Similarity Searching in Files of Three-Dimensional Chemical Structures: Comparison of Fragment-Based Measures of Shape Similarity†

Peter A. Bath, Andrew R. Poirrette, and Peter Willett*

Krebs Institute for Biomolecular Research and Department of Information Studies, University of Sheffield, Sheffield S10 2TN, U.K.

Frank H. Allen

Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, U.K.

Received June 22, 1993⊕

This paper compares several fragment-based measures that can be used to quantify the degree of similarity between pairs of three-dimensional (3-D) chemical structures. The fragments that are considered contain two, three, or four atoms and encode distance information, angular information or both but do not involve chemical information such as atomic type. The effectiveness of the various measures is compared using eight literature datasets for which biological-activity data and calculated 3-D structures are available, and a set of carbohydrate structures from the Cambridge Structural Database that have been classified into eight distinct groups. Similarity searches on these datasets suggest that the four-atom fragments are the most effective.

1. INTRODUCTION

The advent of techniques for the rapid generation of low-energy, three-dimensional (3-D) structures for small molecules has encouraged the development of substructure-searching methods for databases of 3-D structures.[1-3] We are currently engaged in a project to identify measures of structural resemblance that will permit similarity searching to be carried out in large 3-D databases.[4] The *similar property principle* of Johnson and Maggiora[5] states that structurally-similar molecules exhibit similar properties. Thus, given a target molecule that is known to be active in a biological test system of interest, a 3-D similarity search may serve to identify other molecules that exhibit the same activity and that may hence act as leads in the discovery of new drugs or pesticides.

There are many ways of calculating the degree of similarity between a pair of 3-D molecules. A recent review of work in this area[6] notes that the available similarity measures differ drastically in their computational requirements. In particular, the many measures that are based on quantum-mechanical properties or on the precise alignment of molecular fields are far too slow when a target compound needs to be searched against a database containing tens or hundreds of thousands of structures. In this paper, we describe and evaluate procedures that can be used to search such large databases for molecules that are similar in shape to a user-defined target structure. The similarity measures make use only of geometrical information, i.e., distances and angles between the constituent atoms of the pairs of molecules that are being compared, and do not involve chemical information such as atomic type, polarity, or hydrophobicity, etc. Instead, an attempt is made to describe the shape of each molecule using only information that can be rapidly generated from a 3-D structure (this being a distance matrix resulting from the use of the CONCORD program[7] or from experimental X-ray data contained in the Cambridge Structural Database[8]). The similarity measures considered here use fragments containing pairs, triplets or quadruplets of atoms, and the measures are thus analogous to those used for similarity searching in databases of two-dimensional (2-D) chemical structures.[5,9,10] The fragment-based nature means that the measures reported here are not appropriate for the production of molecular alignments, which may be produced from more-detailed, but more time-consuming, database-searching procedures, e.g., the atom-mapping and SPERM measures described by Pepperrell *et al.*[11] and by van Geerestein *et al.*,[12] respectively.

Since the earliest studies of 3-D searching,[13,14] the bulk of the work that has been carried out has used interatomic distance information as the basis for substructure searching, and the last few years have seen such information starting to be used also for 3-D similarity searching.[11,12,15,16] In this paper, we describe 3-D similarity measures that are based on angular information, specifically on the sets of four atoms that comprise a torsion angle, and then compare the results of similarity searches using these measures with the results that are obtained using several distance-based similarity measures.

