monoamine oxidase inhibitors, and the second a set of 101 437 compounds from an in-house enzyme assay. We compare the performance of binary kernel discrimination with a simple procedure which we call "merged similarity search", and also with a feedforward neural network. Binary kernel discrimination is shown to perform robustly with varying quantities of training data and also in the presence of noisy data. We conclude by highlighting the importance of the judicious use of general pattern recognition techniques for compound selection.

1. INTRODUCTION

The prediction of biological activity in structurally diverse datasets is difficult using conventional techniques, particularly when the prediction is based upon tens or hundreds of thousands of screening results. Yet the prediction of biological activity based on screening data is useful in various practical scenarios. It may be that some proportion of a company's compound collection have been screened, and we wish to know which further compounds to test from the remaining compounds in the collection. Perhaps we have screened most or all of a company's compound collection, but we wish to know which compounds to acquire from external suppliers. It could be that we are proposing to design a focused library of compounds for rapid synthesis in a therapeutic area of which we already have some screening experience but want to know which of the potential libraries that we could make would be most likely to yield active compounds. Wherever we have screening data available, and a set of available compounds to select from, methods for predicting activity can be of great value.

Some traditional approaches to structure−activity relationships are based on linear regression[1,2] or other linear methods such as partial least squares.[3] Real-life relationships between structure and activity are often highly nonlinear, particularly when considered across very diverse sets of chemical structures. In an attempt to increase the range and flexibility of relationships between structure and activity which can be modeled, many recent approaches to structure−activity modeling have borrowed techniques from the machine learning literature.[4−7] Such models can be very flexible (a feed forward neural network with linear output units and a single hidden layer can approximate any continuous function with arbitrary accuracy[8]); however, there are often attendant worries such as over-fitting and sensitivity to noise.

Other approaches are based on estimating (either implicitly[9,10] or explicitly[11]) the probability of activity for compounds under the assumption that descriptor variables are stochastically independent. Such an assumption is not generally valid, and the ability of such methods to model complex relationships over diverse sets of molecules is likely to be compromised.

There has recently been some success in applying recursive partitioning methods to problems involving large numbers of diverse compounds.[12,13] Recursive partitioning methods have the advantage of interpretability via the set of rules by which one arrives at any given node in the resulting tree. However not all problems will be amenable to such approaches. Moreover, even when a complex dataset is amenable to analysis, it is unlikely that all possible structure−activity relationships are exploited by any single analytical technique. Although it seems likely that recursive partitioning will prove a useful method for the interpretation of high-throughput screening data, there is a clear need for further algorithms.

A recent paper[14] suggests the use of nonparametric kernel regression methods for structure−activity relationships. The purpose of kernel regression methods is to predict the value of a continuous response variable given the values of continuous descriptors. In this paper, we consider binary kernel discrimination, a method that is closely related to kernel regression. However, when we use binary kernel discrimination, we are interested in predicting the class (active or inactive) of a molecule given a binary representation of its structure.

* Corresponding author phone: +44 (0)1438 768416; e-mail: gh75680@ggr.co.uk.

† Current address: Daylight CIS, Sheraton House, Castle Park, Cambridge CB3 0AX, U.K.

GlaxoSmithKline Research and Development, Medicines Research Centre.

§ University of Oxford.

10.1021/ci000397q CCC: $20.00
Published on Web 05/05/2001

The purpose of this paper is to introduce binary kernel discrimination as a useful tool for activity prediction. We show that it provides a useful method for using the known activity or inactivity of some molecules (a training set) in a screen, to predict in advance of screening which other molecules are most likely to prove active on that screen. We demonstrate the method on two datasets—one well-known small set from the literature and a rather larger proprietary dataset—under a variety of conditions and compare its performance to that of two other methods. It is important to stress, however, that we believe that using different methods for activity prediction can often give a broader understanding of the data than using any one method alone and that we would advocate the use of several methods wherever possible. Hence we see binary kernel discrimination as a useful addition to the computational chemist's toolbox, rather than being in direct competition with the other tools available.

