Comparison of Fingerprint-Based Methods for Virtual Screening Using Multiple Bioactive Reference Structures

Jérôme Hert, Peter Willett,* and David J. Wilton

Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, UK

Pierre Acklin, Kamal Azzaoui, Edgar Jacoby, and Ansgar Schuffenhauer

Novartis Institutes for BioMedical Research, Discovery Technology Centre, Compound Logistics and Properties Unit, CH-4002 Basel, Switzerland

Received October 24, 2003

Fingerprint-based similarity searching is widely used for virtual screening when only a single bioactive reference structure is available. This paper reviews three distinct ways of carrying out such searches when multiple bioactive reference structures are available: merging the individual fingerprints into a single combined fingerprint; applying data fusion to the similarity rankings resulting from individual similarity searches; and approximations to substructural analysis. Extended searches on the *MDL Drug Data Report* database suggest that fusing similarity scores is the most effective general approach, with the best individual results coming from the binary kernel discrimination technique.

INTRODUCTION

Accessing databases of chemical structures is one of the most important facilities in modern systems for chemical information management. Early systems focused on the provision of facilities for substructure searching, but these were complemented from the late 1980s by similarity searching.[1,2] Here, rather than retrieving those molecules that contain the query substructure (either in 2D or, latterly, in 3D), the aim is to identify those database molecules that are most similar to a user-defined *reference structure*, using some quantitative definition of intermolecular structural similarity.[3,4] The reference structure is characterized by one or more structural descriptors, and this set is compared with the corresponding sets of descriptors for each of the molecules in the database. These comparisons enable the calculation of a measure of similarity between the reference structure and each of the database structures, and the latter are then sorted into order of decreasing similarity with the target. The output from the search is a ranked list in which the structures that are calculated to be most similar to the reference structure, the *nearest neighbors*, are located at the top of the list. These neighbors form the initial output of the search and will be those that have the greatest probability of being of interest to the user, given an appropriate measure of intermolecular structural similarity.

At the heart of any similarity searching system is the measure that is used to quantify the degree of resemblance between two molecules. There are many such measures, but the first, and still the most widely used, is based on the numbers of 2D fragment substructures common or not-common to a pair of molecules, these numbers being used to calculate a similarity coefficient such as the Tanimoto Coefficient.[3] This approach provides an efficient and an effective basis for quantifying the degree of structural resemblance between two molecules, and such fragment-based measures of chemical similarity are now available in most operational systems for chemical information management. They continue to be the focus of significant research and development,[5−12] despite the availability of more sophisticated measures of similarity that are based on, e.g., 3D fragments, chemical graphs, and text.[13−15] Much of the current interest in similarity searching arises from its widespread use for virtual screening, where the reference structure is a known bioactive molecule, such as a competitor's compound or a weak lead from an HTS experiment. The relationship that is known to exist between structural similarity and biological similarity[16−18] means that the nearest neighbors are priority candidates for biological testing to see whether they also exhibit the bioactivity of interest.

Most of the approaches that have been reported to date for similarity searching involve the use of individual reference structures. Following Schuffenhauer et al.[8] and Xue et al.,[19,20] we consider here searching techniques that can be used when not one but several different bioactive reference structures are available. The normal screening approach when several actives have been identified is pharmacophore mapping followed by 3D database searching.[21] This approach assumes that the active molecules have a common mode of action and that features that are common to all of the molecules describe the pharmacophoric pattern responsible for the observed bioactivity. This is a powerful technique but one that may not be applicable to the structurally heterogeneous hits that characterize typical HTS experiments; in such cases, it is appropriate to consider approaches based on similarity searching: in this paper, we consider multiple reference searching based on 2D fragment bit-strings and the Tanimoto coefficient. Specifically, we describe three very different approaches to the combination of the structural

* Corresponding author e-mail: p.willett@sheffield.ac.uk.

10.1021/ci034231b CCC: $27.50 © 2004 American Chemical Society
Published on Web 02/20/2004

information that can be gleaned from multiple reference structures and then evaluate the effectiveness of these approaches by simulated virtual screening experiments.

METHODS

We have investigated three distinct ways of combining the structure−activity information implicit in a set of bioactive reference structures: in what follows, we refer to these approaches as single fingerprint, data fusion, and substructural analysis. In brief, the first of these involves creating a single, combined fingerprint from the fingerprints of the individual reference structures; the second involves searching each reference fingerprint separately and then combining the search outputs; and the last involves a predictive ranking rule that is developed from the substructures present in the reference structures. All of these approaches have been studied previously, although not necessarily in the forms used here that are appropriate for multiple reference searching. The detailed implementation of these approaches is discussed below.

Single Fingerprint Methods. The single fingerprint approach was first described by Shemetulskis et al. in their work on Stigmata.[22] The method generates a *modal fingerprint* from an input training set of molecules that seeks to capture the common chemical features present in the members of this training set. A bit j is set to "on" in the modal fingerprint if that bit is found in more than a user-defined threshold percentage of the training set molecules. The modal fingerprint is then used as a query and compared to the fingerprints of the compounds in the database. Shemetulskis et al. used two metrics to rank the compounds of the database: the modal percent and the Tanimoto coefficient, and we have used the latter approach here. If a is the number of bits set in the modal fingerprint, b is the number of bits set in the fingerprint of the molecule to be evaluated, and c is the number of bits set in both fingerprints, then the Tanimoto coefficient is defined to be

$$\frac{c}{a+b-c} \tag{1}$$

In what follows, we refer to searches carried out in this way as *modal* searches, with a range of thresholds tested to generate the optimal modal fingerprint for searching. Our second, *weighted*, approach does not require the use of such a threshold value. Instead, the algorithm computes a weighted fingerprint from the set of actives, where the weight of the jth bit is the number of actives with that bit set (this weighted approach was first described by Singh et al.[14] in their work on joint chemical probes). The database molecules are then ranked using the continuous version of the Tanimoto coefficient for continuous variables. If x_{wi} represents the ith descriptor of the weighted fingerprint and x_{ci} is the ith descriptor of the compound to be evaluated, then the coefficient is defined to be

$$\frac{\sum x_{wi}x_{ci}}{\sum x_{wi}^{\,2} + \sum x_{ci}^{\,2} - \sum x_{wi}x_{ci}} \tag{2}$$

