Evaluation of Similarity Measures for Searching the *Dictionary of Natural Products* Database

Martin Whittle* and Peter Willett

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

Werner Klaffke and Paula van Noort

Unilever Research and Development, Unilever Health Institute, Vlaardingen, Oliver van Noortlaan 120, 3133 AT Vlaardingen, The Netherlands

Received August 29, 2002

Similarity searches using combinations of seven different similarity coefficients and six different representations have been carried out on the *Dictionary of Natural Products* database. The objective was to discover if any special methods of searching apply to this database, which is very different in nature from the many synthetic databases that have been the subject of previous studies of similarity searching. Search effectiveness was assessed by a recall analysis of the search outputs from sets of pharmacologically active target structures. The different target sets produce exceptional but contradictory results for the Russell-Rao and Forbes coefficients, which have been shown to be due to a dependence on molecular size; these are the coefficients of choice in the case of large and small structures, respectively. Rankings from these results have been combined using a data fusion scheme and some small gains in performance were normally obtained by using substructural fingerprints and molecular holograms in combination with the Squared Euclidean or Tanimoto coefficients.

INTRODUCTION

Virtual screening methods are being increasingly used to boost the efficiency of lead-discovery programs in the pharmaceutical and agrochemicals industries.[1,2] A range of methods is available for this purpose, including the use of property-based and structurally based filtering systems, similarity searching, 3D pharmacophore matching, and ligand docking: in this paper, we consider similarity searching, which is one of the simplest and most widely used of the current methods. Similarity searching requires the availability of a known bioactive molecule, such as a hit from a high-throughput screening experiment. This molecule, hereafter referred to as the *target structure*, is compared with each of the molecules in a database of 2D or 3D chemical structures by calculating a measure of the degree of structural resemblance between the target structure and the database structure. A wide range of intermolecular structural similarity measures suitable for the purpose have been reported in the literature.[3−5] The database molecules are then ranked in decreasing order of the calculated similarity values, identifying the top-ranked molecules, the *nearest neighbors*, as those that bear the closest structural similarity to the target structure. These are then presumed to have the highest *a priori* probability of activity and are hence candidates for a more sophisticated form of virtual screening or even for biological testing directly.

Similarity-based virtual screening is widely used in the discovery of novel drugs and pesticides, which are typically synthetic compounds with low molecular weights and limited structural complexity in terms of numbers of donors and acceptors, stereocenters, numbers of rotatable bonds, etc. In this paper, we consider the extent to which similarity searching is also applicable to the screening of natural products. These provide a rich source of biologically active molecules that are of increasing importance not only just for pharmaceutical and agrochemical research but also for the development of consumer products such as functional foods.[6−8] Because of the differences in natural and artificial synthetic pathways, natural product collections are known to represent a pool of compounds that is distinct from collections of synthetic molecules and that may be difficult or impossible to generate by artificial means.[9] There has already been some interest in the differences between synthetic compounds and natural products and their possible complementarity when trying to identify novel drug leads molecules.[10−12] However, we are aware of only one previous publication that considers the effect of such structural differences on search performance; here, Sheridan et al. discuss heuristics for the retrieval of nonpeptidic nearest neighbors in similarity searches for bioactive peptidic target structures.[13] In this paper, we report a detailed study of the effectiveness of a wide range of similarity measures for searching databases of natural products. Examples of such databases include the *Dictionary of Natural Products* (DNP),[14] the *Bioactive Natural Product*[15] and *BioScreenNP* files,[16] and the *Marine Natural Product Database* (MNPD).[17] Our experiments have involved the DNP and a wide range of different structural representations and similarity coefficients, so as to identify that combination that would appear to provide the best search performance for this database. The

* Corresponding author e-mail: m.whittle@sheffield.ac.uk.

10.1021/ci025591m CCC: $25.00
Published on Web 01/28/2003

effectiveness of the various searches is assessed using a recall analysis,[18] with data fusion being used to assess the improvements in effectiveness, if any, that can be achieved by combining different similarity measures.[19]

METHODS

Similarity Measures. A similarity measure comprises two major components: a *representation* (which may need to be weighted or standardized in some way) that describes some structural features of each of the molecules that are to be searched and a *similarity coefficient* that quantifies the degree of resemblance between the chosen representations of the target structure and each of the database structures. A recent review of measures for chemical similarity searching is provided by Willett et al.;[5] here, we describe the six representations and seven similarity coefficients that we have used.

Representations. Fragment bit-strings, or fingerprints, are widely used for 2D similarity searching,[5,20,21] and we have used two in our experiments. The Unity fingerprints from Tripos (hereafter denoted by FPR) use both a fragment dictionary and hashed atom-path substructures in a bit-string containing a total of 992 bits.[22] The fingerprints from Barnard Chemical Information (hereafter denoted by BCI) are entirely dictionary-based: we used the standard fragment dictionary that yields a bit-string containing 1052 bits.[23] Molecular holograms are an extension of 2D fingerprints in which each element of the vector encodes the number of times that a fragment occurs, rather than just its presence or absence: we used the Tripos holograms (hereafter denoted by HOL) with a 503-element vector (this was found to be equivalent in performance to longer holograms in pilot studies not reported here).

The fourth representation was based on the Molconn-Z software, which calculates a wide range of topological indices, information indices, and counts of graph paths, atoms, atoms types, etc.[24] A principal component analysis was used to extract a set of 100 orthogonal variables that were found to describe 79% of the variance in the data; this representation is referred to subsequently as MCZ. The final two representations were based on the spatial autocorrelation functions, first described by Broto et al.[25] for 2D structures and subsequently adapted for 3D structures by Schuur et al.,[26] that characterizes the distribution of atom−atom interactions at different interatomic distances. The basic autocorrelation function is

$$A(d) = \frac{1}{k}\sum_{i=1}^{k}\sum_{j=i}^{k} p_i p_j \Delta(d - d_{ij}) \qquad (1)$$

where k is the number of atoms in the molecule, p_i is some atomic property associated with atom i, d_{ij} is the topological distance between atom i and j (i.e. the shortest path measured by the number of bonds), and

$$\Delta a(x) = 1{:}x = 0; \quad \Delta(x) = 0{:}x \neq 0 \qquad (2)$$

is a binning function. In its simplest form, used here, the atomic property function p is set to unity symbolizing the property of existence so that $A(d)$ is a simply proportional to a frequency distribution of bond distances; however, it could also be given values that represent e.g. partial atomic charge. The basic 2D approach (hereafter referred to as CF2) can be extended to 3D by replacing d_{ij} with r_{ij} the (continuous) interatomic separation between atoms i and j (hereafter referred to as CF3).