2. METHODS

Descriptors. We used atom pairs[15] and topological torsions[16] (APTT) as descriptors for all the experiments reported in this paper, using the same algorithm to generate them as described in Rusinko et al.[12] From the presence or absence of the APTT descriptors, a binary structural fingerprint was derived for each molecule. Every APTT descriptor present in at least one molecule of the dataset used to train the algorithms was included in this fingerprint. By using the same descriptors for all algorithms we hoped to avoid confounding the effect of descriptor choice with the effect of algorithm choice. However, we acknowledge that some algorithms may be better suited to certain descriptors than others. It is not our intention to tackle here the question of which is the most appropriate set of descriptors to use.

Algorithms. Kernel Discrimination. Suppose that we choose to describe the physical or structural properties of molecules in some multidimensional descriptor space. Then we can use a family of methods called kernel density estimators to estimate the parent distribution of a sample of molecules from a population (which we call the training set) based on those descriptors. These methods are nonparametric—they make no strong assumptions regarding the underlying distributional form of the data.

Given a set of n points $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ sampled from some underlying population, the *kernel density estimate* at a point $\mathbf{x}$ is given by

$$\hat{p}(x) = \frac{1}{n}\sum_{i=1}^{n} K_\lambda(\mathbf{x} - \mathbf{x}_i)$$

where K is a symmetric (multidimensional) density function (for example the multivariate Normal density function) which normally relies on the value of a smoothing parameter λ. The value of the smoothing parameter affects the range of influence of each point in the sample, which in turn affects the smoothness of the function $\hat{p}$. An example in one dimension using the univariate Normal kernel is shown in Figure 1. In this case, the smoothing parameter is the variance σ^2. We shall discuss how to choose an appropriate value of λ below.

Suppose that a kernel density estimate is made both for the active molecules in a set, and separately for the inactive molecules. Then by looking at the ratio of the estimated distributions of actives and inactives, we may obtain an estimate of how likely any particular molecule is to be active as opposed to inactive. This can be justified by a simple probabilistic argument, and the resulting method is called kernel discrimination. Suppose that π_A and π_I are the probabilities that a molecule selected at random from the population is active or inactive, respectively. Suppose also that $p(\mathbf{x}|\ A)$ and $p(\mathbf{x}|\ I)$ are the probabilities that a molecule with descriptor vector $\mathbf{x}$ occurs given that we select it at random from the active or inactive populations, respectively. Then the estimated probability $\hat{p}(A|\ \mathbf{x})$ that a compound with descriptor vector $\mathbf{x}$ is active is given by

$$\hat{p}(A|\ \mathbf{x}) = \frac{(\pi_A/n_A)\sum_{i\in A} K_\lambda(\mathbf{x} - \mathbf{x}_i)}{(\pi_A/n_A)\sum_{i\in A} K_\lambda(\mathbf{x} - \mathbf{x}_i) + (\pi_I/n_I)\sum_{i\in I}^{n} K_\lambda(\mathbf{x} - \mathbf{x}_i)}$$

In fact, ranking compounds by $\hat{p}(A|\ \mathbf{x})$ is equivalent to ranking them by

$$L(A|\ \mathbf{x}) = \frac{\sum_{i\in A} K_\lambda(\mathbf{x} - \mathbf{x}_i)}{\sum_{i\in I} K_\lambda(\mathbf{x} - \mathbf{x}_i)}$$

The higher the value of $L(A|\ \mathbf{x})$, the more probable we think it is that a compound is active.

There has been much work on kernel density estimation for continuous descriptors[17,18] but relatively little using binary