Data Fusion Methods. Data fusion is the name given to a range of techniques that combine inputs from different sensors, with the expectation that using multiple information sources enables more effective decisions to be made than if just a single sensor is employed.[23] The approach has been used in many different fields; when applied to chemoinformatics applications (where it is sometimes referred to as consensus scoring) the fusion is effected by combining the results of several database searches using different descriptors or scoring functions.[24−26]

Our earlier work involved the fusion of ranks.[24] Assume that a specific database molecule appears at rank position r_i when the ith similarity measure ($1 \le i \le n$) is used to match the reference structure against each of the database structures. Then, given a total of n individual rankings of the database, the final score for that molecule is given by a fusion rule such as the maximum or the sum of the ranks, r_i. For example, Ginn et al. fused rankings generated by similarity measures based on 2D fingerprints, physicochemical properties and 3D electrostatic potential grids.[24] These different similarity measures result in radically different types of similarity score that cannot readily be combined without the introduction of bias. The use of ranks, rather than the scores underlying those rank positions, provides a (rather drastic) way of standardizing the data so that all of the similarity searches are entirely comparable, albeit at the cost of a considerable loss of information. Similar comments apply when searches using different similarity coefficients, rather than different structure representations, are fused.[27]

Our application of data fusion to similarity searching in the present context is rather different. Rather than having a single reference structure that is searched against a database in several different ways, we have several different reference structures that are all searched against a database in exactly the same way (specifically using 2D fingerprints with the Tanimoto coefficient). As the actual similarity measure used is the same there are not the problems of bias noted above, and we have hence considered the fusion not just of the ranks but also of the actual scores, s_i. Specifically, we have used the fusion rules MAX, for the maximum of the similarity scores, s_i (or the minimum of the rank positions in the case of rank-based fusion),

$$\text{Maximum } \{s_1, s_2..s_i..s_{n-1}, s_n\} \tag{3}$$

and SUM, for the sum of the similarity scores (or rank positions)

$$\sum_{i=1}^{n} s_i \tag{4}$$

Substructural Analysis Methods. Substructural analysis was first described by Cramer et al.[28] and seeks to assign a weight to each bit (or substructure) in a fingerprint that describes that bit's differential occurrence in the active and inactive molecules constituting a training set for which biological testing has already taken place. The resulting weights are then used to rank the test database, with molecules at the top of this ranking being candidates for testing. Many different substructural analysis weights have been described in the literature,[29] but they are all based on some or all of the following pieces of information about the training set: A_j and I_j are the number of active and inactive compounds with bit j set; T_j is the total number of compounds

with bit j set (so $T_j = A_j + I_j$); and N_A and N_I are the total number of active and inactive molecules with N_T being the total number of molecules (so $N_T = N_A + N_I$). In the present context we do not have access to all of the necessary information as the training set consists of just active molecules. However, if we restrict our attention to those weighting schemes that do not make explicit use of information about the inactives and also make the assumption that the overall characteristics of the training set are mirrored by those of the entire database that is to be searched, then we can use the R1 weight.[29] This has the form

$$\log\left(\frac{A_j/N_A}{T_j/N_T}\right) \tag{5}$$

In (5), T_j is the total number of molecules in the database with bit j set and N_T is the total number of molecules in the database (rather than the total numbers in the training set, as in conventional substructural analysis). In fact, the term N_T/N_A makes a constant contribution to all of the weights, so that R1 is effectively

$$\log\left(\frac{A_j}{T_j}\right) \tag{6}$$

This is an intuitively reasonable weighting scheme since a large positive weight, i.e., a strongly discriminating bit-position, will result when a particular bit is set in many of the active reference structures but in few of the database structures (in a manner very similar to some of the term weighting schemes that underlie modern text search engines[30]).

A recent comparison of several virtual screening approaches highlighted the general effectiveness of the binary kernel discrimination (BKD) method.[31] This approach, first applied to chemical applications by Harper et al.,[32] is again based on a training set comprising actives and inactives but uses the bit-occurrence data in a far more sophisticated, and complex, way than the simple weighting schemes described above. BKD uses a kernel function that has been developed for handling binary data; specifically, given two molecules i and j represented by fingerprints containing N bits and differing in d_{ij} of those bits then Harper et al. define the following kernel function $K_\lambda(i,j)$

$$K_\lambda(i,j) = (\lambda^{N-dij}(1-\lambda)^{dij})^{k/N} \tag{7}$$

In (7), λ is a smoothing parameter, the value of which must be defined or optimized, and k is a constant. The fingerprint representing a database molecule, j, is matched against the fingerprints for each of the active and inactive molecules in the training set and its score then computed as

$$L_\lambda(i,j) = \frac{\sum_{i\in\, actives} K_\lambda(i,j)}{\sum_{i\in\, inactives} K_\lambda(i,j)} \tag{8}$$

The use of BKD requires calculating these scores using a range of values of the smoothing parameter, λ, and then choosing that value which gives the best results in terms of ranking the actives toward the top of the training set.