Table 1. Similarity Coefficients in Terms of the Expressions Given in Eq 3

coefficient	expression
Tanimoto	$S_T = \frac{c}{a+b-c}$
Cosine	$S_C = \frac{c}{\sqrt{ab}}$
Squared Euclidean	$S_E = \frac{a+b-2c}{m}$
Russell-Rao	$S_R = \frac{c}{m}$
Kulczynski(2)	$S_K = \frac{1}{2}\left[\frac{c}{a} + \frac{c}{b}\right]$
Forbes	$S_F = \frac{cm}{ab}$
Mod	$S_M = \lvert\frac{c}{m} - \frac{a}{m}\rvert + \lvert\frac{c}{m} - \frac{b}{m}\rvert$

Similarity Coefficients. The similarity coefficient quantifies the degree of resemblance between two representations. Many of the coefficients that have been studied in the literature[5,20] can be defined for either continuous or dichotomous (binary) representations. We define the following quantities for two representations of length m

$$a = \sum_{j=1}^{m}(x_{jA})^2; \quad b = \sum_{j=1}^{m}(x_{jB})^2; \quad c = \sum_{j=1}^{m} x_{jA}x_{jB} \qquad (3)$$

where x_{jA} and x_{jB} are the jth elements of the representation vector for compounds A and B. For the binary case, where the elements of the representation vectors are either 1 or zero, the components of the coefficients are as follows: a is the number of bits "on" for the representation A; b is the number of bits "on" for the representation B; and c is the number of bits "on" in both A and B (i.e., the union of the two representations). From these basic definitions, all of the well-known similarity coefficients can be calculated.

For data fusion purposes (as discussed further below) we need to choose coefficients that are not "monotonic" with (i.e. give the same rankings as) each other and are preferably as different as possible to probe different aspects of the similarity. A recent study by Holliday et al.[20] examined the efficacy of 22 coefficients and their combinations when used with 2D bit-strings and grouped them into different clusters that produced comparable rankings. Based on this work, and given that not all the coefficients studied have a continuous nonbinary form,[27] we have chosen, as far as is possible, a representative from each group. The expressions for these coefficients in terms of the definitions of eq 3 are given in Table 1.

The Russell-Rao, Kulczynski(2), and Forbes coefficients have been found effective for similarity searching in our laboratory when used with fingerprints, and they would appear to have a straightforward extension to continuous form. However, an analysis of these coefficients shows that

Table 2. Similarity Overlap. The entries show the percentage of structures in the row database that have a similarity match in the column database (mean and two standard deviations confidence limits) for the stated values of Tanimoto coefficient using BCI fingerprints

$S_T > 0.68$	DNP91s	NCI AIDS	MDDR	CAS	BIOSTER	ID Alert
DNP	100	73.67 ± 1.78	70.23 ± 1.69	46.81 ± 2.49	33.38 ± 0.91	57.80 ± 1.69
NCI AIDS	32.21 ± 1.61	100	41.77 ± 3.04	44.23 ± 1.62	18.85 ± 2.15	22.74 ± 1.68
MDDR	26.27 ± 2.07	31.55 ± 1.46	100	24.5 ± 1.15	38.86 ± 2.67	68.71 ± 1.14
CAS	36.30 ± 1.89	45.66 ± 1.88	36.72 ± 2.59	100	19.78 ± 2.37	22.97 ± 0.97
BIOSTER	45.66 ± 1.49	49.84 ± 1.59	77.15 ± 0.83	44.94 ± 1.33	100	63.94 ± 1.76
ID Alert	37.12 ± 2.30	36.49 ± 1.43	93.84 ± 0.77	30.23 ± 3.04	49.81 ± 2.47	100
$S_T > 0.80$	DNP91s	NCI AIDS	MDDR	CAS	BIOSTER	ID Alert
DNP91s	100	37.53 ± 1.88	33.74 ± 1.70	19.43 ± 1.18	11.88 ± 0.71	24.88 ± 0.90
NCI AIDS	14.75 ± 1.76	100	14.77 ± 1.56	13.34 ± 2.04	5.45 ± 0.76	6.56 ± 0.72
MDDR	11.79 ± 1.85	9.18 ± 1.80	100	4.13 ± 0.85	15.66 ± 1.29	40.88 ± 2.09
CAS	14.37 ± 1.33	13.23 ± 1.44	8.69 ± 2.07	100	4.14 ± 0.78	4.50 ± 0.65
BIOSTER	24.83 ± 0.86	23.41 ± 23.41 ± 1.63	55.29 ± 0.81	15.22 ± 1.63	100	40.22 ± 2.46
ID Alert	20.39 ± 1.12	13.27 ± 1.65	82.37 ± 1.04	7.76 ± 1.52	25.14 ± 1.53	100
$S_T = 1.0$	DNP91s	NCI AIDS	MDDR	CAS	BIOSTER	ID Alert
100	1.92 ± 0.52	2.45 ± 0.34	2.23 ± 0.55	1.11 ± 0.54	1.69 ± 0.76	
NCI AIDS	3.07 ± 0.72	100	1.59 ± 0.25	0.65 ± 0.43	0.59 ± 0.25	0.40 ± 0.43
MDDR	1.71 ± 0.38	0.40 ± 0.36	100	0.04 ± 0.13	0.68 ± 0.33	4.86 ± 0.94
CAS	1.65 ± 0.30	0.56 ± 0.31	0.21 ± 0.23	100	0.13 ± 0.20	0.21 ± 0.15
BIOSTER	8.47 ± 1.52	3.38 ± 1.19	14.34 ± 1.49	1.15 ± 0.49	100	12.16 ± 0.70
ID Alert	6.75 ± 1.28	1.58 ± 0.32	32.83 ± 1.33	0.44 ± 0.44	5.10 ± 1.54	100

they are not appropriate for use with continuous representations, as discussed in the Appendix. Motivated by the success of the Russell-Rao coefficient for binary strings[20] and in an attempt to circumvent the problems detailed in the Appendix, a new coefficient was developed which we label S_M (the "*M*" standing for modulus). This coefficient is monotonic with the Squared Euclidean distance for dichotomous variables but generates different rankings for continuous representations.

