Molecular Similarity Searching Using Atom Environments, Information-Based Feature Selection, and a Naïve Bayesian Classifier

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

Aventis Pharmaceuticals, 1041 Route 202−206, Bridgewater, New Jersey 08807

Received September 15, 2003

A novel technique for similarity searching is introduced. Molecules are represented by atom environments, which are fed into an information-gain-based feature selection. A naïve Bayesian classifier is then employed for compound classification. The new method is tested by its ability to retrieve five sets of active molecules seeded in the MDL Drug Data Report (MDDR). In comparison experiments, the algorithm outperforms all current retrieval methods assessed here using two- and three-dimensional descriptors and offers insight into the significance of structural components for binding.

1. INTRODUCTION

Database searching for compounds that exhibit similarity[1] to a given structure has become popular over the past decade.[2] Molecules that are similar to lead compounds or known drugs may have the desired pharmacological effects at the target biological receptor. They may also possess improved absorption, distribution, metabolism, and excretion (ADME) properties and lower toxicity, just to name a few examples.

Similarity searches are based on the "similar property principle" which states that structurally similar molecules are more likely to have similar properties.[1,3,4] This is one of the major assumptions generally followed in lead optimization. In a simplistic way, similar molecules are commonly molecules with functional groups in a similar spatial arrangement.

Chemical similarity methods generally consist of three steps. The first step is to define a representation of the molecule. Representations of molecules can be as follows:

- one-dimensional,[5] such as volume and log *P*;
- two-dimensional, such as connectivity tables and derived graph indices;[6] or
- three-dimensional, such as CoMFA fields[7] and three-point pharmacophores.[8]

The second step is feature selection or weighting to detect those features that are useful in computing this particular aspect of similarity (e.g. solubility, biological activity at a receptor, etc.). There are several methods for feature selection, e.g. genetic selection[9] and information-gain-based methods.[10] For different "similarities", different features emerge as being important.

The third step is comparison of the selected features of two or more molecules and assignment of a similarity index by calculating a similarity coefficient.[11] Similarity coefficients can be classified into associative coefficients and distance coefficients, examples of which are the commonly used Tanimoto coefficient and the Euclidean distance, respectively. The performance of similarity coefficients has been subject to a number of reviews.[12,13] The similarity index is often seen as giving sensible results when it is statistically correlated with the property under investigation, which can be a physicochemical property (e.g. log *P*) or a biochemical property (e.g. receptor binding affinity).

The method presented in this work belongs to the group of two-dimensional molecular representations. 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.

Section 2 describes the new algorithm. 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. We give our concluding remarks in section 5.

2. MATERIALS AND METHODS

(a) Descriptor Generation/Molecular Representation. We use atom environments[14] as a molecular representation. Atom environments are similar to signature molecular descriptors.[15−18] They are translationally and rotationally invariant. Furthermore they do not depend on a particular conformation as they are calculated from the connectivity table. This makes generating atom environments less difficult compared to alignment-dependent approaches. Another benefit with atom environments is that they are easily interpretable as they resemble the chemical concept of functional groups.

We calculated atom environments in a two-step procedure (see Figure 1):

(1) Sybyl atom types[19] are employed for the derivation of the environments. These are force-field atom types, which

* Corresponding author phone: +44 (1223) 763 073; fax: +44 (1223) 763 076; e-mail: ab454@cam.ac.uk.

10.1021/ci034207y CCC: $27.50
Published on Web 12/24/2003

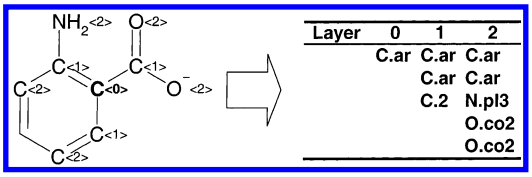


Figure 1. Illustration of descriptor generation step, applied to an aromatic carbon atom. The distances ("layers") from the central atom are given in brackets. In the first step, Sybyl mol2 atom types are assigned to all atoms in the molecule. In the second step, count vectors from the central atom (here C⟨0⟩) up to a given distance (here 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.

implicitly include molecular properties such as geometry. An individual atom fingerprint is calculated for every atom in the molecule. This calculation is performed using distances from 0 up to *n* bonds and keeping count of the occurrences of the atom types. The maximum distance *n* for descriptor generation has been varied from 1 to 3 for parameter optimization; details are given in section 2e.

(2) A count vector is constructed with the vector elements being counts of atom types at a given distance from the central atom. Every atom is described by exactly one count vector resulting in molecular atom environment fingerprints in which the number of atoms in a given molecule equals the number of count vector entries in the fingerprint.

(b) Feature Selection. The information content of individual atom environments was computed using the information gain measure of Quinlan.[10,20] For a particular descriptor, higher information gain is related to better separation between active and inactive structures, for example.

The information gain, I, can be given by

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

where

$$S = -\sum p \log_2 p$$

S is the information entropy; S_v is the information entropy in data subset v; $|D|$ is the total number of data sets; $|D_v|$ is the number of data sets in subset v; and p is the probability that a randomly selected molecule of the whole data set (or subset in the case of D_v) belongs to each of the defined classes.

