Naïve Bayes Classification Using 2D Pharmacophore Feature Triplet Vectors

Paul Watson*

Arena Pharmaceuticals, 6166 Nancy Ridge Drive, San Diego, California 92121

Received August 31, 2007

A naïve Bayes classifier, employed in conjunction with 2D pharmacophore feature triplet vectors describing the molecules, is presented and validated. Molecules are described using a vector where each element in the vector contains the number of times a particular triplet of atom-based features separated by a set of topological distances occurs. Using the feature triplet vectors it is possible to generate naïve Bayes classifiers that predict whether molecules are likely to be active against a given target (or family of targets). Two retrospective validation experiments were performed using a range of actives from WOMBAT, the Prous Integrity database, and the Arena screening library. The performance of the classifiers was evaluated using enrichment curves, enrichment factors, and the BEDROC metric. The classifiers were found to give significant enrichments for the various test sets.

INTRODUCTION

Molecule database searching commonly finds use both in lead generation and lead optimization during the drug development process.[1,2] The results of such searches can be used in many ways. For example, in the early stages of a drug discovery project virtual screening hits can be used to provide chemical starting points or prioritize molecules for biological screening. In the latter stages of a project the results can be used to provide alternate scaffolds or in lead-hopping. There are many methods available for chemical database searching such as 2D[3−7] and 3D[8−12] similarity methods as well as docking approaches.[13−21] These approaches are well documented and are applied routinely with varying degrees of success in most drug development projects. While docking is intellectually the most attractive approach, it is only possible if a crystal structure of the target protein is available. Additionally, it is not clear that docking scores can be used to rank molecules successfully;[13−21] in fact, it has been shown that in many cases ligand-based methods can outperform docking methods in this respect.[22,23] Similarity searching is often considered "a mainstay of computational drug design",[24] and there are many methods in both 2D and 3D for finding a set of molecules that are similar to a single query molecule. If a user wants to find molecules that are similar to a collection of active molecules, the utility of such an approach is diminished as the user has to perform a similarity search using each active molecule as a query.

Machine learning[25] could be viewed as a solution to the problem described above. Typically a machine learning algorithm learns, or generates, a function that maps a collection of input data to a set of observed outcomes. In cheminformatics the input data are typically some description of the molecules, such as fingerprints or fragment counts. The outcome is usually whether a molecule is active or inactive against a given protein or family of proteins. Examples of machine learning algorithms that have been employed in cheminformatics include support vector machines,[26−29] ensemble methods,[30] decision trees,[31,32] neural networks,[33−35] and naïve Bayes classifiers.[25]

A naïve Bayes classifier is generated using a training set of instances where each of the instances is known to belong to a certain class, e.g., good or bad, active or inactive, etc. A set of features or attributes is employed to describe each of the instances. By counting the number of times a given feature appears for each class, the classifier "learns" to distinguish between instances belonging to the different classes. Naïve Bayes classifiers are attractive due to their simplicity, they perform well even in high dimensional space, and the models can be generated in linear time (unlike many other classification methods). A number of papers have been published describing the use of naïve Bayes classifiers for chemical database searching. Xia and co-workers[36] describe the use of the naïve Bayes classifier in Scitegic's Pipeline Pilot (Scitegic Inc. http://www.scitegic.com) to design screening libraries containing molecules that are likely to be active against kinases. Here proprietary functional fingerprints were used to describe the molecules. Bayes classifiers were generated using active and inactive compounds from the Amgen screening library, and these were used to predict whether molecules were likely to be active against kinases. The classifiers were tested both retrospectively and prospectively and showed the approach gave significant enrichments over random selection. Glen and co-workers[37,38] have published a series of papers in which Bayes classification is used in conjunction with 2D (atom environment descriptors) and 3D (interaction energies projected onto molecular surface) descriptors. A naïve Bayes classifier has also been employed to predict multidrug resistance.[39] Here the classifier was trained to predict whether a molecule is an active or inactive MDRR (Multi Drug Resistant Reversal) agent. An atom typing approach was compared with the functional fingerprints offered by Scitegic. Bayes classification has also been employed to predict likely hERG channel blockers as well as cytochrome P450 2D6 inhibition.[40]

* Corresponding author phone: (858)453-7200; fax: (858)453-7210; e-mail: pwatson@arenapharm.com.

10.1021/ci7003253 CCC: $40.75
Published on Web 01/10/2008

Table 1. Pharmacophore Feature Type Assignments

feature	code	description	SMARTS
DON	**D**	hydrogen bond donor	[$([O;H1]),$([#7;H1,H2,H3)]
ACC	**A**	hydrogen bond acceptor	[$([$([#8]);!$([#8](−,:*)−,:*);!$(*=˜N∼O);!$(*∼N=O);X1,X2]),$([#7;v3;!$([nH]);!$([n;D3]);!$(*−a);!$(*C(=O));!$(*C(=[N;D1]));!$(*S(=O)=O);!$(N#C)])]
RIN	**R**	ring atom	[R]
CHA	**C**	chain atom	[R0]
POS	**P**	positively charged atom	[+1,+2,+3]
NEG	**N**	negatively charged atom	[−1,−2,−3]

In this paper the utility of a naïve Bayes classifier in conjunction with 2D feature triplet vectors describing the molecules is investigated. Here each molecule is described using a vector where each element in the vector contains the number of times a particular triplet of atom-based features occurs when separated by a set of bond distances (or topological distances). Using these feature triplet vectors it is possible to generate naïve Bayes classifiers that should predict whether molecules are likely to be active against a given target or family of targets. Two validation experiments are performed. In each case naïve Bayes classifiers were built using the feature triplet vector as the descriptor for the molecules. Training sets and test sets of molecules are created in a number of ways, and the performance of the classifiers generated from the training sets are evaluated using the test sets employing enrichment factors, enrichment curves, and the BEDROC[41] metric.

METHODS

2D Feature Triplet Vectors. The first step of the feature triplet determination is the assignation of feature types for all heavy atoms in the molecule. This is performed using SMARTS[42,43] patterns (Daylight Chemical Information Systems Inc. http://www.daylight.com) that define each of the feature types. These are shown in Table 1. Each atom can have more than one feature assigned to it.

The hydrogen bond donor feature (**D**) is defined as a nitrogen or oxygen atom that has at least 1 attached hydrogen. The hydrogen bond acceptor feature (**A**) is defined as the following: an oxygen atom not including ether oxygen atoms, ester-ether oxygen atoms, aromatic oxygen atoms, and oxygen atoms in nitro or nitroso groups or a 3 valent nitrogen not including 3 coordinate aromatic nitrogen atoms, aliphatic nitrogen atoms bound to aromatic systems, amide and sulfonamide nitrogen atoms, and nitrogen atoms in cyano groups. The reason for excluding ether, aromatic, nitro, and nitroso oxygens is that these groups are weak hydrogen bond acceptors and are usually outcompeted by stronger acceptor features, such as carbonyl oxygens.[44] The ring atom feature (**R**) is simply defined as any atom in a ring. The chain atom feature (**C**) is any atom not in a ring. The positive atom feature (**P**) is any atom with a +1, +2, or +3 charge. The negative atom feature (**N**) is defined as any atom with a −1, −2, or −3 charge.

An example of a molecule with the features assigned is shown in Figure 1. Some of the atoms in the molecule are assigned more than one feature. For example, the oxygen in the hydroxyl group is assigned the **A**, **C**, and **D** features as the hydroxyl group is a hydrogen bond donor, hydrogen bond acceptor, and not in a ring.

In the next step the distances between all pairs of non-hydrogen atoms, and thus the atom features, are calculated in terms of the shortest possible bonded paths between each pair. These topological distances between the atom features are binned in the following fashion [1,2], [3,4], [5,6], [7,8], and [9,10]. Any atom features that are separated by a distance greater than 10 bonds are ignored. All possible triplets of features are identified and combined with the binned distances to form a feature triplet. Each of the feature triplets is canonicalized and is mapped into a vector, where each value in the vector denotes the number of times a particular feature triplet occurs. Given 6 feature types and 5 distance bins, the feature triplet vector has a maximum size of 4096.

The approach of using triplets of features and the distances between them to describe molecules has been used previously, notably for the Similog keys.[45] The main difference here is that the atoms can be represented by multiple features that can generate multiple triplets of features for a single triplet of atoms. An example of this is shown in Figure 2 where a triplet of atoms and the distances between them are shown in an example molecule. In this case the nitrogen in the quinoline ring is an acceptor and ring feature (**A**/**R**), the nitrogen in the amidine group is a donor and chain feature (**D**/**C**), and the carbon in the methyl group is a chain feature (**C**). This generates 4 possible feature triplets all of which are included in the feature triplet vector.