2. ANGLE-BASED 3-D SIMILARITY MEASURES

2.1. Generalized Torsion Angles. Given a set of four atoms *ABCD*, a torsion angle, τ, is the angle between the two three-atom planes, *ABC* and *BCD*. It describes the twist of the vector *AB* relative to the vector *CD* when viewed along the vector *BC*. A torsion angle is very well suited to 3-D searching, since it provides 3-D information, unlike a valence angle, which is 2-D, or a bond, which is 1-D. Sets of torsion angles are commonly used to define conformation and configuration in chemistry; for example, the terms *boat*, *chair*, *extended*, etc., all refer to a specific sequence of torsion angle values that characterize a particular ring shape or side-chain conformation. The importance of torsion angles has led to attempts to characterize them by means of 2-D similarity descriptors.[17]

A torsion angle is normally defined in terms of sets of four connected atoms, *ABCD*. However, there is no reason why the definition should be so restricted, and Poirrette *et al.*[19] have discussed the concept of a *generalized torsion angle*, in which it is not mandatory for the atoms to be connected. Let the presence or absence of a bond between two of the four atoms comprising a torsion angle be characterized by the letters 'B' or 'N', respectively. A conventional, fully-bonded torsion

* To whom all correspondence should be addressed.
† Presented at the Third International Conference on Chemical Structures, Noordwijkerhout, The Netherlands, June 6-10, 1993.
⊕ Abstract published in *Advance ACS Abstracts*, January 15, 1994.

0095-2338/94/1634-0141$04.50/0

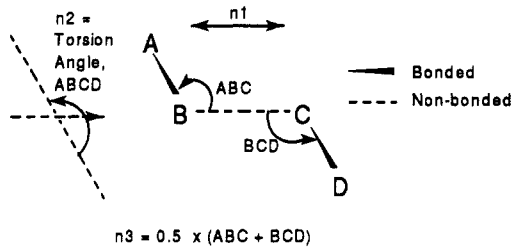

Figure 1. Description of BNB fragment.

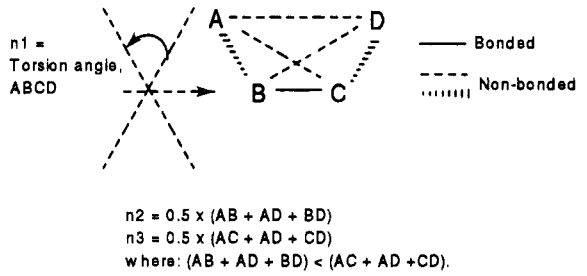

Figure 2. Description of a NBN fragment.

angle (such as that defined in the previous paragraph) can then be described as a *BBB torsion*; in this paper, we shall discuss the use of *BNB torsions* and *NBN torsions* for the calculation of 3-D similarity. A BNB torsion is one in which there are bonds between the two outer pairs of atoms forming the four-atom fragment, i.e., *AB* and *CD*, but in which there is no bond between the central pair of atoms, *BC* (or, for that matter, between *A* and *D*). An NBN torsion is one in which only the two central atoms of the four-atom fragment are bonded, i.e., *BC* in our example. Thus, if we denote the presence or absence of a bond by the symbols '–' and '≈', respectively, then BBB, BNB and NBN torsions are exemplified by the fragments $A - B - C - D$, $A - B \approx C - D$ and $A \approx B - C \approx D$, respectively. These, and other, types of generalized torsion angle have been evaluated for the creation of force fields by Weiner *et al.*[18] and for 3-D substructure searching by Poirrette *et al.*,[19] who have also carried out comparable experiments using a range of *generalized valence angles*.[20]

Generalized torsion angles seem to have been first used by Bartlett *et al.* in the CAVEAT program,[21] which uses BNB torsions as screens in 3-D substructure searches for compounds that have bonded-vectors satisfying specific inter-vector angle relationships. A simplified form of the CAVEAT screens is used in the experimental 3-D searching system that is under development at Chemical Abstracts Service.[22,23] There are very many ways in which we can characterize the geometry of a four-atom fragment *ABCD*. Two such ways are shown in Figures 1 and 2 and discussed further in the remainder of this section; other torsion-angle descriptors will be investigated in future work.