In each run the number of selected features was varied between 10 and 100.

(c) Classification. A naïve Bayesian classifier[21] was employed as a classification tool. The naïve Bayesian classifier provides a simple yet surprisingly accurate machine-learning tool.[21] Trained with a given data set which consists of known feature vectors (**F**) 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

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

where

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

In the naïve Bayesian classifier we assume that

$$P(\mathbf{F}|\mathrm{CL}_v) = \prod_i P(f_i|\mathrm{CL}_v)$$

where f_i are the feature vector elements. Hence, for CL_v, (1) becomes

$$P(\mathrm{CL}_v|\mathbf{F}) = \frac{P(\mathrm{CL}_v)\prod_i P(f_i|\mathrm{CL}_v)}{P(\mathbf{F})}$$

In this work the data are classified into two classes (active and inactive, here referred to as 1 and 2, respectively). Therefore

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

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

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

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

We use this equation to do the 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).

Note that the actual probability $P(\mathrm{CL}_1|\mathbf{F})$ can be easily computed from $\ln((P(\mathrm{CL}_1|\mathbf{F})/P(\mathrm{CL}_2|\mathbf{F}))$ based on the fact that $P(\mathrm{CL}_1|\mathbf{F}) + P(\mathrm{CL}_2|\mathbf{F}) = 1$.

(d) Compilation of Data Set and Preprocessing. For evaluation of the algorithm, 957 ligands extracted from the MDDR database[22] were used.[23] The set contains 49 5HT3 receptor antagonists (from now on referred to as 5HT3), 40 angiotensin converting enzyme inhibitors (ACE), 111 3-hydroxy-3-methylglutaryl-coenzyme A reductase inhibitors (HMG), 134 platelet activating factor antagonists (PAF), and 49 thromboxane A2 antagonists (TXA2). An additional 574 compounds were selected randomly and did not belong to any of these activity classes.

Structures were downloaded in SDF format and converted to Sybyl mol2 format using OpenBabel[24] 1.100.1 with the −d option to delete hydrogen atoms and default mol2 atom typing. Atom environment fingerprints were then calculated directly from mol2 files.

Table 1. Enrichment Factor Averaged over All Five Classes of Active Compounds upon Varying the Number of Selected Features and the Maximum Depth, *n*, Used To Create the Atom Environment Descriptor[a]

no. of selected features	enrichment					
	top 20, $n = 1$	top 50, $n = 1$	top 20, $n = 2$	top 50, $n = 2$	top 20, $n = 3$	top 50, $n = 3$
10	0.00	0.00	9.56	3.82	3.50	1.40
20	0.00	0.00	7.80	3.12	3.50	1.40
30	0.00	0.00	9.95	3.98	3.50	1.40
40	0.00	0.00	11.06	4.42	3.50	1.40
50	0.00	0.00	10.42	4.17	2.92	1.17
70	0.00	0.00	9.45	3.78	3.11	1.24
100	0.00	0.00	6.52	2.61	0.19	0.08
10, $m > 2$, $m <$ max-2	0.00	0.00	7.33	2.93	3.50	1.40
20, $m > 2$, $m <$ max-2	0.00	0.00	8.58	3.43	3.50	1.40
30, $m > 2$, $m <$ max-2	0.00	0.00	9.20	3.68	3.50	1.40
40, $m > 2$, $m <$ max-2	0.00	0.00	10.28	4.11	3.50	1.40
50, $m > 2$, $m <$ max-2	0.00	0.00	10.13	4.05	3.50	1.40
70, $m > 2$, $m <$ max-2	0.07	0.03	8.99	3.59	3.50	1.40
100, $m > 2$, $m <$ max-2	0.07	0.03	4.89	1.96	0.19	0.08

[a] A fixed number of features were selected, and rare and frequent fragments were excluded. This is denoted by $m > 2$, $m <$ max-2, meaning that features had to occur at least three times, but at most as often as the total number of active molecules (max) minus three times. In situations where very low enrichment factors were obtained, many molecules were assigned identical scores, thus producing artifacts (enrichment factors of 0) in this table.

(e) Calculations. Two separate validations of the method presented here were performed. In the first validation, cross-validation with random selection of query molecules was carried out to optimize the parameters related to descriptor generation and feature selection. A 20-fold cross-validation study selecting randomly five query structures for query generation and calculation of the average enrichment factors of the first 20 and 50 molecules of the sorted library has been performed. The selection of five query structures is a realistic number if few ligands of a given target are known. To illustrate the influence of the number of structures chosen to generate the query on search performance, 20-fold random selection of 3, 5, and 10 structures has been performed, selecting 40 features in the feature selection step. An individual hit rate was calculated for each set of compounds based on the number of molecules within its 10 nearest neighbors, which belong to the same activity class as the query compound. To create a query from multiple molecules, individual probabilities (relative frequencies) of features from a set of molecules are calculated and used in the feature selection step described in section 2b and the naïve Bayesian classifier described in section 2c. The maximum bond distance for generation of molecular descriptors, *n*, was varied from 1 to 3. In each run, the number of selected features was set to 10, 20, 30, 40, 50, 70, and 100, starting with the features associated with the highest information gain. To examine the influence of very frequent and very rare features, this series of experiments has been repeated with a slight modification. Using identical settings for maximum bond distance and number of selected features, only features occurring at least three times, but not in more than max-3 molecules (with max being the number of molecules within the positive data set) were selected. To do so, features were chosen starting with those possessing the highest information gain as above, but skipping rare and frequent features as defined here until the preset number of features was selected. For the best performing feature selection, cumulative recall plots were calculated for all five data sets of active compounds.