Recall Analysis. The effectiveness of the searches carried out here was measured using the recall. Assume that there are A molecules in a data set that exhibit some particular biological activity and that $a(n)$ $(1 \le a(n) \le A-1)$ of these A molecules occur in the top-n nearest neighbors retrieved in a similarity search for which one of the active molecules acts as the target structure, then the cumulative recall, often just called the recall, $R(n)$, is defined as the fraction of active compounds that are retrieved among the n nearest neighbors[18]

$$R(n) = \frac{a(n)}{A-1} \tag{4}$$

with $0 \le R(n) \le 1$. Repeating the search for all of the A active compounds gives an average recall value

$$\bar{R}(n) = \frac{1}{A}\sum_{i=1}^{A}\frac{a_i(n)}{A-1} \tag{5}$$

where a_i denotes the nearest neighbors found for the ith target structure. As the number n of retrieved molecules increases so does the number $a(n)$ of retrieved actives and a cumulative recall plot of $a(n)$ versus n has a characteristic shape depending on the effectiveness of retrieval. The relationship is linear for the random recall expectation, $R_0(n) = n/N$, that would be obtained for searches that did no better than random selection, while an effective scheme gives an early retrieval (low n) of active compounds above the random expectation leading to a convex shape.[18]

Data Fusion. Data fusion (also known as consensus scoring in the context of docking experiments[28]) is a process that involves the combination of results from several different sensors or inputs, with the aim of producing a better output than is obtainable using just a single data source.[29] In the context of similarity searching, this involves combining (or fusing) the rankings generated by different similarity measures. As in previous work,[19] we have used a sum-based approach to fusion. Assume that Q different rankings have been generated for the search of a given target structure against some database and that r_i denotes the rank of a particular molecule obtained with the ith similarity measure. We calculate the sum of the ranks, r_s

$$r_s = \sum_{i=1}^{Q} r_i, \tag{6}$$

and then rerank the database according to the resulting set of r_s scores.

Search Database. Version 9.1 of the DNP was filtered to remove duplicates, records that contained no structural information and records containing more than 112 non-hydrogen atoms that could not be processed by the Molconn-Z software. The resulting 106 652 structures were stored in SDF format and formed the basis for the analyses below.

Natural products are distinctly different from synthetic compounds, as demonstrated by studies of the distributions of physicochemical properties such as molecular weight and numbers of heteroatoms.[9] These differences are further studied in Table 2 using the similarity overlap approach described by Voigt et al.[30] Here, we have calculated the similarities between structures from the DNP file and from the following data sets: 5024 compounds from the BIOSTER database; 31 181 compounds from the CAS Registry Service (CAS); 11 603 compounds from the ID-Alert database; 102 444 compounds from the MACCS Drug Data Report (MDDR); and 37 117 compounds from the NCI AIDS

database. A random sample of 2000 compounds was chosen from each database, and its similarity calculated (using the Tanimoto coefficient and BCI fingerprints) with all of the structures in each of the files. The percentage of the target structures that had similarities greater than 0.68 or 0.80, or had a similarity value of 1.0, were noted. The process was repeated with five different samples in each case, and the mean plus two standard deviations are listed in Table 2. It will be seen that the CAS data set is noticeably different from any of the other files, as would be expected as it has been drawn from right across chemical space, whereas all of the other data sets have been chosen according to some explicit criterion. The two drug databases, MDDR and ID-Alert, are very similar to each other and both differ substantially from the DNP; such differences have been noted previously in the detailed study of Henkel et al.[9] and in several subsequent studies.[12,31,32]

Target Structures. Many of the molecules in the DNP have an associated pharmacological activity, and we have used this information for the evaluation of the various similarity measures we have tested. Thirty-one sets of compounds were identified that had a specific pharmacological activity and that occurred in at least eight compounds in the file. The molecules in each activity class were used as target structures for similarity searching, with the aim of retrieving as many as possible of the other members in its class.

The sets of pharmacological compounds contained numerous derivatives with very similar structures; these were filtered to remove very close analogues, leaving 994 possible target structures. Three mutually exclusive sets of target structures were randomly chosen, each containing 300 compounds: these are referred to subsequently as A1, A2, and A3. This method of choosing the target sets reflects the distribution of compounds among the pharmacological classes in the DNP. However, the distribution of class sizes is very disparate with, in particular, very large numbers of antineoplastics and antibiotics. Another test set, referred to subsequently as B, was obtained by choosing 20 compounds from each of the 15 classes that contained at least 20 members; duplicates (i.e., selected compounds that belonged to two or more of the chosen activity classes) were then eliminated to give a set of 276 target structures.

RESULTS AND DISCUSSION

Recall Results. Cumulative recall results for the four test-sets are given in Table 3 for rank $n = 1000$, which represents just under 1% of the full database. We also carried out entirely analogous experiments with $n = 6000$ but have not included the results here as they showed that the relative ratios of recall between the various searches remained effectively unchanged between $n = 1000$ and $n = 6000$.

The focus of this paper is the relative performance of the various measures we have tested. However, it is worth noting that, for this data set, the absolute performance of all the measures is quite low when compared with the results of virtual screening experiments on data sets such as MDDR or NCI AIDS. There are two reasons for this behavior: first, many of the activity classes used here are structurally disparate, with few of the "me too" compounds that characterize drug data sets; second, there is only a small fraction content (about 1%) of the DNP structures for which activity data is available in the database, and thus the absolute recall is inevitably low.

Table 3. Recall at Rank 1000, Expressed as $R(1000)/10^{-2}$, for All Viable Combinations of Representation and Similarity Coefficient[a]

A1	S_T	S_C	S_E	S_M	S_K	S_R	S_F
BCI	1.75	1.74	1.66	1.66	1.70	2.24	1.72
FPR	1.89	1.91	1.85	1.85	1.89	2.33	1.27
HOL	1.61	1.67	1.53	1.35			
MCZ	1.54	1.54	1.30	1.30			
CF2	1.17	1.22	1.17	0.99			
CF3	1.24	1.23	1.28	1.21			
A2	S_T	S_C	S_E	S_M	S_K	S_R	S_F
BCI	1.95	2.02	2.16	2.16	1.99	2.02	2.27
FPR	2.25	2.25	2.26	2.26	2.30	2.46	2.56
HOL	2.17	2.16	2.35	2.17			
MCZ	2.10	2.14	1.73	1.74			
CF2	1.34	1.32	1.36	1.16			
CF3	1.31	1.31	1.91	1.75			
A3	S_T	S_C	S_E	S_M	S_K	S_R	S_F
BCI	1.77	1.76	1.76	1.76	1.75	2.36	1.82
FPR	1.88	1.88	1.80	1.80	1.88	2.99	1.47
HOL	1.69	1.77	1.75	1.59			
MCZ	1.75	1.69	1.41	1.43			
CF2	1.12	1.10	1.09	0.98			
CF3	1.20	1.39	1.76	1.64			
B	S_T	S_C	S_E	S_M	S_K	S_R	S_F
BCI	1.50	1.50	1.82	1.82	1.47	1.18	2.55
FPR	1.54	1.52	1.84	1.84	1.55	1.39	2.38
HOL	1.59	1.5	1.94	1.83			
MCZ	1.62	1.66	1.45	1.44			
CF2	1.28	1.29	1.30	1.18			
CF3	1.38	1.52	1.87	1.72			

[a] Target sets A1−3 and B.