Naïve Bayes Classification Using the Feature Triplet Vector. Bayesian classification is a statistical method that allows the user to categorize instances in a data set based on the equal and independent contributions of their attributes. In real life this assumption is rarely valid as attributes are often unequally weighted and are almost never linearly independent; however, the application of naïve Bayes (as this method is known) has been shown to work well in numerous applications.[25] Bayes' rule of conditional prob-

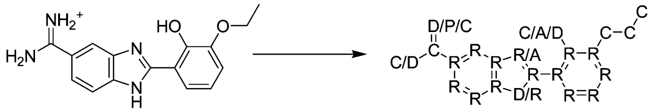

Figure 1. Example molecule showing how atom features are assigned. The atom features are labeled using the codes defined in Table 1. Each atom can be assigned one or more features; for example, the NH in the benzimidazole is both a ring atom and a hydrogen bond donor and, as such, has both the **R** and **D** atom features assigned.

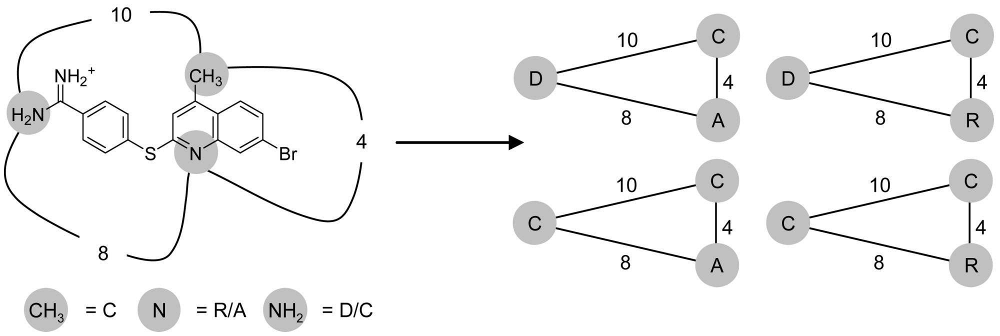

Figure 2. Example feature triplet generation. The chosen atom triplet is highlighted on the left-hand side, and the shortest bond paths are shown. The assigned atom features for the three highlighted atoms are also shown (i.e., the carbon of the CH_3 group is a chain feature (**C**), the nitrogen of the quinoline ring is an acceptor and ring feature (**A**/**R**), and the nitrogen of the highlighted NH_2 group is a donor and chain feature (**D**/**C**)). As each of the atoms can represent one or more atom features, multiple feature triplets can be generated for each atom triplet.

ability is shown in eq 1 for hypothesis H given evidence E.

$$P[H|E] = \frac{P[E|H]P[H]}{P[E]} \tag{1}$$

Here, $P[H|E]$ is the probability the hypothesis H is correct for evidence E. $P[E|H]$ is the product of the probabilities that each piece of evidence occurs for the hypothesis H. $P[H]$ is the probability that the hypothesis H is correct without any knowledge of the evidence E; this is often known as the *prior probability*.

A naïve Bayes classifier is generated using a training set to provide the prior evidence that an instance belongs to a certain class. An example of this would be a training set of B molecules where A of the molecules are known to be active and the remainder are known to be inactive against a given target. These molecules can be used to train the classifier such that it is able to distinguish the active molecules from the inactive molecules. The prior probability of a molecule being active, $P[A]$, is given by eq 2.

$$P[A] = \frac{A}{B} \tag{2}$$

A naïve Bayes classifier can be generated using the feature triplet vector described above. For each molecule in a training set of B molecules, the feature triplet vector is determined. The classifier is built by counting the number of actives and the total number of molecules (both active and inactive) that contain a given feature triplet F_i occurring j times. The uncorrected estimate of activity $P[A|F_i(j)]$ is shown in eq 3.

$$P[A|F_i(j)] = \frac{M_{Fi(j)}}{N_{Fi(j)}} \tag{3}$$

Here $F_i(j)$ is the feature triplet F_i, occurring j times in a molecule. $M_{Fi(j)}$ is the number of active molecules that contain feature triplet F_i, j times. $N_{Fi(j)}$ is the total number of molecules in the database that contain feature triplet F_i, j times. Thus $P[A|F_i(j)]$ is the likelihood that an active molecule will contain feature triplet F_i, j times.

A Laplace estimator is used to correct for undersampled features. If $N_{Fi(j)}$ is low and is always associated with a given class, i.e., $M_{Fi(j)} \approx N_{Fi(j)}$, the uncorrected estimate for the class will be 1.0. This value is probably overconfident given that the feature F_i occurring j times has only been observed in a small number of molecules. Therefore a Laplace estimator is used to estimate the effect of additional sampling of the feature. This is shown in eq 4.

$$P_{\mathrm{lap}}[A|F_i(j)] = \frac{M_{Fi(j)} + 1}{N_{Fi(j)} + P[A]^{-1}} \tag{4}$$

This ensures that as $N_{Fi(j)}$ approaches zero the corrected estimate of activity will approach $P[A]$. The corrected estimators are normalized by dividing by the prior probability, $P[A]$, to give the normalized estimator, $P_{\mathrm{norm}}[A|F_i(j)]$, which is shown in eq 5. The log of the normalized estimator provides a clear indication of how common features are in active molecules. If $\log P_{\mathrm{norm}}[A|F_i(j)] > 0$, then the feature triplet F_i appearing j times in a molecule is more common in actives. If $\log P_{\mathrm{norm}}[A|F_i(j)] < 0$, then the feature triplet F_i appearing j times in a molecule is less common in actives.

$$P_{\mathrm{norm}}[A|F_i(j)] = \frac{P_{\mathrm{lap}}[A|F_i(j)]}{P[A]} \tag{5}$$

To calculate the estimate that a molecule is active, $P_{\mathrm{mol}}[A]$, the log values of $P_{\mathrm{norm}}[A|F_i(j)]$ are summed for each of the features, F_i, present in the molecule j times (eq 6). Note that only the features present in the molecule are used in the summation.

$$\log P_{\mathrm{mol}}[A] = \sum_i \log P_{\mathrm{norm}}[A|F_i(j)] \tag{6}$$

The program to create the 2D feature triplet vectors, train the Bayes classifier, and perform the validation of the resultant model is written in Java (http://sun.java.com) using the OpenEye OEChem programming toolkit (OpenEye Scientific Software http://www.eyesopen.com). Typically the classifier can be trained and tested at a rate of around 300 molecules per second. The timings were obtained using Java version 1.5.0.6 on the Centos v4.0 Linux operating system (http://www.centos.org) running on 2GHz AMD Opteron 64 bit processors with 2GB of RAM available.

Validation of the 2D Feature Triplet Vector Naïve Bayes Classifier. In order to validate the use of the 2D feature triplet vector as a descriptor in conjunction with a naïve Bayes classifier, databases of molecules can be used where the activities of the molecules against protein targets are already known. In such an experiment the molecules are split into a training set and a test set. The classifier is built (or trained) using the training set, and the performance of the classifier is measured by analyzing the predictions made on the test set. The performance of the classifiers is measured in three ways. The first of these uses enrichment curves. Here the percentage of active molecules found in the ranked list is plotted against the percentage progress through the ranked list. The second metric for measuring the performance of the classifiers is the enrichment factor. The enrichment factor demonstrates whether using the classifier to rank the molecules according to their likelihood of belonging to a given activity class is better than choosing the molecules at random. The following method is used to calculate the enrichment factors. The number of active molecules that would be expected in the top 10% of the ranked list of test set molecules, provided the active molecules are distributed randomly, n_{exp}, is given by eq 7

$$n_{\mathrm{exp}} = 0.1 \times n_{\mathrm{a}} \qquad (7)$$

where n_{a} is the number of active molecules in the test set. Using the number of active molecules that actually appear in the top 10% of the ranked list, n_{obs}, the enrichment factor can be calculated using eq 8.

$$E = \frac{n_{\mathrm{obs}}}{n_{\mathrm{exp}}} \qquad (8)$$

If the enrichment factor is greater than 1.0, then the method of ranking the molecules using the naïve Bayes classifier is better than choosing the actives from a randomly ordered list.