In all calculations presented here, the inactive data set containing all structures except those of the active class in each calculation was split in two subsets of equal size to create independent training and test sets. Each similarity calculation was carried out twice, using the active query and each of the two subsets and scoring the remaining active compounds and the inactive compounds not used to generate the model. The average score of the active structures from both runs was calculated. Both subsets of scored inactive structures and the set of active structures with associated average scores were concatenated to give the complete scored list of compounds used for further processing. As an example, for one validation run using a sample of the ACE inhibitor data set, we have drawn the query molecules, selected fragment features, and highest scoring molecules.

In the second validation, for each of the 383 active compounds of the five classes of active compounds its 10 nearest neighbors were calculated on the basis of the similarity measure proposed in sections 2a−c. The maximum distance for descriptor calculation was set to 2 as it produced the best results in the first validation run as well as in additional validations which were performed. Identical values for selection of features as mentioned above have been applied. Exclusion of frequent and rare atom environments was not applied due to the use of single-query molecules. An individual hit rate was calculated for each compound on the basis of the number of molecules within its 10 nearest neighbors, which belong to the same activity class as the query compound. Enrichment is observed when the hit rate among the nearest neighbors is higher than the fraction of the activity class under consideration in the whole data set. Enrichments have been averaged over all classes of active compounds, and the result was compared to that of other methods.

Note that the nearest neighbor protocol of Briem and Lessel[23] has been followed in this validation to make it easy to compare the performance of our algorithm with commonly used methods.

3. RESULTS

For the first validation, the influence of the maximum bond distance for creating the atom environment descriptor and

Table 2. Average Hit Rates among the 10 Nearest Neighbors in a Cross-Validation Study[a]

no. of query structures	5HT3	std dev	ACE	std dev	HMG	std dev	PAF	std dev	TXA2	std dev	mean	mean std dev
1	5.65	4.26	6.40	2.96	7.90	2.75	7.15	2.25	6.40	3.27	6.70	3.10
3	8.55	1.73	6.70	2.64	9.30	0.92	9.15	1.57	8.30	1.13	8.40	1.60
5	9.25	1.02	9.10	0.64	9.50	0.83	9.15	0.82	8.15	1.04	9.03	0.87
10	9.30	1.03	8.80	1.51	9.70	0.57	9.25	0.72	8.95	0.76	9.20	0.92

[a] Shown here are the hit rates and the standard deviations among different data set sizes used to generate the query.

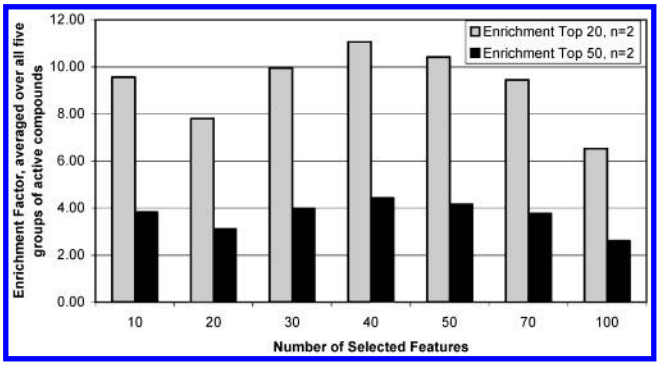


Figure 2. Enrichment factor, averaged over all five groups of active compounds, using atoms up to two bonds apart from the central atom to construct the atom environment descriptor and a variable number of selected features for classification.

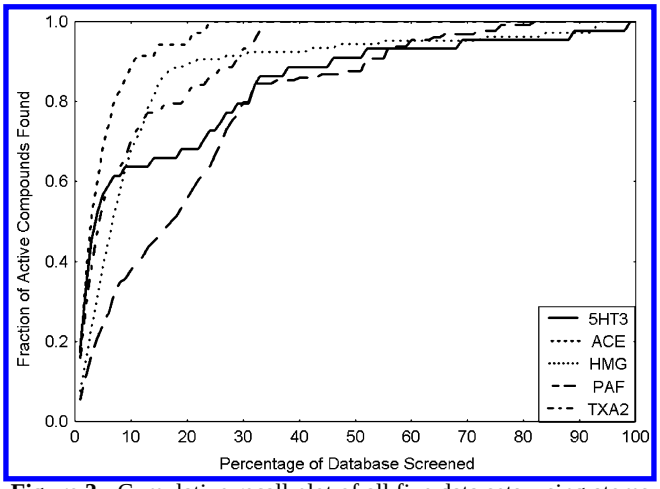


Figure 3. Cumulative recall plot of all five data sets, using atoms up to two bonds apart from the central atom for descriptor generation and 40 features associated with the highest information gain for classification.