Table 1. Activity Classes Used in the Study

activity name	similarity mean	similarity SD
5HT3 antagonists	0.351	0.116
5HT1A agonists	0.343	0.104
5HT reuptake inhibitor	0.345	0.122
D2 antagonists	0.345	0.103
renin inhibitor	0.573	0.106
angiotensin II AT1 antagonists	0.403	0.101
thrombin inhibitor	0.419	0.127
substance P antagonist	0.399	0.106
HIV protease inhibitor	0.446	0.122
cyclooxygenase inhibitor	0.268	0.093
protein kinase C inhibitor	0.323	0.142

The success of the BKD approach[31,32] has led us to consider how it might be used given just a set of active reference structures, rather than a full training set. The approach we have taken is to make the assumption that the characteristics of the inactives are approximated with a high degree of accuracy by the characteristics of the entire database that is to be searched. If this assumption is accepted, then a training set can be generated by taking the set of reference structures and adding to it molecules randomly selected from the database (subject, in our experiments, to the qualification that none of the resulting pairs of molecules had a similarity coefficient greater than 0.80 using Unity fingerprints and the Tanimoto coefficient), with the expectation that most, if not all, of these added molecules are inactive. This expectation is not unreasonable given that actives are inherently very rare.

It is possible to regard BKD as a substructural analysis method, since it is based on a training set of known active and known inactive molecules; however, we prefer to regard it as a score-based fusion method, since it is based on weighted sums of the similarities to these two sets of training set molecules. We have, however, discussed it in this section since its application to multiple reference searching requires an approximation based on the whole data set, in a manner similar to our approximation of the R1 weight for substructural analysis.

Database Searches. We have evaluated the various approaches above by means of simulated virtual screening searches on the *MDL Drug Data Report* (MDDR) database.[33] After removal of duplicates and molecules that could not be processed using local software, a total of 102 535 molecules were available for searching. These were represented by three types of 2D fingerprint: 988-bit Tripos Unity fingerprints;[34] 1052-bit Barnard Chemical Information Ltd. (BCI) fingerprints;[35] and 2048-bit Daylight Chemical Information Inc. fingerprints.[36] These molecules were searched using the 11 sets of active compounds detailed in Table 1.

A rough guide to the diversity of each of the chosen sets of bioactives is provided by matching each compound with every other in its activity class, calculating similarities using the Unity fingerprint and Tanimoto coefficient, and computing the mean and standard deviation for these intraset similarities. The resulting similarity scores are listed in the second column of Table 1, where it will be seen that the renin inhibitors are the most homogeneous and the cyclooxygenase inhibitors are the most heterogeneous.

For each of the 11 activity classes, 10 active compounds were selected for use as the training set. The selections were

Table 2. Comparison of the Average Percentage of Active Compounds Retrieved over the Top 5% of the Ranked Test Set Using Different Threshold Values To Generate the Modal Unity Fingerprint

	30%		40%		50%		60%		70%		80%		90%		100%	
activity name	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
5HT3 antagonists	24.15	4.88	30.31	5.21	32.26	6.00	27.76	3.15	22.78	4.57	13.11	5.82	7.90	4.66	3.85	0.87
5HT1A agonists	16.41	4.12	21.85	2.29	25.09	3.10	25.02	4.12	24.04	3.84	21.92	3.32	19.71	1.98	17.60	1.94
5HT reuptake inhibitor	33.98	4.14	39.63	3.67	38.65	3.97	31.35	6.37	25.70	4.20	21.09	1.42	20.20	2.15	16.68	3.69
D2 antagonists	21.51	6.76	27.12	5.35	27.69	5.62	24.03	6.41	19.06	5.58	15.71	3.29	12.03	1.96	9.40	2.35
renin inhibitor	88.12	2.97	88.77	3.15	88.16	3.93	86.46	4.94	80.46	4.10	74.38	4.38	63.40	10.64	33.03	10.83
angiotensin II AT1 antagonists	62.36	5.50	73.63	5.64	75.16	5.62	70.62	6.08	57.51	14.65	37.01	14.41	11.75	10.42	2.85	3.19
thrombin inhibitor	41.21	4.85	49.43	7.31	47.99	5.78	44.44	6.40	38.92	6.47	31.87	9.96	20.37	11.37	11.56	10.23
substance P antagonist	34.19	4.38	36.80	5.08	34.05	5.92	33.74	5.66	28.09	2.49	22.80	3.33	17.22	3.80	9.60	3.00
HIV protease inhibitor	51.47	5.64	53.53	4.45	48.72	2.62	42.93	4.84	37.62	5.27	29.50	8.91	21.74	10.70	11.01	10.37
cyclooxygenase inhibitor	10.19	3.33	10.96	3.07	7.67	2.55	7.00	2.87	6.76	1.93	5.93	1.58	5.40	1.37	5.19	1.14
protein kinase C inhibitor	31.83	7.38	35.60	10.33	30.65	10.59	22.23	10.57	14.49	8.17	9.75	6.46	7.65	4.15	9.82	1.56
average over all classes	37.77	4.90	42.51	5.05	41.46	5.06	37.78	5.58	32.31	5.57	25.73	5.72	18.85	5.75	11.87	4.47

Table 3. Comparison of the Average Percentage of Active Compounds Retrieved over the Top 5% of the Ranked Test Set by Combining Either the Scores or the Ranks of Different Single Similarity Searches Using the MAX and SUM Fusion Rules with Unity Fingerprints

	ranking				scoring			
	MAX		SUM		MAX		SUM	
activity name	mean	SD	mean	SD	mean	SD	mean	SD
5HT3 antagonists	7.14	2.30	12.16	3.01	49.03	5.43	38.42	4.57
5HT1A agonists	8.47	1.48	11.20	2.63	37.15	4.06	26.93	2.82
5HT reuptake inhibitor	14.67	3.12	21.81	3.37	49.68	5.45	42.98	4.05
D2 antagonists	6.86	2.12	9.43	3.59	37.40	4.92	31.53	4.63
renin inhibitor	70.61	8.45	83.94	2.67	88.62	1.90	90.50	1.65
angiotensin II AT1 antagonists	23.75	7.38	43.17	9.22	80.44	6.08	79.04	5.28
thrombin inhibitor	14.17	6.00	26.46	7.30	58.58	8.98	52.77	5.88
substance P antagonist	14.05	4.30	22.69	5.31	47.14	5.16	43.12	3.73
HIV protease inhibitor	19.59	10.07	34.53	8.99	61.62	7.85	60.15	6.05
cyclooxygenase inhibitor	2.09	1.42	1.04	0.93	26.52	7.15	14.27	4.26
protein kinase C inhibitor	6.68	2.31	10.84	5.19	48.01	8.99	37.22	8.47
average over all classes	17.10	4.45	25.21	4.75	53.11	6.00	46.99	4.67