2.2. BNB Measure. Similarity searching in a database of 2-D chemical structures is effected using the fragment screens that are conventionally used for substructure searching. Given a target molecule, *T*, and a dataset molecule, *D*, the similarity, *SIM*(*T*,*D*), is calculated by means of an association coefficient, normally the Tanimoto coefficient.[9] If *T* and *D* contain $N(T)$ and $N(D)$ screens, respectively, and *COMMON* of these are in common, then this coefficient is given by

$$SIM(T,D) = \frac{COMMON}{N(T) + N(D) - COMMON}$$

We now describe how a comparable similarity measure can be calculated by characterizing a target structure and a dataset structure by sets of codes that describe their constituent BNB torsion angles.

We have chosen to characterize a BNB torsion, $A - B \approx C - D$, by the integer code

$$n_1 + 180 \times n_2 + 180^2 \times n_3$$

where: n_1 is the arithmetic mean of the two intervector angles *ABC* and *BDC*; n_2 is the absolute value of the torsion angle τ; and n_3 is the interatomic distance $B - C$ in Å after multiplying by 100. Both of the angular components, n_1 and n_2, are rounded to the nearest degree. They are absolute torsion angles, i.e., $0 \le n_1, n_2 \le 180$, and n_3 (before multiplication by 100) is typically in the range $2 \le n_3 \le 25$. The use of the 180-based coefficients ensures that different combinations of values n_1 n_2, n_3 will result in different values for the resulting codes. This fragment description is shown in Figure 1.

A molecule can be characterized by generating the integer codes for all of the possible BNB fragments in that molecule, and then taking the set of distinct codes. The sets of codes describing two molecules can then be used for the calculation of a Tanimoto measure analogous to that defined above for 2-D fragment data. Assume that a target molecule has $N(T)$ BNB codes and a dataset molecule has $N(D)$ such codes, *COMMON* of which match with the target codes; then the similarity, *SIM*(*T*,*D*), will again be given by the expression above.

2.3. NBN Measure. The NBN measure is calculated in a similar manner, by generating integer codes for all of the possible NBN fragments in a molecule. The code here is

$$n_1 + 10 \times n_2 + 1000 \times n_3$$

where: n_1 is the integer obtained by rounding the absolute value of the torsion angle *ABCD* after dividing by 20 (this corresponding to the allowable tolerance in degrees if two fragments are to be matched); and n_2 and n_3 ($n_2 \le n_3$) are the integers obtained by rounding the sums of the triangle sides *ABD* and *ACD* after multiplying by 0.5 (this corresponding to the allowable tolerance in Å if two fragments are to be matched). The values for n_1 are, necessarily, in the range $0 \le n_1 \le 9$, while those for n_2 and n_3 were found to be in the range $0 < n_2, n_3 \le 99$. This fragment definition is shown in Figure 2.

No account was taken of the central separation in a fragment, i.e., the distance $B - C$, since this varies little in comparison with the other distances in the quadrilateral, *ABCD*, that defines a torsion angle. As with the BNB measure, a molecule was characterized by the set of its distinct integer codes, and pairs of these sets of codes were used to calculate a Tanimoto coefficient.

3. DISTANCE-BASED 3-D SIMILARITY MEASURES

In this section, we summarize several approaches that have been suggested previously for 3-D similarity searching using interatomic distance information.

3.1. Bemis–Kuntz Measure. Bemis and Kuntz have described molecular shape in terms of the distances in all distinct sets of three non-hydrogen atoms that can be generated from a 3-D structure.[15] The distances are used to generate a distribution that summarizes the frequencies of occurrence of these three-atom sets, and a hash-coding procedure is then used to produce a single numeric descriptor that provides a highly efficient way of characterizing molecular shape. Unfortunately, the use of a hash code means that similarities can be readily calculated only between molecules that have the same molecular formula,[15] which reduces the attractiveness

of this approach as a general mechanism for 3-D similarity searching. However, this restriction can be removed if the frequency distributions themselves, rather than the hash codes generated from them, are used to calculate inter-molecular similarities.