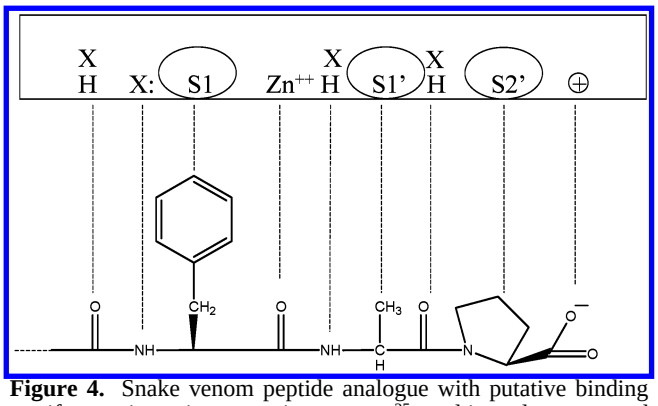


Figure 4. Snake venom peptide analogue with putative binding motif to angiotensin converting enzyme[25] used in early compound design.

the influence of the number of selected features on the average enrichment factor among the first 20 and the first 50 compounds of the ranked database are given in Table 1. Using atoms up to two bonds from the central atom for generating atom environment descriptors ($n = 2$) produces best results with enrichment factors between 11 and 6.5 in the first 20 compounds and between about 4 and 2 in the first 50 compounds. Using three layers for construction of the descriptor still gives enrichment of more than 3 in most cases of feature selection, whereas using only the first layer adjacent to the central atom produces virtually no enrichment, independent of the method used for feature selection.

A visualization of enrichment factors, which depend on the number of selected features, is given in Figure 2. In this case, the bond level for descriptor generation, n, has been set to $n = 2$ because it performed best, as shown in Table 1. Exclusion of frequent and rare features does not perform as well as selection of a fixed number of features, and it is not shown in the figure.

We have found that feature selection has its optimum at a selection of 40 features with respect to enrichment factors observed among the first 20 and among the first 50 highest scoring structures of the sorted library. If fewer or more features are selected, performance of the algorithm continuously decreases.

The influence of the number of structures chosen to generate the query on search performance is shown in Table 2. Results using single structures to generate the active query are presented later and are included here for completeness. In every case except one (going from 5 to 10 query structures using ACE inhibitors), performance improves as the number of compounds used for query generation increases. The average deviation in performance between different sets of query compounds decreases if the size of the query data set is increased. Again, the only exception is if the number of ACE inhibitors used to generate the query is increased from 5 to 10 structures.

For the best performing method using 40 features with the highest information gain, cumulative recall plots are given in Figure 3. These plots were calculated using the 20-fold random selection of five queries for ranking of the library and screening for the remaining active compounds. The five data sets can be classified into two groups: The 5HT3, HMG, and PAF data sets belong to one group as some of their active molecules are found only after evaluating half of the sorted library. ACE and TXA2 belong to the second group with all active molecules found well within the first 40% of the sorted library.

To gain an insight into the algorithm, query molecules, selected features, and the highest scoring structures of the sorted library have been plotted for a sample run using angiotensin converting enzyme inhibitors[25] (ACE inhibitors). The design of ACE inhibitors originally followed the hypothesis that ACE had binding site homology with carboxypeptidase-A. A number of interaction sites were

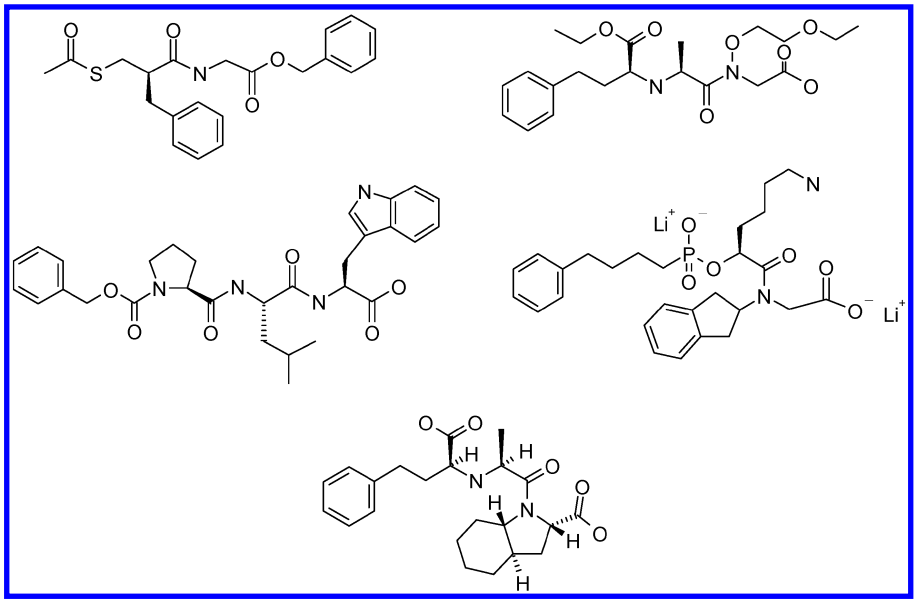


Figure 5. Five active molecules from the data set of ACE inhibitors, used to construct the query and perform feature selection.

proposed on the basis of analogue design, shown in Figure 4.

