Similarity Searching of Chemical Databases Using Atom Environment Descriptors (MOLPRINT 2D): Evaluation of Performance

Andreas Bender, Hamse Y. Mussa, and Robert C. Glen*

Unilever Centre for Molecular Science Informatics, Chemistry Department, University of Cambridge, Cambridge CB2 1EW, United Kingdom

Stephan Reiling

Lead Generation Informatics, Drug Innovation & Approval (DI&A), Aventis, 1041 Route 202-206, Bridgewater, New Jersey 08807

Received April 16, 2004

A molecular similarity searching technique based on atom environments, information-gain-based feature selection, and the naive Bayesian classifier has been applied to a series of diverse datasets and its performance compared to those of alternative searching methods. Atom environments are count vectors of heavy atoms present at a topological distance from each heavy atom of a molecular structure. In this application, using a recently published dataset of more than 100000 molecules from the MDL Drug Data Report database, the atom environment approach appears to outperform fusion of ranking scores as well as binary kernel discrimination, which are both used in combination with Unity fingerprints. Overall retrieval rates among the top 5% of the sorted library are nearly 10% better (more than 14% better in relative numbers) than those of the second best method, Unity fingerprints and binary kernel discrimination. In 10 out of 11 sets of active compounds the combination of atom environments and the naive Bayesian classifier appears to be the superior method, while in the remaining dataset, data fusion and binary kernel discrimination in combination with Unity fingerprints is the method of choice. Binary kernel discrimination in combination with Unity fingerprints generally comes second in performance overall. The difference in performance can largely be attributed to the different molecular descriptors used. Atom environments outperform Unity fingerprints by a large margin if the combination of these descriptors with the Tanimoto coefficient is compared. The naive Bayesian classifier in combination with information-gain-based feature selection and selection of a sensible number of features performs about as well as binary kernel discrimination in experiments where these classification methods are compared. When used on a monoaminooxidase dataset, atom environments and the naive Bayesian classifier perform as well as binary kernel discrimination in the case of a 50/50 split of training and test compounds. In the case of sparse training data, binary kernel discrimination is found to be superior on this particular dataset. On a third dataset, the atom environment descriptor shows higher retrieval rates than other 2D fingerprints tested here when used in combination with the Tanimoto similarity coefficient. Feature selection is shown to be a crucial step in determining the performance of the algorithm. The representation of molecules by atom environments is found to be more effective than Unity fingerprints for the type of biological receptor similarity calculations examined here. Combining information prior to scoring and including information about inactive compounds, as in the Bayesian classifier and binary kernel discrimination, is found to be superior to posterior data fusion (in the datasets tested here).

1. INTRODUCTION

Molecular similarity searching is based on the "similar property principle" [1] and attempts to predict properties of molecules on the basis of knowledge derived from measured properties of another set of molecules.[2,3] It is the underlying principle of many current in silico structure-based drug design efforts, and there are several papers that discuss the problems with this principle and the related issue of neighborhood behavior.[4,5] In particular, medicinal chemistry is generally based on the principles of molecular similarity where small changes, maintaining overall similarity of structure, often result in maintenance of biological activity.[6,7]

Comparison of two items, in this case the comparison of molecules, involves their meaningful representation, a selection of features deemed to be important, and the actual distance metric to define similarity or dissimilarity.

Generally, the representation of structures is often classified by the dimensionality of data used to calculate the descriptors. One-dimensional descriptors use overall properties such as volume and log *P*,[8] two-dimensional descriptors may be derived from the connectivity table,[9] and three-dimensional descriptors use geometrical information from points in 3D space.[10,11] A more comprehensive list may

* Corresponding author phone: +44 (1223) 336 432; fax: +44 (1223) 763 076; e-mail: rcg28@cam.ac.uk.

10.1021/ci0498719 CCC: $27.50
Published on Web 08/20/2004

Table 1. Activity Classes, MDDR Activity IDs, and Sizes of Active Datasets Derived from the MDDR

activity name	MDDR activity ID	dataset size	activity name	MDDR activity ID	dataset size
5HT3 antagonists	06233	752	thrombin inhibitors	37110	803
5HT1A agonists	06235	827	substance P inhibitors	42731	1246
5HT reuptake inhibitors	06245	359	HIV protease inhibitors	71523	750
D2 antagonists	07701	395	cyclooxygenase inhibitors	78331	636
renin inhibitors	31420	1130	protein kinase C inhibitors	78374	452
angiotensin II AT1 antagonists	31432	943			

include one-dimensional descriptors,[8] topological indices,[9] fragment-based descriptors,[12,13] field-based descriptors,[10,14] subshape descriptors (which are based on parts of the overall shape, e.g., multiple-point pharmacophores[11]), surface-derived descriptors (e.g., autocorrelation[15]), affinity fingerprints,[16,17] spectrum-derived descriptors,[18] and back-projectable descriptors.[14,19] These lists partially overlap.

Selection of features is an important step of the method presented here. However, in many molecular similarity algorithms reported, fingerprints are often calculated and similarity coefficients are applied to the entire fingerprint sets, so that feature selection does not take place. This may introduce noise into the system, because some (or even many) of the features used to establish similarity or dissimilarity are irrelevant to the question asked. One may instead want to focus on those features deemed to be important. These features are often found to be different between different types of calculated "similarities".

The third step, comparison of molecules, usually employs similarity or dissimilarity coefficients of which the Jacquard/Tanimoto[20] coefficient and the Manhattan distance are well-known. Several dozen similarity coefficients exist, and they have been extensively reviewed.[21] The Tanimoto coefficient, as well as binary bitstrings, was shown to possess some inherent properties.[22−24] The size dependence[22,23] of the Tanimoto coefficient was recently addressed by a modified Tanimoto coefficient.[25] If information from more than one molecule is given, the problem of merging information can be problematic. One approach, data fusion, can be utilized after a series of single-query similarity searches has been carried out.[21,26] In this paper we conclude that data fusion prior to scoring, forming one query derived from multiple structures, is superior to data fusion after scoring (see section 5).

The descriptor used in this work belongs to the group of two-dimensional molecular representations and has been presented previously.[27,28] It is based on the connectivity table of a molecular structure and describes the environment of each of its atoms by taking into account the atom types of its neighbors. In this approach the selection of features is carried out using information-gain-based feature selection. In a fashion similar to that of binary kernel discrimination,[29,30] a type of data fusion is performed prior to scoring using the naive Bayesian classifier.

Here, we have further investigated the combination of atom environments, information-gain-based feature selection, and a Bayesian classifier by applying the method to additional datasets and by exploring its behavior with respect to adjustable parameters.

The method has been applied to a recently published dataset[31] derived from the MDL Drug Data Report (MDDR).[32] This dataset comprises more than 100,000 structures, which include 11 sets of active structures. The active datasets range in size from 349 to 1236 structures and are currently some of the largest datasets with associated activity data available. (The precise dataset sizes are given in Table 1.) In a recent publication, similarity searching results using Unity fingerprints and various data fusion methods as well as binary kernel discrimination were published on this dataset,[31] which makes it a comprehensive benchmark for any new method with respect to size and diversity of the structures.

The method was also applied to a monoaminooxidase (MAO) dataset[29,33] which comprises 1650 structures. This dataset was used as a test set for the investigation of the binary kernel discrimination method applied to chemical similarity searching.[29] In this study, binary kernel discrimination was used in connection with topological torsion and atom pair descriptors.

Finally, the dataset used in the original presentation of the algorithm[28] has been reevaluated. This dataset contains 957 structures in 5 activity classes[17] and is also derived from the MDDR database.[32] Previously, information-gain-based feature selection was used throughout. Here, the Tanimoto coefficient is investigated as a similarity index in combination with atom environments. The objective is to investigate how the descriptor itself compares to other similar types of descriptors. We also investigate bypassing the information-gain-based feature selection in favor of selection of the most frequent features from the active and inactive sets. This is compared to skipping the feature selection step completely to gauge its influence on search performance.