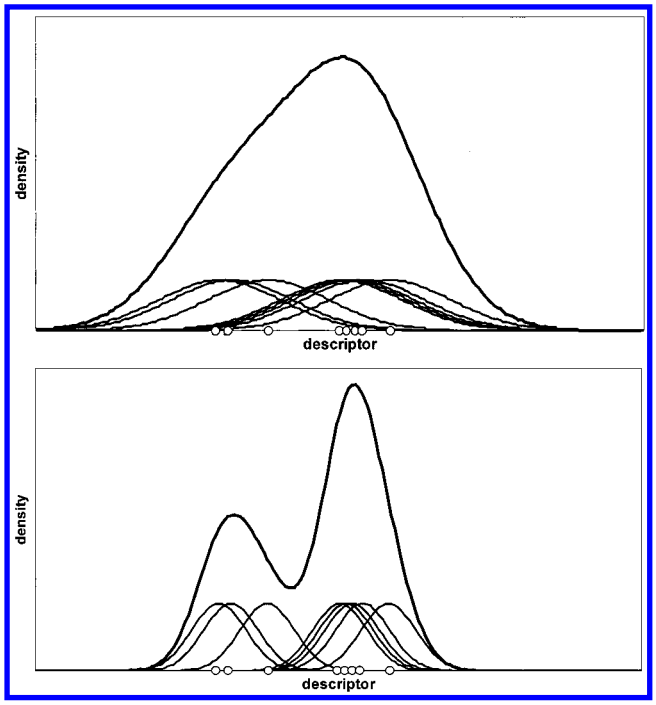

Figure 1. Normal kernel density estimates using a sample of 8 points on one descriptor variable. The results when the kernel function has variance $\sigma^2 = 1$ (top) and $\sigma^2 = 0.2$ (bottom) are shown. The contributions to the density estimate from each point in the sample are also plotted and are summed to give the density estimate. The top plot suggests a single peak to the distribution, while the bottom one suggests a bimodal distribution.

descriptors. However, Aitchison and Aitken[19] proposed a form of kernel function which is based on binary descriptors. This methodology is of particular interest in the current paper, since descriptors based on chemical structure are often binary, representing the presence or absence of substructural fragments.

Suppose that **w** and **x** are binary vectors of length M. Suppose also that the vectors differ in value at d of their M entries. Then Aitchison and Aitken[19] suggest the kernel function

$$K_\lambda(\mathbf{x},\mathbf{w}) = \lambda^{M-d}(1-\lambda)^d$$

If λ is 0.5, this is just a uniform density. As λ approaches 1, the density estimate will become less smooth—analogous to the effect of allowing the value of σ^2 to approach 0 in Figure 1—with a peak at $d = 0$. There is no need to have the same value of smoothing parameter for estimating the distribution of actives and estimating the distribution of inactives; however, it is a convenient simplification (and results in much greater computational efficiency) which we adopt here. We choose the value of the smoothing parameter λ as follows.

FOR each value of λ (0.50, 0.51, 0.52, ..., 0.99) DO 1 to 3.

1. For each point in the training set, calculate $\hat{p}(A|\mathbf{x}_i)$ based on all the other points in the training set.

2. Rank the training set points in descending order of $\hat{p}(A|\mathbf{x}_i)$ as calculated in 1.

3. Sum the ranks of the true actives, and call this sum S_λ.

Choose the smoothing parameter λ that minimizes S_λ.

This is an extremely crude way of choosing λ, and superior methods are given by Aitchison and Aitken[19] and Tutz[20,21] among others. Such methods tend to be computationally intensive, hence our pragmatic approach here, since we intend to apply the algorithm to very large datasets.

Merged Similarity Search. To be of practical use, a method for compound selection must be able to perform at least as well as the most trivial approaches to compound selection. For this reason we introduce as a benchmark a simple extension to the similarity searching capabilities found in commercial software packages in order to produce a standard algorithm for compound selection based on similarity searching.

In this paper we will use the Tanimoto coefficient of similarity $S(\mathbf{x},\mathbf{y})$. It is defined by the ratio of the number of bits set to 1 in both **x** and **y** to the number of bits set to 1 in either **x** or **y**. We define the "merged similarity score"[22] by

$$M(\mathbf{x}) = \max_{i\in A} S(\mathbf{x},\mathbf{x}_i)$$

In other words, the merged similarity score of a new molecule is its similarity to the nearest (most similar) active compound in the training set. The higher the value of $M(\mathbf{x})$, the more likely we judge it to be that a molecule is active.

Note that all that we are doing is carrying out a similarity search for each active in the training set and ranking the molecules that we find by the similarity to their nearest active neighbor. This is a plausible and consistent course of action for a computational chemist to undertake manually.