Enrichment factors are attractive because the user can set a cutoff that is of practical interest, and the metric rewards a classification method that places active molecules within the bounds of the chosen cutoff. However, enrichment factors equally weight the actives within the chosen cutoff (e.g., the top 10%). It makes no distinction between a well ordered list of actives, where all the actives are ranked at the very beginning of the top 10%, and a poorly ordered list, where the actives are at the bottom of the top 10%. As such, the value chosen for the cutoff can have a large impact on the enrichment factor. With this in mind a third method for analyzing the performance of the classifier is employed known as the BEDROC[41] metric (Boltzmann Enhanced Discrimination of Receiver Operator Characteristics). This metric has the advantage of removing the arbitrary cutoff point involved in the enrichment factor as well as rewarding better rankings of the active compounds and early recognition. The BEDROC metric is given in eq 9.

$$B = \frac{\sum_{i=1}^{n} e^{-\alpha r_i/N}}{\frac{n}{N}\left(\frac{1-e^{-\alpha}}{e^{\alpha/N}-1}\right)} \times \frac{R_{\mathrm{a}}\sinh(\alpha/2)}{\cosh(\alpha/2) - \cosh(\alpha/2 - \alpha R_{\mathrm{a}})} + \frac{1}{1-e^{\alpha(1-R_{\mathrm{a}})}} \qquad (9)$$

Here r_i is the rank of the ith active in the ordered list, N is the number of molecules in the test set, n is the number of active molecules in the test set, R_{a} is the ratio of actives in the test set (i.e., n/N), and α is a constant that can be chosen to represent the weighting given to the early part of an ordered list. A value of 16.1 is used for the α parameter which means that the first 10% of the ranked list contributes 80% of the BEDROC score. If the actives are uniformly distributed in the ordered list (i.e., random), then the BEDROC score is given by eq 10. For $\alpha = 16.1$, $B_{\mathrm{r}} = 0.062$.

$$B_{\mathrm{r}} = \frac{1}{\alpha} + \frac{1}{1-e^{\alpha}} \quad \text{if } \alpha R_{\mathrm{a}} \ll 1 \qquad (10)$$

There are sources of error in any metric used to validate a virtual screening approach. The first of these is intrinsic variance in the metric that is usually solved by having a large number of actives in the validation experiment. However, a large proportion of actives in the test set is hardly realistic and can lead to a second source of error or "saturation effect". The saturation effect can lead to large differences in the values for these metrics when the ratio of actives in the test set is different for the same "constant true performance".[41] It is therefore impossible to compare metrics across different test sets (or different methods) unless these sources of error are minimized.

It has been suggested by Truchon et al.[41] that at least 50 actives are required to avoid intrinsic variance. The saturation effect can be avoided by having a small fraction of the test set made up of actives. The exact fraction depends on the metrics used to evaluate the classifier's performance. Here, as we are using the BEDROC metric with $\alpha = 16.1$, the fraction of actives in the test set is set to 0.0062. This has the added advantage of bounding the enrichment factors by 0 and 10 as the ratio of actives is smaller than the cutoff used (e.g., 10%).

In many of the test sets employed in the validation experiment the number of actives for the given set of inactives is too high. As such, a bootstrap average with active replacement is employed in which a number of actives are randomly selected that match the ratio of actives given above for the test set size. This not only has the advantage of avoiding sources of variance but also allows the calculation of standard deviations in the metrics used.

Validation Experiment 1. In this experiment data sets of molecules are retrieved from the WOMBAT database.[46] The WOMBAT database is a collection of molecules and their associated biological activities from popular chemical literature sources. The 2006.2 version of the database contains 170 406 entries (149 451 unique isomeric SMILES), with biological activities for 1525 targets from 7570 references.

The activity classes and the number of molecules active against the proteins encompassed in the given class are listed in Table 2. An activity class is defined as a set of molecules that are active against one or more targets that make up a family or subfamily of proteins. There are 3 activity classes for enzymes: aspartyl proteases, kinases, and serine proteases. The GPCR activity classes are hierarchical in nature. At the top level there are 3 activity classes representing the amine, nucleotide, and peptide families of GPCR receptors.

Table 2. Activity Classes and the Numbers of Active and Inactive Compounds Used in the First Validation Experiment[a]

	total			training set		test set		
activity class	n_a	n_i	split	n_a	n_i	n_a	n_i	n_{ba}
			Enzymes					
aspartyl protease	2851 (628)	111359 (56379)	1:4	570	22272	2281	89087	567
			1:1	1425	55680	1426	55679	354
			4:1	2280	89088	571	22271	141
kinase	8725 (2417)	105485 (54590)	1:4	1745	21097	6980	84388	567
			1:1	4362	52743	4363	52742	354
			4:1	6980	84388	1745	21097	141
serine protease	5850 (3553)	108350 (53454)	1:4	1162	21670	4688	86680	567
			1:1	2920	54175	2930	54175	354
			4:1	4678	86680	1172	21670	141
			GPCR (Level 1)					
amine GPCR	13338 (11699)	100872 (45308)	1:4	2667	20175	10671	80697	567
			1:1	6669	50436	6669	50436	354
			4:1	10670	80698	2668	20174	141
nucleotide GPCR	3931 (506)	110279 (56501)	1:4	786	22056	3145	88223	567
			1:1	1965	55140	1966	55139	354
			4:1	3144	88224	787	22055	141
peptide GPCR	11750 (8149)	102460 (48858)	1:4	2350	20492	9400	81968	567
			1:1	5875	51230	5875	51230	354
			4:1	9400	81968	2350	20492	141
			GPCR (Level 2)					
adrenoceptors	3553 (2978)	110657 (54029)	1:4	710	22132	2843	88525	567
			1:1	1776	55329	1777	55328	354
			4:1	2842	88526	711	22131	141
dopamine GPCR	4289 (4250)	109921 (52757)	1:4	857	21985	3432	87936	567
			1:1	2144	54961	2145	54960	354
			4:1	3431	87937	858	21984	141
serotonin GPCR	5769 (5547)	108441 (51460)	1:4	1153	21689	4616	86752	567
			1:1	2884	54221	2885	54220	354
			4:1	4615	86753	1154	21688	141

[a] The number of active compounds, n_a, and the number of inactive compounds, n_i, are shown for the total WOMBAT database and for the different training test set splits. n_{ba} is the number of active molecules used in the bootstrapping and is calculated such that the number of actives is a constant fraction of the total number of molecules in the test set (0.0062). The numbers in parentheses denote the number of molecules in the set that carry some form of charge at biological pH.

The adrenoceptor, dopamine, and serotonin activity classes are subclasses of the amine GPCR activity class.

Molecules are only included in an activity class if they have a measured IC_{50}, EC_{50}, or K_i of <10 μM against a target encompassed in the activity class. If a molecule is not found to be active against any of the proteins in the activity class, then it is considered inactive. It should be noted that a weakness of this approach is that some of the molecules classified inactive for a particular activity class may not in fact be inactive; rather they have not been tested against any of the proteins in the activity class. This may have a negative impact on the measured performance of the classifier when applied to the test sets.

Active and inactive molecules were split into training and test sets by the random selection of molecules for each activity class. Three different splits of the data were employed to examine the effect of the size of the training set. These splits were 1:4, 1:1, and 4:1. The number of actives and inactives used in the training and test sets for this validation is shown in Table 2 for each activity class and each split.

Each of the splits was performed 5 times, selecting molecules at random for the training and test sets, resulting in 15 data sets for each activity class. This was done to test that the performance of the classifier is consistent across differing sets of molecules for the same activity class and the same training test set ratio. As described above, bootstrap average with active replacement was employed for each of these data sets to provide a sensible ratio of active molecules in the test sets. This was performed 20 times for each of the 15 data sets for each activity class. The number of active molecules used in the bootstrapping is shown in Table 2 for all activity classes and splits.

A factor often overlooked in cheminformatics data mining techniques is the fact that many druglike molecules are charged at biological pH. In an approach where atoms are approximated by pharmacophoric features this would seem to be a critical factor given that the nature of an atomic feature can vary depending on whether it is charged or not. An example of this is an aliphatic secondary nitrogen atom that, in many cases, at biological pH, would be protonated. If this atom is not protonated, then the secondary nitrogen would be considered a hydrogen bond donor/acceptor; however, upon protonation the nitrogen atom is only a hydrogen bond donor. As such, the test sets described above have been duplicated, and the molecules are charged at biological pH. A rule-based in-house developed program was used to do this.