Section 2 briefly summarizes the new algorithm. (For a more detailed description see ref 28, where additional particulars are given on the datasets as well as computational details.) Section 3 presents the results, which are discussed fully in section 4. These sections also give a comparison of the performance of the algorithm to those of established methods. Conclusions are in section 5.

2. MATERIALS AND METHODS

a. Descriptor Generation/Molecular Representation. Translationally and rotationally invariant atom environments are used[27,28] as molecular representation. Atom environments are calculated directly from the molecular connectivity table. They are calculated in a two-step procedure (see Figure 1): (1) Sybyl atom types[34] are assigned to every heavy atom in the hydrogen-depleted structure of the molecule. (2) An individual atom fingerprint is calculated for every heavy atom in the molecule. A count vector is constructed with the vector elements being counts of atom types at a given distance from the central atom. This calculation is performed using distances from zero up to two bonds and keeping count of the occurrences of the atom types.

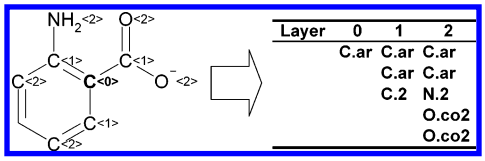

Figure 1. Illustration of the descriptor generation step, applied to an aromatic carbon atom. The distances ("layers") from the central atom are given in angular brackets. In the first step, Sybyl mol2 atom types are assigned to all atoms in the molecule. In the second step, count vectors from the central atom (here C⟨0⟩) up to a given distance (two bonds from the central atom apart) are constructed. Molecular atom environment fingerprints are then binary presence/absence indicators of count vectors of atom types.

Atom environments are stored as binary presence/absence features for each molecule. Since they are calculated for every heavy atom of the structure, as many descriptors are calculated as there are heavy atoms present in the structure. They are similar to signature molecular descriptors[35,36] and resemble augmented atoms.[37] SciTegic extended connectivity fingerprints (ECFPs)[38] are constructed in a similar fashion, but a major difference is the atom-type definition used (where we use Sybyl atom types).

b. Feature Selection. The information content of individual atom environments was computed using the information gain measure of Quinlan.[39,40] This was originally introduced to choose the best features for nodes of a decision tree, but the underlying concept of information entropy is generally applicable. Higher information gain is related to lower information entropy of the subsets defined by the presence and absence of a particular feature. This effectively describes better separation between active and inactive structures by that particular feature.

The information gain, I, is given by

$$I = S - \sum_{v} \frac{|D_v|}{|D|} S_v$$

where

$$S = -\sum_{i} p_i \log_2 p_i$$

S is the information entropy, S_v is the information entropy in data subset v (e.g., v = active, inactive), $|D|$ is the total number of datasets, $|D_v|$ is the number of data points in subset v, and p is the probability that a randomly selected molecule of the whole dataset (or subset in the case of D_v) belongs to each of the defined classes. Index i in the latter formula can be illustrated as follows. If a split of the whole dataset with respect to the presence and absence of a feature is performed, $i = 1$ or 2. Then, p_1 and p_2 represent the proportion of the dataset containing the feature under consideration and the size of the dataset not containing the feature under consideration, respectively.

c. Classification. A naive Bayesian classifier[41] was employed as a classification tool. Its underlying assumption is the independence of features, although it appears to perform surprisingly effectively where features are not strictly independent. Trained with a given dataset which consists of known feature vectors (**F**) containing features f_i and their associated known classes (CL), a Bayesian classifier predicts the class that a new feature vector belongs to as the one with the highest probability of $P(\mathrm{CL}_v|\mathbf{F})$, which is given by

$$P(\mathrm{CL}_v|\mathbf{F}) = \frac{P(\mathrm{CL}_v)\, P(\mathbf{F}|\mathrm{CL}_v)}{P(\mathbf{F})} \quad (1)$$

where $P(\mathrm{CL}_v)$ is the probability of class v, $P(\mathbf{F})$ is the feature vector probability, $P(\mathbf{F}|\mathrm{CL}_v)$ is the probability of **F** given CL_v, and v is the class.

For two datasets, after application of the assumption of the independence of features, the resulting binary naive Bayesian classifier is given by

$$\frac{P(\mathrm{CL}_1|\mathbf{F})}{P(\mathrm{CL}_2|\mathbf{F})} = \frac{P(\mathrm{CL}_1)}{P(\mathrm{CL}_2)} \prod_i \frac{P(f_i|\mathrm{CL}_1)}{P(f_i|\mathrm{CL}_2)} \quad (2)$$

This equation is used to perform classification; i.e., all molecules are represented by their feature vectors **F**, and the resulting ratios $P(\mathrm{CL}_1|\mathbf{F})/P(\mathrm{CL}_2|\mathbf{F})$ are sorted in decreasing order. Molecules with the highest probability ratios are most likely to belong to class 1 (here the class of active molecules). Molecules with the lowest values are most likely to belong to class 2 (the class of inactive molecules). The ratio $P(\mathrm{CL}_1)/P(\mathrm{CL}_2)$ in formula 2 is characteristic for Bayesian methods, and its optimum value is generally not known in advance. It is set to achieve the best classification on the training set, and in our case, as we only use relative probability scores, it is set to the ratio of the sizes of the training subsets.

If a given feature from a new molecule is not present in one of the data subsets CL_1 or CL_2, the probability of class membership for that class would drop to zero immediately. This would cause problems in particular for the denominator in formula 2 as the probability ratio becomes infinite. We use a Laplacian correction to avoid this problem. Features which are not present in data subset v are assumed to be present in a dataset of twice the size. The term $P(f_i|\mathrm{CL}_v)$ in this case assumes the value $1/(2|D_v|)$.

d. Compilation of Datasets and Preprocessing. Three different datasets were used for validation of the algorithm. The first and largest one was presented recently by Hert et al.,[31] who compiled a dataset comprising virtually the whole MDDR database. Eleven sets of active structures were defined, ranging in size from 349 to 1236 structures. (Full details of the dataset sizes are given in Table 1.). From the 102535 structures of the original dataset, 102524 could be retrieved from our local MDDR database. Conversion of the dataset to mol2 format gave 102513 valid structures. This corresponds to 99.98% of the original dataset. All of the structures not retrieved belong to the inactive dataset. All structures in mol2 format could be converted to atom environment fingerprints.

This dataset spans a variety of activities as well as a very large number of compounds with defined end points (some of which are however ambiguous), which provides a useful benchmark for a similarity searching method. One has to be aware of the occurrence of close analogues though, which favors 2D methods, and of the fact that the MDDR does not include explicit information about the inactivity of com-

pounds. Nonetheless, relative values can be used to judge the relative performance of different molecular similarity searching methods.

The second dataset used was the Abbott monoaminooxidase dataset,[29,33] comprising 1650 structures. Of the total number of compounds, 1360 were inactive and 290 were active. In the dataset, activities were given in three categories, 1, 2, and 3, which were merged as in the binary kernel discrimination application to give one dataset containing the active molecules.

The third dataset has previously been reported using this method.[28] It comprises 957 ligands[17] extracted from the MDDR database. The set contains 49 5HT3 receptor antagonists, 40 angiotensin-converting-enzyme inhibitors, 111 3-hydroxy-3-methylglutaryl coenzyme A reductase inhibitors, 134 platelet-activating-factor antagonists, and 49 thromboxane A2 antagonists. An additional 574 compounds were selected randomly and did not belong to any of these activity classes.

Salts and solvent were removed, if present. Structures were converted to Sybyl mol2 format using OpenBabel[42] 1.100.2 with the −d option to delete hydrogen atoms and default mol2 atom typing. Atom environment fingerprints were then calculated directly from the mol2 files.

e. Calculations. For all calculations using the large MDDR dataset, the performance measure was the fraction of active compounds found within the first 5% of the sorted library. Sorting was performed by descending Tanimoto similarity or decreasing probability of compounds being active, if the naive Bayesian classifier was employed.