The atoms of a 3-D structure are represented by an interatomic distance matrix. Each distinct subset of three atoms is then generated from this structure: for a molecule containing N non-hydrogen atoms, there are $N(N-1)(N-2)/6$ such subsets. As in Bemis and Kuntz's original experiments, a 64-element integer frequency distribution is created and all of the elements initialized by setting them to zero. If the three-atomic distances for a three-atom subset are n_1, n_2 and n_3, then an increment of one is made to that element of the distribution which corresponds to the sum

$$n_1^2 + n_2^2 + n_3^2$$

The resulting distribution provides a simple, fixed-format characterization of the shape of the dataset molecule under consideration, and a dataset structure's similarity with a 3-D target molecule can then be calculated by the degree of resemblance between the two structures' frequency distributions. The similarity coefficient used here is again the Tanimoto Coefficient. Let *T_FD* and *D_FD* be the frequency distributions for the target structure and a dataset structure, respectively, and let B be the number of elements in each frequency distribution. Then the Tanimoto coefficient was calculated as

$$\frac{\sum T\_FD(I) \times D\_FD(I)}{\sum T\_FD(I)^2 + \sum D\_FD(I)^2 - \sum T\_FD(I) \times D\_FD(I)}$$

where all of the summations are for $1 \leq I \leq B$.

Bemis and Kuntz discussed their approach only in the context of sets of three atoms.[15] However, it can be generalized to describe 3-D molecules in terms of sets containing any number of atoms, and we have evaluated their approach using sets of two, three and four atoms; in what follows, we shall refer to these similarity measures as BK-2, BK-3 and BK-4, respectively. In general, a molecule containing N non-hydrogen atoms will produce a total of

$$\frac{N!}{M!(N-M)!}$$

subsets of M atoms. For consistency with Bemis and Kuntz's original description, we have used the same ranges of values for each of the elements in the frequency distribution, with the following two exceptions: in the case of the BK-2 measure, one element was added at the lower end of the distribution to encompass sums of squared distances that totalled less than 6.0; in the case of the BK-4 measure, each of the sums was halved before the distribution was incremented to avoid overpopulating the high-valued end of the distribution. Thus, the ranges for the BK-2 measure, for example, are

$$\leq 5.99,\ 6.00\text{–}6.99,\ 7.00\text{–}7.99 \ldots > 871.00$$

as discussed by Bemis and Kuntz.[15] A more systematic procedure would be to generate ranges corresponding to the actual distribution of squared-distance sums in a particular dataset, using a partitioning algorithm such as that described by Cringean *et al.*[25] When the BK-2 measure is used, the approach is closely related to one of the similarity measures described by Pepperrell and Willett,[24] which they called the *distance-distribution* measure.

3.2. Lederle Measure. Nilakantan *et al.* at Lederle Laboratories have described a similarity measure that is closely related to the BK-3 measure, since it also is based on the three interatomic distances that are generated from a set of three non-hydrogen atoms.[16] Their procedure is in two stages. In the first stage, each set of three atoms, which they refer to as an *atom-triplet*, is used to generate an integer that characterizes the interatomic distances within the triplet. The integers resulting from the first stage are then used as the seeds for a hashing procedure that addresses individual bits in a bit-string representation of molecular shape. In our experiments, we have used just the first stage, *viz* the generation of the integer descriptors.

Given an atom-triplet, the three interatomic distances, n_1, n_2 and n_3, are calculated and sorted into increasing order. Assume that the distances are such that

$$n_1 \leq n_2 \leq n_3$$

then the integer code is calculated as

$$n_1 + 1000 \times n_2 + 1000000 \times n_3$$

A target structure or a dataset structure is represented by its list of codes, and the similarity between them calculated by a simple comparison of the list of codes using the Dice coefficient (an association measure that gives rankings that are monotonic with the Tanimoto coefficient). In addition to using the Dice coefficient,

$$\frac{2 \times COMMON}{N(T) + N(D)}$$

(where *COMMON*, $N(T)$ and $N(D)$ refer to sets of the triplet descriptors described above), Nilakantan *et al.* also reported experiments that used an alternative similarity measure, the *Asymmetric coefficient*,[16] which is given by

$$\frac{COMMON}{N(T)}$$

In what follows, we shall refer to these two similarity measures as LT and LA (for Lederle Tanimoto and Lederle Asymmetric).