Inspection of Table 3 shows that the three sets A1−A3 show similar trends and demonstrate that Unity fingerprints with the Russell-Rao coefficient performs twice as well as the random expectation while the 2D correlation function results are often little better than random (the expected recall value of a random search is 0.934×10^{-2}). A clearer view of the relative performance of the various representations and coefficients is obtained if the raw recall data is normalized. This has been done by dividing by the average value for each set, so that values above 1.0 correspond to relatively effective combinations of representation and similarity coefficient; the normalized values are then averaged over the three target sets to give the mean normalized recall values, $R(1000)^*$, shown in Table 4. These figures show that the Russell-Rao coefficient using Unity fingerprints gives the best performance by a considerable margin. The Russell-Rao also performs well with BCI bit-strings, with the other coefficients giving a remarkably uniform response for these representations. Unity bit-strings are the best overall representation using the remaining coefficients with BCI and the holograms coming next. The 3D autocorrelation functions were the best of the other representations, showing some preference for the squared Euclidean coefficient.

Inspection of the corresponding results for test-set B, which was chosen to give a more even distribution of activities, reveals a different picture. For this group of target structures,

Table 4. Mean Normalized Recall *R*(1000)* for Sets A1−A3 and Normalized Recall for Set B, for All Viable Combinations of Representation and Similarity Coefficient

A1−A3	ST	SC	SE	SM	SK	SR	SF
BCI	1.07	1.07	1.08	1.08	1.06	1.30	1.12
FPR	1.17	1.17	1.15	1.15	1.18	1.52	1.01
HOL	1.06	1.09	1.08	0.98			
MCZ	1.04	1.04	0.86	0.87			
CF2	0.71	0.71	0.70	0.61			
CF3	0.73	0.77	0.96	0.89			

B	ST	SC	SE	SM	SK	SR	SF
BCI	0.94	0.94	1.14	1.14	0.92	0.74	1.59
FPR	0.96	0.95	1.15	1.15	0.97	0.87	1.49
HOL	0.99	0.94	1.21	1.14			
MCZ	1.01	1.04	0.91	0.90			
CF2	0.80	0.81	0.81	0.74			
CF3	0.86	0.95	1.17	1.07			

Table 5. Mean Normalized Recall *R*(1000)* for Sets A1−A3 and Normalized Recall for Set B, Using BCI and Unity Fingerprints

	BCI				UNITY			
A1−A3	S_T	S_E	S_F	S_R	S_T	S_E	S_F	S_R
mw1	0.89	1.27	1.60	0.26	0.88	1.21	1.58	0.31
mw2	1.04	0.84	0.39	1.57	1.19	0.99	0.22	1.76
full	0.91	0.92	0.96	1.10	0.99	0.97	0.86	1.29

	BCI				UNITY			
B	S_T	S_E	S_F	S_R	S_T	S_E	S_F	S_R
mw1	0.83	1.27	1.68	0.16	0.78	1.27	1.80	0.21
mw2	1.11	1.02	0.57	1.35	1.20	1.08	0.28	1.40
full	0.85	1.03	1.44	0.66	0.87	1.04	1.34	0.78

the Russell-Rao is rather a poor coefficient, with the combination of the Forbes coefficient and BCI fingerprints performing best. Of the other combinations the Euclidean distance does well, particularly with holograms or the 3D autocorrelation function but also giving above-average performance with the BCI or Unity bit-strings. A clue to the underlying reason for this is the observation that the variations from search to search in the recall associated with the Russell-Rao and Forbes results are rather larger than those of the coefficients. This suggests that the variation in the target structures is affecting the rankings obtained using these coefficients much more strongly than the others. It is known that molecular size, and hence the density of the bits set in a bit-string, can affect the performance of similarity coefficients in chemical database searching.[21,33,34] We have hence split the sets of target structures on the basis of molecular weight, as discussed in the next section.

Effect of Molecular Weight. For this part of the study each of the four test-sets was split equally at the distribution median into a low (mw1) and high molecular weight (mw2) fraction. Normalized recall analyses were then carried out on the resulting eight sets using BCI and Unity fingerprints and the Tanimoto, Euclidean, Forbes, and Russell-Rao coefficients, with the results shown in Table 5 (where "full" corresponds to the complete test-set, as in Table 3). Inspection of this table demonstrates clearly that the Forbes coefficient gives enhanced recall results in test-sets A1−A3 for the low molecular weight group, while the Russell-Rao coefficient gives a good performance for high molecular weight. Also, results for the Squared Euclidean coefficient fall relative to the Tanimoto for high molecular weight. These

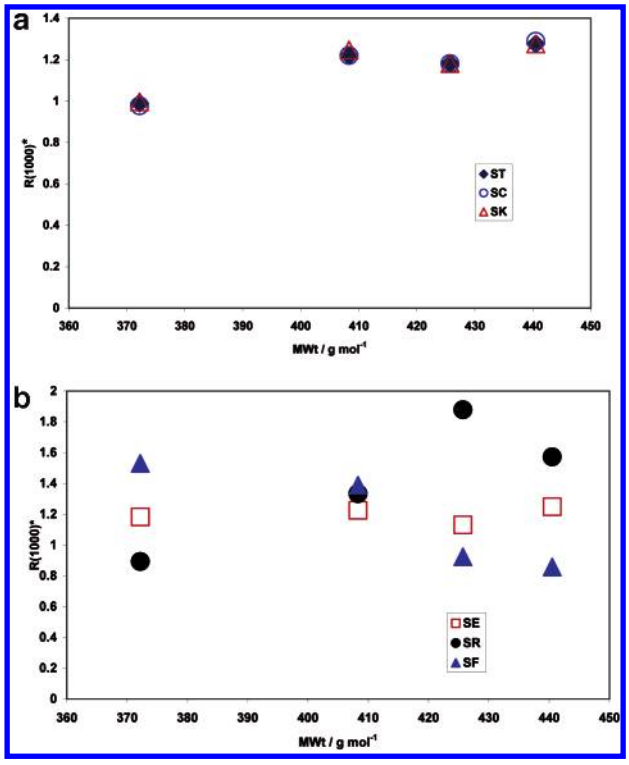

Figure 1. Relationship between molecular size of the target structures and cumulative recall for different similarity coefficients: (a) Tanimoto (ST), cosine (SC) and Kulczynski(2) (SK) coefficients and (b) Russell-Rao (SR), Euclidean (SE) and Forbes (SF) coefficients.

patterns are repeated for group B although the dominance of the Russell-Rao for the high molecular weight set is reduced.