In the first part of the validation runs, 10 active compounds were selected randomly from each of the 11 classes of active compounds and the Tanimoto coefficient was employed. This similarity measure was used by Hert et al.[31] in combination with Unity fingerprints (results are only reproduced here), so it allows us to compare the descriptors only, atom environments vs Unity fingerprints, since the similarity coefficients as well as the datasets used were identical. For each selected compound, retrieval rates were calculated as given above and group and overall averages were calculated from the individual values.

In the next calculation, the Tanimoto coefficient was replaced by a naive Bayesian classifier and information-gain-based feature selection was added to the algorithm. A 10-fold selection of active compounds from each of the 11 datasets was performed, and the retrieval rates of the active compounds were calculated as described above. As in the application of the binary kernel discriminator by Hert et al.,[31] the number of inactive compounds was set to 10 and all calculations were repeated, selecting 100 inactive compounds for each run. Totals of 150, 250, and 500 features were selected in the information-gain based feature selection step when 10 active and 10 inactive structures were used. Totals of 250, 500, and 1000 features were selected using the dataset of 10 active and 100 inactive structures. Both runs were repeated using all available features for classification.

For the MAO dataset,[29,33] a 10-fold cross validation with a 50/50 random split of both active and inactive structures was performed. (In the paper introducing the binary kernel discrimination,[29] 5-fold cross validation was performed. The number of runs was increased due to the minimal computational demand of the algorithm, using only seconds of CPU time.) As done by Harper,[29] a smaller training set was also used, consisting of a total of 200 randomly selected compounds. The fraction of active compounds found was analyzed depending on the fraction of the sorted library screened. The number of features selected was set to 100, 200, and 500.

For the dataset published by Briem[17] containing 957 structures from the MDDR, the earlier performance measure was kept for comparison. For this dataset, the hit rate (which is equal to the number of active compounds) at the first 10 positions of the ranked library was calculated. For all five sets of active compounds, 1, 2, 3, 5, or 10 compounds each were chosen randomly and the rest of the library was ranked according to score. The negative compounds were used in a 50/50 split as described earlier,[28] giving several hundred inactive structures in each of the two subsets. This was based on the assumption that knowledge about inactive structures is generally cheaper and more readily available than knowledge about active structures. Comparison of structures was carried out using the Tanimoto coefficient as well as the Bayesian classifier. In the case of the Bayesian classifier, the selection of features showing the highest information gain was compared to the selection of simply the most frequent features and to using no feature selection at all. The number of selected features was set to 10, 20, 50, 100, 250, 500, and 1000; this number of features was selected according to information gain as well as starting from the features with the highest relative frequency. A selection of that high a number of features was possible because of the utilization of a higher number of readily available inactive structures. Using a very small number of active compounds is clearly not ideal when using the Bayesian classifier (which needs conditional probabilities), but in the protocol followed here it was performed to investigate its behavior in those extreme cases.

3. RESULTS

Results from the run employing the large dataset derived from the MDDR database are given in Tables 2 and 3 and visualized in Figure 2. All results from this dataset employing Unity fingerprints are reproduced from Hert et al.[31]

First, we compared retrieval rates of atom environments and Unity fingerprints used in combination with the Tanimoto similarity coefficient. The percentage of active compounds found in the top 5% of the ranked database varies greatly between the datasets, from 95.04% in the case of renin inhibitors to 13.16% in the case of cyclooxygenase inhibitors, if atom environment descriptors are employed. For Unity fingerprints, retrieval rates range between 80.54% in the case of renin inhibitors and 9.39% in the case of cyclooxygenase inhibitors. Retrieval rates averaged over all datasets using single queries give better results for atom environments, with 37.44% of the active compounds retrieved, compared to 30.58% if Unity fingerprints are used.

In the next series of runs, we used information from multiple molecules for similarity searching. Atom environments, information-gain-based feature selection, and the naive Bayesian classifier were compared to results reported earlier using Unity fingerprints and data fusion as well as binary kernel discrimination.[31] Results are again given in Tables 2 and 3 and visualized in Figure 2.

Table 2. Mean Percentage of Active Compounds Found in the Top 5% of the Ranked Library[a]

dataset	Unity, single queries	atom environments, single queries	data fusion max	BKD 10a10i, $k = 100$	BKD 10a100i, $k = 100$	AE 10a10i, 500 features	AE 10a100i, 250 features
5HT3 antagonists	21.15 (7.36)	25.40 (10.46)	49.03 (5.43)	47.79 (4.28)	52.32 (8.27)	61.67 (6.55)	66.58 (8.65)
5HT1A agonists	18.43 (5.32)	27.73 (6.63)	37.15 (4.06)	30.78 (5.71)	38.19 (7.03)	44.19 (6.77)	57.05 (6.41)
5HT reuptake antagonists	24.02 (10.08)	22.75 (8.62)	49.68 (5.45)	37.28 (4.56)	45.82 (7.93)	41.40 (4.26)	46.07 (3.99)
D2 antagonists	17.35 (6.60)	23.24 (11.09)	37.40 (4.92)	33.30 (7.70)	38.65 (7.38)	49.14 (6.94)	53.69 (7.52)
renin inhibitors	80.54 (13.83)	95.04 (2.83)	88.62 (1.90)	89.84 (5.95)	93.34 (1.35)	94.66 (0.86)	95.71 (0.68)
angiotensin II AT1 antagonists	48.04 (17.95)	68.01 (11.19)	80.44 (6.08)	82.19 (4.59)	84.47 (6.59)	93.62 (1.98)	95.05 (2.49)
thrombin inhibitors	33.51 (14.72)	34.79 (19.98)	58.58 (8.98)	54.48 (9.20)	63.06 (7.66)	60.14 (11.04)	66.15 (7.27)
substance P antagonists	26.87 (10.47)	31.03 (14.09)	47.14 (5.16)	44.79 (6.47)	58.39 (8.27)	59.16 (9.51)	68.43 (5.48)
HIV protease inhibitors	37.60 (13.82)	49.56 (25.04)	61.62 (7.85)	59.07 (9.73)	68.45 (8.31)	72.26 (11.41)	76.00 (4.60)
cyclooxygenase inhibitors	9.39 (4.76)	13.16 (6.49)	26.52 (7.15)	30.51 (6.58)	33.15 (4.68)	25.66 (4.85)	34.70 (4.47)
protein kinase C inhibitors	19.42 (13.43)	21.13 (16.14)	48.01 (8.99)	47.47 (9.84)	49.37 (10.84)	50.50 (4.25)	54.61 (10.13)
average	30.58 (10.76)	37.44 (12.05)	53.11 (6.00)	50.68 (6.78)	56.84 (7.12)	59.31 (6.22)	64.91 (5.61)

[a] Results using single Unity fingerprints, data fusion, and binary kernel discrimination are taken from Hert et al.[31] Numbers in parentheses are standard deviations of the mean values.

