J Comput Aided Mol Des (2007) 21:53–62
DOI 10.1007/s10822-006-9096-5

ORIGINAL PAPER

Evaluation of machine-learning methods for ligand-based virtual screening

Beining Chen · Robert F. Harrison · George Papadatos · Peter Willett · David J. Wood · Xiao Qing Lewell · Paulette Greenidge · Nikolaus Stiefl

Received: 2 November 2006 / Accepted: 4 December 2006 / Published online: 5 January 2007

Abstract Machine-learning methods can be used for virtual screening by analysing the structural characteristics of molecules of known (in)activity, and we here discuss the use of kernel discrimination and naive Bayesian classifier (NBC) methods for this purpose. We report a kernel method that allows the processing of molecules represented by binary, integer and real-valued descriptors, and show that it is little different in screening performance from a previously described kernel that had been developed specifically for the analysis of binary fingerprint representations of molecular structure. We then evaluate the performance of an NBC when the training-set contains only a very few active molecules. In such cases, a simpler approach based on group fusion would appear to provide superior screening performance, especially when structurally heterogeneous datasets are to be processed.

Keywords Group fusion · Kernel discrimination · Ligand-based virtual screening · Machine learning · Naive Bayesian classifier · Similarity searching · Virtual screening

B. Chen
Department of Chemistry, University of Sheffield, Western Bank, Sheffield S3 7HF, UK

R. F. Harrison
Department of Automatic Control and Systems Engineering, University of Sheffield, Mappin Street, Sheffield S1 3JD, UK

G. Papadatos · D. J. Wood
Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, 211 Portobello Street, Sheffield S10 2TN, UK

P. Willett (✉)
Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, 211 Portobello Street, Sheffield S1 4DP, UK
e-mail: p.willett@sheffield.ac.uk

X. Q. Lewell
GlaxoSmithKline Research and Development, Stevenage SG1 2NY, UK

P. Greenidge · N. Stiefl
Novartis Pharma AG, CH-4056 Basel, Switzerland

Introduction

The discovery of novel chemical entities (NCEs) in the pharmaceutical industry is becoming increasingly difficult, costly and time-consuming. Many approaches have been suggested to increase the cost-effectiveness of discovery programmes, one of them being the use of virtual, or in silico, screening methods to complement the more traditional chemical and biological approaches [1–3]. Virtual screening involves the computational filtering of a large body of molecules (e.g., those comprising a corporate database) to identify those that have a high probability of activity in the biological test system of interest. Thus a virtual screening method takes as input all those molecules that might be acquired (or synthesised) and tested, and then outputs those few that should be tested. Similar techniques are also used for the discovery of NCEs in the agrochemicals industry [4], but we shall restrict ourselves here to the problems of drug discovery.

There are two basic approaches to virtual screening. The popular structure-based approaches require the availability of a 3D structure for the biological target of interest, this information permitting the use of methods based on de novo design or ligand-protein docking

[5–7], these methods becoming widely used with the continuing growth in the availability of 3D protein data [8, 9]. If the necessary 3D information is not available then it may be possible to identify the structural requirements for activity by analysis of those structures that have already been shown to be active. There are several such ligand-based approaches that can be used for virtual screening. If just a single active is available, e.g., a literature or competitor compound, then a similarity search can be used to identify molecules that are structurally similar and that might thus be expected to exhibit similar activity characteristics [10–12]. If several actives are available then a pharmacophore mapping procedure can be used to determine whether they possess some common substructural feature(s) [13, 14]; if this proves to be the case then the common features can be used as the query for a 3D substructure search [14–16]. However, the initial hits from a high-throughput screening (HTS) programme are often structurally diverse, in which case it may not be possible to identify any structural commonalities, even if the molecules exhibit the same mode of action. This is an increasingly common situation for which machine-learning approaches are suitable.

A machine-learning method takes as input a training-set of objects that have previously been classified into two or more classes [17]; in the virtual screening context, this would be a set of molecules that had previously been tested and shown to be either active or inactive. These training-set molecules are then analysed to develop a decision rule that can be used to classify new molecules (the test-set) into one of the two classes. The first application of machine learning in computer-aided molecular design (CAMD) was probably substructural analysis, which was introduced by Cramer et al. in the early Seventies as a tool for the automated analysis of biological screening data [18, 19]. Machine learning is now a very active area of research in computer science, with the increasing availability of large data repositories of all sorts spurring interest in the development of novel tools for data mining [20, 21]. Such tools are now starting to be applied to the analysis of chemical datasets, with CAMD applications in the last few years involving decision trees [22], support vector machine (SVMs) [23], recursive partitioning [24] and binary kernel discrimination (BKD) [25], inter alia.

The concept of molecular similarity lies at the heart of such methods, since no machine-learning method can reasonably be expected to discriminate between active and inactive test-set molecules unless there are some structural commonalities (in terms of the descriptors available) between the training-set actives and/or structural dissimilarity between the training-set actives and inactives. The relationship between structure and property was first enunciated explicitly by Johnson and Maggiora [26], whose Similar Property Principle states that molecules that are structurally similar are likely to have similar properties. Whilst there are many exceptions to the Principle [27, 28], its appropriateness seems self-evident: if there was not some form of relationship between chemical structure and biological activity then it would be difficult to develop rational approaches for drug discovery. The implications of this for database processing were first analysed in the mid-Eighties [29, 30] and molecular similarity continues to be the focus of a range of methods for CAMD [11, 12, 31, 32].