done at random, subject to the constraint that no pairwise similarity in a group exceeded 0.80 (using Unity fingerprints and the Tanimoto coefficient). Each searching method (modal, weighted, fusion of ranks, fusion of scores, R1, and BKD) was repeated 10 times using 10 different training sets, and each training set was represented by each type of fingerprint (Unity, BCI, and Daylight). For each search, a note was made of the percentage of the active molecules (i.e., those in the same class as those in the training set) that occurred in the top 1% and the top 5% of the ranking resulting from that search. The results presented below are the mean and standard deviations for these recall values, averaged over each set of 10 searches. To save space in what follows, we have included only the top 5% experiments using the Unity fingerprints; the conclusions that can be drawn from these results are the same as those that can be drawn from the top 1% experiments and from the use of the BCI and Daylight fingerprints.

RESULTS AND DISCUSSION

Initial Experiments. Our initial experiments were carried out to identify the best settings when there was some variable parameter associated with a search method.

The key feature in the modal approach is the user-defined threshold associated with the generation of the modal fingerprint. For example, when this threshold is set to be 80%, all of the bits set to "on" in the modal fingerprint are set in at least 80% of the molecules in the training set. Searches were carried out with the threshold increasing from 30% to 100% in steps of 10%. In all cases, the recall increases from that at 30%, reaches a maximum at 40% or 50% (around the bit-density of the fingerprints of typical individual molecules), and then decreases steadily as the threshold is raised, as is illustrated by the Unity results shown in Table 2. The recall at the highest threshold values (90% and 100%) was sometimes little better than random, which is not too surprising since such a high threshold implies that very few bits are likely to be set, making it very difficult for the modal fingerprint to discriminate between active and inactive molecules in the search database. We hence adopted 40% as the default threshold value for the main searches discussed below.

With the data fusion methods, the principal choices are, first, what is being fused (here, fusion by ranks or fusion by scores) and, second, what fusion rule to use (here, SUM or MAX). Inspection of the results, such as those shown in Table 3, suggest that fusion of the scores is much more effective than fusion of the ranks, with the difference in performance being greatest for the more heterogeneous activity classes. SUM is the better fusion rule when fusion is by ranks, but MAX is the fusion rule of choice for fusion by scores, and we hence chose this latter combination of parameters to represent the data fusion approach. The superiority of MAX here supports some of the results obtained in a previous study by Schuffenhauer et al. of search methods for multiple reference structures.[8] These authors found that a consistently high level of performance was

Table 4. Comparison of the Average Percentage of the Active Compounds Retrieved over the Top 5% of the Ranked Test Set Using Various Parametrizations of the BKD Method with Unity Fingerprints

	100 inactives, $k = 100$		100 inactives, $k = 988$		10 inactives, $k = 100$	
activity name	mean	SD	mean	SD	mean	SD
5HT3 antagonists	52.32	8.27	49.47	8.70	47.79	4.28
5HT1A agonists	38.19	7.03	37.21	7.41	30.78	5.71
5HT reuptake inhibitor	45.82	7.93	45.67	7.34	37.28	4.56
D2 antagonists	38.65	7.38	38.00	7.06	33.30	7.70
renin inhibitor	93.34	1.35	91.98	2.26	89.84	5.95
angiotensin II AT1 antagonists	84.47	6.59	84.27	4.68	82.19	4.59
thrombin inhibitor	63.06	7.66	61.95	8.20	54.48	9.20
substance P antagonist	58.39	8.27	56.91	8.40	44.79	6.47
HIV protease inhibitor	68.45	8.31	67.34	8.44	59.07	9.73
cyclooxygenase inhibitor	33.15	4.68	33.31	4.89	30.51	6.58
protein kinase C inhibitor	49.37	10.84	48.76	10.92	47.47	9.84
average over all classes	56.84	7.12	55.90	7.12	50.68	6.78

Table 5. Comparison of the Average Percentage of Active Compounds Retrieved by the Various Methods over the Top 5% of the Ranked Test Set Using Unity Fingerprints

	substructural analysis			modal		weighted		data-fusion		BKD	
activity name	upper-bound	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
5HT3 antagonists	67.79	29.27	5.07	30.31	5.21	2.99	2.14	49.03	5.43	52.32	8.27
5HT1A agonists	48.23	30.13	2.21	21.85	2.29	1.05	0.79	37.15	4.06	38.19	7.03
5HT reuptake inhibitor	65.62	33.12	4.72	39.63	3.67	10.14	6.19	49.68	5.45	45.82	7.93
D2 antagonists	61.56	27.51	3.06	27.12	5.35	4.83	3.47	37.40	4.92	38.65	7.38
renin inhibitor	90.98	52.94	6.67	88.77	3.15	73.84	4.92	88.62	1.90	93.34	1.35
angiotensin II AT1 antagonists	89.07	43.40	6.66	73.63	5.64	51.55	3.58	80.44	6.08	84.47	6.59
thrombin inhibitor	73.90	35.64	7.69	49.43	7.31	22.50	4.24	58.58	8.98	63.06	7.66
substance P antagonist	70.23	36.52	6.60	36.80	5.08	14.51	3.12	47.14	5.16	58.39	8.27
HIV protease inhibitor	73.78	34.05	6.22	53.53	4.45	40.88	8.45	61.62	7.85	68.45	8.31
cyclooxygenase inhibitor	61.18	19.20	3.48	10.96	3.07	5.30	2.53	26.52	7.15	33.15	4.68
protein kinase C inhibitor	64.79	35.58	6.69	35.60	10.33	21.67	5.25	48.01	8.99	49.37	10.84
average over all classes	69.74	34.31	5.37	42.51	5.05	22.66	4.06	53.11	6.00	56.84	7.12

achieved using a search strategy (called 1NN in their paper) that is identical to the MAX strategy used here, when searches were carried out for database molecules with the same biological target as the reference structures (which is the situation here); an alternative, centroid strategy was found to be more appropriate in chemogenomics searches for closely related biological targets.