Table 3. Influence of the Number of Selected Features on the Mean Percentage of Active Compounds Found in the Top 5% of the Ranked Library, Applied to the Atom Environment Similarity Searching Algorithm[a]

	10 active compounds, 10 inactive compounds				10 active compounds, 100 inactive compounds			
training dataset	150 features	250 features	500 features	all features	250 features	500 features	1000 features	all features
5HT3 antagonists	60.49 (9.41)	59.65 (8.10)	61.67 (6.55)	59.53 (11.48)	66.58 (8.65)	59.72 (5.72)	53.88 (8.58)	45.93 (3.77)
5HT1A agonists	42.73 (10.76)	43.44 (10.15)	44.19 (6.77)	47.92 (5.76)	57.05 (6.41)	49.84 (5.83)	38.24 (5.00)	32.20 (5.05)
5HT reuptake antagonists	35.64 (6.80)	35.50 (4.94)	41.40 (4.26)	37.22 (6.47)	46.07 (3.99)	42.09 (6.27)	35.73 (4.63)	26.50 (8.02)
D2 antagonists	44.78 (11.85)	46.55 (9.90)	49.14 (6.94)	48.96 (7.03)	53.69 (7.52)	52.18 (5.45)	41.17 (6.70)	37.22 (5.32)
renin inhibitors	94.18 (0.83)	94.44 (0.78)	94.66 (0.86)	94.09 (1.51)	95.71 (0.68)	95.00 (0.67)	93.36 (1.15)	92.29 (1.04)
angiotensin II AT1 antagonists	91.51 (4.95)	92.45 (4.03)	93.62 (1.98)	92.44 (3.51)	95.05 (2.49)	94.39 (2.11)	92.35 (2.25)	88.23 (2.81)
thrombin inhibitors	55.40 (5.88)	50.84 (11.50)	60.14 (11.04)	59.51 (12.23)	66.15 (7.27)	62.62 (11.50)	57.39 (13.29)	55.86 (7.75)
substance P antagonists	64.34 (5.65)	49.47 (11.19)	59.16 (9.51)	57.10 (11.29)	68.43 (5.48)	62.10 (7.34)	55.60 (9.51)	56.12 (5.51)
HIV protease inhibitors	71.66 (8.88)	72.89 (9.56)	72.26 (11.41)	68.31 (6.34)	76.00 (4.60)	73.19 (4.39)	73.43 (7.06)	64.80 (8.05)
cyclooxygenase inhibitors	22.60 (8.41)	20.93 (5.95)	25.66 (4.85)	22.72 (5.98)	34.70 (4.47)	26.84 (7.48)	20.94 (4.80)	14.50 (4.29)
protein kinase C inhibitors	46.65 (6.15)	46.13 (6.86)	50.50 (4.25)	47.85 (15.54)	54.61 (10.13)	51.11 (12.11)	46.20 (6.54)	44.03 (7.74)
average	57.27 (7.23)	55.66 (7.54)	59.31 (6.22)	57.79 (7.92)	64.91 (5.61)	60.83 (5.81)	55.30 (6.08)	50.70 (5.40)

[a] Numbers in parentheses are standard deviations of the mean values.

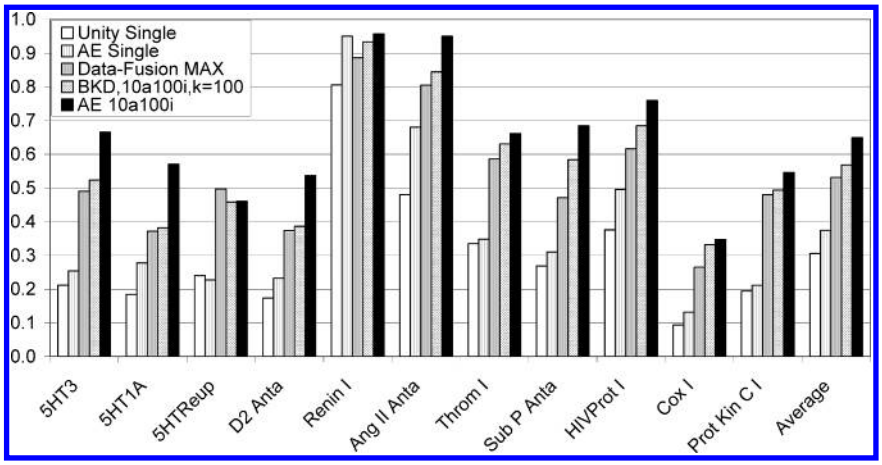

Figure 2. Fraction of active compounds retrieved in the top 5% of the sorted libraries. Compared are Unity fingerprints (white) and atom environments (light gray) using single-query structures; both are used in combination with the Tanimoto coefficient. Data fusion (dark gray) and binary kernel discrimination (BKD, medium gray) are used to merge information from Unity fingerprints, and atom environments are for that purpose combined with the naive Bayesian classifier (AE, black). Results using data fusion and binary kernel discrimination are taken from Hert et al.[31] Binary kernel discrimination and the atom environment method use 10 active and 100 inactive structures (AE 10a100i); the parameter k for the binary kernel method is set to 100. Information-gain-based feature selection in the case of the Bayesian classifier was set to select 250 features.

Combining information outperforms results using single queries in all cases. The percentage of active compounds found in the top 5% of the ranked database again varies greatly between the datasets, from greater than 90% in the case of renin inhibitors and angiotensin antagonists to about 30% in the case of cyclooxygenase inhibitors. Retrieval rates averaged over all datasets among the methods used give the lowest results for binary kernel discrimination ($k = 100$) using 10 active and 10 inactive structures at 50.68% of the active compounds retrieved, followed by data fusion at 53.11% actives found, binary kernel discrimination using 10 active and 100 inactive structures ($k = 100$) at 56.84% actives

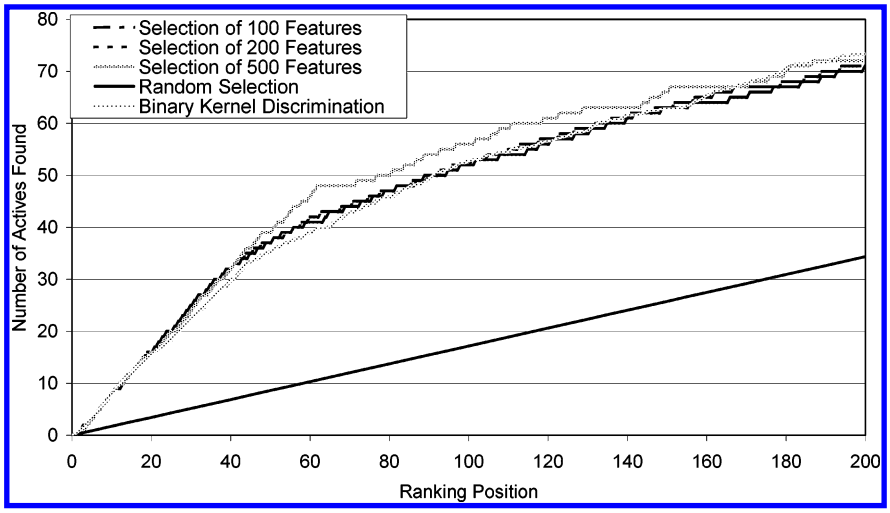

Figure 3. Number of MAO inhibitor compounds found at the top 200 positions of the ranked library when a 50/50 split of training and test data is used. The Bayesian classifier and binary kernel discrimination show comparable performance, with the Bayesian classifier using 500 features showing slightly better performance from ranking position 40 to ranking position 160.

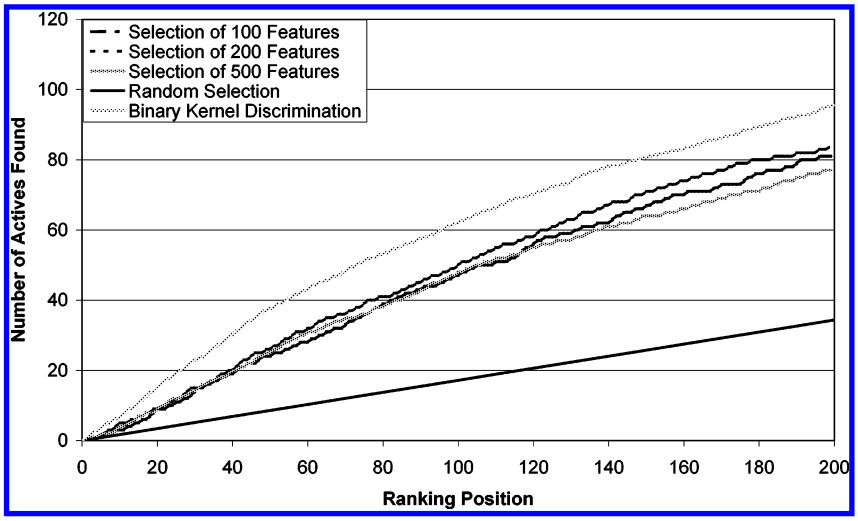

Figure 4. Number of MAO inhibitor compounds found at the top 200 positions of the ranked library when 200 training data points are used. Binary kernel discrimination outperforms the naive Bayesian classifier on sparse training data when using this dataset.

found, atom environments and the naive Bayesian classifier using 10 active and 10 inactive structures at 59.31% actives found, and the same method using 10 active and 100 inactive structures at 64.91% actives found. Atom environments and the Bayesian classifier retrieve on average nearly 10% more active structures than the next best method, binary kernel discrimination.