Validation Experiment 2. A second validation experiment was carried out in order to provide a more challenging test of this approach of classifying molecules. This validation focused on GPCR activity classes. Active molecules were extracted from the Prous Integrity database (Prous Science http://www.integrity.com) for the following classes: amine, nucleotide, peptide, dopamine, and serotonin GPCR activity classes. Active molecules were extracted from the Arena

Table 3. Activity Classes and Numbers of Active and Inactive Compounds Used in the Second Validation Experiment Unique Data Sets[a]

		total		training set		test set		
activity class	source	n_a	n_i	n_a	n_i	n_a	n_i	n_{ba}
GPCR (Level 1)								
amine GPCR	Prous	11126 (9773)	100872 (45308)	6669	50436	4457	50436	340
	Arena	9250 (8396)	100872 (45308)	6669	50436	2581	50436	329
nucleotide GPCR	Prous	2861 (435)	110279 (56501)	1965	55140	896	55139	347
peptide GPCR	Prous	9582 (6660)	102460 (48858)	5875	51230	3707	51230	341
GPCR (Level 2)								
dopamine GPCR	Prous	3101 (3059)	109921 (52757)	2144	54961	957	54960	347
serotonin GPCR	Prous	4828 (4613)	108441 (51460)	2884	54221	1944	54220	348
	Arena	4494 (4355)	108441(51460)	2884	54221	1610	54220	346

[a] The number of actives, n_a, and the number of inactives, n_i, are shown in total and for the training and test sets. n_{ba} is the number of active molecules used in the bootstrapping and is calculated such that the number of actives is a constant fraction of the total number of molecules in the test set (0.0062). The source column refers to the source of the active molecules in the test sets. The numbers in parentheses denote the number of molecules in the set that carry some form of charge at biological pH.

Table 4. Activity Classes and Numbers of Active and Inactive Compounds Used in the Second Validation Experiment Low Similarity Data Sets[a]

		total		training set		test set		
activity class	source	n_a	n_i	n_a	n_i	n_a	n_i	n_{ba}
GPCR (Level 1)								
amine GPCR	Prous	8390 (7305)	100872 (45308)	6669	50436	1721	50436	323
	Arena	8337 (7694)	100872 (45308)	6669	50436	1868	50436	324
nucleotide GPCR	Prous	2304 (310)	110279 (56501)	1965	55140	339	55139	344
peptide GPCR	Prous	5875 (5034)	102460 (48858)	5875	51230	1463	51230	327
GPCR (Level 2)								
dopamine GPCR	Prous	2553 (2614)	109921 (52757)	2144	54961	409	54960	343
serotonin GPCR	Prous	3765 (3583)	108441 (51460)	2884	54221	881	54220	342
	Arena	4166 (4030)	108441(51460)	2884	54221	1282	54220	344

[a] The number of actives, n_a, and the number of inactives, n_i, are shown in total and for the training and the test sets. n_{ba} is the number of active molecules used in the bootstrapping and is calculated such that the number of actives is a constant fraction of the total number of molecules in the test set (0.0062). The source column refers to the source of the active molecules in the test sets. The numbers in parentheses denote the number of molecules in the set that carry some form of charge at biological pH.

screening library for the amine and serotonin GPCR activity classes.

The Prous Integrity database contains over 258 000 bioactive compounds, 5600 biological targets from over 1500 journals, and 300 conferences in the areas of medicinal chemistry, organic synthesis, experimental pharmacology, clinical pharmacology, and genomics. Active molecules were extracted from the Prous Integrity database if the "mechanism of action" for the molecules includes one of the proteins encompassed by the activity class. Active molecules were taken from the Arena screening library for both the amine GPCR and serotonin GPCR activity classes if they had a measured IC_{50} <10 μM against a protein encompassed by either of the activity classes. In both cases active molecules were removed from these sets if they are contained in the WOMBAT database. The data sets including these molecules will be referred to as the "unique" sets.

Classifiers were created using data from the WOMBAT database and then tested on the active molecules from the data sources described above. Classifiers for the aforementioned activity classes were trained using half the available actives and half the available inactives from WOMBAT. For example, the classifier for the serotonin activity class was created using 2885 randomly selected actives and 54 221 randomly selected inactives. The classifier was then tested using the actives from either the Prous Integrity database or the Arena screening library in conjunction with the remaining inactives from WOMBAT.

To make this validation experiment more challenging, the similarities of the actives from either Prous or the Arena screening library were calculated with the corresponding actives from WOMBAT using the MDL320 keys[47] and the Tanimoto coefficient. Molecules were then removed from the test set actives (Prous or Arena) if they had a Tanimoto similarity ≥ 0.7 to any molecule in the training set actives (WOMBAT). The data sets including these somewhat dissimilar test set actives will be referred to as the "low similarity" sets. The performance of each classifier was then tested against the low similarity sets in the same manner as the unique sets. Table 3 shows the number of training and test sets, actives and inactives, used in the unique data sets. Table 4 shows the same data for the low similarity data sets.

As in the previous validation experiment these test sets were charged at biological pH to investigate the effect of charging molecules on the performance of the classifier. A bootstrap average with active replacement was performed 20 times to provide the correct ratio of actives to inactives in the test sets. The number of active molecules used in the bootstrapping is shown in Tables 3 and 4 for the unique and low similarity molecule sets, respectively.

RESULTS AND DISCUSSION

Validation Experiment 1. Figure 3 shows the mean enrichment factors and BEDROC scores calculated. Confidence intervals were calculated for all the metrics using the

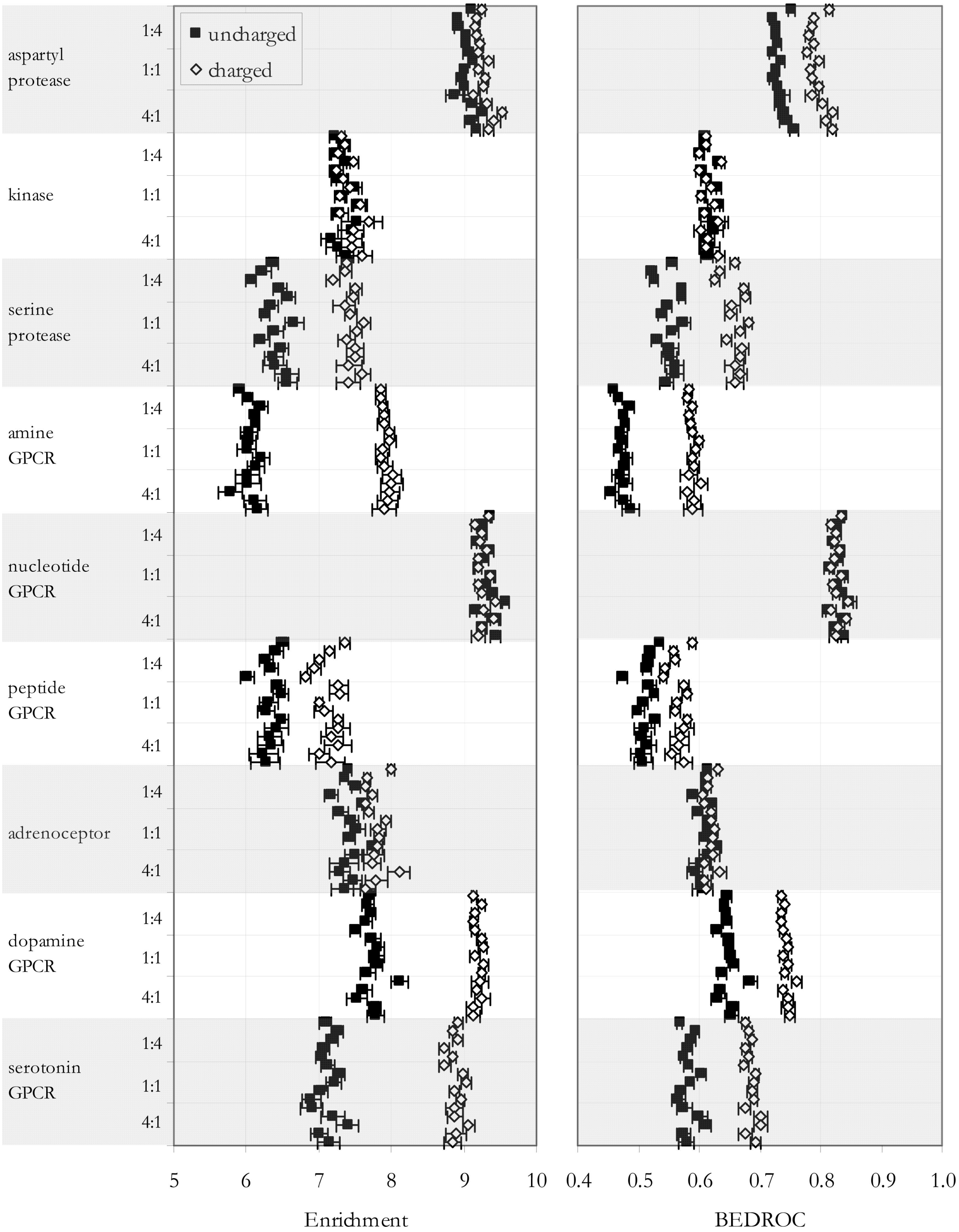

Figure 3. Dot plot showing the enrichment factors (LHS) and the BEDROC scores (RHS) for each of the activity classes for both the uncharged and charged molecule sets from WOMBAT. The error bars show the confidence intervals calculated assuming 95% confidence.

standard deviations from the bootstrapping (at the 95% confidence level) and are shown on the dot-plot as error bars.