Figure 1 plots the normalized recall from each of the Unity-based similarity coefficients against the average molecular weight for each of the sets A1, A2, A3, and B: these average weights are 439.5, 408.3, 425.7, and 372.2, respectively. The results of this analysis are split into two plots for clarity. Figure 1a shows that the Tanimoto, Cosine, and Kulczynski(2) coefficients behave almost identically, increasing gradually in value with molecular weight. In Figure 1b the Euclidean distance results are almost flat, while the Forbes coefficient results show a marked decrease, and the Russell-Rao coefficient results an equally marked increase, with respect to molecular weight. These observations explain the difference between the results for A1−A3 and for B: the latter set has a lower average molecular weight and the recall using Forbes is enhanced relative to the Russell-Rao, which is depressed. The Euclidean distance is also successful with set-B, giving results that are some 20% better than the Tanimoto coefficient with the hologram and fingerprint representations. Other dependencies can also be extracted the results but are less significant for the present argument and may also be statistically marginal; for example, the Molconn-Z and 3D autocorrelation functions appear to be more effective with low molecular weight target structures.

The dependencies can be rationalized by considering the bit-string overlap for a molecule *A*. Ranking always takes place with respect to the value of a coefficient at "perfect similarity", *S*(*A*,*A*), that is for comparison of a molecule with itself. For this situation, in the notation of eq 3, $b = a$ and $c = a$. For the Tanimoto, Cosine, and Kulzynski(2) coef-

ficients the value at perfect similarity is 1, while for the Squared Euclidean and the Mod coefficients the value is zero; in each of these cases it is a simple constant. However, for the Russell-Rao coefficient $S_R(A,A) = a$ and for the Forbes $S_F(A,A) = 1/a$; both forms introduce a dependence on the number of bits set for A and thus its size. For dissimilar molecules, the rank position depends on the difference $S(A,A) - S(A,B)$ or equivalently, when defined, the ratio $S(A,B)/S(A,A)$ of the coefficient values. For the Russell-Rao coefficient this ratio is c/a and notably does not depend on b, the number of bits set for molecule B (and thus roughly proportional to its size). If A is a small molecule, and B is a larger molecule that contains A as a substructure, then the Russell-Rao coefficient will identify B as identical with A irrespective of its size. By a parallel argument, if B contains a substructure closely similar to A, then it will be identified as a near neighbor even if it is much larger. The Russell-Rao tends to identify a relative abundance of larger molecules in a similarity search. The Forbes is unusual in that the ratio of its value with that at perfect similarity is c/b and does not depend on a. In this case if A is large, then the coefficient cannot distinguish between a comparably large molecule that is a substructure ($c = b$) and a much smaller molecule that is also a substructure since both give same ratio with perfect similarity $c/b = 1$. By a parallel argument, the Forbes coefficient does not distinguish well between large and small near-neighbors and thus tends to identify a relative abundance of smaller molecules when used for similarity searching. For all of the other coefficients in Table 1 the value at perfect similarity is a simple constant, and thus the ratios or differences that determine the rankings each depend on all of the quantities a, b, and c.

A recent modification to the Tanimoto coefficient[35] has been described that considers not just the bits that are set but also those that are unset and that might hence be expected to alleviate these size dependencies. The modified coefficient takes the form

$$S_{MT} = \left(\frac{2-p}{3}\right) S_T + \left(\frac{1+p}{3}\right) S_{T0} \qquad (7)$$

where S_{T0} is the Tanimoto coefficient for absent features and p is the proportion of set bits. Using $p = 0.333$ (which is the maximum density for the Unity strings used here) or $p = 0.176$ (which is the corresponding average density), this coefficient gave recall values that were insignificantly different from the standard Tanimoto results given in Table 3.

Data Fusion Results. As noted previously, the Kulczynki-(2) and Russell-Rao coefficients cannot be used in their continuous form, and S_M is monotonic with S_T for the two fingerprints. This leaves 28 sets of independent rankings that can be generated for the target structures in each of the four test-sets. The number of possible combinations that must be considered in fusing Q individual results is a rapidly increasing function, and it would be impractical to fuse all of our rankings. However, Wang and Wang[36] have suggested that the performance enhancements that might be expected from data fusion level off after the inclusion of three or four separate rankings, and we have hence chosen to fuse up to $Q = 6$ rankings at a time, giving a total of 62 possible combinations for investigation.

The choice of combinations is somewhat arbitrary: results for several of our choices are presented in Table 6. In this table, for each of the four sets of target compounds, the results of data fusion at rank 1000 are summarized for various combinations indicated under the column "comb" as follows. The similarity coefficients used are identified by their subscripts as used in the definitions given in Table 1: thus, T is Tanimoto, C is Cosine, E is Euclidian distance, K is Kulczynski(2), R is Russell-Rao, F is Forbes, and M is the new coefficient. These symbols are in boldface in the table if they form part of the best fusion combination, and they are boldface italic if they represent the best individual result. The number of sets making up the best combination is given in the column "Q". In the penultimate columns, "S_{max}" is the maximum individual recall result given as $R(1000)/10^{-2}$ and "F_{max}" is the maximum recall obtained using fusion, also given as $R(1000)/10^{-2}$. This is in boldface if it represents the best result in a group. The final column, "I%", gives the improvement as a percentage of the individual result.

In making these choices of fusion combination we have aimed to include the full range of results, but an obvious starting point is fusion of the most successful individual results. In the fusion combination labeled fuse1 (Table 6) we include the Russell-Rao pair that gave the best results for the sets A1−A3 along with the popular Tanimoto coefficient and the Squared Euclidean coefficient that proved useful for set B. Some improvement is obtained by combining the Russell-Rao results for groups A1 and A2, but group A3 showed no improvement over the individual result for Russell-Rao using Unity fingerprints. Group B showed some improvement by including squared Euclidean results with the Russell-Rao for this combination. The combination fuse2 mixes the Russell-Rao pair with results using molecular holograms, these being a different, but relatively successful representation. There is no unambiguous winner over the four active sets. Combinations fuse3 and fuse4 include all six representations using the Tanimoto and Squared Euclidean coefficients, respectively. Some successful blends are found but, again, there is no agreement between the four active groups. We have shown that the Russell-Rao and Forbes coefficient have opposing tendencies with respect to molecular weight, and these are combined in the set fuse5. There is no consistent evidence for any synergy between these methods. The Russell-Rao and Forbes are excluded from combination fuse6, which uses only Unity fingerprints and the four remaining nonmonotonic coefficients. There is little perceptible improvement seen for this grouping. The combination fuse7 includes the Forbes with BCI, which gave the best results for group B, along with the squared Euclidean results for the BCI and Unity fingerprints. Only one set, A2, produced successful fusion results in this group, individual results using the Forbes giving the best result for other sets. Finally set fuse8 is comparable with fuse6 in that it excludes the Russell-Rao and Forbes coefficients but in this case uses molecular holograms as the representation (i.e., a subset of the results for fuse2). In this case, fusion only gave any improvement for the set A1, while for the others a single result was the best.