Some of the most important evidence for the applicability of the Principle has been provided by Yvonne Martin, in three much-cited papers involving retrospective analyses of large datasets. Studies in the mid-Nineties explored the extent to which different types of structural descriptor, both 2D and 3D, encoded sufficient structural information to enable the prediction of a range of physicochemical properties [33, 34]; more recent work has demonstrated the extent to which one such representation, Daylight 2D fingerprints, enables the prediction of bioactivity in HTS systems [35]. Here in Sheffield, we have had a long-standing interest in the use of similarity and machine-learning methods, principally using 2D fingerprint representations, and in this paper, we discuss briefly some of our recent work in the latter area. We first describe the use of kernel discrimination for the analysis of datasets where the molecules are characterised by non-binary variables, and then the use of a naive Bayesian classifier (NBC) when only limited amounts of training-data are available.

Use of kernel discrimination methods with non-binary molecular data

The basic model

Introduced by Parzen [36], kernel density estimators have found widespread use in pattern recognition applications. They provide a non-parametric method for building probability distributions and can be used to estimate the individual likelihood functions of different categories from sample data. Knowledge of these likelihoods can then be used in a variety of ways to produce scores (e.g., probability of category membership) or rankings (e.g., likelihood ratio) for novel objects with the ultimate aim of classifying them correctly [20]. The

technique is strongly related to the recent surge of interest in so-called kernel methods [37].

The great majority of non-chemical datasets are characterised by integer-valued or real-valued (i.e., continuous) descriptors and the kernel methods that have been developed have hence been designed to handle this sort of representation. The most common type of representation for molecules in CAMD studies is the 2D fingerprint, a binary vector in which the bits denote the presence or absence of topological substructures in a molecule. Harper et al. demonstrated that there was one kernel method, appropriately named BKD, that could be used with such data, and described the successful application of BKD to several pharmaceutical datasets [25]. Spurred by this study, we have reported the application of BKD to both pharmaceutical and agrochemical data and investigated some of the inherent characteristics of the approach [38–41]; here, we describe a kernel discrimination method for the analysis of non-binary molecular representations.

The *k*-nearest neighbour (*k*NN) algorithm is arguably the simplest machine-learning approach, and typically assumes that all objects correspond to points in multivariate real numbered space $\Re^n$. The nearest neighbours of a query object are defined in terms of the standard Euclidean distance, i.e., if an object x is described by a feature vector $\langle a_1(x), a_2(x), \ldots a_n(x)\rangle$, where $a_r(x)$ describes the value of the rth attribute of object x, then the Euclidean distance between two objects x_i and x_j is defined as $\mathrm{d}(x_i, x_j)$, where

$$\mathrm{d}(x_i, x_j) = \sqrt{\sum_{r=1}^{n} \left(a_r(x_i) - a_r(x_j)\right)^2} \tag{1}$$

Assume that a training-set molecule x has associated with it a category $f(x)$. For a binary classification task, such as the classification as active or inactive of a test-set molecule q, the value $\hat{f}(x_q)$ returned by the *k*NN algorithm as an estimate of the compound's activity $f(x_q)$ is the most common value of f among the k nearest neighbours, i.e. if $k = 1$, the algorithm returns the category of the nearest neighbour, and if $k = 5$, the algorithm returns the most common category of the five nearest neighbours.

A variation of the simple *k*NN algorithm involves weighting the influence of the near neighbours according to their distance from the query object, thus increasing the influence of those training-set examples that are closest to the query object. Typical weighting schemes are the inverse square of the Euclidean distance from the query object, or functions based on Gaussian distributions around the data point. The classification of a query molecule x_q therefore becomes

$$\hat{f}(x_q) = \max \sum_{i=1}^{k} w_i \delta(v, f(x_i)) \tag{2}$$

where

$$w_i = \frac{1}{\mathrm{d}(x_q, x_i)^2}, \tag{3}$$

and where $\delta(a, b) = 1$ if $a = b$ and $\delta(a, b) = 0$ otherwise. If the query point exactly matches a training-set example, i.e., $\mathrm{d}(x_q, x_i) = 0$, then $\hat{f}(x_q)$ is assigned to be $f(x_i)$; if there are several such training-set examples, the majority class is assigned. BKD can be considered to be a variation of the distance weighted *k*NN algorithm, with Aitchison and Aitkin's kernel function [42] providing a distance-based weighting for molecules represented as points in multivariate binary descriptor space [25].

The distance-weighted *k*NN algorithm described above can be modified to allow predictions of real-valued target functions, as in Hirst's QSAR studies of kernel-based regression [43, 44]. Here, we consider the Parzen window method for kernel density estimation [20, 36], which is an example of a distance-weighted *k*NN procedure that uses a Gaussian-based kernel to weight the influences of the training set neighbours. This kernel is defined to be

$$K_h(x_i, x) = \frac{1}{h\sqrt{2\pi}} e^{-\mathrm{d}(x_i, x)^2 / 2h^2} \tag{4}$$

where h is the bandwidth of the Gaussian, its value being determined by an analysis of the training-set data. This kernel (Eq. 4) provides a distance-based weighting for training-set points in multivariate continuous descriptor space, that can be used in just the same way as Aitchison and Aitkin's kernel is used in BKD [25]. Thus, given training data comprising a set, A, of n active molecules and a set, I, of m inactive molecules, the test-set molecules can be ranked in decreasing probability of activity by means of the function

$$KD\_SCORE(x|A) = \frac{\frac{1}{n}\sum_{i=1}^{n} K(x, A_i)}{\frac{1}{m}\sum_{i=1}^{m} K(x, I_i)} \tag{5}$$

where the kernels $K(x, A_i)$ and $K(x, I_i)$ are computed using (4) above. For brevity in what follows, we will refer to this scoring function based on the Parzen window method as continuous kernel discrimination (CKD), by analogy with BKD.