Neural Network. To provide a potentially more challenging benchmark than merged similarity search, we also make a comparison with a feedforward neural network with a single hidden layer, implemented using the SNNS program.[23] A method similar to the one we use has previously been successfully applied to a classification problem with two classes, where the task was to discriminate between drugs and nondrugs.[5] Our neural network has a single hidden layer containing 50 neurons, a single output neuron, and full connection between adjacent layers. The number of input neurons is equal to the number of distinct APTT descriptors present in the training set. Of the two datasets we examine in Section 3, the first (monoamine oxidase) contains 6158 distinct APTT descriptors, and the second (HTS) contains 19 795. The training of the neural network is performed over 100 cycles with a learning rate of 0.2 and a momentum term of 0.1 using the "backpropogation with momentum scheme" as implemented in SNNS. The order of presentation of the training set to the network is reshuffled before each cycle.

3. RESULTS

In this section, we study the performance of binary kernel discrimination and merged similarity search on two datsets. The first is a small publicly available dataset. The second is a larger set, which comes from an in-house high-throughput screening project.

Monoamine Oxidase Data. The first dataset studied is a set of 1650 compounds tested for monoamine oxidase (MAO) inhibition.[24] Activity is recorded on a four-point scale—inactive compounds are recorded as having activity 0, while the values 1, 2, and 3 correspond to increasing levels of activity. For the purposes of our analyses, we do not distinguish between different levels of activity, so that all class 1, 2, and 3 compounds (290 compounds in all) are treated simply as "actives" and class 0 compounds as "inactives". We undertake a sequence of three analyses, for each of which we partition the full set of 1650 compounds into two parts. The first part, which we call the "training set", is of fixed size and made up of compounds chosen at random from the 1650. The three algorithms described in Section 2 are trained using this set. The "test set" is made up of the remaining compounds. The algorithms attempt to rank the compounds in the test set according to the chances of them proving to be active. We use this ranking to choose a subset of 200 compounds from the test set with the intention of maximizing the number of actives in this subset. For each of the three analyses reported below, we repeat the experiment five times, choosing a new training set each time. This is to counteract the possibility of a particularly atypical training set occurring by chance and giving a biased picture of algorithmic performance, as might well happen if we were only to undertake a single replicate of the experiment. The results which we report are averaged over the five replicates and are given in Figure 2.

In the first MAO experiment, the training set was made up of half (825) of the compounds, leaving a further 825 in the test set. The cumulative number of actives found as we select in order the 200 compounds predicted most likely to be active by each algorithm is given by the first plot in Figure 2. From this plot, it is possible to read off the average number of actives that are found if we only pick 50 or 100 compounds rather than 200 by observing y-values at the corresponding value of the x-axis. There is little difference