The enrichment factors for all the data sets are significantly higher than 1.0. Furthermore the BEDROC scores are all significantly higher than 0.062 (the BEDROC score for a random distribution of actives). This indicates that using the naïve Bayes classifier to predict the likelihood of molecules belonging to a given activity class is better than choosing these molecules at random.

In all cases the enrichment factors and the BEDROC scores across the randomly sampled sets for a given activity class and training test set ratio are relatively constant. These scores are also consistent across the different training test set splits for a given activity class. This indicates that provided the diversity of the molecules in the training set is large enough, the classifier can perform well given a limited set of training data. The confidence intervals appear to get larger as the ratio of training set to test set molecules is increased. This is most clearly observed for the charged molecule sets for the amine GPCR activity class in Figure 3. WOMBAT is a database comprised of small congeneric sets of molecules often obtained from medicinal chemistry research papers. As such, the increase observed in the confidence intervals could be due to overfitting if there is a

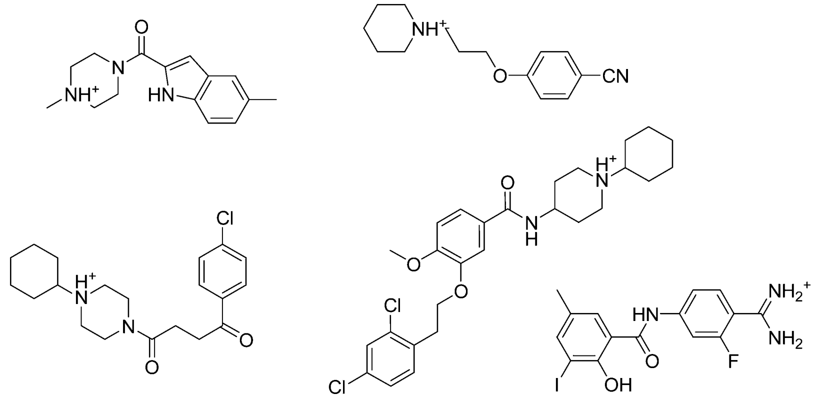

Figure 4. Examples of molecules containing protonatable nitrogens from the first validation experiment.

large number of very similar active molecules used in the training sets. This seems to indicate that a large nonredundant diverse set of molecules should be used to train the classifiers for optimum performance.

It is interesting to note that the enrichment factors and BEDROC scores are generally higher when the molecules are charged as opposed to uncharged. A charged enrichment factor or BEDROC score is considered statistically significantly higher than its uncharged counterpart if the minimum value for the charged metric (calculated by subtracting the confidence interval from the mean value) is greater than the maximum value for the uncharged metric (calculated by adding the confidence interval to the mean value). One hundred one out of the 135 enrichment factors are statistically significantly higher for the charged molecule sets compared to the uncharged molecules sets. Ninety-four of the 135 calculated BEDROC scores are statistically significantly higher for the charged molecule sets than the uncharged sets. There are no cases where the enrichment factor is significantly higher for the uncharged molecule sets and only 3 cases where the BEDROC score is significantly higher. The largest differences in the metrics for the charged and uncharged sets of molecules occur in the serine protease, amine GPCR, dopamine GPCR, and serotonin GPCR activity classes. The serotonin and dopamine GPCR activity classes both belong to the amine GPCR class. Many of the molecules that are active against the proteins making up these activity classes contain protonatable nitrogens. Some example molecules are a shown in Figure 4. The numbers of molecules that carry some form of charge at biological pH are shown in parentheses in Table 2 for both the active and inactive molecule sets. In the charged data sets these nitrogens would be protonated, and as such the atoms that would have previously been ascribed an acceptor (**A**) feature (and possibly a donor feature (**D**) for primary and secondary amines) would now only be ascribed a donor feature (**D**). Charging the molecules at biological pH in this case reduces the number of possible feature triplets in the molecule and would thus better differentiate the molecule from others that have an acceptor feature that cannot be protonated. A second reason for the enhanced enrichments is that by charging the molecules the use of the positive (**P**) and negative (**N**) feature types are employed. This means that additional triangles will be created for molecules containing atoms that would be charged at biological pH. This will further differentiate these molecules from those that do not have these features. Given that many of the molecules in the aforementioned activity classes contain these features it is no surprise the enrichments and BEDROC scores are greatly enhanced when the molecules are charged.

Example enrichment curves are shown for the kinase, serine protease, amine GPCR, and peptide GPCR activity classes in Figure 5. In each case the first randomly generated training–test set split is shown for each ratio (i.e., 1:4, 1:1, and 4:1) for both the uncharged and the charged data sets for the first random selection of actives in the bootstrap active replacement. The uncharged enrichment curves are shown with dotted lines, and the charged enrichments are shown with solid lines. Enrichment curves for a random distribution of active molecules in the test set are also shown for comparison purposes. The points at which 10% and 20% of the ranked list of molecules in the test set have been screened are shown with vertical gray dashed lines.

In all cases the enrichments curves are above the random curve indicating an improvement over random selection using this technique of classifying molecules. The enrichment curves for the charged data sets are significantly higher than the uncharged data sets in the serine protease and amine GPCR activity classes. This corresponds with the increase in enrichment factors and BEDROC scores previously observed. It is interesting to note the effect of charging the molecules on early retrieval. In the case of the amine GPCR activity class, early on in the percentage of the database screened, the percentage of actives found for the charged molecule sets is very similar to the uncharged sets. However, after 10% of the ranked list of molecules has been screened, there is a much larger difference in the percentage of actives retrieved (i.e., ~55% in the uncharged case compared to ~80% in the charged case). In the case of the serine protease activity class the effect of charging the molecules is observed much earlier on in the screening of the ranked list, i.e., after screening 1% of the ranked list of molecules ~55% of the active molecules are found in the charged case compared to ~35% in the uncharged case.

While the enrichment factors, BEDROC scores, and enrichment curves look impressive, they are probably overestimating the performance of the classifier. The problem here is the classic confederate or "me too" problem where the active molecules comprise small groups of highly similar molecules. It may be that the overall diversity of the active molecules in each activity class is high; however, when the active molecules are randomly partitioned into training and test sets it is likely that there will be one or more example molecules from each of these small groups in both the