There are two parameters with the BKD method: the value of k in the exponent of the modified kernel function and the number of inactives in the training set. Experiments were carried out using $k = N$ (where N is the number of bits in the fingerprint representation that is being used) and $k \approx 0.1N$ (i.e., $k = 988$ and $k = 100$ in the case of the Unity fingerprints), and with either 10 inactives (i.e., the same as the number of actives) or 100 inactives. The results for the Unity fingerprints are shown in Table 4, from which it will be seen that the value of k has only a slight effect on the overall level of performance, whereas the use of the larger number of inactives has a marked effect.

Main Experiments. The parameter values that performed best in the initial experiments were then used in the main set of searches, the results of which are shown in Table 5; the figures here are again the top 5% experiments with Unity fingerprints, but similar levels of relative performance are obtained when just the top 1% of the ranking is considered, or when the other two types of fingerprint are used to represent the molecules.

It will be seen that we have included two sets of values in the substructural analysis column. The first is the simple R1 weight as defined previously; the other, upperbound value is one we have calculated to determine the maximum possible level of performance that could be achieved given these data sets and fingerprints. Specifically, we have calculated the R1 weights in each case using the entire data set for the T_j and N_T values and using the entire set of actives for the A_j and N_A values. Hardly surprisingly, the rankings obtained using these values are markedly better than those obtained using the realistic, simplified R1 weight based on the set of reference structures (or, indeed, those obtained using any of the other methods). It would, of course, also be possible to calculate an analogous upperbound for the BKD approach by using the entire data set as the training set for the optimization of the scoring scheme; however, this would be extremely demanding of computational resources,[31] whereas the R1 upperbound can be calculated at negligible cost.

Returning now to the practical methods, data fusion of similarity scores and BKD are clearly the methods of choice, consistently outperforming the other approaches. Of the two, BKD is the more demanding of computational resources. Assume that the training set contains N_A actives and N_I inactives and that there are N molecules in the whole database, then the data fusion method requires NN_A similarity calculations while the BKD method requires $N(N_A+N_I)$ kernel function calculations, each of which is rather more complex than the Tanimoto coefficients used in the data fusion method. In our experiments, there were 10 times as many inactives as actives in the BKD training set, meaning that BKD is at least an order of magnitude slower than data fusion, even if we leave aside time spent in training to determine the optimal value of λ. That said, the computational requirements are not that large for the data sets used here, with a C implementation of a typical BKD search of

Table 6. Comparison of the Rankings in Decreasing Order of Self-Similarity and the Rankings in Decreasing Order of the Recall at 5% Using Unity Fingerprints

activity name	self-similarity	SSA-upperbound	SSA	modal	weighted	data fusion	BKD
5HT3 antagonists	6	6	9	8	10	6	6
5HT1A agonists	9	11	8	10	11	10	10
5HT reuptake inhibitor	7.5	7	7	5	7	5	8
D2 antagonists	7.5	9	10	9	9	9	9
renin inhibitor	1	1	1	1	1	1	1
angiotensin II AT1 antagonists	4	2	2	2	2	2	2
thrombin inhibitor	3	3	4	4	4	4	4
substance P antagonist	5	5	3	6	6	8	5
HIV protease inhibitor	2	4	6	3	3	3	3
cyclooxygenase inhibitor	11	10	11	11	8	11	11
protein kinase C inhibitor	10	8	5	7	5	7	7
Spearman's ρ		0.911	0.698	0.866	0.711	0.848	0.916

MDDR taking ca. 380 s on an R12000 processor running under IRIX 6.5.

With some minor exceptions, the performance of all of the methods tends to increase as the self-similarity of the active molecules increases. This is illustrated in Table 6, in which the activity classes have been ranked in decreasing order of self-similarity and compared to the effectiveness of ranking (as detailed in Table 5). The extent of the correlation between the performance ranking and the degree of intraclass similarity for each of the methods has been calculated using the Spearman rank correlation coefficient, ρ. These values are listed in the bottom row of the table, demonstrating clearly that the absolute level of performance is strongly correlated with the degree of self-similarity of the activity class of interest. There is also a clear degree of similarity between the performance rankings of the various methods, with a value for Kendall's W of 0.676 ($p \leq 0.001$). The correlation with intraclass similarity is not unexpected; what is of importance here is that a good virtual screening performance is obtained even with quite diverse activity classes (such as the protein kinase C inhibitors and the D2 agonists). The worst results are obtained with the most diverse set of actives, i.e., the cyclooxygenase inhibitors; even here, however, the data fusion runs represent a 5.3-fold enrichment over a random ranking of the data set (the average enrichment factor with this approach is 10.62 and the largest—for the renin inhibitors—an impressive 17.72).