Experimental details

Our experiments used data from the MDL Drug Data Report (MDDR) database (available from MDL Information Systems Inc. at http://www.mdli.com) Specifically, training-sets and test-sets were chosen from eleven activity classes that have been studied previously by Hert et al. [39]. These activity classes were as follows: 5HT3 antagonists, 5HT1A agonists, 5HT reuptake inhibitors, D2 antagonists, renin inhibitors, angiotensin II AT1 antagonists, thrombin inhibitors, substance P antagonists, HIV protease inhibitors, cyclooxygenase inhibitors, and protein kinase C inhibitors. These classes include both structurally homogeneous sets of molecules (e.g., the renin and HIV protease inhibitors) and structurally diverse sets of molecules (e.g., the cyclooxygenase inhibitors). We have noted previously that machine-learning methods are likely to work best when there are clear structural similarities between the training-set and test-molecules, and thus additional experiments were undertaken using a further eight activity classes that are among the most diverse in the MDDR database: muscarinic (M1) agonists, NMDA receptor antagonists, nitric oxide synthase inhibitors, aldose reductase inhibitors, reverse transcriptase inhibitors, aromatase inhibitors, phospholipase A2 inhibitors, and lipoxygenase inhibitors. The concept of ''most diverse'' was quantified by computing the Tanimoto similarity between each pair of molecules comprising an activity class and then ranking the various activity classes in the MDDR database in increasing order of the resulting mean pair-wise similarity (MPS) values [45].

Five training-sets were generated for each of the activity classes using the following procedure. The MDDR was split into two non-overlapping groups: those molecules that had been noted as belonging to the current activity class (i.e., were regarded as active) and the remaining molecules that had not been so noted (i.e., were regarded as inactive). One hundred molecules were randomly selected from the active group to form the training-set actives, subject to the criterion that no selected molecule had a similarity >0.80 (based on the Tanimoto coefficient and Unity 2D fingerprints (available from Tripos Inc. at http://www.tripos.com)) with any other selected molecule. A total of 4000 training-set inactives were randomly selected from the inactive group. The remaining 98435 active and inactive molecules formed the test-set for that activity class for the experiments.

Three different types of non-binary descriptor were used to characterise the molecules in the MDDR activity classes. First, sets of physicochemical property descriptors were generated with the SciTegic Pipeline Pilot software (available from SciTegic Inc. at http://www.scitegic.com). The descriptors included both simple integer counts (e.g., of numbers of atoms, bonds and ring assemblies) and computed molecular features that could be either integer or real (e.g., polar surface area, Alog *P*, and molecular weight). All descriptors were normalised, with zero mean and unit variance, and then submitted to a principal components analysis (PCA) routine that returned 29 components describing 100% of the variance in the data. Second, the molecules were described by SYBYL holograms (available from Tripos Inc. at http://www.tripos.com), 997-element integer vectors that encode the numbers of occurrences of hashed 2D substructural features. Some elements had values of 50 or more, and superior results were obtained if each element x in a molecular vector was replaced by $\log_{10}x$, effectively reducing the size of the space containing the molecules. Third, the molecules were described by Molconn-Z descriptors, a set of topological indices of molecular structure that have been used extensively for QSAR and QSPR analyses (available from eduSoft LC at http://www.edusoft-lc.com). A PCA analysis identified 300 components that described 100% of the variance in the data.

Running the CKD algorithm requires the optimisation of the bandwidth, h, of the Gaussians that are used to estimate the probability distributions. The bandwidth has a function analogous to the smoothing parameter, λ, in BKD, and the optimisation here was carried out with the same leave-one-out cross validation algorithm used previously for the optimisation of λ in BKD [25]. Once the bandwidth had been optimised, CKD was used to rank the test-set compounds, and then the screening performance computed as the mean percentage of the test-set actives that occurred in the top-1% of the ranked list, when averaged over the five different training-sets for each activity class.

Experimental results

The results that were obtained are detailed in Table 1, using the Pipeline Pilot, Hologram and Molconn-Z structural representations. It will be seen in all cases that CKD results in a significant clustering of the actives at the top of the ranked test-sets, as a purely random selection of molecules for biological screening would have identified just 1% of the actives. The results hence demonstrate the potential of CKD for virtual screening purposes when sets of non-binary descriptors are available to characterise the molecules that are being considered for screening. If we consider the results in Table 1a, the hologram representation

Table 1 Mean percentage of test-set actives retrieved in the top-1% of the ranked test-sets, using CKD with different types of descriptor: (a) eleven MDDR activity classes used by Hert et al. [39]; (b) eight structurally diverse MDDR activity classes used by Hert et al. [45]