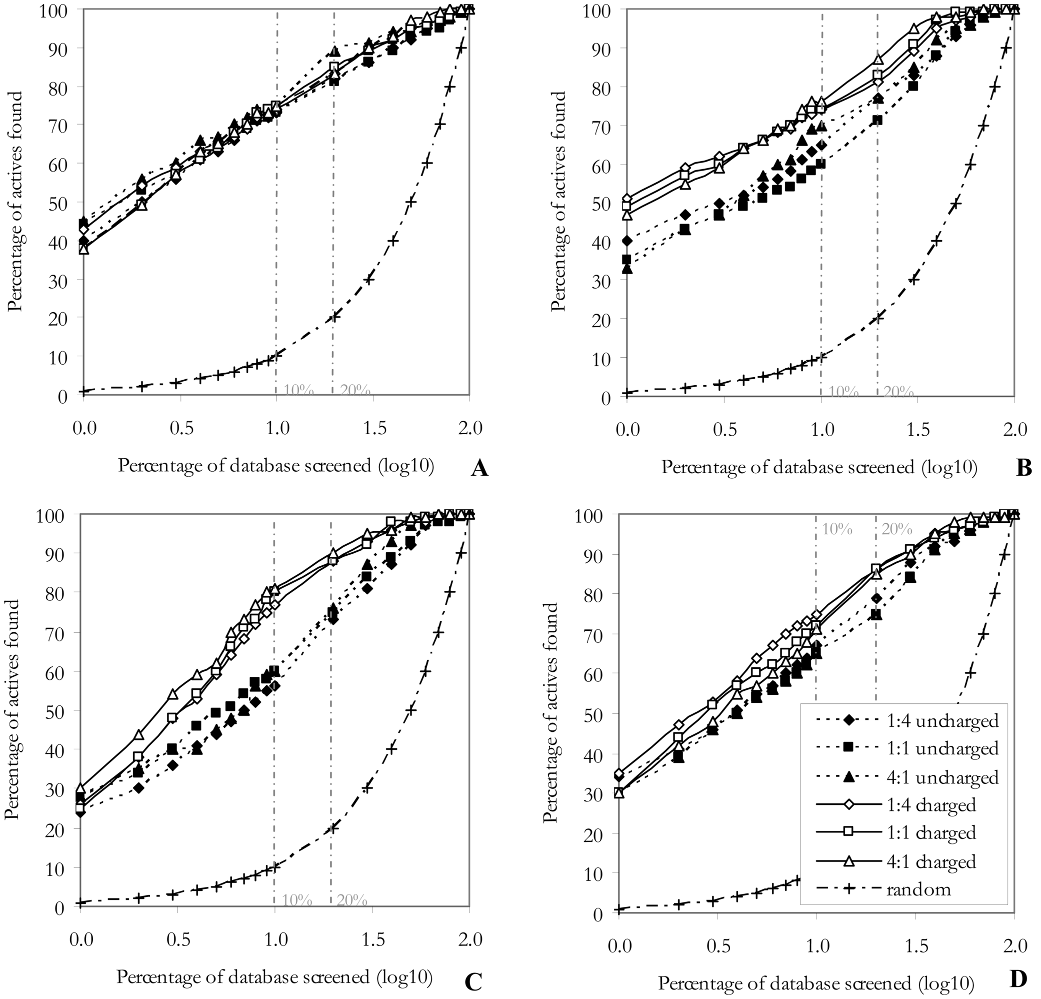

Figure 5. A selection of enrichment curves from the first validation experiment for the kinase (**A**), serine protease (**B**), amine GPCR (**C**), and peptide GPCR (**D**) activity classes. In each case the curve for the first randomly generated training test set split is shown for each ratio (i.e., 1:4, 1:1, and 4:1) for both the uncharged and the charged data sets for the first random selection of actives in the bootstrapping.

training and the test sets. Therefore enrichment studies using random selections of molecules to generate the training and test sets should always perform well.

Validation Experiment 2. The enrichment factors and BEDROC scores for the second validation experiment are shown in Figure 6. Figure 6A shows the enrichment and BEDROC scores for the unique test sets, and Figure 6B shows the enrichments and BEDROC scores for the low similarity test sets. Confidence intervals are shown on the plots calculated at 95% confidence. These are calculated using the standard deviations from the bootstrap average.

As in the first validation experiment the enrichment factors for all the sets of molecules are all significantly higher than 1.0, and the BEDROC scores are all significantly higher than 0.062. This means that using the naïve Bayes classifier to predict the likelihood that molecules belong to a given activity class is better than choosing molecules at random for these data sets. Generally speaking the enrichment factors and BEDROC scores for this validation experiment are lower than that of the first, which is to be expected given the nature of the test. The performance of the classifier is naturally worst for the low similarity test sets as these only contain actives that are somewhat different from the actives used to train the classifiers. However, significant enrichments are still obtained, particularly when the molecules in the training and test sets are charged. The confidence intervals themselves are very small across all the test sets showing that the variance in the enrichments from the bootstrap average is small. This observation demonstrates that the performance of a given classifier is consistent even though the active molecules in the test sets are varied.

The confidence intervals show (as in the first validation experiment) that in nearly all cases the metrics for the charged data sets are significantly higher than the uncharged data sets. This is not true for the nucleotide GPCR activity class for the unique test set from the Prous database, where the enrichment factors and the BEDROC scores are nearly identical, and the amine GPCR activity class for the low similarity test set from the Arena database. In the latter case the BEDROC score for the uncharged data set is significantly higher than the charged data set. This seems to indicate that although the percentage of actives retrieved in the top 10% is increased on charging the molecules, the early retrieval is actually reduced in this case. Table 5 shows the absolute and percentage differences in the enrichment factors and BEDROC scores between the charged and uncharged molecule sets for both the unique and low similarity test sets. It is clearly seen that the percentage increase in the enrichment factors is much larger in all cases than the percentage increase in the corresponding BEDROC score. Therefore, early retrieval (i.e., at the beginning of the top 10%) is probably very similar for the uncharged and charged

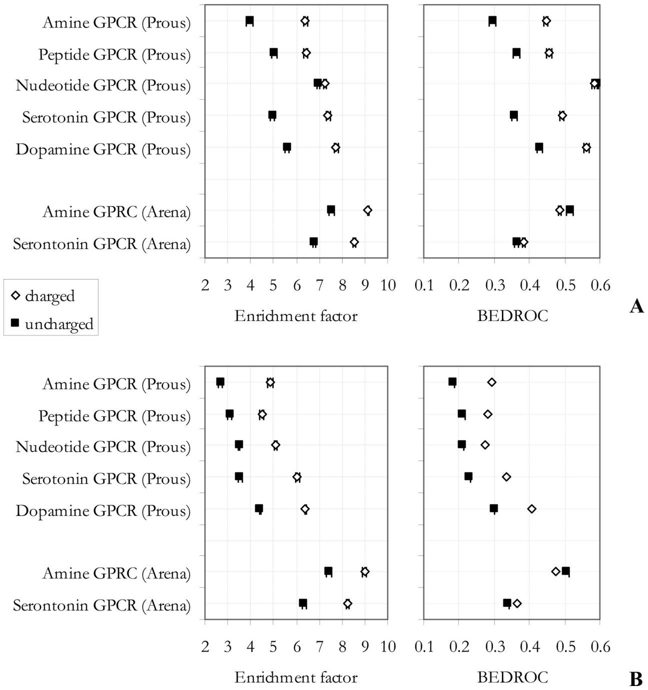

Figure 6. Enrichment factors and BEDROC scores for the second validation experiment. **A** shows the performance of the classifier against the test sets with only the duplicate actives removed (the "unique" test sets). **B** shows the performance of the classifiers against the test sets with actives removed that had a Tanimoto similarity ≥0.7 with any active molecules in the training set (the "low similarity" test sets). The performance is evaluated when the molecules are charged at biological pH (white diamonds) and when they are uncharged (black squares).

sets. The increased enrichment factors are due to an increase in the number of actives ranked at the end of the top 10%. Therefore, if this method were used to cherry pick a small number of compounds from a large ranked list, then the retrieval performance would be unaffected by the charged state of the molecules. However, if a larger number of molecules were selected from the ranked list, then the retrieval would be improved if the molecules were charged at biological pH.

The numbers of molecules that carry some form of charge at biological pH are shown in Tables 3 and 4 for both the active and inactive molecules making up the data sets. As can be seen, a high percentage of the active molecules for the amine GPCR (Prous + Arena), peptide GPCR (Prous), dopamine GPCR (Prous), and serotonin GPCR (Prous + Arena) activity classes would be charged at biological pH. Therefore the increase in the retrieval performance of the classifier for these data sets is easily explained. In contrast the nucleotide GPCR (Prous) activity class is dominated by neutral active molecules. It is interesting to note that there is a significant increase in both the enrichment factor and BEDROC metric for this activity class in the case of the low similarity set. This suggests that the improvement in retrieval performance when the molecules are charged does not only occur if the active molecules are dominated by charged entities.

Example enrichment curves from this validation experiment are shown in Figure 7. Figure 7A shows the enrichments for the nucleotide GPCR (Prous) test set, Figure 7B shows the peptide GPCR (Prous) test set, Figure 7C shows the dopamine GPCR (Prous) test set, and Figure 7D shows the amine GPCR (Arena) test set. In each case 4 enrichment curves are shown. These are for the uncharged and charged versions of the unique and low similarity data sets. From the shapes of the curves it is clear in all cases the enrichment is highest for the charged unique test sets where 71%, 71%, 77%, and 93% of the actives are found in the top 10% of the ranked list for the aforementioned test sets. As in the first validation experiment it is likely that these test sets contain "me too" compounds, so it is expected that the performance of the classifier would be best for these test sets. The retrievals for the actives in the uncharged unique test sets are 67%, 51%, 59%, and 75%, respectively, for the top 10% of the ranked list. These values are all lower than the corresponding charged case, although the magnitude by which they differ does vary. As described in the first validation experiment, this is due to the nature of the molecules that make up these test sets. The percentage retrievals of actives for the top 10% of the ranked list in the charged low similarity test sets are 50%, 50%, 65%, and 91%, respectively. As expected these are all less than (or equal to in the case of the amine GPCR (Arena) test set) the corresponding retrievals for the unique test sets.