A recent crystallographic study[26] of an ACE inhibitor, lisinopril (N^2-[(*S*)-1-carboxy-3-phenylpropyl]-L-lysyl-L-proline), has revealed the binding site interactions in some detail. Much of the originally deduced binding site topology is seen in the crystal structure with some notable differences such as the absence of the C-terminal carboxylate arginine interaction. The selection of features associated with a significant information gain in separating the classes of ACE and non-ACE inhibitors can be compared with the crystallographically determined binding motif. It may be expected that those interactions that are seen crystallographically may also emerge from the analysis of the analogues as being important.

The 5 molecules used to construct the query are shown in Figure 5, the 10 selected features giving the highest information gain are given in Table 3 , and the 10 highest ranked structures from the sorted library are shown in Table 4.

The selected features, given in Table 3, possess carbon, nitrogen, and oxygen atoms as central atoms. Some analysis of the selected features with respect to the experimentally determined interaction of ligands within the ACE binding site is given in the Discussion.

In the second validation, the number of active molecules among the 10 nearest neighbors of each individual active molecule from each data set was calculated following the protocol of Briem and Lessel.[23] Feature selection was performed selecting 10, 20, 30, 40, 50, 70, and 100 features. Hit rates, averaged over all five classes of active compounds, are given in Figure 6.

An optimum can be seen at the selection of 20 features. If more or less features are used for classification, performance declines continuously. The individual hit rates for each group, using the best-performing selection of 20 features, are given in Table 5. The average number of active compounds among the top 10 ranked compounds varies from about 5.65 (5HT3) to about 7.90 (HMG), with an overall average of 6.70. These numbers, taking into account the variable number of active structures in each active subset, result in enrichment factors between 5.14 (PAF) and 15.6 (ACE). The overall average enrichment factor calculates to 8.48, which is significantly higher than the value of 1 that would be achieved in a random selection.

The nearest neighbor protocol of Briem and Lessel[23] has been followed in this validation to enable ease of comparison of the algorithm performance with established methods. The methods used for comparison are feature trees,[27] ISIS MOLSKEYS,[28] Daylight fingerprints,[29] SYBYL hologram QSAR fingerprints,[30] and FLEXSIM-X,[31] FLEXSIM-S,[32] and DOCKSIM[33] virtual affinity fingerprints. Feature trees represent molecules as trees (acyclic graphs), which are subsequently matched for comparison. In current versions, FlexX interaction profile and van der Waals radii have been 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 bits long bit 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 compared to established methods is shown in Figure 7.

Shown here are mean sample hit rates as averaged over all five classes of active compounds. Using one query structure, this method 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; however, the atom environments approach additionally deduces those fragments having the greatest influence on similarity and is significantly faster than feature trees and therefore of utility in searching larger databases.

Using five query structures, the atom environment approach achieves a mean sample hit rate of greater than 90%.

Table 3. Set of 10 Features Associated with the Highest Information Gains from a Sample Run Using 5 Inhibitors from the ACE Data Set

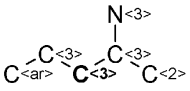
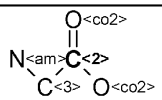
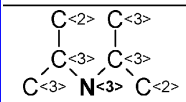
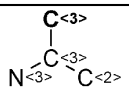
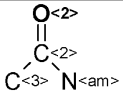
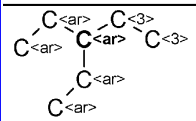
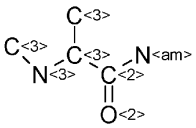
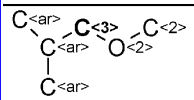
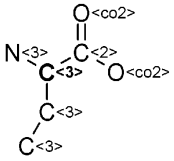
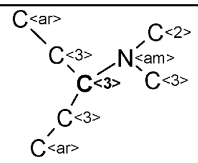
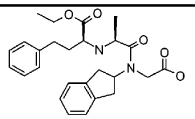
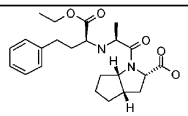
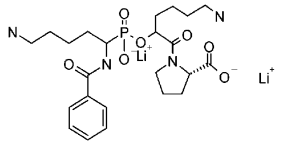
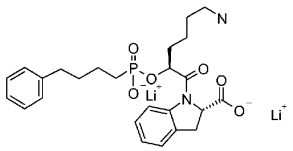
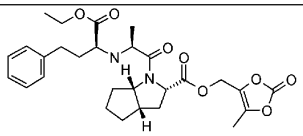
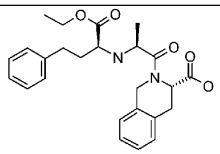
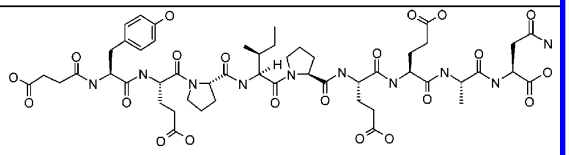
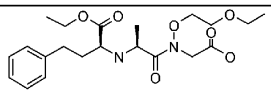
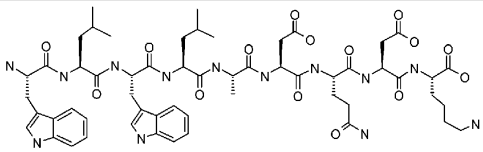
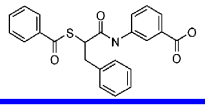
Selected Feature	Information Gain Associated with this Feature	Putative Interaction Site on ACE
	0.017	S1
	0.0141	+
	0.0127	XH/S1′
	0.0118	S1
	0.0114	XH/Zn^{++}
	0.0104	S1
	0.0096	S1′
	0.0086	S1
	0.0083	S2′/+
	0.0083	S2′