Activity class	Structure representation			
	Pipeline Pilot	Holograms	Molconn-Z	ECFP4
(a)				
5HT3 antagonists	56.4	63.5	43.6	79.8
5HT1A agonists	36.9	42.7	30.4	58.6
5HT reuptake inhibitors	50.7	59.5	50.6	75.6
D2 antagonists	44.7	51.4	38.1	74.4
Renin inhibitors	81.8	87.8	57.8	93.2
Angiotensin II AT1 antagonists	46.9	65.2	36.5	70.4
Thrombin inhibitors	61.9	67.9	59.7	85.8
Substance P antagonists	61.5	73.2	49.7	82.2
HIV protease inhibitors	67.1	85.4	61.0	88.3
Cyclooxygenase inhibitors	43.4	52.2	45.3	68.1
Protein kinase C inhibitors	67.3	75.5	70.5	82.5
Average	56.2	65.9	49.4	78.1
(b)				
Muscarinic (M1) agonists	62.4	65.1	48.2	79.8
NMDA receptor antagonists	49.9	50.7	23.5	73.4
Nitric oxide synthase inhibitors	69.3	53.2	70.7	88.2
Aldose reductase inhibitors	55.8	67.8	51.5	85.0
Reverse transcriptase inhibitors	56.9	71.6	38.7	80.4
Aromatase inhibitors	68.7	71.5	55.1	94.1
Phospholipase A2 inhibitors	49.2	56.1	44.1	74.0
Lipoxygenase inhibitors	47.5	64.8	27.9	79.3
Average	57.5	62.6	45.0	81.8

used here is consistently superior to the Pipeline Pilot representation, which in turn normally out-performs the Molconn-Z representation. The superiority of the hologram representation, based on the frequencies of occurrence of 2D substructures, is far less for the structurally diverse activity classes in Table 1b. This is not particular surprising given the structure-based nature of the hologram representation, whereas the property-based representation is less likely to be adversely affected by the heterogeneous natures of the molecules in these classes.

The CKD scores (i.e., the $KD\_SCORE(x|A)$ values computed from Eq. 5) can be used to visualise the

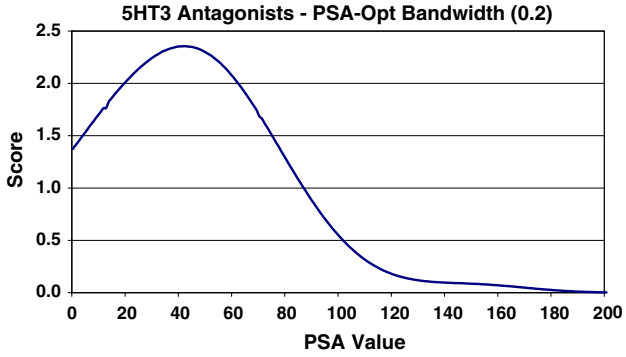

Fig. 1 The effect of the polar surface area (PSA) values on the CKD scores of a training-set containing 1000 5HT3 antagonists and 4000 inactives

effect of the various descriptors that characterise the molecules. For example, Fig. 1 shows the variation in CDK score for a training-set containing 100 5HT3 antagonists and 4000 inactive molecules. Here, the higher the score, the greater the likelihood of activity, with a score of unity indicating that a molecule with that specific descriptor value is no more likely to be active than it is to be inactive. The figure shows that a PSA value of about 40 corresponds to a score of just <2.5, and that molecules with PSA values greater than about 80 are relatively unlikely to be active. Analyses of the corresponding plots for other Pipeline Pilot descriptors shows that molecules are more likely to be 5HT3 antagonists if they are: quite small (with 10–30 atoms and with molecular weights <450); rigid (containing 3–8 rings and with <6 rotatable bonds); and not very lipophilic, with few H-bond acceptors and a low log *P* value. 5HT3 antagonists are CNS drugs that must be able to pass through the blood-brain barrier (BBB) to exhibit activity, and the analysis of the scores above mirrors closely the physicochemical properties that are known to be necessary for a molecule to pass through the BBB [46].

Although CKD has been designed to handle non-binary data, it can also be used with binary fingerprints, and we have hence carried out experiments with the molecules represented by SciTegic ECFP4 circular fingerprints (available from SciTegic Inc. at http://www.scitegic.com). This fingerprint type encodes circular substructures centred on each non-hydrogen atom in a molecule by a string of extended connectivity values that are calculated using a modification of the Morgan Algorithm. The results obtained using this representation are shown in the right-hand column of Table 1, where it will be seen that a consistently high level of performance is obtained, as has been noted in our previous experiments using BKD [38, 40, 41]. In fact, there is very little difference between the results

from BKD and CKD: for example, the average BKD recall for the datasets in Table 1a is 79.7%, just slightly more than the 78.1% for CKD. A mathematical analysis shows that the BKD and CKD kernels are entirely equivalent to within a scaling factor that is determined by the values of the variable parameters that are optimised by analysis of the training data: the smoothing parameter, λ, in BKD and the bandwidth of the Gaussian, h, for CKD. If the two kernels can be shown to be equivalent then the reader may well ask why there is any difference at all in the BKD and CKD recall values. The reason is the optimisation stage in the two procedures. The value of the smoothing parameter in BKD is always between 0.5 and 1, whereas the bandwidth in CKD can take any value greater than zero; indeed, with the hologram datasets, values of up to 200 were obtained in some cases. This means that CKD requires a more coarse-grained assessment of the range of possible optimal values for the bandwidth if an appropriate value is to be identified without excessive computation.