The highest number of active compounds is found in 10 out of 11 classes of active compounds by the atom environment approach if 10 active and 100 inactive compounds are used. Data fusion excels in one case, when applied to 5HT reuptake inhibitors. If 10 active and 10 inactive structures are used, the atom environment approach comes first in 9 out of 11 classes of active compounds, with data fusion being superior in the case of 5HT reuptake inhibitors. Binary kernel discrimination and data fusion are superior in the case of cyclooxygenase inhibitors.

The influence of the number of selected features on similarity searching performance is given in Table 3. Feature selection in the case of 10 active and 10 inactive molecules gives approximately comparable results for any number of selected features, on average between 55% and 60% of the active compounds retrieved. If no feature selection is employed, 57.79% of all active structures are retrieved, so skipping feature selection in this case does not decrease search performance. If 10 active and 100 inactive structures are used for query generation, average retrieval rates vary between about 65% and 55%, with the highest number of 1000 features showing relatively the worst results. If no feature selection is employed, the overall average retrieval rate falls even lower, to 50.70%. This result is clearly inferior to that of Unity fingerprints in combination with binary kernel discrimination and shows the importance of using feature selection in combination with the Bayesian classifier.

Using the MAO dataset, as shown in Figures 3 and 4, classification is comparable to binary kernel discrimination in the case of a 50/50 split of training and test sets (Figure 3). In the experiment with the smaller training set (Figure 4), comprising about 12% of the whole dataset, binary kernel discrimination is superior to the Bayesian classifier. Whereas 50 active compounds are found at ranking position 100 in the case of the Bayesian classifier, the binary kernel discriminator has already found 60 of the active compounds at this position.

The individual hit rates for each group of active compounds from the small dataset derived from the MDDR

Table 4. Performance of the Atom Environment Approach for Groups of Active Compounds: Hit Rates and Enrichment Factors among the Top 10 Compounds of the Sorted Library Using the Tanimoto Coefficient and Using the Naive Bayesian Classifier, Selection of 10 Features by Information Gain[a]

	5HT3	ACE	HMG	PAF	TXA2	overall
expected hit rate	0.50	0.41	1.15	1.39	0.50	0.79
av no. of active compds among top 10 ranked compds, using the Tanimoto coeff	7.4 (2.2)	7.8 (2.6)	8.6 (2.1)	7.7 (2.3)	6.6 (2.2)	7.5 (2.3)
enrichment factor	14.8	19.0	7.5	5.5	13.2	9.5
av no. of active compds among top 10 ranked compds, using the Bayesian classifier and 10 features	6.0 (3.2)	8.6 (1.8)	8.1 (2.8)	7.0 (2.2)	6.0 (3.0)	7.1 (2.6)
enrichment factor	12.0	21.0	7.0	5.0	12.0	9.0

[a] Numbers in parentheses are standard deviations of the mean values.

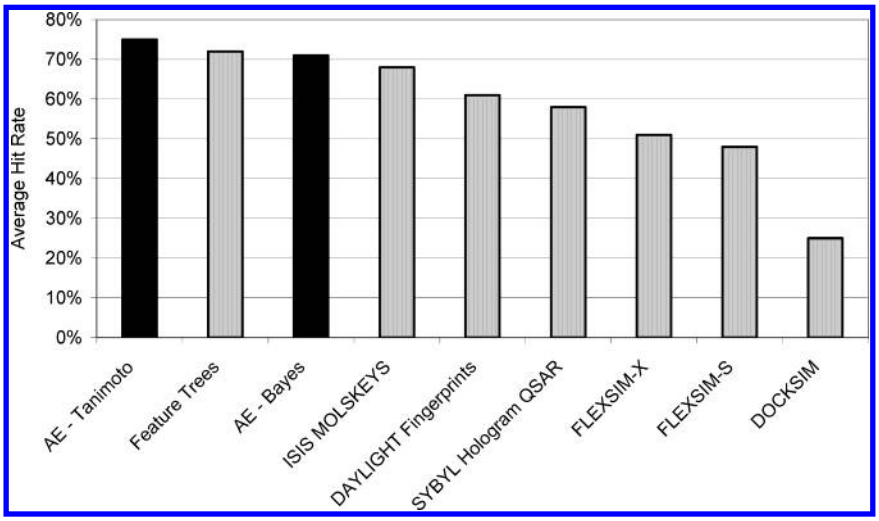

Figure 5. Mean sample hit rates of the atom environment (AE) approach (black), in comparison to the methods applied by Briem and Lessel (gray). The performance of the AE approach is on one hand shown using single queries combined with information-gain-based feature selection and on the other hand using the Tanimoto coefficient instead of the Bayesian classifier.

database are given in Table 4. Using the Tanimoto coefficient, the average number of active compounds among the top 10 ranked compounds varies from 6.6 (TXA2) to 8.6 (HMG), with an overall average of 7.5. These numbers result in enrichment factors between 5.5 (PAF) and 19.0 (ACE) with an average value of 9.5. These results are slightly better than those using the naive Bayesian classifier and single-query structures.

The nearest neighbor protocol of Briem[17] has been followed in this validation to enable ease of comparison of the algorithm performance with established methods, taken from an earlier publication by Briem and Lessel.[17] The methods used for comparison are feature trees,[43] ISIS MOLSKEYS,[44] Daylight fingerprints,[45] Sybyl hologram QSAR fingerprints,[46] and FLEXSIM-X and FLEXSIM-S[48] and DOCKSIM[49] virtual affinity fingerprints. Feature trees represent molecules as trees (acyclic graphs), which are subsequently matched for comparison. In current versions, the FlexX interaction profile and van der Waals radii are used as descriptors and a size-weighted ratio of fragments is used to calculate a similarity index. ISIS MOLSKEYS use 166 predefined two-dimensional fragments for describing a structure. Daylight fingerprints are algorithmically generated and describe atom paths of variable length: they are commonly folded, and a 1024-bit-long string is used. Hologram QSAR is an extension of 2D fingerprints and additionally includes branched and cyclic fragments as well as stereochemical information. For all 2D and 3D descriptors, Euclidean distances were calculated for each possible combination of test ligands. The performance of the algorithm presented here is compared to those of established methods and is shown in Figure 5.

Shown are mean sample hit rates averaged over all five classes of active compounds. Using one query, information-gain-based feature selection and the naive Bayesian classifier, the method presented here outperforms all three virtual affinity fingerprint algorithms as well as two of the two-dimensional methods, Daylight fingerprints and Sybyl hologram QSAR fingerprints. It performs as well as ISIS MOLSKEYS fingerprints and is only (marginally) outperformed by the feature tree approach. The top three methods are of comparable performance. If atom environments in combination with the Tanimoto coefficient are used, all commonly employed two-dimensional methods are outperformed slightly.

The influence of the number of active structures used to generate the query, the number of features selected, and the feature selection method on search performance is shown in Table 5. Performance at 10, 50, 500, and all selected features is shown in Figure 6.