4. EXPERIMENTAL DETAILS

4.1. QSAR Data Sets. An inherent assumption in the practical use of similarity methods is that structurally-similar molecules are likely to possess similar properties, and thus that the effectiveness of a similarity procedure may be assessed by determining the extent to which this occurs in practice. Following previous studies in Sheffield,[9,24] the effectiveness of each measure described in Sections 2 and 3 was compared using a 'leave-one-out' approach on datasets for which both 3-D structural and biological-activity data were available.

An active molecule in a dataset was selected and its similarity calculated with each of the other molecules in that dataset, using one of the similarity measures. The molecules were then ranked in decreasing order of the calculated similarity, and cut-offs applied to retrieve some fixed number of the top-ranked compounds. These molecules were then checked to determine whether they were active or inactive in the particular biological test system associated with that dataset. The overall utility of a similarity measure was calculated by taking the mean number of active molecules at the cut-off position, when averaged over all of the active compounds that had been used to generate a ranking. Thus, if there were ACT actives in a dataset and if the use of the I-th active as the target identified AI actives above the chosen cut-off, which

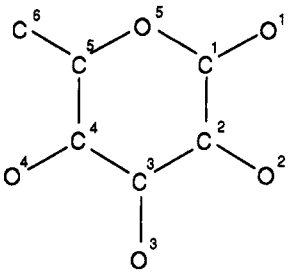

Figure 3. Hydrogen-depleted 2-D diagram of a 6-deoxyhexopyranose carbohydrate.

was taken to be 5, 10 or 20 compounds in the experiments reported below, then the overall effectiveness of the current similarity measure was given by

$$\frac{1}{ACT}\sum_{I=1}^{ACT} AI$$

The relative merits of the various similarity measures were then assessed on the basis of these mean numbers of actives.

In all, we have used this approach with eight datasets from the QSAR literature, as follows:

A: 209 9-anilinoacridines of which 150 showed anti-tumor activity and 59 were inactive.[26]

B: 141 aromatic amines of which 98 were carcinogenic and 43 were non-carcinogenic.[27]

C: 147 barbiturates for which duration of action data were available. After previous studies by Stuper and Jurs[28] the 110 compounds with a duration of action of less than 200 minutes were classified as inactives and the remaining 37 as actives.

D: 112 nitrobenzenes of which 53 were musk odorants and 59 not musk odorants.[29]

E: 109 cyclic nitrogeneous compounds of which 63 were mutagenic and 46 were non-mutagenic.[30]

F: 145 nitrosamines of which 112 are carcinogenic and 33 non-carcinogenic.[31]

G: 196 heterogeneous compounds of which 121 are carcinogenic and 75 non-carcinogenic.[32]

H: 113 steroids of which 69 were potent and 44 non-potent in the Mackenzie-Stoughton human vasoconstrictor assay for anti-inflammatory behavior.[33]

CONCORD 3-D structures were generated for each molecule and each dataset, and these structures then used for the generation of the various types of structural feature that have been described in Sections 2 and 3 above. The datasets are not ideal for the present purpose given that shape similarity is not necessarily the main factor in determining activity in each case: other factors (such as lipophilicity or electronic effects) may be of importance. Again, we have used just the CONCORD-derived structures, and this may well not correspond to the biologically-active conformation. However, the datasets do cover a range of types of activity, include both homogeneous and heterogeneous sets of compounds, and have also been used in several previous studies in Sheffield (see, e.g., ref 24), thus facilitating the comparison of the results obtained here with those obtained previously as is discussed in Section 5.3 below.