Some additional runs were carried out, using the set of eleven activity classes (as in Table 1a) and the Pipeline Pilot descriptors, in which the CKD results were compared with the results obtained from runs using five different types of SVM provided by the popular SVM^{light} software (available from http://svmlight.joachims.org/). These SVMs included a radial basis function (RBF) SVM with the default parameters, a linear SVM with the default parameters, and three polynomial SVMs with second-, third- and fourth-order kernels. The results of these experiments are shown in Table 2 (with the CKD results being repeated from Table 1a to facilitate comparison); here, we have listed just the RBF results as this SVM gave by some considerable way the best performance of the five different types of SVM that were tested. Even so, the SVM is slightly outperformed by CKD, the former giving the better performance for five of the activity classes and the latter for the other six classes. Similar results were obtained when these experiments were repeated using the set of eight structurally diverse activity classes.

Table 2 Mean percentage of test-set actives retrieved in the top-1% of the ranked test-sets, using CKD and SVM-RBF with Pipeline Pilot descriptors

Activity class	Pipeline Pilot	
	CKD	SVM-RBF
5HT3 antagonists	56.4	62.1
5HT1A agonists	36.9	43.9
5HT reuptake inhibitors	50.7	52.9
D2 antagonists	44.7	46.4
Renin inhibitors	81.8	68.9
Angiotensin II AT1 antagonists	46.9	45.6
Thrombin inhibitors	61.9	54.2
Substance P antagonists	61.5	44.4
HIV protease inhibitors	67.1	54.6
Cyclooxygenase inhibitors	43.4	44.7
Protein kinase C inhibitors	67.3	58.3
Average	56.2	52.4

Current work with CKD is focusing on the similarities and differences between it and BKD, with the hope that it may be possible to increase screening performance by combining the two approaches by means of data fusion; we also intend to analyse the effect of errors in the training data, as we have shown previously that even quite small numbers of false positives can severely affect the ability of BKD to predict accurately the (in)activities of test-set molecules [41]. Even so, the results we have obtained to date suggest that CKD may provide a powerful tool for database analysis, irrespective of the type of molecular representation that is available.

Use of a naive Bayesian classifier with minimal training data

Substructural analysis, naive Bayesian classifiers and group fusion

As noted in the Introduction, substructural analysis was the first application of machine learning in CAMD. The basic idea of the substructual analysis approach is a very simple one. For each fragment or bit in the binary fingerprints that characterise 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 that particular bit set to one. This weight reflects the probability that a molecule having that bit set (and thus containing some particular substructural fragment(s)) will be active so that, e.g., a bit that is set in many of the training-set actives and few of the training-set inactives will be assigned a much larger weight than will be a bit where the converse applies. 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, this sum representing the overall probability of activity for that molecule given that it contains a particular pattern of bits. Substructural analysis was studied in considerable detail by workers at the National Institutes of Health in an extended programme to develop novel anti-cancer agents [47–49], and also by workers at Lederle [29] and Sheffield [50–52]. However, it is only in the last few years that this general approach has become widely used [53–62].

Although perhaps not recognised when the approach was first introduced, substructural analysis is an example of an NBC [17, 63, 64]. An NBC is a simple classification algorithm that is based on the use of Bayes' theorem and on strong assumptions as to the statistical independence of the descriptors characterising the objects that are to be classified. The use of ''naive'' arises from the independence assumptions: these are often demonstrably incorrect, but this has not seemed to affect the performance of the classifier in many application domains [63].

Machine learning methods, such as substructural analysis or NBCs, are generally used when considerable amounts of training data are available, and this is often the case in CAMD applications when, e.g., HTS experiments have been carried out on large numbers of compounds. Here, we evaluate the performance of an NBC when very little training data is available, specifically when just a few actives are known: this is often the case at the start of a lead-discovery programme where the only information available may be a few literature and/or competitor compounds.

We have compared the NBC results with those obtained from similarity-based screening, the normal approach when limited numbers of actives are available. Conventional similarity searching is the screening method of choice when just a single active is available [11, 12, 31, 32], and Hert et al. have recently demonstrated the utility of group fusion for similarity searching when a few actives are available [39, 45, 65]. Group fusion involves the use of a single representation and a single similarity coefficient (SciTegic ECFP4 fingerprints and the Tanimoto coefficient in the work reported here), and combines the similarity-search outputs obtained from matching several different reference structures against the database that is to be screened [66–69] Specifically, assume that some particular database molecule yields similarity scores of s_1, $s_2..s_k$ with k different reference structures; then an effective similarity search can be obtained by ranking the database molecules on the basis of the largest of these scores, i.e., $\max\{s_1, s_2...s_k\}$.

Experimental details

As noted above, the experiments here were designed to evaluate the effectiveness of an NBC when used with limited amounts of training data, specifically training-set active molecules (as there is normally never any shortage of training-set inactives). In addition, we were interested in the extent to which the structural diversity of the training data affected the ability of the NBC to produce good predictive models. To this end, 14 activity classes were identified in the MDDR database, the version used here containing a total of 133,809 unique molecules. Each of the classes was categorised as being of high, medium or low structural diversity using the MPS criterion described previously. The three groups of classes were then as follows: muscarinic (M1) agonists, NMDA receptor antagonists, dopamine β-hydroxylase inhibitors, reverse transcriptase inhibitors, aromatase inhibitors, cyclooxygenase inhibitors and phospholipase A2 inhibitors (high diversity); CRF antagonists, 5HT3 antagonists, 5HT2 antagonists and oxytocin antagonists (medium diversity); cephalosporins, carbapenems and vitamin D analogues (low diversity).