The application of data fusion, using combination fuse2 from Table 6 with target-set A1, is shown in Figure 2a. Here, each point represents one search: thus there are six points when a combination of size-1 is used, 15 for size-2, 20 for

Table 6. Summary Results from Data Fusion Experiments[a]

target	comb	Q	BCI	FPR	HOL	MCZ	CF2	CF3	S_{max}	F_{max}	I%
A1	fuse1	2	T E **R**	T E ***R***					2.26	**2.55**	12.8
	fuse2	2	**R**	***R***	T C E M				2.26	**2.55**	12.8
	fuse3	2	T	***T***	**T**	T	T	T	1.88	1.93	2.7
	fuse4	4	**E**	***E***	**E**	E	E	**E**	1.83	1.93	2.7
	fuse5	2	F **R**	F ***R***					2.26	**2.55**	12.8
	fuse6	2		T ***C*** E **K**					1.97	1.99	1.0
	fuse7	1	E ***F***	E					1.72	1.72	0.0
	fuse8	2			**T *C*** E M				1.67	1.72	3.0
A2	fuse1	3	T **E** R	T **E** ***R***					2.41	2.47	2.0
	fuse2	2	R	***R***	T C **E** M				2.41	**2.68**	11.2
	fuse3	2	T	***T***	T	T	T	T	2.24	2.45	9.4
	fuse4	2	E	***E***	E	E	E	E	2.30	2.50	8.7
	fuse5	2	F R	F ***R***					2.26	2.55	12.8
	fuse6	1		T C E ***K***					2.37	2.37	0.0
	fuse7	2	**E** F	***E***					2.31	2.38	3.0
	fuse8	1			T C ***E*** M				2.30	2.30	0.0
A3	fuse1	1	T E R	T E ***R***					2.91	**2.91**	0.0
	fuse2	1	R	***R***	T C E M				2.91	**2.91**	0.0
	fuse3	2	**T**	***T***	T	T	T	T	1.85	1.94	4.9
	fuse4	3	**E**	***E***	**E**	E	E	E	1.79	1.86	3.9
	fuse5	1	F R	F ***R***					2.91	**2.91**	12.8
	fuse6	1		T ***C*** E K					1.95	1.95	0.0
	fuse7	1	E ***F***	E					1.82	1.82	0.0
	fuse8	1			T ***C*** E M				1.75	1.75	0.0
B	fuse1	3	T ***E*** R	T **E R**					1.82	1.90	4.4
	fuse2	3	R	**R**	T C ***E M***				1.90	1.94	2.1
	fuse3	3	T	**T**	***T***	T	T	T	1.55	1.68	8.4
	fuse4	2	**E**	E	**E**	E	E	E	1.90	2.02	6.3
	fuse5	1	F R	*F* R					2.55	**2.55**	0.0
	fuse6	1		T C ***E*** K					1.88	1.88	0.0
	fuse7	1	E ***F***	E					2.55	**2.55**	0.0
	fuse8	1			T C ***E*** M				1.90	1.90	0.0

[a] See text for column descriptions in this table.

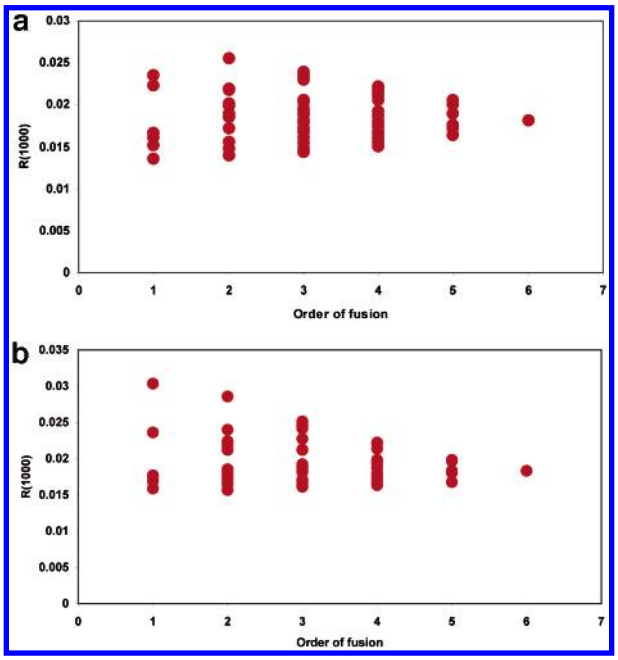

Figure 2. Effect of data fusion on cumulative recall for different numbers of fused rankings: (a) target-set A1 and (b) target-set A2.

size-3, 15 for size-4, 6 for size-5, and just one for size-6. The figure shows that it is possible to obtain an improvement in performance when two rankings are combined but that the inclusion of further sets leads to a reduction in recall. In all cases where there was any improvement using fusion, we found that this maximum occurred with two, three, or four rankings (see column *Q* in Table 6): there was never any advantage gained by the inclusion of further sets, a finding that is in accordance with the theoretical model proposed by Wang and Wang.[36] The same combination using a different set of actives may not always be successful, as shown in Figure 2b for fuse2 with set A2.

CONCLUSIONS

In this paper, we have evaluated a range of similarity measures when used for searching the natural product molecules in the *Dictionary of Natural Products* (DNP) database. Cumulative recall results for three independent, but similarly distributed, sets of target structures (sets A1−A3) chosen to represent the spectrum of pharmacological activity found in the DNP demonstrate a good level of reproducibility. For these sets of actives, the best single combination was the Russell-Rao coefficient and Unity fingerprints, and this fingerprint was also the best overall representation. The Tanimoto coefficient was no better at retrieving actives than any of the other coefficients, whereas this has been used very extensively and very successfully for similarity searching in synthetic databases such as the AIDS and MDDR files. A fourth set of actives (set B) was chosen in an effort to moderate the influence of those pharmacological activity groups that had many members. Here, the Forbes coefficient with both types of fingerprint gave the best recall by a considerable margin, the Euclidean distance also gave

relatively good results for most of the representations, and the Russell-Rao coefficient was poor. Analysis of the sizes of the molecules in the four sets of target structures suggests that the differences observed are due to inherent biases in the Russell-Rao and Forbes coefficient, the former preferentially retrieving large molecules and the latter preferentially retrieving small molecules.

Data fusion gave a worthwhile increase in recall for some combinations of rankings, but the results are far from consistent. If we ignore the Russell-Rao and the Forbes coefficients, then combinations of one of the fingerprints with molecular holograms using the squared Euclidean or Tanimoto coefficient would normally appear to give a small advantage over the individual results.