4.2. Carbohydrate Data Set. The results that were obtained with the QSAR datasets were then compared with those obtained with a further, highly homogeneous dataset comprising the orthogonalized X-ray coordinates of 249 6-deoxyhexopyranose carbohydrates, which have the general form shown in Figure 3. Each of these structures contains 11 non-hydrogen atoms (six carbons and five oxygens) and each ring carbon is a stereogenic center. There are thus 2^5 (=32) possible stereoisomers existing as 16 enantiomeric pairs. Further, since the shape descriptors used in this paper do not distinguish atomic types, the structure shown in Figure 3 becomes symmetric about the O^5–C^3 vector, and the number of possible shape groups that we can describe uniquely is reduced to 8. Note that the hexose ring adopts a $^4C_1$ conformation in all of the structures studied here.

The 249 entries in this dataset have recently been classified into 14 of the possible 32 stereochemical (shape) classes using numerical clustering methods based on torsional dissimilarity coefficients.[34] After taking enantiomers and topological symmetry into account, all 8 of the possible shape groups mentioned above are represented in the dataset. These shape groups are dictated entirely by configurational, rather than conformational, effects and, hence, represent discrete and distinct shape classes. The availability of this 8-part preclassification of the dataset represents a major difference from the two-class activity scale that characterizes the QSAR datasets.

A second difference lies in the method used to evaluate the rankings that result from the application of each of the similarity measures. As before, the dataset was ranked in decreasing similarity order with a target molecule, with the expectation that the nearest-neighbor molecules near to the top of the ranking would come from the same class as the target molecule. If molecules were ranked at random, then one would expect that the median-ranked compound in the target-molecule's class would occur at rank 125 (for a dataset containing 249 molecules), irrespective of the number of molecules in that class. It is hence possible to evaluate the effectiveness of a similarity measure by seeing how far the median-ranked molecule comes above position 125. Let there be *CLASS* members of a particular class of carbohydrates, and let *MI* be the median-ranked position for the *I*-th member of that class; then, by analogy with the performance measure described for the previous eight datasets, the mean effectiveness (when averaged over similarity searches using each of the *CLASS* members as the target compound) is given by

$$\frac{1}{CLASS}\sum_{I=1}^{CLASS} MI$$

Thus, the smaller the effectiveness value calculated in this way the better. The best possible result for a similarity measure used on a class containing *CLASS* compounds is 1 + mod(*CLASS*/2) (for odd *CLASS*) or the mean of 1 + mod(*CLASS*/2) and mod(*CLASS*/2) (for even *CLASS*).

We have noted above that there are eight classes of structure in this dataset; however, searches were carried out using only the members of the four most-frequently occurring classes to ensure some degree of precision in the results; these classes contained 95, 74, 28 and 19 members.

The very high degree of homogeneity in this dataset provides an excellent test of the discriminatory abilities of the various similarity measures. However, it does mean that the intermolecular similarities will all tend to be large. In particular, the 3-D structure of the central ring is known to be very similar in all of the carbohydrates, with what differences there are amongst the molecules in the four classes being almost exclusively due to the orientation of the ring substituents. The most-discriminating fragments are thus expected to be those that lie on the exterior of each molecule, and it was hence decided, in the BNB experiments, to use just those fragments that were associated with the largest interatomic distances, the n_3 term in the expression given previously in Section 2.2.