Table 4. Top 10 Ranking Molecules of the Sorted Library[a]

Rank number /Activity	Structure
1 / Active	
2 / Active	
3 / Inactive	
4 / Active	
5 / Active	
6 / Active	
7 / Inactive	
8 / Active	
9 / Inactive	
10 / Active	

[a] Out of these, seven are active ACE inhibitors and three are inactive molecules in this respect.

The computation of molecular fingerprints was implemented in C programming language and was able to process about 1000 molecules/s on a Pentium III, 1 GHz workstation. Feature selection and scoring was implemented in Perl and was able to evaluate one molecule against the 956 remaining compounds of the data set in 1 s, using identical hardware.

4. DISCUSSION

The first series of runs was performed to optimize parameters of the algorithm for typical database screenings where several active compounds are known. As Table 1 shows, the algorithm only gives sensible results when the atom environment descriptor is constructed using atoms up to two bonds apart from the central atom. If less than two bonds are considered, atom environments are ambiguous and do not capture enough information about the atom environment. If more than two bonds are considered, they tend to become unique so no generalization capability is acquired. This result is in agreement with the results found by Faulon et al.[17,18] Optimum performance is found with the selection of 40 features. This is the result for queries derived from five query structures and applies across the five different sets of active molecules used. Fewer features do not allow the classification of each molecule reliably (by recognizing

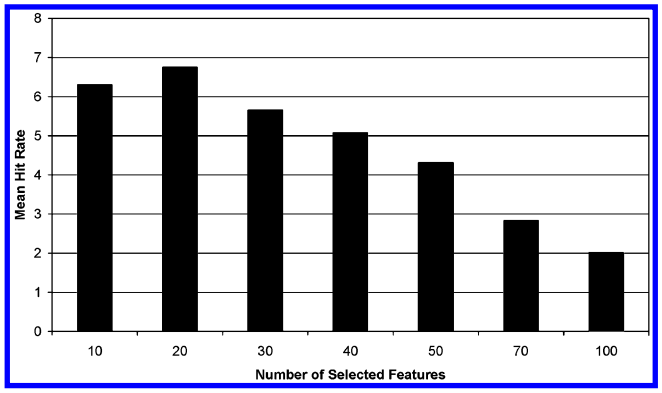


Figure 6. Mean hit rates among the 10 nearest neighbors of the Briem data set, averaged over all five classes of active compounds, depending on the number of selected features.

Table 5. Performance of the Atom Environment Approach by Measuring Mean Sample Hit Rates of the 10 Top-Scored Compounds in the Sorted Hit List[a]

group of active compds	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	5.65	6.40	7.90	7.15	6.40	6.70
enrichment factor	11.0	15.6	6.87	5.14	12.8	8.48

[a] Feature selection was performed selecting 20 features associated with the highest information gain.

a certain number of its atom environments), and more features appear to introduce noise into the system thus reducing its classification ability.

The performance of the algorithm generally increases if more and more structures are used to generate the query (Table 2), as well as the standard deviation in performance between different sets of query structures decreases. Using five query structures, atom environments outperform other methods by a large margin (Figure 7), giving mean sample hit rates of about 90%. These hit rates are not directly comparable, because information from multiple structures is used to formulate the query. Nonetheless, it shows that the algorithm is capable of handling information from multiple molecules reliably. For real-world applications, it appears that all active molecules across the range of structural diversity could be used in order to train the classifier used in this method.

The five data sets used can be classified into two groups. In one group, comprising the 5HT3, HMG, and PAF data sets, hits are still found among lower ranked molecules (Figure 3). Apparently, there are molecules in this group of data sets which do not possess close analogues in the training and the test sets. In the other group, comprising ACE and TXA2, all active molecules are easily found in the first half of the focused library. The molecules in these classes of active compounds seem to be more similar to each other.