ACKNOWLEDGMENT

We thank Unilever Research for funding, and Barnard Chemical Information, the Royal Society, Tripos, and the Wolfson Foundation for hardware, software, and laboratory support. We thank Rita Azevedo, Scott Lusher, and Inge Muszynski for assistance with the preparation of the database and John Holliday, Naomie Salim, David Wilton, and Eleanor Gardiner for helpful comments on this work. The Krebs Institute for Biomolecular Research is a Biomolecular Sciences Centre of the Biotechnology and Biological Sciences Research Council.

APPENDIX

Initial experiments demonstrated that the Russell-Rao and Kulczynski(2) coefficients performed very poorly when used with continuous-data representations. This appendix reveals why and extends the reasoning to show that the Forbes coefficient also suffers from a similar problem. We also describe the rationale behind the new coefficient S_M.

The proofs are all based on the following observation. Using the definitions of a, b and c given earlier at (3), the inequalities

$$0 \le c \le a; \quad 0 \le c \le b \tag{A1}$$

must hold for binary data since c is just the overlap of a and b. However, for continuous representations c may take values greater than a or b and these inequalities are no longer always true. One counter-example is sufficient to prove this. Suppose that for structure B the representation is a simple multiple of that for structure A, i.e., $x_{jB} = \alpha x_{jA}$ where $0 < \alpha < \infty$. For molecular holograms, for example, if B was a dimer of the structure A we might have a close approximation of this scenario with $\alpha \sim 2$. Then, using the definitions of eq 3, the value of the union is $c = \sum_{j=1}^{m} \alpha x_{jA} x_{jA} = \alpha a$ and $b = \sum_{j=1}^{m} \alpha x_{jA} \alpha x_{jA} = \alpha^2 a$. For the range of possible α, c may now take on values $c > a$ or perhaps $c > b$.

The Russell-Rao is the simplest similarity coefficient (Table 2) based on the overlap normalized with the vector length m. For perfect similarity molecule B is the same as A so that, using the definition given in Table 1 and dropping the m which is superfluous for descriptors of the same length

$$S_R(A,B) = c; S_R(A,A) = a. \tag{A2}$$

Then, for binary data with the inequality Eq. (A1) we have $0 \le S_R(A,B) \le S_R(A,A)$and ranking of the coefficient values gives a sensible measure of similarity. However, for continuous variables, where the inequality Eq. (A1) is not obeyed, $S_R(A,B)$ may be greater or less than the value that represents perfect similarity: $S_R(A,A) = a$. Ranking with respect to the magnitude of $S_R(A,A)$ becomes impossible and the Russell-Rao factor cannot, therefore, be considered a similarity coefficient for continuous representations. In fact, the full range of the Russell-Rao factor for continuous representations that admit negative elements is $-\infty < S_R(A,B) < +\infty$ while the value of $S_R(A,A)$ must always be positive.

A related valid measure appears to be the absolute difference $|S_R(A,B) - S_R(A,A)|$ which is zero for maximum similarity. To rank descending we could use $-|S_R(A,B) - S_R(A,A)|$. This would be a new coefficient $S_{NEW}(A,B) = -|c - a|$ but turns out to be biased toward small molecules. If many values of x_{jB} are zero, but the nonzero values are contained within a larger x_{jA}, then $\sum_{j=1}^{m} x_{jA} x_{jB} = \sum_{j=1}^{m} (x_{jA})^2$ and $S(A,B) = 0$ even though molecules A and B are quite dissimilar. Attempting to overcome this problem we have used instead

$$S_M(A,B) = -\left|\frac{c}{m} - \frac{a}{m}\right| - \left|\frac{c}{m} - \frac{b}{m}\right| \tag{A3}$$

This new coefficient is comparable with Squared Euclidean distance

$$S_E(A,B) = \frac{a + b - 2c}{m} = \frac{(a - c) + (b - c)}{m}. \tag{A4}$$

Indeed, for bit-strings, we have found it to be monotonic with the mean square Euclidian distance which in turn is monotonic with the Mean Manhatten distance. However, for the more general continuous vectors S_M does give different rankings and we have therefore retained it.

The Kulczynski(2) coefficient (Table 1), S_K, suffers from a similar problem when used with continuous variables. For binary data with the inequality eq A1 it is easily shown that $0 \le S_K(A,B) \le 1$ and ranking is possible. However, for more general vectors this is not true and c may take values greater than a or b. Suppose again that $x_{jB} = \alpha x_{jA}$ where $0 < \alpha < \infty$ so that $c = \alpha a$and $b = \alpha^2 a$. Then

$$S_K(A,B) = \frac{1}{2}\left[\frac{\alpha a}{a} + \frac{\alpha a}{\alpha^2 a}\right] = \frac{1}{2}\left[\alpha + \frac{1}{\alpha}\right] \tag{A5}$$

and $S_K(A,B) \ge 1$ for $\alpha \ne 1$. But, as for binary representations, other differences between A and B may lead to $0 \le S_K(A,B) \le 1$. Hence $S_K(A,B)$ may take values either side of the value that represents perfect similarity: $S_K(A,A) = 1$ and, like the Russell-Rao, the Kulczynski(2) is not a valid similarity coefficient for continuous variables. In fact, the range for continuous variables that admit negative elements is $-\infty < S_K(A,B) < +\infty$.

For the Forbes coefficient (Table 2) we can again drop the dependence on m, which is superfluous for comparisons of descriptors of the same length. The value at perfect similarity is then given by

$$S_F(A,A) = \frac{a}{a^2} = \frac{1}{a} \tag{A6}$$

For binary data, with the inequality eq A1 we must have $0 \leq S_F(A,B) \leq S_F(A,A)$ for dissimilar structures since $c/ab < 1/a$. However, for continuous representations that break eq A1

$$S_F(A,B) = \frac{\alpha a}{\alpha^2 a^2} = \frac{1}{\alpha a} \tag{A7}$$

For the range of possible α, $S_F(A,B)$ can therefore take on values greater or less than the value representing perfect similarity, $S_F(A,A)$. In this sense ranking with respect to the magnitude of $S_F(A,A)$ can bear no relation to similarity. The Forbes is not a similarity coefficient for continuous representations in the sense that its magnitude alone does not give a measure of similarity.

Applying the same procedure to the other coefficients used here shows that a necessary condition for a successful continuous representation coefficient is that it should either be independent of α, as is the Cosine coefficient, or have an extremum at $\alpha = 1$, the value that represents perfect similarity. For example, the Tanimoto coefficient gives

$$S_T(A,B) = \frac{\alpha}{\alpha^2 - \alpha + 1} \quad \text{and}$$

$$\frac{d}{d\alpha} S_T(A,B) = \frac{1 - \alpha^2}{(\alpha^2 - \alpha + 1)^2} \tag{A8}$$

However, as we have seen for the Kulczynski(2) coefficient, one must also show that this extremum is of the same type (maximum or minimum) as the variation arising from binary overlap.