By comparing the enrichment curves for the uncharged and charged low similarity test sets, the effect of charging the molecules on early and late retrieval within the top 10% of the ranked list can be more clearly observed. As noted in the comparison of the differences of the enrichment factors and BEDROC scores, the retrieval rates for the uncharged and charged molecule sets are very similar at the beginning of the top 10% of the ranked list. Toward the end of the top

Table 5. Absolute and Percentage Differences in Enrichment Factors and BEDROC Scores for the Unique and Low Similarity Test Sets between Uncharged and Charged Data Sets[a]

		unique				low similarity			
activity class	source	ΔE	ΔB	E_{pi}	B_{pi}	ΔE	ΔB	E_{pi}	B_{pi}
				GPCR (Level 1)					
amine GPCR	Prous	2.4	0.15	62	51	2.2	0.11	81	60
	Arena	1.6	−0.03	22	−5	1.6	−0.03	21	−6
nucleotide GPCR	Prous	0.3	−0.01	4	−1	1.6	0.07	45	32
peptide GPCR	Prous	1.3	0.09	27	25	1.4	0.07	45	33
				GPCR (Level 2)					
dopamine GPCR	Prous	2.1	0.13	39	31	2.0	0.11	45	36
serotonin GPCR	Prous	2.4	0.14	49	38	2.5	0.11	71	48
	Arena	1.8	0.02	26	6	1.9	0.03	30	8

[a] ΔE and ΔB are the difference between the charged and uncharged enrichment factors and BEDROC scores, respectively. E_{pi} and B_{pi} are the percentage increase in the enrichment factors and BEDROC scores when the molecules are charged.

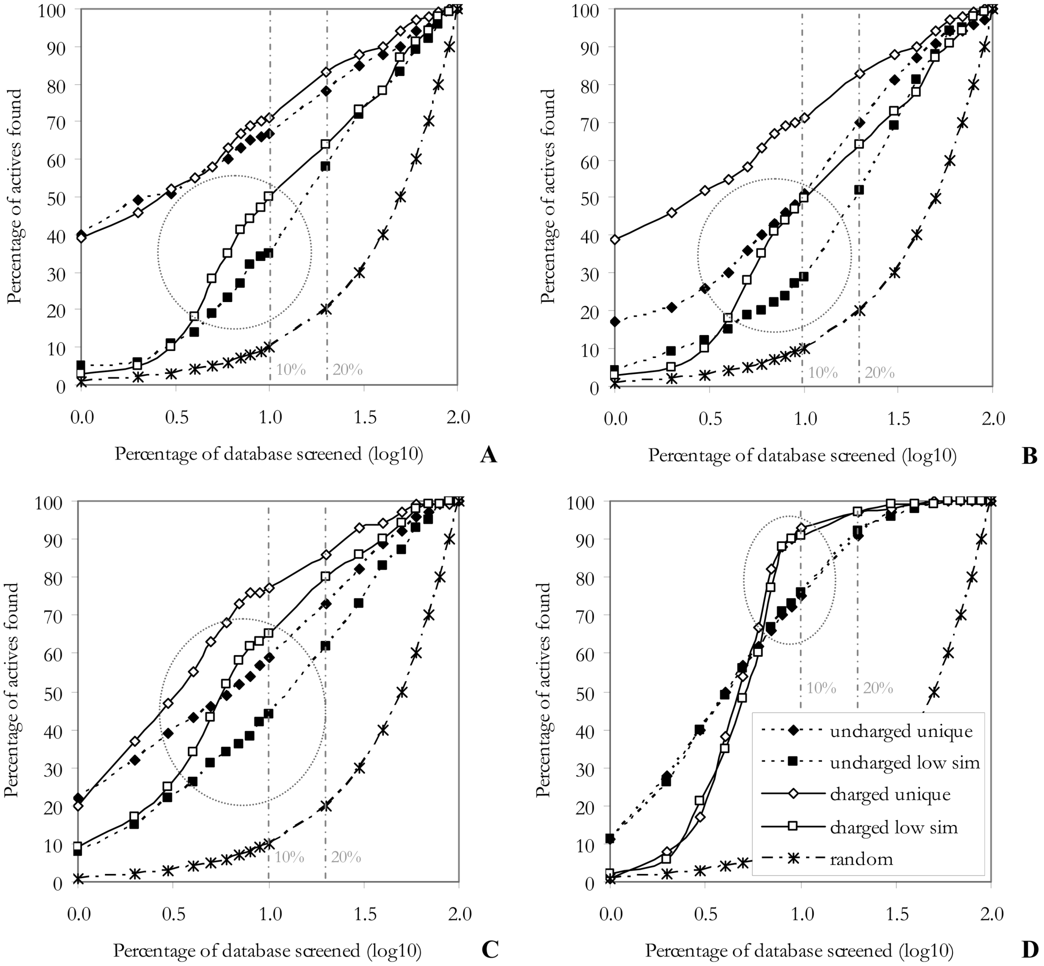

Figure 7. A selection of enrichment curves from the second validation experiment for the nucleotide GPCR (**A**), peptide GPCR (**B**), dopamine GPCR (**C**), and amine GPCR (**D**) activity classes. **A**, **B**, and **C** show the enrichment curves using the test set actives from the Prous database, and **D** shows the enrichment curves using the test set actives from the Arena screening library. In all cases both the "unique" and "low similarity" test sets are shown where the molecules in the data sets are charged and uncharged.

10% however, the rates of the retrieval are significantly different. This difference is clearly illustrated in all the enrichment graphs shown in Figure 7 where the pertinent areas of the charts are circled to show the increase in late retrieval in the top 10% for the charged molecule sets.

CONCLUSIONS

In this paper the use of a naïve Bayes classifier in conjunction with 2D feature triplet vectors has been assessed using two retrospective validation experiments. The retrieval rates and enrichment factors across a broad spectrum of targets and target families suggest that this method could be used to find novel molecules for screening or as chemical starting points.

In the first validation experiment the enrichment factors, BEDROC scores, and retrieval rates were very high. On reflection these results probably overestimate the performance of the classifier due to "me too" or confederate molecules. The second validation experiment was designed to try to account for this by selecting active test set molecules from different data sources, such as the Prous Integrity database and the Arena screening library, that were measurably dissimilar from the actives making up the training sets. It was found that the enrichment factors and BEDROC scores, while being lower than the first validation experiment, indicate significant improvement over random selection. It is therefore suggested that this method of ranking molecules as to their likelihood of being active against a given class of proteins could be of use in drug design, particularly in the creation of targeted screening libraries.

In nearly all cases the performance of the classifier is improved upon charging the molecules at biological pH. It was suggested there are two reasons for this. First, by charging the molecules donor (**D**) and acceptor (**A**) features can be more accurately assigned, i.e., a donor/acceptor feature may only be a donor feature at biological pH. Second, by charging the molecules new descriptors (i.e., the positive (**P**) and negative (**N**) atom features) are effectively introduced into the classifier. Intuitively the molecules appear better described by charging at biological pH and this is born out in the improved performance of the classifier.

By comparing the enrichment factors and BEDROC metrics for the uncharged and charged data sets in the second validation experiment, it was observed that the increases in the enrichment factors were larger than the increases in the corresponding BEDROC scores. This observation suggested that the increase in retrieval upon charging the molecules was occurring late in the top 10% of the ranked list. This was also observed in the enrichment curves from the second validation experiment. Therefore, if this method were used to select a small number of very highly ranked molecules (cherry picking), then the retrieval performance would probably be very similar regardless of whether the molecules were charged or uncharged. However, if a larger percentage of the database is to be selected (e.g., for creating focused libraries for screening purposes), then the success rate is likely to be increased if the molecules were charged.

ACKNOWLEDGMENT

The author thanks Gareth Jones and Carleton Sage for the very helpful comments during the preparation of this manuscript.

REFERENCES AND NOTES

(1) Willett, P. Searching techniques for databases of two- and three-dimensional chemical structures. *J. Med. Chem.* **2005**, *48*, 4183−4199.

(2) Bender, A.; Glen, R. C. Molecular similarity: a key technique in molecular informatics. *Org. Biomol. Chem* **2004**, *2*, 3204−3218.

(3) Willett, P. Similarity-based virtual screening using 2D fingerprints. *Drug Discovery Today* **2006**, *11*, 1046−1053.

(4) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Similarity searching of chemical databases using atom environment descriptors (MOLPRINT 2D): evaluation of performance. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1708−1718.

(5) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of topological descriptors for similarity-based virtual screening using multiple bioactive reference structures. *Org. Biomol. Chem* **2004**, *2*, 3256−3266.

(6) Hert, J.; Willett, P.; Wilton, D. J.; Acklin, P.; Azzaoui, K.; Jacoby, E.; Schuffenhauer, A. Comparison of fingerprint-based methods for virtual screening using multiple bioactive reference structures. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 1177−1185.

(7) Vidal, D.; Thormann, M.; Pons, M. LINGO, an efficient holographic text based method to calculate biophysical properties and intermolecular similarities. *J. Chem. Inf. Model.* **2005**, *45*, 386−393.

(8) Rhodes, N.; Willett, P.; Calvet, A.; Dunbar, J. B.; Humblet, C. CLIP: similarity searching of 3D databases using clique detection. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 443−448.

(9) Rush, T. S., III; Grant, J. A.; Mosyak, L.; Nicholls, A. A shape-based 3-D scaffold hopping method and its application to a bacterial protein-protein interaction. *J. Med. Chem.* **2005**, *48*, 1489−1495.