Overall the selection of fragments of ACE inhibitors seems consistent with the binding information deduced crystallographically.[26] The five fragments associated with the highest information gain given in Table 3 correspond to the binding motif of enalapril and captopril including the zinc binding site and the + and S1′ sites in the top rank. Among the 10 highest scoring molecules of the sorted library listed in Table 4, seven are known active ACE inhibitors while three are not tested with respect to ACE inhibitor activity. The inactives (which of course, may be active—the data on these molecules in MDDR do not include ACE assay results) are peptidic, larger than small molecule analogues, and contain many peptidic environments common to the natural substrates. Elimination of such peptidic moieties would give (in this case) an ideal result. A penalty factor for molecules larger than the probe molecules (a scoring relative to size) could be used.

When calculating the hit rate of the 10 nearest neighbors of each individual active molecule (i.e. using one molecule to retrieve its neighbors from the remaining database), an optimum in classification is obtained if 20 features are selected (Figure 6). This is a different result from that observed in those runs where five molecules are used to

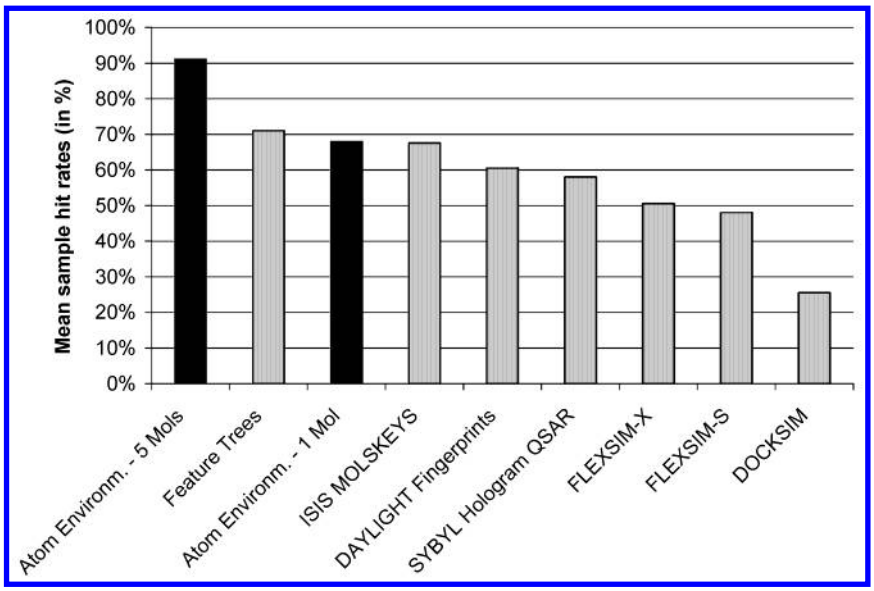


Figure 7. Mean sample hit rates of the atom environment approach (black), in comparison to the methods applied by Briem (light gray). The performance of the atom environment approach is shown using single queries and randomly selected subsets of five query molecules.

derive the query. In those cases, the optimum feature number was seen to increase to 40 features. In single molecule queries, one active molecule containing only a small number of atom environments is used. An increase in the number of features thus exceeds the number of environments present in many molecules. Therefore, there is no gain in including additional features.

As described in Table 5, enrichment factors have been found to be between 5.14 (PAF) and 15.6 (ACE). The overall average enrichment factor is 8.48, showing general validity of the approach.

The method presented here and all top-performing algorithms it is compared to are two-dimensional approaches. Two-dimensional similarity searching algorithms often lead to surprisingly good results. However, one has to be careful that this is not simply due to the libraries used which often contain analogue structures. Analogue design is commonly successful in finding new active molecules, and analogue molecules often contain identical substructures. Two-dimensional algorithms, which are based on connectivity tables, easily detect these identical substructures. This is a general problem of compiling databases for evaluating database retrieval performance and affects all of the algorithms employed in this work.

Using single queries, feature trees, atom environments, and ISIS MOLSKEYS perform considerably better than Daylight fingerprints and hologram QSAR on the test sets. The latter group has in common that it includes information in addition to local subgraph features, whereas the former group only uses local information. This is the case because feature trees are commonly repeatedly cut before matching, ISIS MOLKEYS use predefined fragments, and atom environments only consider an atom and its neighbors at a maximum of two bonds apart. Restricting molecular representation to local information might therefore be a useful feature.

In addition, ISIS MOLSKEYS and atom environments employ feature selection. ISIS MOLSKEYS considers only fragments occurring in a library, whereas, in the case of atom environments, fragments are explicitly selected. Daylight fingerprints, on the other hand, consider every atom path in a certain distance range and then fold the information to give uniform length descriptors. The lack of feature selection or the hashing and folding process seems to worsen the performance of this type of descriptor.

All three virtual affinity fingerprint methods perform worse than any of the two-dimensional methods when applied to the test data sets. Virtual affinity methods consider the three-dimensional structure of the ligand and also take the structure of the receptor into account. Probably because of currently used strategies of library design, as mentioned above, the performance of three-dimensional virtual affinity fingerprint methods is generally seen to be lower than the performance of two-dimensional methods. Nonetheless, it is reported that three-dimensional similarity measures are able to detect similarities which two-dimensional methods are unable to pick up.[23] This would be true in particular in the case of conformationally labile molecules which can achieve pharmacophoric patterns that are important for activity or stereochemically important combinations which are not encoded in the 2D representation.