Molecules were selected for inclusion in a training-set in one of two ways: clustered selection and random selection (which has been used in our previous studies of group fusion). In the clustering approach, the molecules of the activity class were clustered using Pipeline Pilot non-hierarchic clustering routines, and then 5 or 10 of the cluster centres were selected at random to form the set of actives for training. Two such routines were tested. The first (A1) resulted in an average of 20 compounds being allocated to each cluster; the second (A2) in no cluster member having a similarity <0.40 to the cluster centre (i.e., the molecule closest to the cluster's centroid). Both of these selection methods, A1 in particular, result in sets of actives that are structurally more diverse than the sets of actives that are obtained from simple random selection (A3), which was implemented by randomly selecting 5 or 10 compounds from the whole activity class to serve as the training-set actives.

The sets of actives, selected using A1, A2 or A3, were used for the training-sets for both the NBC and group fusion experiments. The former additionally requires training-set inactives, and three selection methods were again used for this purpose; in each case, a total of 200 inactive molecules was selected from the MDDR database that did not belong to the activity class under study. In the first and the third selection methods (I1 and I3), the inactives remaining after removal of all of the members of the chosen activity class were filtered to remove those with a Tanimoto similarity >0.25 to any active training-set molecule; 200 compounds were then selected from those remaining after filtering: either randomly (I1); or using the Diverse Molecules component of Pipeline Pilot (a dissimilarity-based compound selection procedure) (I3). In the second selection method (I2), 25% of the remaining inactives were selected at random, and then 200 molecules selected using the Diverse Molecules

component (as in I3). Of the three methods, I3 yielded the most diverse sets of inactives, then I2 and then I1.

The MDDR molecules were characterised by their ECFP4 fingerprints for both the NBC and group fusion experiments, and all similarity calculations were based on the Tanimoto coefficient. The former experiments involved the Laplacian-modified NBC in the Pipeline Pilot implementation produced by SciTegic Inc. whilst the latter involved user software. All experiments were run ten times and, as before, the effectiveness of the searches estimated by noting the mean percentage of the actives in a particular class that were retrieved in the top-1% of the ranked test-set.

Experimental results

We have summarised the many experiments that were carried out by averaging across the searches for the high, medium and low diversity activity classes, as shown in Table 3. The comments below relate to these averaged results; that said, the figures in the tables are averages and there are particular combinations of activity class and selection method that provide counter-examples to the generalisations discussed below. We can draw three principal conclusions from the results in Table 3.

Table 3 Mean percentage of test-set actives retrieved in the top-1% of the ranked test-set using group fusion and naive Bayesian classification (NBC), with different ways of selecting the active and inactive training-set molecules. The results are averaged across all activity classes of a given level of structural diversity: (**a**) high diversity activity classes; (**b**) medium diversity activity classes; (**c**) low diversity activity classes

Selection of actives	Reference structures	Group fusion	NBC: Selection of inactives		
			I1	I2	I3
(a)					
A1	5	18.1	14.0	15.6	15.3
	10	23.1	17.3	21.5	20.8
A2	5	19.3	17.0	18.6	18.4
	10	27.2	22.9	26.7	26.2
A3	5	20.3	18.7	20.0	19.4
	10	31.1	25.5	29.3	28.7
(b)					
A1	5	33.7	37.1	39.6	39.4
	10	52.7	49.3	55.6	57.2
A2	5	33.2	37.9	39.4	39.2
	10	49.3	48.8	55.2	56.3
A3	5	36.7	38.9	40.2	40.8
	10	56.0	57.8	61.4	62.2
(c)					
A1	5	52.0	79.9	80.8	81.5
	10	56.7	88.1	88.3	88.9
A2	5	62.3	82.0	82.6	82.9
	10	77.9	88.4	89.5	89.7
A3	5	83.9	90.3	90.0	90.3
	10	85.4	90.7	90.2	90.6

First, group fusion, which uses just the active reference structures, consistently out-performs the NBC in terms of the numbers of active test-molecules retrieved in the top-1% of the rankings in the case of the high-diversity (low similarity) activity classes. The converse applies for the low-diversity activity classes, where the NBC is markedly and consistently superior to group fusion; this is also the case, but to a far less marked extent, for the medium-diversity activity classes, where the two approaches offer more comparable levels of performance. We believe that the reason for this behaviour is the NBC's need for some degree of structural commonality amongst the actives to enable it to ''learn'' the structural features that contribute to activity. The level of discrimination provided by the training-set actives will hence be maximised if these molecules are structurally similar to each other: accordingly, NBC achieves its highest level of performance with the low-diversity activity classes, both in absolute terms and relative to group fusion. With the high-diversity activity classes, there is much less structural overlap across the training-set actives, thus providing only a limited amount of information that the classifier can use when it is applied to the test-set molecules. Group fusion is less adversely affected in that it focuses on the active molecules themselves without taking explicit account of the substructural relationships between them.

Second, random selection of the training-set actives was consistently superior to the use of either of the two clustered selection methods (A1 and A2). The latter methods provide more diverse sets of active training-set actives, and it is thus not surprising from the comments immediately above that the availability only of heterogeneous structures reduces the performance of the NBCs; the results in the table suggest that this is also the case for group fusion (which is perhaps surprising given group fusion's ability better to handle the more diverse activity classes).

Third, the use of the more diverse inactive training-sets (I2 and I3) gave better results than did I1, with the exception of the low-diversity activity classes where all three types of NBC search achieved comparably high levels of search performance. Thus, the NBC is better able to ''learn'' the features of the inactive compounds when they are structurally diverse than when they are structurally related, as noted also by Glick et al. [60]. NBC can hence be expected to provide high levels of screening performance with highly similar training-sets actives and highly dissimilar training-set inactives (indeed, Hert et al. have suggested that the inactives

can be approximated by the entire database that is being searched [39]).