(10) Rhodes, N.; Clark, D. E.; Willett, P. Similarity searching in databases of flexible 3D structures using autocorrelation vectors derived from smoothed bounded distance matrices. *J. Chem. Inf. Model.* **2006**, *46*, 615−619.

(11) Ballester, P. J.; Richards, W. G. Ultrafast shape recognition to search compound databases for similar molecular shapes. *J. Comput. Chem.* **2007**, *28*, 1711−1723.

(12) Good, A. C.; Hodgkin, E. E.; Richards, W. G. Similarity screening of molecular data sets. *J. Comput.-Aided Mol. Des.* **1992**, *6*, 513−520.

(13) Ewing, T. J.; Makino, S.; Skillman, A. G.; Kuntz, I. D. DOCK 4.0: search strategies for automated molecular docking of flexible molecule databases. *J. Comput.-Aided Mol. Des.* **2001**, *15*, 411−428.

(14) Friesner, R. A.; Banks, J. L.; Murphy, R. B.; Halgren, T. A.; Klicic, J. J.; Mainz, D. T.; Repasky, M. P.; Knoll, E. H.; Shelley, M.; Perry, J. K.; Shaw, D. E.; Francis, P.; Shenkin, P. S. Glide: a new approach for rapid, accurate docking and scoring. 1. Method and assessment of docking accuracy. *J. Med. Chem.* **2004**, *47*, 1739−1749.

(15) Goodsell, D. S.; Morris, G. M.; Olson, A. J. Automated docking of flexible ligands: applications of AutoDock. *J. Mol. Recognit.* **1996**, *9*, 1−5.

(16) Halgren, T. A.; Murphy, R. B.; Friesner, R. A.; Beard, H. S.; Frye, L. L.; Pollard, W. T.; Banks, J. L. Glide: a new approach for rapid, accurate docking and scoring. 2. Enrichment factors in database screening. *J. Med. Chem.* **2004**, *47*, 1750−1759.

(17) Jones, G.; Willett, P.; Glen, R. C.; Leach, A. R.; Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **1997**, *267*, 727−748.

(18) McGann, M. R.; Almond, H. R.; Nicholls, A.; Grant, J. A.; Brown, F. K. Gaussian docking functions. *Biopolymers* **2003**, *68*, 76−90.

(19) Morris, G. M.; Goodsell, D. S.; Huey, R.; Olson, A. J. Distributed automated docking of flexible ligands to proteins: parallel applications of AutoDock 2.4. *J. Comput.-Aided Mol. Des.* **1996**, *10*, 293−304.

(20) Rarey, M.; Kramer, B.; Lengauer, T.; Klebe, G. A fast flexible docking method using an incremental construction algorithm. *J. Mol. Biol.* **1996**, *261*, 470−489.

(21) Taylor, R. D.; Jewsbury, P. J.; Essex, J. W. A review of protein-small molecule docking methods. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 151−166.

(22) McGaughey, G. B.; Sheridan, R. P.; Bayly, C. I.; Culberson, J. C.; Kreatsoulas, C.; Lindsley, S.; Maiorov, V.; Truchon, J. F.; Cornell, W. D. Comparison of topological, shape, and docking methods in virtual screening. *J. Chem. Inf. Model.* **2007**, *47*, 1504−1519.

(23) Hawkins, P. C.; Skillman, A. G.; Nicholls, A. Comparison of shape-matching and docking as virtual screening tools. *J. Med. Chem.* **2007**, *50*, 74−82.

(24) Goldman, B.; Walters, P. Machine Learning In Computational Chemistry. *Annu. Rep. Comput. Chem.* **2006**, *2*, 127−140.

(25) Witten, I. H.; Frank E. *Data mining: practical machine learning tools and techniques with Java implementations*, 1st ed.; Academic Press: San Diego, CA, 2000.

(26) Muller, K. R.; Ratsch, G.; Sonnenburg, S.; Mika, S.; Grimm, M.; Heinrich, N. Classifying 'drug-likeness' with kernel-based learning methods. *J. Chem. Inf. Model.* **2005**, *45*, 249−253.

(27) Byvatov, E.; Fechner, U.; Sadowski, J.; Schneider, G. Comparison of support vector machine and artificial neural network systems for drug/nondrug classification. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1882−1889.

(28) Byvatov, E.; Schneider, G. SVM-based feature selection for characterization of focused compound collections. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 993−999.

(29) Byvatov, E.; Schneider, G. Support vector machine applications in bioinformatics. *Appl. Bioinformatics* **2003**, *2*, 67−77.

(30) Svetnik, V.; Liaw, A.; Tong, C.; Culberson, J. C.; Sheridan, R. P.; Feuston, B. P. Random forest: a classification and regression tool for compound classification and QSAR modeling. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1947−1958.

(31) Buontempo, F. V.; Wang, X. Z.; Mwense, M.; Horan, N.; Young, A.; Osborn, D. Genetic programming for the induction of decision trees to model ecotoxicity data. *J. Chem. Inf. Model.* **2005**, *45*, 904−912.

(32) Li, H.; Yap, C. W.; Ung, C. Y.; Xue, Y.; Cao, Z. W.; Chen, Y. Z. Effect of selection of molecular descriptors on the prediction of blood-brain barrier penetrating and nonpenetrating agents by statistical learning methods. *J. Chem. Inf. Model.* **2005**, *45*, 1376−1384.

(33) Bazeley, P. S.; Prithivi, S.; Struble, C. A.; Povinelli, R. J.; Sem, D. S. Synergistic use of compound properties and docking scores in neural network modeling of CYP2D6 binding: predicting affinity and conformational sampling. *J. Chem. Inf. Model.* **2006**, *46*, 2698−2708.

(34) Bernazzani, L.; Duce, C.; Micheli, A.; Mollica, V.; Sperduti, A.; Starita, A.; Tine, M. R. Predicting physical-chemical properties of compounds from molecular structures by recursive neural networks. *J. Chem. Inf. Model.* **2006**, *46*, 2030−2042.

(35) Selzer, P.; Ertl, P. Applications of self-organizing neural networks in virtual screening and diversity selection. *J. Chem. Inf. Model.* **2006**, *46*, 2319−2323.

(36) Xia, X.; Maliski, E. G.; Gallant, P.; Rogers, D. Classification of kinase inhibitors using a Bayesian model. *J. Med. Chem.* **2004**, *47*, 4463−4470.

(37) Bender, A.; Mussa, H. Y.; Gill, G. S.; Glen, R. C. Molecular surface point environments for virtual screening and the elucidation of binding patterns (MOLPRINT 3D). *J. Med. Chem.* **2004**, *47*, 6569−6583.

(38) Bender, A.; Mussa, H. Y.; Glen, R. C. Screening for dihydrofolate reductase inhibitors using MOLPRINT 2D, a fast fragment-based method employing the naive Bayesian classifier: limitations of the descriptor and the importance of balanced chemistry in training and test sets. *J. Biomol. Screening* **2005**, *10*, 658−666.

(39) Sun, H. A naive bayes classifier for prediction of multidrug resistance reversal activity on the basis of atom typing. *J. Med. Chem.* **2005**, *48*, 4031−4039.

(40) O'Brien, S. E.; de Groot, M. J. Greater than the sum of its parts: combining models for useful ADMET prediction. *J. Med. Chem.* **2005**, *48*, 1287−1291.

(41) Truchon, J. F.; Bayly, C. I. Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model.* **2007**, *47*, 488−508.

(42) Weininger, D. SMILES: A Chemical Language and Information System: 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31−36.

(43) Weininger, D.; Weininger, J. L. SMILES 2: Algorithm for Generation of Unique SMILES Notation. *J. Chem. Inf. Comput. Sci.* **1989**, *29*, 97−101.

(44) Bruno, I. J.; Cole, J. C.; Lommerse, J. P.; Rowland, R. S.; Taylor, R.; Verdonk, M. L. IsoStar: a library of information about nonbonded interactions. *J. Comput.-Aided Mol. Des.* **1997**, *11*, 525−537.

(45) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 391−405.

(46) Olah, M.; Mracec, M.; Ostopovici, L.; Rad, R.; Hadaruga, N.; Olah, I.; Banda, M.; Simon, Z.; Mracec, M.; Oprea, T. I. WOMBAT: World of Molecular Bioactivity. In *Chemoinformatics in Drug Discovery*; Oprea, T. I., Ed.; Wiley-VCH: New York, 2004; pp 223−239.

(47) Durant, J. L.; Leland, B. A.; Henry, D. R.; Nourse, J. G. Reoptimization of MDL keys for use in drug discovery. *J. Chem. Inf. Comput. Sci* **2002**, *42*, 1273−1280.

CI7003253