Inspection of Table 5 shows that the weighted results are very variable: some of the data sets (such as the renin inhibitors, the angiotensin II AT1 antagonists and the HIV protease inhibitors) give poor, but not totally unacceptable, results, but others (such as the 5HT3 antagonists and the 5HT1A agonists) give results little better than would be expected from random selection. It is noticeable that these very bad results are associated with the most diverse data sets. If the set of active reference structures is diverse, then very many of the bits in the weighted fingerprint will have an associated nonzero weight, and such a noisy representation is likely to be less discriminating than one resulting from a homogeneous set of reference structures. Some support for this belief is provided by the figures in Table 7, which details the mean numbers of Unity bits set for the original set of active structures and for the 10 weighted fingerprints, together with the ratio of the weighted to the original figures. The value of this ratio mirrors the self-similarity values in Table 1 and hence also (as exemplified by Table 6) the relative search performance. Thus the values of the ratio tend to be smaller for those data sets with the best weighted search performance and tend to be larger for those with the worst performance, with the lowest (2.18) and highest (3.20) ratios corresponding to the renin inhibitors and the cyclooxygenase inhibitors, respectively.

Table 7. Numbers of Unity Bits Set for the Original Structures and for the Ten Weighted Fingerprints, Together with the Ratio of the Weighted to the Original Figures

	whole set		weighted		weighted
activity name	mean	SD	mean	SD	whole set
5HT3 antagonists	205.3	35.9	578.3	58.3	2.82
5HT1A agonists	186.3	40.8	548.1	55.3	2.94
5HT reuptake inhibitor	189.7	46.7	585.9	59.1	3.09
D2 antagonists	203.6	48.7	598.6	60.3	2.94
renin inhibitor	254.1	36.9	553.2	55.8	2.18
angiotensin II AT1 antagonists	290.1	54.7	703.3	70.9	2.42
thrombin inhibitor	236.7	51.4	602.0	60.7	2.54
substance P antagonist	221.6	49.4	602.7	60.8	2.72
HIV protease inhibitor	251.1	46.3	614.1	61.9	2.45
cyclooxygenase inhibitor	208.2	60.2	665.9	67.1	3.20
protein kinase C inhibitor	226.6	72.8	654.6	66.0	2.89

Combination of Methods. Thus far, we have considered the performance of the various methods in isolation. Table 8 summarizes the extent to which the methods retrieve similar sets of compounds. Specifically, we have looked at the percentage of the top 5% of the rankings that are common to pairs of methods. These results, averaged over all of the 11 activity classes, are listed in Table 8, which details not only just the percentage of the molecules in the top 5% of the rankings that are in common but also the percentage of the active molecules in the top 5% of the rankings that are in common. It will be clear that while there are significant variations in the former figure, there is a high level of commonality in the actives that are identified, with the highest degree of commonality (90.55%) occurring for the two best-performing methods, BKD and data fusion. Given this level of similarity in the search outputs, it is to be expected that fusion of these two rankings is unlikely to bring about noticeable improvements in performance, when compared to using either of the methods on their own. This was found to be the case in practice; for example, rank-based fusion of the BKD and data fusion results gave a mean recall (top 5% of the Unity rankings) of 57.08, as against 57.37 for BKD on its own.

Xue et al. have recently described a single fingerprint approach called *fingerprint profiling*, which monitors the relative occupancy of fingerprint bit positions.[19,20] These profiles are then used to create a *consensus pattern*, which

Table 8. Percentage of the Molecules, and Percentage of Active Molecules, in the Top 5% of the Unity Rankings that Are Common to a Pair of Retrieval Methods

	data fusion				modal				substructural analysis				weighted			
	molecules		actives		molecules		actives		molecules		actives		molecules		actives	
	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
BKD	52.03	7.34	90.55	4.58	38.30	10.43	83.36	10.87	23.80	5.653	80.60	7.71	19.29	5.52	83.45	15.59
data fusion					50.63	7.63	84.51	7.30	31.11	3.99	83.44	5.86	16.71	4.73	78.33	16.14
modal									23.86	4.36	69.33	8.28	23.64	5.74	77.43	15.69
substructural analysis													3.74	0.90	42.90	15.54

Table 9. Average Percentage of Active Compounds Retrieved over the Top 5% of the Ranked Test Set Using Unity Fingerprints and Fingerprint Scaling

	scores MAX						scores SUM					
	40%		80%		100%		40%		80%		100%	
activity name	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD	mean	SD
5HT3 antagonists	45.43	6.39	45.26	6.50	48.89	4.29	24.41	6.17	35.07	8.34	37.52	4.65
5HT1A agonists	31.09	4.42	32.77	4.91	37.58	3.89	12.46	3.74	23.79	4.65	31.21	2.70
5HT reuptake inhibitor	45.62	3.99	45.90	7.37	48.57	7.12	33.27	4.54	36.62	6.07	38.11	5.67
D2 antagonists	34.91	6.03	35.58	5.77	37.09	4.00	20.70	5.18	26.62	4.41	29.17	4.84
renin inhibitor	91.99	1.16	92.23	1.13	92.83	2.25	90.38	1.11	91.11	1.11	92.39	2.38
angiotensin II AT1 antagonists	82.14	5.16	82.86	6.08	80.08	4.01	72.93	5.42	78.16	4.58	77.61	3.34
thrombin inhibitor	56.10	9.15	57.24	6.05	56.85	10.84	47.00	3.38	51.31	5.76	53.53	9.07
substance P antagonist	42.39	5.43	49.64	4.88	51.29	5.77	34.35	5.33	44.39	4.90	45.01	2.70
HIV protease inhibitor	64.99	6.20	68.93	6.24	64.86	8.43	60.35	4.65	65.04	6.05	64.15	8.75
cyclooxygenase inhibitor	24.38	5.71	22.43	6.79	24.73	5.69	10.32	2.92	9.82	3.30	11.69	1.61
protein kinase C inhibitor	47.13	9.31	45.19	9.25	46.30	9.87	32.28	9.17	31.33	11.01	34.47	9.57
average over all classes	51.47	5.72	52.55	5.91	53.55	6.01	39.86	4.69	44.84	5.47	46.81	5.03