The finding that the simple group-fusion approach can be competitive with the more sophisticated machine-learning approach may appear surprising. However, previous work by Bender et al. [55] and Hert et al. [65] provides some support for this view. Bender et al. have used the 11 MDDR activity classes listed in Table 1a to test an NBC that uses information-gain-based feature selection, and atom environment fingerprints that encode, like the SciTegic fragments, circular substructures centred on each of the heavy atoms in a molecule [55]. Their experiments used the same version of the MDDR database as that employed by Hert et al. in their experiments on group fusion using SciTegic fingerprints [65], and it is hence possible to compare the results obtained by the two procedures. The data are shown in Table 4: they are based on the use of 10 actives (and 100 inactives in the case of the NBC procedure) and on the mean percentage of actives retrieved in the top-5% of the database when averaged over ten randomly-selected training-sets (rather than the top-1% as in the other experiments reported in this paper). It must be emphasised that there are some differences in the experimental set-ups here, most notably in the precise nature of the circular-substructure fingerprints, and the two sets of results are thus not directly comparable. Even so, these results provide further support for the view that group fusion is not inferior to NBC when very limited training data is available.

We hence conclude that an NBC is effective given appropriate training data, specifically homogenous active molecules and heterogenous inactive molecules, since it is then able to compute the probabilities that underlie the approach with some degree of accuracy. When only limited numbers of heterogeneous actives are available, this sophisticated machine-learning approach would appear to be less effective than the simple, similarity-based group-fusion approach.

Table 4 Mean percentage of test-set actives retrieved in the top-5% of the ranked test-sets, using NBC and group fusion on eleven MDDR activity classes used by Hert et al. [39]; the NBC results are from Bender et al. [55] and the group fusion results from Hert et al. [65]

Activity class	NBC	Group fusion
5HT3 antagonist	66.6	70.4
5HT1A agonist	57.1	63.2
5HT Reuptake inhibitor	46.1	49.5
D2 antagonist	53.7	54.7
Renin inhibitor	95.7	96.9
Angiotensin II AT1 antagonist	95.1	97.1
Thrombin inhibitor	66.2	72.5
Substance P antagonist	68.4	61.4
HIV protease inhibitor	76.0	79.4
Cyclooxygenase inhibitor	34.7	38.8
Protein kinase C inhibitor	54.6	57.0
Average over all classes	64.9	67.4

Conclusions

The last few years have seen intense interest in the application of structure-based virtual screening for lead discovery programmes in the pharmaceutical industry. This is a very powerful CAMD technique but there are still many circumstances, particularly at the commencement of a programme, when ligand-based virtual screening can be of value; here, we have summarised some of our current studies of the use of machine learning for this purpose. In the first part of the paper, we focus on kernel discrimination methods and describe a kernel that is suitable for use with binary, integer and real-valued representations of molecule structure, and that appears to be competitive with a previously described kernel that can be used only with binary fingerprint representations. In the second part of the paper, we consider the extent to which NBCs can provide an acceptable level of screening performance when just a very few active structures are available as training data. Our results suggest that such classifiers perform well if the active molecules that are being sought are structurally homogeneous; if this is not the case, then a simpler approach based on group fusion provides an effective alternative.

Acknowledgements We thank the following: the Alexander S. Onassis Public Benefit Foundation, the Engineering and Physical Sciences Research Council and the Novartis Institutes for Biomedical Research for funding George Papadatos; the Biotechnology and Biological Sciences Research Council and GlaxoSmithKline for funding David Wood; MDL Information Systems Inc. for provision of the *MDL Drug Data Report* database; and the Royal Society, SciTegic Inc., Tripos Inc. and the Wolfson Foundation for hardware, laboratory and software support.

References

1. Böhm H-J, Schneider G (eds) (2000) Virtual screening for bioactive molecules, Wiley-VCH, New York
2. Klebe G (ed) (2000) Virtual screening: an alternative or complement to high throughput screening, Kluwer, Dordrecht
3. Bajorath J (2002) Nature Rev Drug Discov 1:882
4. Delaney J, Clarke E, Hughes D, Rice M (2006) Drug Discov Today 11:839
5. Kitchen DB, Decornez H, Furr JR, Bajorath J (2004) Nature Rev Drug Discov 3:935