We see a number of consistent trends in Figure 6. Performance generally increases with the number of active molecules used for query generation. If 10 features are selected, performance only marginally depends on the number of active structures. If more and more features are selected, performance increases with the number of active compounds available for training. Information-gain-based feature selection generally outperforms selection of the most

Table 5. Average Hit Rates over All Five Groups of Active Compounds Using Different Numbers of Active Compounds for Query Generation and Different Feature Selection Methods[a]

no. of selected features	selected by information gain (IG) or relative frequency (F)	av hit rates for different nos. of active molecules for query generation				
		1	2	3	5	10
10	IG	7.1 (2.6)	7.8 (1.8)	8.1 (1.7)	7.7 (2.2)	8.3 (1.2)
10	F	3.4 (2.1)	4.8 (2.4)	5.8 (2.0)	6.0 (2.1)	6.8 (1.4)
20	IG	7.1 (2.5)	7.8 (1.6)	8.2 (1.9)	8.7 (1.1)	8.6 (1.0)
20	F	4.7 (2.6)	6.1 (2.3)	6.4 (2.6)	7.0 (1.8)	8.2 (1.2)
50	IG	5.9 (2.8)	7.7 (1.9)	8.2 (1.4)	8.3 (1.3)	9.1 (0.8)
50	F	3.9 (2.4)	6.0 (2.1)	6.9 (2.0)	7.5 (1.7)	8.4 (1.1)
100	IG	3.4 (2.0)	5.5 (2.0)	7.6 (1.7)	8.2 (1.3)	8.6 (0.9)
100	F	2.3 (1.6)	5.1 (1.7)	6.8 (1.7)	7.4 (1.4)	8.7 (0.8)
250	IG	0.9 (0.7)	2.8 (1.7)	5.0 (1.8)	7.1 (1.4)	7.0 (0.9)
250	F	0.5 (0.5)	2.1 (1.5)	4.2 (1.6)	6.5 (1.4)	8.3 (1.0)
500	IG	0.4 (0.5)	1.4 (0.8)	3.3 (1.1)	5.1 (1.3)	8.1 (1.1)
500	F	0.4 (0.3)	1.3 (0.8)	2.8 (1.3)	4.9 (1.2)	7.8 (0.9)
1000	IG	0.2 (0.2)	0.9 (0.7)	1.7 (0.8)	3.4 (1.1)	7.2 (1.2)
1000	F	0.3 (0.2)	0.8 (0.7)	1.7 (0.8)	3.3 (0.9)	6.8 (1.0)
all		0.4 (0.0)	0.6 (0.5)	1.2 (0.5)	2.1 (0.8)	5.2 (1.3)

[a] For comparison, performance using all features is given. Numbers in parentheses are standard deviations of the mean values.

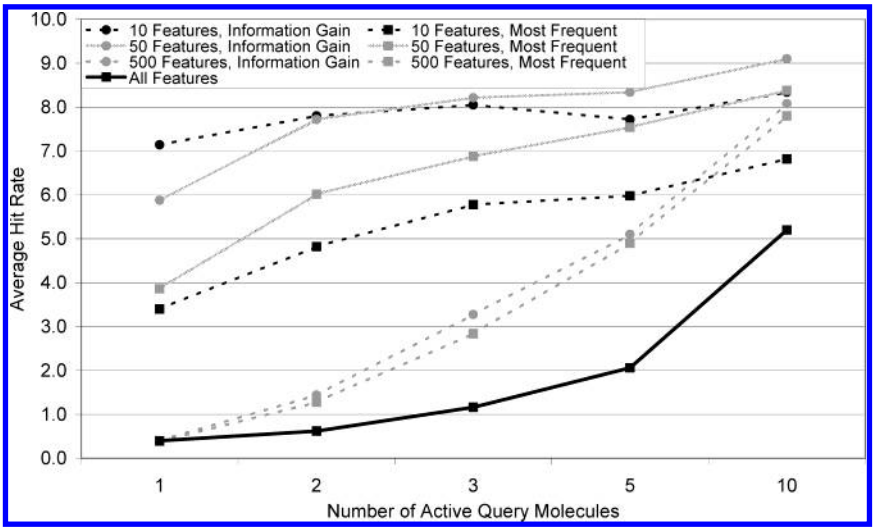

Figure 6. Influence of the number of active query molecules, the number of selected features, and the feature selection method (information gain, marked by circles, vs the selection of the most frequent features, marked by squares) on the performance of the similarity searching algorithm. For comparison, results using all features are shown. Information-gain-based feature selection performs better than the selection of the most frequent features.

frequent features as well as using the Bayesian classifier without any feature selection.

4. DISCUSSION

A consistent picture emerges in the validation using the large MDDR library, using single-query structures with the Tanimoto coefficient. In 10 out of 11 cases (with the exception of 5HT reuptake inhibitors), atom environments outperform Unity fingerprints with respect to compound retrieval rates, with 37.44% vs 30.58% of active compounds found. In absolute numbers, the performance difference is 6.86%. This calculates to a relative performance difference of 22.4%. Thus, atom environments capture more information that is relevant to the similarity searching task performed here.

If multiple-query molecules are used, atom environments in combination with the naive Bayesian classifier outperform Unity fingerprints in combination with data fusion as well as binary kernel discrimination in most cases (9 or 10 out of 11 cases depending on the number of inactive compounds chosen; see the Results). This observation is discussed in more detail below.

Comparing the performance of atom environments and the naive Bayesian classifier to that of Unity fingerprints and binary kernel discrimination, we find that in the case of 10 active and 10 inactive structures relative retrieval rates are 17.0% better if atom environments and the naive Bayesian classifier are used. If 10 active and 100 inactive structures are used, relative retrieval rates are on average 14.2% larger. In absolute numbers, atom environments and the naive Bayesian classifier are superior by 8.63% (8.08%).

The absolute difference of retrieval rates between atom environments and Unity fingerprints is retained at about 7−8% if the naive Bayesian classifier and binary kernel discrimination replace the Tanimoto coefficient, respectively. The relative performance gain drops from about 22% to 14−17%. This may be partly because retrieval performance gets close to the theoretical optimum in some of the datasets (renin inhibitors and angiotensin II inhibitors). Overall, we see that increased performance is in good part due to the

atom environment descriptor used and that binary kernel discrimination and the naive Bayesian classifier show comparable performance (absolute performance differences with respect to the Tanimoto coefficient are constant).

Performance of the naive Bayesian classifier comparable to that of binary kernel discrimination contrasts with the findings by Harper,[30] who found binary kernel discrimination to perform superior to "ordinary weighting" (what we call the naive Bayesian classifier) of features derived from MACCS-II and Daylight fingerprints. There are at least three possible reasons for this behavior.

First, the representation of structures has to be meaningful and has to provide "sensible" data to allow the classification method to do its job. It may be the case that database-tailored atom environments capture more information than Daylight and MACCS-II fingerprints, which were used by Harper.[30] This would of course not account for the different performance between both methods found by Harper, because a consistent representation was used throughout. But as discussed above, atom environments outperform Unity fingerprints for molecular similarity searching on the datasets we used. This explanation of performance differences is corroborated by the fact that atom environments in combination with information-gain-based feature selection and the naive Bayesian classifier perform worse only in the case of 5HT reuptake inhibitors—which is exactly the dataset where atom environments also perform worse than Unity fingerprints in combination with the Tanimoto coefficient.

This finding underlines the importance of having suitable different descriptors for different tasks at hand. While atom environments perform well with all other datasets, Unity fingerprints are superior in this particular case of 5HT reuptake inhibitors. In this dataset, crucial information seems not to be present in the atom environment descriptor (possibly long-distance information), so neither the Tanimoto coefficient nor the Bayesian classifier is able to give superior results, compared to Unity fingerprints. This gives hints that, for further improvements of the atom environment descriptor, longer distance nonbonded information could be included.

In addition to bioactivity prediction, similarity calculations can be applied to a much wider field of property predictions (see the Introduction), making a diverse collection of descriptors desirable. These descriptors have to capture as much information as possible that is relevant to the "similarity" under consideration. Applied to biological receptor-type similarity calculations, atom environments appear to fulfill this task sufficiently well to be useful additions in, e.g., database searching.