consists of all bits that are always set "on" in a data set, and Xue et al. have demonstrated that different activity classes have different consensus patterns. Database molecules are ranked against a reference structure using the simple (binary) version of the Tanimoto coefficient but with a scaling factor applied to the bits present in the consensus pattern, so that bits that are in common and that are in the consensus pattern make a greater contribution to the overall similarity than nonconsensus bits. This technique, called *fingerprint scaling*, improved the retrieval of active compounds without increasing the number of false positives and provides an elegant way of using a training set of active molecules. The approach is not directly applicable here as it still assumes the use of individual reference structures, rather than a set of reference structures as in our work. We have hence carried out individual scaled fingerprint searches for each of the training set actives, using three thresholds for the creation of the consensus pattern (40%, 80%, and 100%) and then combined the results using the MAX and SUM fusion rules investigated previously. The results for the Unity fingerprints are shown in Table 9, where it will be seen that there is a very slight improvement in performance using MAX and 100% threshold when compared with the simple data fusion runs in Table 5 that do not involve fingerprint scaling; there was no difference in the case of the Daylight fingerprints and a slight decrease with the BCI fingerprints. We hence conclude that fingerprint scaling does not materially affect the performance of multiple-reference similarity searching.

Single Reference Structures. In the final set of experiments, we sought to quantify the benefit that can be achieved using multiple reference structures, rather than single reference structures as in conventional similarity searching. This was done by using every single active molecule in each of the 10 chosen activity classes as the reference structure and recording the minimum, mean, and maximum performance,

Table 10. Percentage of Active Compounds Retrieved over the Top 5% of the Ranked Test Set with Single Similarity Searches Using Unity Fingerprints

activity name	mean	SD	max	min
5HT3 antagonists	21.15	7.36	40.97	1.89
5HT1A agonists	18.43	5.32	39.29	2.45
5HT reuptake inhibitor	24.02	10.08	42.69	1.43
D2 antagonists	17.35	6.60	35.58	0.26
renin inhibitor	80.54	13.83	93.21	2.95
angiotensin II AT1 antagonists	48.04	17.95	81.67	3.64
thrombin inhibitor	33.51	14.72	63.56	0.63
substance P antagonist	26.87	10.47	57.69	0.57
HIV protease inhibitor	37.60	13.82	63.65	1.89
cyclooxygenase inhibitor	9.39	4.76	21.09	0.32
protein kinase C inhibitor	19.42	13.43	46.05	0.68
average over all classes	30.57	10.76	53.22	1.52

as detailed in Table 10. The mean values correspond to the performance that might be expected using a single reference structure and are clearly much lower than the figures reported in Table 5 for the BKD and data fusion methods (30.57% as against 57.37% and 53.11%, respectively). Thus, the use of 10 actives, rather than just one, results in an increase of over two-thirds in the numbers of actives retrieved. Perhaps the most interesting figures in Table 10 are those for the best possible single similarity search, i.e., the figures listed under maximum. These represent the best single similarity searches possible from the many hundreds of individual bioactive molecules (this number ranges from 349 for the 5HT reuptake inhibitors up to 1236 for the substance P antagonists). If we consider the average over all activity classes, it will be seen that this upperbound is only fractionally better than the data fusion result in Table 5 and is actually worse than the BKD figure. Thus, on average, picking any 10 active reference structures and combining them using the best approaches discussed in this paper will enable searches to be carried out that are comparable to even

the best possible conventional similarity search using a single active reference structure.

CONCLUSIONS

Similarity searching using 2D fragment bit-strings and the Tanimoto coefficient provides an effective and an efficient technique for virtual screening when a single reference structure is available. In this paper we have reviewed the use of similarity searching when multiple reference structures are available, using both published and novel approaches to the combination of information from the various reference structures. Experiments with the MDDR database demonstrate that data fusion based on similarity scores and an approximate form of the binary kernel discrimination method are by far the best of the approaches that we have considered. The kernel discrimination method is slightly the more effective, but notably the less efficient, of the two. Their use is hence recommended for 2D virtual screening applications when more than a single reference structure is available.

There are many ways in which this work could be extended. One way, which is currently under investigation in our laboratories, is the use of a wider range of types of fingerprint. Those studied here—BCI, Daylight, and Tripos—are all based on small, atom-centered and bond-centered substructures, and it is thus appropriate to consider other types that use more chemical fragment definitions, such as the topological pharmacophore keys described by Schuffenhauer et al.[8] As another alternative, we have considered only 2D fingerprint representations, and there is now much interest in 3D-based and graph-based approaches to similarity searching (see, e.g., refs 13 and 15) that could be applied here; again, the use of similarity coefficients other than the Tanimoto coefficient is the subject of continuing interest in this department,[9,27] and such coefficients could be investigated for use with multiple-reference searching. Finally, our work has involved specific activity classes thus far, and it would be appropriate to consider the extent to which our results are also applicable to chemogenomics applications in which searches are carried out for ligands of related biological targets.[7,8]

ACKNOWLEDGMENT

We thank the following: Novartis Institutes for BioMedical Research for funding J.H.; MDL Information Systems Inc. for the provision of the MDDR database; Barnard Chemical Information Ltd., Daylight Chemical Information Systems Inc., the Royal Society, Tripos Inc., and the Wolfson Foundation for software and laboratory support. The Krebs Institute for Biomolecular Research is a designated biomolecular sciences center of the Biotechnology and Biological Sciences Research Council.