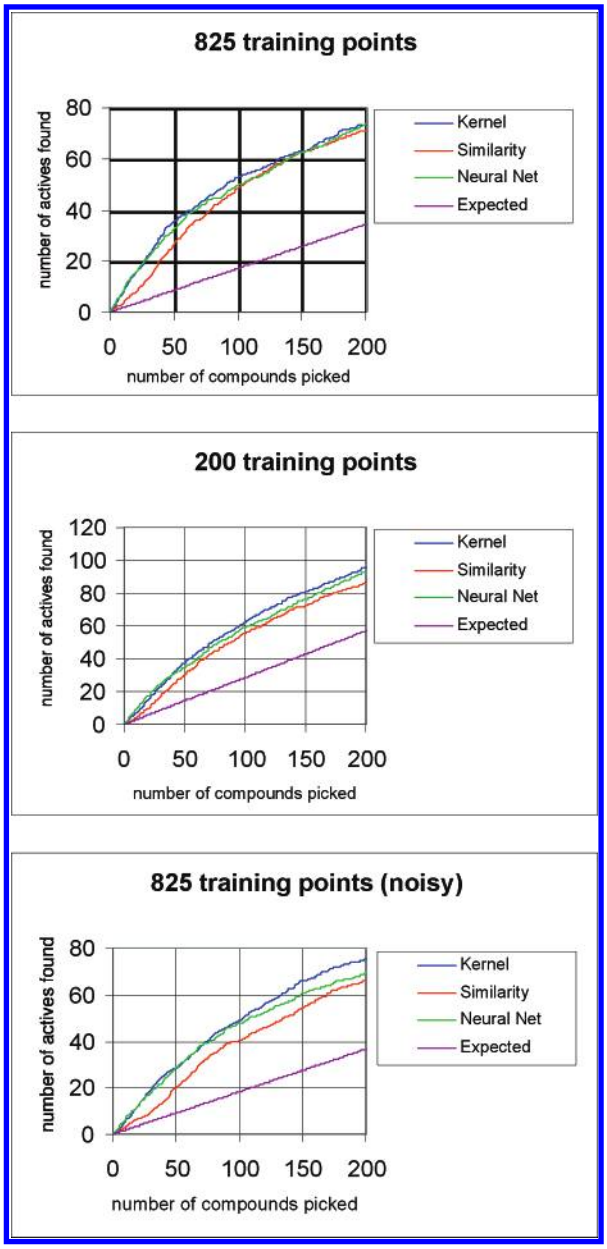

Figure 2. MAO data. The average (over five replicates of the experiment) number of actives found when we pick the top-ranked compounds as predicted by different algorithms are plotted against the number of compounds picked. The number of actives that we would expect to find by random selection is also plotted under the heading "expected". The three different scenarios that lead to the three plots are described in the text.

between the performance of the three algorithms on this dataset, so neither binary kernel discrimination nor the neural network is able to clearly outperform the simple merged similarity search algorithm. All three algorithms greatly outperform the expected number of actives when compounds are selected at random.

In the second MAO experiment, the training set was made up of just 200 compounds, leaving 1450 compounds in the test set. The training set data here is sparser than in the first experiment, and one would expect it to be difficult to predict activity based on this training set. Nonetheless, all three algorithms do quite well (compared to what would be expected if compounds were selected at random), suggesting that there are some very obvious structural features which can be picked out in this dataset and define activity in some subset of the actives. Binary kernel discrimination has performed quite reasonably, which may give some comfort that the algorithm does not behave pathologically when training data is sparse. This is consistent with its nonparametric nature. The neural network performs similarly.

In the third MAO experiment, the training set was made up of the same 825 compounds as in the first experiment, but this time with 100 of the training set inactives selected at random and reclassified as active. This was to simulate the effect of an increased noise level, with a number of false positives being identified in the training set. In this experiment, binary kernel discrimination performs better than merged similarity search. We would suggest that this is because binary kernel discrimination uses the information that there exist inactives in the neighburhood of the false positives and so will tend not to select so many compounds in these regions of space. Since merged similarity search has no way of using the information about inactive compounds, it fares comparatively poorly (although it is still far better than would be expected if compounds were selected at random). The neural network performs better than merged similarity search but more poorly than binary kernel discrimination.

The differences in performance between algorithms are relatively small in all three experiments with the MAO dataset. The good performance of merged similarity search suggests that the dataset gives a relatively "easy" problem. Analysis of a larger, noisier dataset would give a much better idea of how well these methods might work in practice with data from high-throughput screening.

HTS Dataset. The second dataset studied is a larger set of data from an in-house enzyme assay at Glaxo Wellcome. This set contained 106 437 compounds, of which 2332 were classified as active on the basis of percentage inhibition, and the rest as inactive. Mirroring the analysis with the MAO data, we carried out five replicates of three experiments. The aim for these experiments was to select 5000 compounds from the test set. The results are shown in Figure 3.

In the first experiment there were 5000 training points which were selected at random from the full 106 437. The test set was made up of the remaining 101 437 compounds. Binary kernel discrimination clearly outperforms merged similarity search in this experiment. The neural network also performs well but not as well as binary kernel discrimination.

In the second experiment, we allowed only 500 training compounds, which were again selected at random. Despite the relative sparseness of the training set data, once again binary kernel discrimination clearly outperforms merged similarity search. However, the neural network is the most effective of the three methods in this case.