There are also two other important points which deserve being mentioned. Harper used the NCI dataset,[50] which includes structures with associated 50% growth inhibition data in a variety of cell lines. Growth inhibition of cells can be caused by a wide variety of molecular mechanisms. In contrast, more clearly defined end points are used here which are inhibition or agonist or antagonistic effects on clearly defined enzymes or receptors. It may thus be possible that the naive Bayesian classifier performs better in the case of single activity peaks, whereas binary kernel discrimination as a local method performs better in the case of multiple bioactivities. On the other hand, the set of serotonin ligands (5HT) includes ligands of all subtypes of this receptor, and it is thus quite a diverse dataset as well. Atom environments and the naive Bayesian classifier still perform better than binary kernel discrimination on this dataset. Given a relatively better performance of the descriptor by 5.2% (4.3% in absolute terms) in combination with the Tanimoto coefficient, if information from 10 active and 100 inactive molecules is used, atom environments and the naive Bayesian classifier outperform Unity fingerprints and binary kernel discrimination by an even larger margin, which is 27.3% in relative terms (and 14.3% in absolute numbers). Although this is not a general proof, applied to the diverse dataset of 5HT ligands, the naive Bayesian classifier outperforms binary kernel discrimination with respect to classification ability.

Third, feature selection seems to be a crucial step for the performance of the naive Bayesian classifier. Feature selection was not used in Harper's "ordinary weighting" scheme,[30] possibly contributing to the relative superiority of binary kernel discrimination. In our calculations feature selection did not have a big impact on search performance if 10 active and 10 inactive compounds were selected (first part of Table 3). Indeed, classification performance was as good if the feature selection step was skipped. But if 10 active and 100 inactive compounds were selected (second part of Table 3), the average retrieval rate continuously fell from 64.91% at 250 selected features to 60.83% at 500 selected features to 55.30% at 1000 selected features to 50.70% if feature selection was skipped completely and all features were used for classification. This illustrates the importance of the feature selection step, and it is a possible explanation for the relative superiority of binary kernel discrimination found in earlier work.[30] While it seems to be counterintuitive that classification performance decreases as more knowledge becomes available (using 100 instead of 10 inactive structures), one has to be aware that the signal/noise ratio is the quantity of importance, and from the results it seems that this ratio deteriorates if more inactive structures are used without employing feature selection. We discuss the importance of feature selection on search performance using the naive Bayesian classifier further below.

Applied to the MAO dataset, atom environments perform comparably or in a slightly superior way to binary kernel discrimination where the (large) training set comprises 50% of the library. If a smaller number of 200 training data points are used, binary kernel discrimination performs superior to the naive Bayesian classifier on this dataset. This is in contrast to the results obtained from the large MDDR dataset, where also small training datasets (10 active and 10 inactive structures or 10 active and 100 inactive structures) were used. Thus, it may be attributed to the particular nature of the dataset.

Employment of the Tanimoto similarity measure gives performance slightly superior to those of all other employed similarity methods (average hit rate of 0.75 over all five classes of active molecules). We can conclude that atom environments capture chemically meaningful information for similarity searching of bioactive molecules and that they also perform well in combination with a commonly used (Tanimoto) similarity coefficient.

Most of the variations in performance with the number of active structures and the number and type of features selected (Table 5 and Figure 6) are intuitively explicable.

Performance increases with the number of active structures because meaningful probabilities for features can now be

calculated. They are less likely to be random probabilities and more likely to be drawn from the underlying distribution of features of all active compounds. In addition, knowledge about different chemotypes becomes available, depending on the exact query structures used for training.

If more features are selected, we see that more active structures have to be available for training to achieve the best performance. To calculate meaningful information gains for a larger and larger number of features, knowledge about relative frequencies of occurrence has to be known for a larger number of features. This is only given if more training structures and thus features are given. If features are selected by relative frequency, performance increases rapidly with the number of available training structures, finally achieving performance of information-gain based feature selection. This is due to a property inherent in the feature selection step: Features have to occur with a certain frequency to possess information content relevant to the classification task and high information gain. Thus, the features possessing the highest information gain are also present near the top of the list of most frequent features. The set of the most frequent features and the set of the most meaningful features overlap.

Information-gain-based feature selection generally outperforms selection of the most frequent features. It is also superior to employing the Bayesian classifier without any feature selection, in particular with a small number of training structures.

Corroborating the results obtained on the large dataset, this clearly demonstrates the importance of feature selection for the algorithms presented.

5. CONCLUSIONS

The similarity searching algorithm based on atom environments, information-gain-based feature selection, and the naive Bayesian classifier shows very strong performance (measured as retrieval rates of bioactive compounds) on the datasets examined here. Due to the size and variety of the datasets used, this may imply that the results—as well as the method itself—are also transferable to other datasets at high performance levels (a thorough proof of which would be a true, prospective screening).

Applied to a recently published dataset using more than 100000 molecules from the MDDR database, the atom environment approach consistently outperforms fusion of ranking scores (in 10 out of 11 cases) as well as binary kernel discrimination in combination with Unity fingerprints (in 10 out of 11 cases if 10 active and 10 inactive structures are used for training and in all cases if 10 active and 100 inactive structures are available for training). Performance is particularly superior in the case of more diverse datasets, such as the 5HT3 set (which contains ligands of all receptor subtypes). Retrieval rates among the top 5% of the sorted library averaged over all active subsets are nearly 10% better (more than 14% better in relative numbers) than those of the second best method, binary kernel discrimination. Better performance is shown to be in good part due to the atom environment descriptor used, which captures information relevant to the similarity examined here (bioactivity, or target affinity). Using single queries in combination with the Tanimoto coefficient, atom environments outperform Unity fingerprints in all but one case.

When used on a monoaminooxidase dataset, the method performs as well as binary kernel discrimination using atom pairs and topological torsion in the case of a 50/50 split of training and test compounds. In the case of sparse training data, binary kernel discrimination is superior. This may be due to the particular dataset, as it partly contradicts the finding on the first large and diverse MDDR dataset. Another possible explanation is that binary kernel discrimination performs better with atom pairs and topological torsion than with Unity fingerprints. Upon varying the number of selected features in the method presented, information-gain-based feature selection is shown to be a crucial step for the performance of the naive Bayesian classifier.

Atom environments, which are adapted to the dataset under consideration, are superior to Unity fingerprints for similarity searching. Combining information prior to scoring and including information about inactive compounds, as is performed by the Bayesian classifier and binary kernel discrimination, is found here to be superior to posterior data fusion.

ACKNOWLEDGMENT

We thank Unilever, The Gates Cambridge Trust, and Tripos Inc. for support. Gavin Harper is thanked for providing us with the binary kernel data and for interesting discussions. Uta Lessel and Jérome Hêrt are thanked for providing us with their MDDR datasets. Yvonne C. Martin is thanked for providing us with the MAO dataset. A number of anonymous reviewers are thanked for valuable input on the manuscript.

REFERENCES AND NOTES

(1) *Concepts and Applications of Molecular Similarity*; Johnson, A. M., Maggiora, G. M., Eds.; Wiley: New York, 1990.

(2) Walters, W. P. Virtual Screening—an Overview. *Drug Discov. Today* **1998**, *3*, 160−178.

(3) Nikolova, N.; Jaworska, J. Approaches to measure chemical similarity—A review. *QSAR Comb. Sci.* **2004**, *22*, 1006−1026.

(4) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighbourhood behaviour: A useful concept for validation of "molecular similarity" descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.

(5) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.

(6) Kubinyi, H. Similarity and dissimilarity: A medicinal chemist's view. *Perspect. Drug Discovery Des.* **1998**, *9*−*11*, 225−252.

(7) Kubinyi, H. Chemical similarity and biological activities. *J. Braz. Chem. Soc.* **2002**, *13*, 717−726.

(8) Downs, G. M.; Willett, P.; Fisanick, W. Similarity searching and clustering of chemical-structure databases using molecular property data. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1094−1102.

(9) Estrada, E.; Uriarte, E. Recent Advances on the Role of Topological Indices in Drug Discovery Research. *Curr. Med. Chem.* **2001**, *8*, 1573−1588.