Ref 23 gives more details about variations in performance among virtual affinity fingerprint based techniques.

5. CONCLUSIONS

In this paper we introduced the combination of atom environments, information-gain-based feature selection, and a naïve Bayesian classifier to describe the similarity of molecules. On average, our algorithm achieved an enrichment factor of about 8 when calculating the 10 nearest neighbors of five data sets containing active structures. In addition to this encouraging result, the algorithm was compared to several two- and three-dimensional methods. Using single queries, it performs as well as the best commonly used 2D algorithms while outperforming all 3D methods. Using multiple queries, close-to-ideal hit rates are obtained. The technique described in this paper can also be useful in identifying key functional groups in active molecules and is computationally efficient. There is ongoing research in substituting the Sybyl atom types used by other descriptors (e.g., to include stereochemistry), employing fuzzy matching, and other machine learning techniques that are expected to further improve the performance.

ACKNOWLEDGMENT

We thank Unilever, The Gates Cambridge Trust, and Tripos Inc. for support.

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 Discovery Today* **1998**, *3*, 160−178.
(3) 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.
(4) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do Structurally Similar Molecules Have Similar Biological Activity? *J. Med. Chem.* **2002**, *45*, 4350−4358.
(5) 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*, 4, 1094−1102.
(6) Estrada, E.; Uriarte, E. Recent Advances on the Role of Topological Indices in Drug Discovery Research. *Curr. Med. Chem.* **2001**, *8*, 1573−1588.
(7) 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.
(8) Mason, J. S.; Good, A. C.; Martin, E. J. 3D-Pharmacophores in Drug Discovery. *Curr. Pharm. Des.* **2001**, *7*, 567−597.
(9) Wegner, J. K.; Zell, A. Prediction of Aqueous Solubility and Partition Coefficient Optimized by a Genetic Algorithm Based Descriptor Selection Method. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 1077−1084.
(10) Quinlan, J. R. Induction of Decision Trees. *Machine Learning* **1986**, *1*, 81−106.
(11) *An Introduction to Chemoinformatics*; Leach, A. R., Gillet, V. J., Eds.; Kluwer: Dordrecht, The Netherlands, 2003.
(12) Willett, P.; Winterman, V. A. Comparison of Some Measures for the Determination of Intermolecular Structural Similarity Measures of Intermolecular Structural Similarity. *Quant. Struct.-Act. Relat.* **1986**, *5*, 18−25.
(13) 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.
(14) Xing, L.; Glen, R. C. Novel Methods for the Prediction of log *P*, pK_a, and log *D*. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 796−805.
(15) 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.
(16) Visco, D. P., Jr.; Pophale R. S.; Rintoul M. D.; Faulon, J. L. Developing a Methodology for an Inverse Quantitative Structure−Activity Relationship Using the Signature Molecular Descriptor. *J. Mol. Graphics Modell.* **2002**, *20*, 429−438.

(17) Faulon, J. L.; Visco, D. P., Jr.; 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.
(18) Faulon, J. L.; Churchwell, C. J.; Visco, D. P., Jr. The Signature Molecular Descriptor. 2. Enumerating Molecules from Their Extended Valence Sequences. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 721−734.
(19) 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.
(20) 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.
(21) *Machine Learning*; Mitchell, T. M., Ed.; McGraw-Hill: New York, 1997.
(22) MDL Drug Data Report; MDL ISIS/HOST software, MDL Information Systems, Inc.: San Leandro, CA.
(23) 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.
(24) OpenBabel, http://openbabel.sourceforge.net/.
(25) Cushman, D. W.; Cheung, H. S.; Sabo, E. F.; Ondetti, M. A. Design of Potent Competitive Inhibitors of Angiotensin-Converting Enzyme. Carboxyalkanoyl and Mercaptoalkanoyl Amino Acids. *Biochemistry* **1977**, *16*, 5484−5491.
(26) Natesh, R.; Schwager, S. L. U.; Sturrock, E. D.; Acharya, K. R. Crystal Structure of the Human Angiotensin-Converting Enzyme-Lisinopril Complex. *Nature* **2003**, *421*, 551−554.
(27) 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.
(28) *ISIS*, Version 2.1.4; Molecular Design Ltd.: San Leandro, CA, June 1, 1998.
(29) *Daylight*, Version 4.62; DAYLIGHT Inc.: Mission Viejo, CA, March 5, 1999.
(30) *Sybyl*, Version 6.5.3, HQSAR Module; Tripos Inc.: St. Louis, MO, June 1999.
(31) 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.
(32) Lemmen, C.; Lengauer, T.; Klebe, G. FLEXS: A Method for Fast Flexible Ligand Superposition. *J. Med. Chem.* **1998**, *41*, 4502−4520.
(33) Briem, H.; Kuntz, I. D. Molecular Similarity Based on DOCK-Generated Fingerprints. *J. Med. Chem.* **1996**, *39*, 3401−3408.

CI034207Y