In the third experiment, we chose at random 500 inactive training set compounds from the first experiment and reclassified them as active. Despite the added noise from these misclassifications, kernel discrimination performs nearly as well as in the first experiment. The performance of merged similarity search and the neural network is rather poor, with the performance of the neural network in particular suffering in comparison with the first experiment.

4. DISCUSSION

The superiority of binary kernel discrimination to merged similarity search is backed up by our further experience of

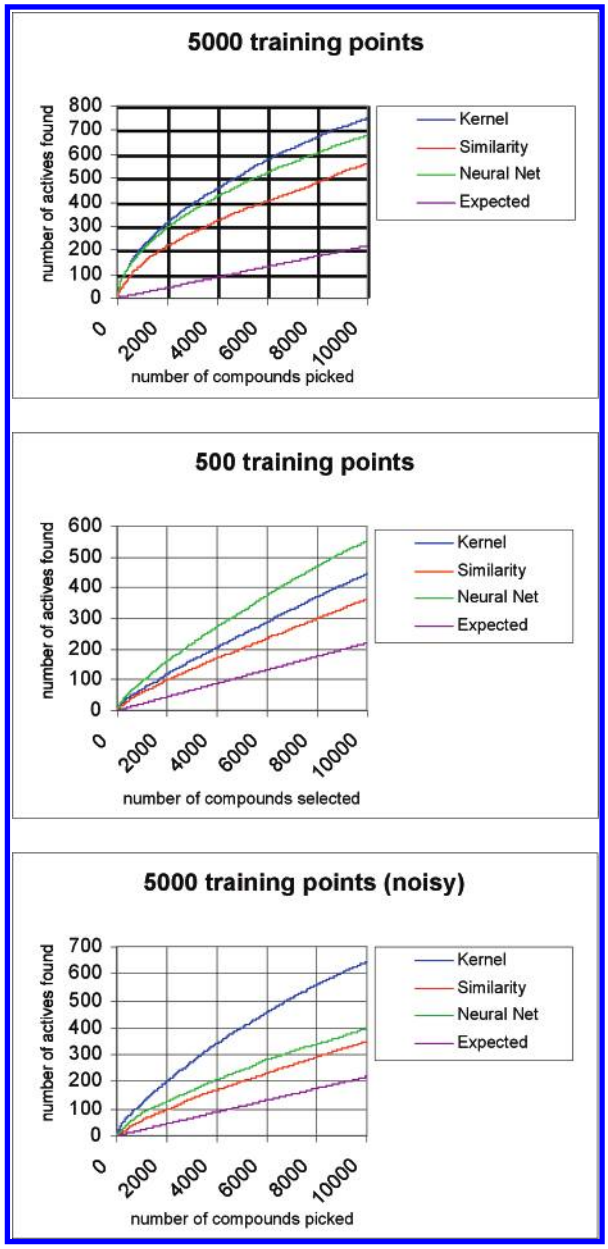

Figure 3. In-house enzyme assay data. The average (over five replicates of the experiment) number of actives found when we pick the top-ranked compounds as predicted by different algorithms are plotted against the number of compounds picked. The number of actives that we would expect to find by random selection is also plotted under the heading "expected". The three different scenarios that lead to the three plots are described in the text.

using these methods on high-throughput screening data. Merged similarity search performs well on the MAO data largely because of the highly related nature of many of the compounds in the dataset, allowing a simple "near match" identification of actives in the test set given those in the training set. It is quite likely that experimenting with the parameters of the neural network would improve its performance. However, the comparison of binary kernel discrimination with the neural network implemented does suggest that the method has a performance which is comparable to other state-of-the-art techniques. In our implementation of binary kernel discrimination, the smoothing parameter is chosen automatically, and this is a strength of the method, since it makes its use straightforward for the nonspecialist.

Data from high-throughput screening programs often produce a high degree of noise in the assay results and include the assaying of a structurally highly diverse range of chemical compounds. Pattern recognition methods such as binary kernel discrimination and neural networks can help to identify further compounds to screen using this kind of data. However, such methods need to be used judiciously and not considered a complete replacement for chemists' knowledge.