(10) Cramer, R. D.; Patterson, D. R.; Bunce, J. D. Comparative Molecular Field Analysis (CoMFA). 1. Effect of Shape on Binding of Steroids to Carrier Proteins. *J. Am. Chem. Soc.* **1988**, *110*, 5959−5967.

(11) Mason, J. S.; Good, A. C.; Martin, E. J. 3D-Pharmacophores in Drug Discovery. *Curr. Pharm. Des.* **2001**, *7*, 567−597.

(12) Free, S. M., Jr.; Wilson, J. W. A Mathematical Contribution to Structure-Activity Studies. *J. Med. Chem.* **1964**, *7*, 395−399.

(13) Kubinyi, H. QSAR: Hansch Analysis and Related Approaches. *Methods and Principles in Medicinal Chemistry*; Mannhold, R., et al., Eds.; VCH: Weinheim, Germany, 1993; Vol. 1.

(14) Pastor, M.; Cruciani, G.; McLay, I.; Pickett, S.; Clementi, S. GRid-INdependent descriptors (GRIND): a novel class of alignment-independent three-dimensional molecular descriptors. *J. Med. Chem.* **2000**, *43*, 3233−3243.

(15) Wagener, M.; Sadowski, J.; Gasteiger, J. Autocorrelation of Molecular-Surface Properties for Modeling Corticosteroid-Binding Globulin and Cytosolic Ah Receptor Activity by Neural Networks. *J. Am. Chem. Soc.* **1995**, *117*, 7769−7775.
(16) Kauvar, L. M.; Higgins, D. L.; Villar, H. O.; Sportsman, J. R.; Engquist-Goldstein, A. E.; Bukar, R.; Bauer, K. E.; Dilley, H.; Rocke, D. M. Predicting ligang binding to proteins by affinity fingerprinting. *Chem. Biol.* **1995**, *2*, 107−118.
(17) Briem, H.; Lessel, U. F. In vitro and in silico affinity fingerprints: Finding similarities beyond structural classes. *Perspect. Drug Discovery Des.* **2000**, *20*, 231−244.
(18) Baumann, K. Uniform-length molecular descriptors for quantitative structure-property relationships (QSPR) and quantitative structure-activity relationships (QSAR): classification studies and similarity searching. *Trends Anal. Chem.* **1999**, *18*, 36−46.
(19) Stiefl, N.; Baumann K. Mapping Property Distributions of Molecular Surfaces: Algorithm and Evaluation of a Novel 3D Quantitative Structure-Activity Relationship Technique. *J. Chem. Inf. Comput. Sci.* **2003**, *46*, 1390−1407.
(20) Jaccard, P. Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bull. Soc. Vaudoise Sci. Nat.* **1901**, *37*, 547−579.
(21) Holliday, J. D.; Hu, C.-Y.; Willett, P. Grouping of Coefficients for the Calculation of Inter-Molecular Similarity and Dissimilarity using 2D Fragment Bit-Strings. *Comb. Chem. High Throughput Screening* **2002**, *5*, 155−166.
(22) Holliday, J. D.; Salim, N.; Whittle, M.; Willett, P. Analysis and Display of the Size Dependence of Chemical Similarity Coefficients. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 819−828.
(23) Dixon, S. L.; Koehler, R. T. The hidden component of size in two-dimensional fragment descriptors: Side effects on sampling in bioactive libraries. *J. Med. Chem.* **1999**, *42*, 2887−2900.
(24) Flower, D. R. On the Properties of Bit String-Based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 379−386.
(25) Fligner, M. A.; Verducci, J. S.; Blower, P. E. A modification of the Jaccard-Tanimoto similarity index for diverse selection of chemical compounds using binary strings. *Technometrics* **2002**, *44*, 110−119.
(26) Ginn, C. M. R.; Willett, P.; Bradshaw, J. Combination of molecular similarity measures using data fusion. *Perspect. Drug Discovery Des.* **2000**, *20*, 1−16.
(27) Xing, L.; Glen, R. C. Novel methods for the prediction of logP, pK_a and logD. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 796−805.
(28) Bender, A.; Mussa, H. Y.; Glen, R. C.; Reiling, S. Molecular similarity searching using atom environments, information-based feature selection, and a naive bayesian classifier. *J. Chem. Inf. Comput. Sci.* **2004**, *44*, 170 −178.
(29) Harper, G.; Bradshaw, J.; Gittins, J. C.; Green, D. V. S.; Leach, A. R. Prediction of biological activity for high-throughput screening using binary kernel discrimination. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1295−1300.
(30) Harper, G. The Selection of Compounds for Screening in Pharmaceutical Research. Ph.D. Thesis, University of Oxford,, U.K., 1999.
(31) 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.
(32) MDL Drug Data Report, MDL ISIS/HOST software, MDL Information Systems, Inc., San Leandro, CA.
(33) Brown, R. D.; Martin, Y. C. Use of structure Activity data to compare structure-based clustering methods and descriptors for use in compound selection. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 572−584.
(34) Clark, R. D.; Cramer, R. D.; Van Opdenbosch, N. Validation of the General Purpose Tripos 5.2 Force Field. *J. Comput. Chem.* **1989**, *10*, 982−1012.
(35) Faulon, J. L. Stochastic generator of chemical structure: 1. Application to the structure elucidation of large molecules. *J. Chem. Inf. Comput. Sci.* **1994**, *34*, 1204−1218.
(36) Faulon, J. L.; Visco Jr., D. P.; Pophale, R. S. The Signature Molecular Descriptor. 1. Using Extended Valence Sequences in QSAR and QSPR Studies. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 707−720.
(37) Adamson, G. W.; Cowell, J.; Lynch, M. F.; McLure, A. H. W.; Town, W. G.; Yapp, A. M. Strategic Considerations in the Design of a Screening System for Substructure Searches of Chemical Structure Files. *J. Chem. Doc.* **1973**, *13*, 153−157.
(38) SciTegic, Inc., San Diego, CA.
(39) Quinlan, J. R. Induction of Decision Trees. *Mach. Learn.* **1986**, *1*, 81−106.
(40) Glen, R. C.; A-Razzak, M. Applications of Rule-Induction in the Derivation of quantitative structure-activity relationships. *J. Comput.-Aided Mol. Des.*, **1992**, *6*, 349−383.
(41) *Machine Learning*; Mitchell T. M., Ed.; McGraw-Hill: New York, 1997.
(42) OpenBabel, http://openbabel.sourceforge.net/.
(43) Rarey, M.; Dixon, J. S. Feature Trees: A new molecular similarity measure based on tree matching. *J. Comput.-Aided Mol. Des.* **1998**, *12*, 471−490.
(44) ISIS, Version 2.1.4, Molecular Design Ltd., San Leandro, CA.
(45) Daylight, Version 4.62, Daylight Inc., Mission Viejo, CA.
(46) Sybyl, Version 6.5.3, HQSAR Module, Tripos Inc., St. Louis, MO.
(47) Lessel, U. F.; Briem, H. Flexsim-X: A Method for the Detection of Molecules with Similar Biological Activity. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 246−253.
(48) Lemmen, C.; Lengauer, T.; Klebe, G. FLEXS: A Method for Fast Flexible Ligand Superposition. *J. Med. Chem.* **1998**, *41*, 4502−4520.
(49) Briem, H.; Kuntz, I. D. Molecular similarity based on DOCK-generated fingerprints. *J. Med. Chem.* **1996**, *39*, 3401−3408.
(50) Monks, A., Scudiero, D., Skehan, P., Shoemaker, R., Paull, K., Vistica, D., Hose, C., Langley, J., Cronise, P., Vaigro-Wolff, A., Gray-Goodrich, M., Campbell, H., Mayo, J.; Boyd, M. Feasibility of a high-flux anticancer drug screen using a diverse panel of cultured human tumor cell lines. *J. Nat. Cancer Inst.* **1991**, *83*, 757−766.

CI0498719