REFERENCES AND NOTES

(1) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure−activity studies: definition and applications. *J. Chem. Inf. Comput. Sci*. **1985**, *25*, 64−73.
(2) Willett, P.; Winterman, V.; Bawden, D. Implementation of nearest neighbour searching in an online chemical structure search system. *J. Chem. Inf. Comput. Sci.* **1986**, *26*, 36−41.
(3) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical similarity searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(4) Sheridan, R. P.; Kearsley, S. K. Why do we need so many chemical similarity search methods? *Drug Discov. Today* **2002**, *7*, 903−911.
(5) Flower, D. R. On the properties of bit-string based measures of chemical similarity. *J. Chem. Inf. Comput. Sci*. **1998**, *38*, 379−386.
(6) Godden, J. W.; Xue, L.; Bajorath, J. Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and Tanimoto coefficients. *J. Chem. Inf. Comput. Sci*. **2000**, *40*, 163−166.
(7) Schuffenhauer, A.; Zimmermann, J.; Stoop, R.; van der Vyver, J.-J.; Lecchini, S.; Jacoby, E. An ontology for pharmaceutical ligands and its application for *in silico* screening and library design. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 947−955.
(8) Schuffenhauer, A.; Floersheim, P.; Acklin, P.; Jacoby, E. Similarity metrics for ligands reflecting the similarity of the target proteins. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 391−405.
(9) 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.
(10) Sheridan, R. P. Finding multiactivity substructures by mining databases of drug-like compounds. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 1037−1050.
(11) Chen, X.; Reynolds, C. H. Performance of similarity measures in 2D fragment-based similarity searching: Comparison of structural descriptors and similarity coefficients. *J. Chem. Inf. Comput. Sci*. **2002**, *42*, 1407−1414
(12) 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.
(13) Mason, J. S.; Morize, I.; Menard, P. R.; Cheney, D. L.; Hulme, C.; Labaudiniere, R. F. A new 4-point pharmacophore method for molecular similarity and diversity applications: overview of the method and applications, including a novel approach to the design of combinatorial libraries containing privileged substructures. *J. Med. Chem*. **1999**, *42*, 3251−3264.
(14) Singh, S. B.; Sheridan, R. P.; Fluder, E. M.; Hull, R. D. Mining the chemical quarry with joint chemical probes: an application of latent semantic indexing (LaSSI) and TOPOSIM (Dice) to chemical database mining. *J. Med. Chem*. **2001**, *44*, 1564−1575.
(15) Raymond, J. W.; Willett, P. Effectiveness of graph-based and fingerprint-based similarity measures for virtual screening of 2D chemical structure databases. *J. Comput.-Aided. Mol. Des.* **2002**, *16*, 59−71.
(16) *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; Wiley: New York, 1990.
(17) Patterson, D. E.; Cramer, R. D.; Ferguson, A. M.; Clark, R. D.; Weinberger, L. E. Neighbourhood behaviour: a useful concept for validation of "molecular diversity" descriptors. *J. Med. Chem.* **1996**, *39*, 3049−3059.
(18) Martin, Y. C.; Kofron, J. L.; Traphagen, L. M. Do structurally similar molecules have similar biological activity. *J. Med. Chem*. **2002**, *45*, 4350−4358.
(19) Xue, L.; Stahura, F. L.; Godden, J. W.; Bajorath, J. Fingerprint scaling increases the probability of identifying molecules with similar activity in virtual screening calculations. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 746−753.
(20) Xue, L.; Godden, J. W.; Stahura, F. L.; Bajorath, J. Profile scaling increases the similarity search performance of molecular fingerprints containing numerical descriptors and structural keys. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 1218−1225.
(21) *Pharmacophore Perception, Development and Use in Drug Design*; Guner, O., Ed.; International University Line: La Jolla, CA, 2000.
(22) Shemetulskis, N. E.; Weininger, D.; Blankey, C. J.; Yang, J. J.; Humblet, C. Stigmata: an algorithm to determine structural commonalities in diverse data sets. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 862−871.
(23) Klein, L. A. *Sensor and Data Fusion Concepts and Applications*, 2nd ed.; SPIE The International Society for Optical Engineering: Bellingham, WA, 1999.
(24) Ginn, C. M. R.; Willett, P.; Bradshaw, J. Combination of molecular similarity measures using data fusion. *Perspect. Drug Discovery Des.* **2000**, *20*, 1−16.
(25) Charifsen, P. S.; Corkery, J. J.; Murcko, M. A..; Walters, W. P. Consensus scoring: a method for obtaining improved hit rates from docking databases of three-dimensional structures into proteins. *J. Med. Chem*. **1999**, *42*, 5100−5109.
(26) Wang, R.; Wang, S. How does consensus scoring work for virtual library screening? An idealized computer experiment. *J. Chem. Inf. Comput. Sci*. **2001**, *41*, 1422−1426.
(27) Salim, N.; Holliday, J. D.; Willett, P. Combination of fingerprint-based similarity coefficients using data fusion. *J. Chem. Inf. Comput. Sci.* **2003**, *43*, 435−442.
(28) Cramer, R. D.; Redl, G.; Berkoff, C. E. Substructural analysis. A novel approach to the problem of drug design. *J. Med. Chem.* **1974**, *17*, 533−535.

(29) Ormerod, A.; Willett, P.; Bawden, D. Comparison of fragment weighting schemes for substructural analysis. *Quant. Struct.-Act. Relat.* **1989**, *8*, 115−129.
(30) *Readings in Information Retrieval*; Spark Jones, K., Willett, P., Eds.; Morgan Kaufman: San Francisco, 1997.
(31) Wilton, D. J.; Willett, P.; Lawson, K.; Mullier, G. Comparison of ranking methods for virtual screening in lead-discovery programs. *J. Chem. Inf. Comput. Sci*. **2003**, *43*, 469−474.
(32) 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.
(33) The MDL Drug Data Report database is available from MDL Information Systems Inc. at http://www.mdli.com.
(34) Tripos Inc. is at http://www.tripos.com.
(35) Barnard Chemical Information Ltd.is at http://www.bci.gb.com/.
(36) Daylight Chemical Information Systems Inc. is at http://www.daylight.com.

CI034231B