In this paper we have presented methods for predicting further active molecules given the results of a biological screen. In complex problems, these methods will typically identify actives at a higher rate than would be likely to occur if compounds were chosen for screening purely at random. At the same time, it can be seen that many active compounds are not found using these methods. This is an example of a trait that is common when using predictive methods for highly complex problems—they may successfully model some of the more obvious relationships between structure and activity in the data but are unlikely to fully exploit all such relationships.

Although it is convenient, we note that "hit rate" is by itself an overly simplistic criterion of algorithmic success when in practice, often the objective is to identify several distinct structural classes of active compounds. Implied by this objective is the existence of a precise and absolute description of what constitutes distinct structural classes, itself a difficult problem. In whatever way such classes are defined, it is clear from this objective that a good "hit rate" does not justify the blind use of an algorithm for compound selection. The algorithms presented here predict the likely activity of compounds that have yet to be screened, but the structural diversity of the set of compounds selected for further screening should also be taken into consideration. If different structural classes of active molecules are to be found, then a balance must be maintained between focusing on molecules with a strong predicted chance of activity and continuing to explore less promising avenues. The correct balance between focus and diversity is difficult to pin down without very specific ideas about what is a desirable criterion to optimize (for example, the expected number of actives or the probability of finding at least some minimum number of actives). However some preliminary work of a theoretical nature has been undertaken which examines this question.[22,25] It is a question that is of huge strategic importance and has a clear practical impact upon the effectiveness with which pattern recognition techniques are integrated into a screening regime.

We regard the comparison of an algorithm for predicting activity with the random selection of compounds to be an insufficient justification for its use. That is why in this paper we have chosen to compare the performance of binary kernel discrimination with merged similarity search, which provides a baseline measure of performance that mimics a realistic selection procedure based on similarity searching, and a neural network, which is a more sophisticated method. Limiting evaluation of the performance of algorithms for compound selection to a comparison with the hit rate achieved by the random selection of compounds provides an unrealistic assessment of their utility. This message is particularly borne out by the MAO dataset examined above.

The use of a binary split into active and inactive compounds is controversial, and it is inevitable that some compounds will be incorrectly classified by such a procedure. However, the compounds which show up as weakly active in preliminary screening can often be developed into compounds with higher potency at a later stage in the development process and have the potential to be novel leads. Such compounds are considered on an equal footing to more potent actives when a simple binary split is used. Methods that use the magnitude of a biological response rather than a binary activity value are less influenced by weakly active compounds than they are by compounds showing particularly strong activity.

Binary kernel discrimination is a "black box" method. No attempt is made to explicitly find the rules that render a compound active. Hence binary kernel discrimination can only be used to estimate the chances of a molecule being active—not to suggest what structural changes to make a priori to improve its activity. This is clearly a disadvantage of the method. However, the method has many attractive properties. Very complicated relationships, including the implicit modeling of interactions,[19] can be achieved. We have shown that the behavior of the algorithm is good using sparse data, in contrast to recursive partitioning[12] (although binary kernel discrimination is also most effective when there is a large quantity of training data, and the neural network provides a better performance on the HTS dataset presented here). We have also shown that the method performs particularly well when the dataset is noisy. Binary kernel discrimination builds a model in which the estimated probability of activity of a compound is heavily influenced by structurally close compounds in the training set. Hence the probability estimate is local in nature, which allows binary kernel discrimination to deal with very structurally diverse compounds acting through different mechanisms.

5. CONCLUSIONS

Binary kernel discrimination is a method that has several attractive features for using screening data to predict the biological activity of further compounds. It can deal with both noisy data and sparse data. It is possible to apply the method to datasets with hundreds of thousands of compounds and to use a range of different binary descriptors to describe the data. This allows the potential application of the method to a wide variety of problems. At present, neither binary kernel discrimination nor other pattern recognition methods correctly predict all active molecules. Hence it is important to see pattern recognition algorithms as a complement to, rather than a replacement for, other compound selection techniques and sampling strategies.