6. Leach AR, Shoichet BK, Peishoff CE (2006) J Med Chem 49:5851
7. Schneider G, Fechner U (2005) Nature Rev Drug Discov 4:649
8. Berman HM, Battistuz T, Bhat TN, Blum WF, Bourne PE, Burkhardt K, Feng Z, Gilliland GL, Iype L, Jain S, Fagan P, Marvin J, Padilla D, Ravichandran V, Schneider B, Thanki N, Weissig H, Westbrook JD, Zardecki C (2002) Acta Cryst D 58:899
9. Congreve M, Murray CW, Blundell TL (2005) Drug Discov Today 10:895
10. Willett P (1987) Similarity and clustering in chemical information systems. Research Studies Press, Letchworth
11. Sheridan RP, Kearsley SK (2002) Drug Discov Today 7:903
12. Bender A, Glen RC (2004) Org Biomol Chem 2:3204
13. Martin YC, In: Martin YC, Willett P (eds) (1998) Designing bioactive molecules: three-dimensional techniques and applications. American Chemical Society, Washington, pp 121–148
14. Güner O (ed) (2000) Pharmacophore perception, development and use in drug design. International University Line, La Jolla CA
15. Martin YC (1992) J Med Chem 35:2145
16. Willett P (1995) J Mol Recognit 8:290
17. Mitchell TM (1997) Machine learning. McGraw-Hill, New York, NY
18. Cramer RD, Redl G, Berkoff CE (1974) J Med Chem 17:533
19. Redl G, Cramer RD, Berkoff CE (1974) Chem Soc Rev 3:273
20. Duda RO, Hart PE, Stork DG (2000) Pattern classification. 2nd ed., Wiley Interscience, New York
21. Hand D, Mannila H, Smyth P (2001) Principles of data mining. MIT Press Cambridge MA
22. Wagener M, van Geerestein VJ (2000) J Chem Inf Comput Sci 40:280
23. Saeh JC, Lyne PD, Takasaki BK, Cosgrove DA (2005) J Chem Inf Model 45:1122
24. Hawkins DM, Young SS, Rusinko A (1997) Quant Struct-Active Relat 16:296
25. Harper G, Bradshaw J, Gittins JC, Green DVS, Leach AR (2001) J Chem Inf Comput Sci 41:1295
26. Johnson MA, Maggiora GM (eds) (1990) Concepts and applications of molecular similarity. John Wiley, New York
27. Kubinyi H (1998) Perspect Drug Discov Design 9–11:225
28. Stahura FL, Bajorath J (2002) Drug Discov Today 7:S41
29. Carhart RE, Smith DH, Venkataraghavan R (1985) J Chem Inf Comput Sci 25:64
30. Willett P, Winterman V, Bawden D (1986) J Chem Inf Comput Sci 26:36
31. Willett P, Barnard JM, Downs GM (1998) J Chem Inf Comput Sci 38:983
32. Nikolova N, Jaworska J (2003) QSAR Combin Sci 22:1006
33. Brown RD, Martin YC (1996) J Chem Inf Comput Sci 36:572
34. Brown RD, Martin YC (1997) J Chem Inf Comput Sci 37:1
35. Martin YC, Kofron JL, Traphagen LM (2002) J Med Chem 45:4350
36. Parzen E (1962) Ann Math Stat 33:1065
37. Christianini N, Shawe-Taylor J (2000) An introduction to support vector machines and other kernel-based learning methods. Cambridge Universtity Press, Cambridge
38. Wilton D, Willett P, Lawson K, Mullier G (2003) J Chem Inf Comput Sci 43:469
39. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A (2004) J Chem Inf Comput Sci 44:1177
40. Wilton DJ, Harrison RF, Willett P, Delaney J, Lawson K, Mullier G (2006) J Chem Inf Model 46:471
41. Chen B, Harrison RF, Pasupa K, Wilton DJ, Willett P, Wood DJ, Lewell XQ (2006) J Chem Inf Model 46:478
42. Aitchison J, Aitken CGG (1976) Biometrika 63:413
43. Constans P, Hirst JD (2000) J Chem Inf Comput Sci 40:452
44. McNeany TJ, Hirst JD (2005) J Chem Inf Comput Sci 45:768
45. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A (2006) J Chem Inf Comput Sci 46:462
46. Clark DE (2003) Drug Discov Today 8:927
47. Hodes L, Hazard GF, Geran RI, Richman S (1977) J Med Chem 20:469
48. Hodes L (1981) J Chem Inf Comput Sci 21:132
49. Hodes L (1981) J Chem Inf Comput Sci 21:128
50. Ormerod A, Willett P, Bawden D (1989) Quant Struct-Active Relat 8:115
51. Ormerod A, Willett P, Bawden D (1990) Quant Struct-Active Relat 9:302
52. Cosgrove DA, Willett P (1998) J Mol Graph Model 16:19
53. Anzali S, Barnickel G, Cezanne B, Krug M, Filimonov D, Poroikov V (2001) J Chem Inf Comput Sci 44:2432
54. Bender A, Mussa HY, Glen RC, Reiling S (2004) J Chem Inf Comput Sci 44:170
55. Bender A, Mussa HY, Glen RC, Reiling S (2004) J Chem Inf Comput Sci 44:1708
56. Glick M, Klon AE, Acklin P, Davies JW (2004) J Biomol Screen 9:32
57. Klon AE, Glick M, Davies JW (2004) J Med Chem 47:4356
58. Xia XY, Maliski EG, Gallant P, Rogers D (2004) J Med Chem 47:4463
59. Rogers D, Brown RD, Hahn M (2005) J Biomol Screen 10:682
60. Glick M, Jenkins JL, Nettles JH, Hitchings H, Davies JW (2006) J Chem Inf Model 46:193
61. Capelli AM, Feriani A, Tedesco G, Pozzan A (2006) J Chem Inf Model 46:659
62. Eckert H, Bajorath J (2006) J Med Chem 49:2284
63. Domingos P, Pazzani M (1997) Machine Learn 29:103
64. Hand DJ, Yu K (2001) Int Stat Rev 69:385
65. Hert J, Willett P, Wilton DJ, Acklin P, Azzaoui K, Jacoby E, Schuffenhauer A (2004) Org Biomol Chem 2:3256
66. Whittle M, Gillet VJ, Willett P, Alex A, Loesel J (2004) J Chem Inf Comput Sci 44:1840
67. Zhang Q, Muegge I (2006) J Med Chem 49:1536
68. Williams C (2006) Mol Divers 10:311
69. Willett P (2006) QSAR Combin Sci 25:1143