REFERENCES AND NOTES

(1) Walters, W. P.; Stahl, M. T.; Murcko, M. A. Virtual Screening − an Overview. *Drug Discov. Today* **1998**, *3*, 160−178.
(2) *Virtual Screening for Bioactive Molecules*; Bohm, H.-J., Schneider, G., Eds.; Wiley-VCH: Weinheim, 2000.
(3) *Concepts and Applications of Molecular Similarity*; Johnson, M. A., Maggiora, G. M., Eds.; Wiley: New York, 1990.
(4) *Molecular Similarity in Drug Design*; Dean, P. M., Ed.; Chapman and Hall: Glasgow, 1994.
(5) Willett, P.; Barnard, J. M.; Downs, G. M. Chemical Similarity Searching. *J. Chem. Inf. Comput. Sci.* **1998**, *38*, 983−996.
(6) Cragg, G. M. Natural Products in Drug Discovery and Development. *J. Nat. Prod.* **1997**, *60*, 52−60.
(7) Lawrence, R. N. Rediscovering Natural Product Biodiversity. *Drug Discovery Today* **1999**, *4*, 449−451.
(8) Harvey, A. Strategies for Discovering Drugs from Previously Unexplored Natural Products. *Drug Discovery Today* **2000**, *5*, 294−300.
(9) Henkel, T.; Brunne, R. M.; Müller, H.; Reichel, F. Statistical Investigation into the Structural Complementarity of Natural Products and Synthetic Compounds. *Angew. Chem., Int. Ed. Engl.* **1999**, *38*, 643−647.
(10) Stahura, F. L.; Godden, J. W.; Xue, L.; Bajorath, J. Distinguishing between Natural Products and Synthetic Molecules by Descriptor Shannon Entropy Analysis and Binary QSAR Calculations. *J. Chem. Inf. Comput. Sci.* **2000**, *40*, 1245−1252.
(11) Godden, J. W.; Bajorath, J. Chemical Descriptors with Distinct Levels of Information Content and Varying Sensitivity to Differences between Selected Compound Databases Identified by SE-DSE Analysis. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 87−93.
(12) Lee, M.-L.; Schneider, G. Scaffold Architecture and Pharmacophoric Properties of Natural Products and Trade Drugs: Application in the Design of Natural Product-Based Combinatorial Libraries. *J. Comb. Chem.* **2001**, *3*, 284−289.
(13) Sheridan, R. P.; Singh, S. B.; Fluder, E. M.; Kearsley, S. K. Protocols for Bridging the Peptide to Nonpeptide Gap in Topological Similarity Searches. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1395−1406.
(14) The *Dictionary of Natural Products* database is available from Chapman & Hall/CRC at URL http://www.crcpress.com/.
(15) The *Bioactive Natural Product* database is available from Szenzor Management Consulting Company, H-1134 Budapest, Lehel u. 11, Hungary.
(16) The *BioScreenNP* database is available from InterBioScreen at URL http://www.ibscreen.com/natural.shtml
(17) Lei, J.; Zhou, J. A Marine Natural Product Database. *J. Chem. Inf. Comput. Sci.* **2002**, *42*, 742−748.
(18) Edgar, S. J.; Holliday, J. D.; Willett, P. Effectiveness of Retrieval in Similarity Searches of Chemical Databases: a Review of Performance Measures. *J. Mol. Graph. Mod.* **2000**, *18*, 343−357.
(19) Ginn, C. M. R.; Willett, P.; Bradshaw, J. Combination of Molecular Similarity Measures Using Data Fusion. *Perspec. Drug Discov. Design* **2000**, *20*, 1−16.
(20) Holliday, J. D.; Hu, C.-Y.; Willett, P. Grouping of Coefficients for the Calculation of Intermolecular Similarity and Dissimilarity Using 2D Fragment Bit-Strings. *Combin. Chem. High-Through. Screening* **2002**, *5*, 155−166.
(21) Flower, D. R. On the Properties of Bit String Based Measures of Chemical Similarity. *J. Chem. Inf. Comput. Sci.* **1988**, *38*, 379−386.
(22) The Unity software packages are available from Tripos Inc. at URL http://www.tripos.com.
(23) The BCI software is available from Barnard Chemical Information Ltd. at URL http://www.bci.gb.com/.
(24) The Molconn-Z software is available from EduSoft at URL http://www.eslc.vabiotech.com/.
(25) Broto, P.; Moreau, G.; Vandycke, C. Molecular Structures: Perception, Autocorrelation Descriptor and SAR Studies. *Eur. J. Med. Chem.* **1984**, *19*, 66−70.
(26) Schuur, J. H.; Selzer, P.; Gasteiger, J. The Coding of the Three-Dimensional Structure of Molecules by Molecular Transforms and Its Application to Structure-Spectra Correlations and Studies of Biological Activity. *J. Chem. Inf. Comput. Sci.* **1996**, *36*, 334−344.
(27) Ellis, D.; Furner-Hines, J.; Willett, P. Measuring the Degree of Similarity between Objects in Text Retrieval Systems. *Perspect. Inf. Manag.* **1994**, *3*, 128−149.
(28) Charifson, 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.
(29) Klein, L. A. *Sensor and Data Fusion Concepts and Applications*, 2nd ed.; SPIE The International Society for Optical Engineering: 1999.
(30) Voigt, J. H.; Bienfait, B.; Wang, S.; Nicklaus, M. C. Comparison of the NCI Open Database with Seven Large Chemical Structural Databases. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 702−712.
(31) Bajorath, J.; Chemoinformatics Methods for Systematic Comparison of Molecules from Natural and Synthetic Sources and Design of Hybrid Libraries. *J. Comput.-Aided Mol. Des.* **2002**, *16*, 431−439.
(32) Feher, M.; Schmidt, J. M. *J. Chem. Inf. Comput. Sci.* In press.
(33) Dixon, S. L.; Koehler, R. T. The Hidden Component of Size in Two-Dimensional Fragment Descriptors: Side Effects on Sampling in Bioactive Libraries. *J. Med. Chem.* **1999**, *42*, 2887−2900.
(34) Salim, N.; Holliday, J. D.; Willett, P. Combination of Fingerprint-based Similarity Coefficients Using Data Fusion. *J. Chem. Inf. Comput. Sci.* In press.
(35) Fligner, M. A.; Verducci, J. S. A Modification of the Jaccard-Tanimoto Similarity Index for Diverse Selection of Chemical Compounds Using Binary Strings. *Technometrics* **2002**, *44*, 110−119.
(36) 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.

CI025591M