ACKNOWLEDGMENT

We thank Chris Keefer and Stan Young of GlaxoSmith-Kline, Research Triangle Park, North Carolina, for supplying us with the computer code to generate the atom pairs and topological torsions descriptors, and Yvonne Martin of Abbott Laboratories for providing us with the MAO dataset.

REFERENCES AND NOTES

(1) Free, S. M.; Wilson, J. W. A mathematical contribution to structure−activity studies. *J. Med. Chem.* **1964**, *7*, 395−399.
(2) Hansch, C. A quantitative approach to biochemical structure−activity relationships. *Acc. Chem. Res.* **1969**, *2*, 232−239.
(3) Wold, S. In *Systems under Indirect Observation − Volume II*; Joreskog, K.-G., Wold, H., Eds.; North-Holland, Amsterdam, 1982.
(4) Burden, F. R. Using artificial neural networks to predict biological activity from simple molecular structural considerations. *Quant. Struct.-Act. Relat.* **1996**, *15*, 7−11.
(5) Sadowski, J.; Kubinyi, H. A scoring scheme for discriminating between drugs and nondrugs. *J. Med. Chem.* **1998**, *41*, 3325−3329.
(6) King, R. D.; Srinivasan, A. The discovery of indicator variables for QSAR using inductive logic programming. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 571−580.
(7) Bolis, G.; Di Pace, L.; Fabrocini, F. A machine learning approach to computer-aided molecular design. *J. Comput.-Aided Mol. Des.* **1991**, *5*, 617−628.
(8) Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics of Control Signals and Systems* **1989**, *2*, 303−314,
(9) Cramer, R. D., III.; Redl, G.; Berkoff, C. E.. Substructural analysis. A novel approach to the problem of drug design. *J. Med. Chem.* **1974**, *17*, 7, 533−535.
(10) Ormerod, A.; Willett, P.; Bawden, D. Comparison of fragment weighting schemes for substructural analysis. *Quant. Struct.-Act. Relat.* **1989**, *8*, 115−129.
(11) Gao, H.; Williams, C.; Labute, P.; Bajorath, J. Binary quantitative structure−activity relationship (QSAR) analysis of estrogen receptor ligands. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 164−168.
(12) Rusinko, A., III.; Farmen, M. W.; Lambert, C. G.; Brown, P. L.; Young, S. S. Analysis of a large structure/biology activity data set using recursive partitioning. *J. Chem. Inf. Comput. Sci.* **1999**, *39*, 1017−1026.
(13) Cho, J. C.; Shen, F.; Hermsmeier, M. A. Binary formal inference-based recursive modeling using multiple atom and physicochemical property class pair and torsion descriptors as decision criteria. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 668−680.
(14) Constans, P.; Hirst, J. D. Nonparametric regression applied to quantitative structure−activity relationships. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 452−459.
(15) 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.
(16) Nilakantan, R.; Baumann, N.; Dixon, J. S.; Venkataraghavan, R. Topological torsions: a new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82−85.
(17) Scott, D. W. *Multivariate Density Estimation: Theory, Practice, and Visualization*; John Wiley: New York, 1992.
(18) Simonoff, J. S. *Smoothing Methods in Statistics;* Springer-Verlag: Berlin, 1996.
(19) Aitchison, J.; Aitken, C. G. G. Multivariate binary discrimination by the kernel method. *Biometrika.* **1976**, *63*, 413−420.
(20) Tutz, G. An alternative choice of smoothing for kernel-based density estimates in discrete discriminant analysis. *Biometrika.* **1986**, *73*, 405−411.
(21) Tutz, G. On cross-validation for discrete kernel estimates in discrimination. *Commun. Statistics − Theory Methods* **1989**, *18*, 4145−4162.
(22) Harper, G. *The Selection of Compounds for Screening in Pharmaceutical Research;* D.Phil. thesis, University of Oxford, 1999.
(23) SNNS: Stuttgart Neural Network Simulator; Version 4.2, University of Stuttgart, 1998.
(24) Brown, R. D.; Martin, Y. C. Use of structure−activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.
(25) Gittins, J. C.; Harper, G. Bounds on the Performance of a Greedy Algorithm for Probabilities. To appear in *Math. Operations Res.*.

CI000397Q