Table 1. Mean Number of Actives in the Top-5 Molecules of the QSAR Data Sets When Ranked Using a Range of Similarity Measures

			Bemis–Kuntz			Lederle		
data set	BNB	NBN	BK-2	BK-3	BK-4	LT	LA	ID
A	4.47	4.59	4.32	4.22	4.33	4.47	4.14	4.43
B	3.92	3.91	4.03	4.10	3.93	4.03	3.16	4.27
C	2.92	2.89	3.05	2.78	2.81	2.92	1.62	2.68
D	3.68	3.60	3.47	3.26	3.30	3.57	2.60	3.62
E	4.33	4.38	4.33	4.19	3.87	4.30	3.49	4.40
F	4.25	4.21	4.22	4.24	4.23	4.21	3.35	4.46
G	3.94	3.85	3.91	3.79	3.83	3.82	3.26	4.01
H	4.25	4.09	4.05	4.17	4.22	4.16	4.45	4.26

Table 2. Mean Number of Actives in the Top-10 Molecules of the QSAR Data Sets When Ranked Using a Range of Similarity Measures

			Bemis–Kuntz			Lederle		
data set	BNB	NBN	BK-2	BK-3	BK-4	LT	LA	ID
A	8.75	8.89	8.44	8.24	8.38	8.75	8.45	8.33
B	7.67	7.73	7.82	7.72	7.81	7.66	6.52	7.94
C	4.70	4.95	5.54	5.14	4.97	5.03	2.54	4.86
D	6.38	6.23	6.17	6.47	6.00	6.45	5.19	6.30
E	7.95	8.08	7.92	7.57	6.97	7.32	6.60	8.38
F	8.39	8.19	8.38	8.30	8.12	8.17	6.67	8.64
G	7.38	7.45	7.17	7.15	7.07	7.22	6.86	7.45
H	7.97	8.07	7.94	7.75	7.83	7.84	8.90	7.79

Table 3. Mean Number of Actives in the Top-20 Molecules of the QSAR Data Sets When Ranked Using a Range of Similarity Measures

			Bemis–Kuntz			Lederle		
data set	BNB	NBN	BK-2	BK-3	BK-4	LT	LA	ID
A	16.87	17.49	16.31	16.13	16.16	16.77	16.79	16.53
B	15.13	15.26	15.43	14.91	14.76	15.30	14.07	15.19
C	7.89	8.97	9.84	9.49	9.11	8.84	4.24	8.53
D	11.08	11.49	11.17	12.36	11.77	11.83	10.49	11.64
E	14.97	15.52	14.62	13.70	13.32	14.03	11.76	15.19
F	16.68	16.02	16.32	16.33	16.17	16.18	13.67	16.66
G	14.04	15.04	12.70	12.96	13.56	13.76	13.46	14.17
H	15.25	15.30	13.68	14.85	13.90	14.57	16.80	15.03

Table 4. Median Position of Carbohydrate Classes When Ranked Using a Range of Similarity Measures

size of	best			Bemis–Kuntz			Lederle		
class	possible	BNB	NBN	BK-2	BK-3	BK-4	LT	LA	ID
19	10	33.21	15.53	55.42	23.21	15.89	17.11	20.42	11.89
74	37	96.81	40.73	65.36	47.43	44.46	52.00	66.14	42.95
95	48	60.09	52.19	60.49	55.58	52.61	52.85	66.58	50.00
28	14	60.54	21.14	80.61	33.82	20.75	19.11	17.25	20.14

The results listed here are those obtained with the 10 fragments with the largest interatomic distance components in each molecule in the calculation of the intermolecular similarities; other results are presented by Bath.[35]

5. RESULTS AND DISCUSSION

5.1. QSAR Data Sets. Similarity searches were carried out on the QSAR datasets in which the top-5, the top-10 and the top-20 structures were retrieved. The results of these searches are detailed in Tables 1–3. The last column of these tables (and also of Table 4), headed ID, will be discussed in Section 5.3.

The most obvious feature of the figures in Tables 1–3 is the very poor performance of the LA measure (the one based on atom-triplets and the Asymmetric coefficient), which gives the worst results of all of the measures except in the case of Dataset H (the steroids) where it does best of all. The generally-poor performance with this measure may arise from the way that it is calculated. As noted earlier, the coefficient is given by

$$\frac{COMMON}{N(T)}$$

$N(T)$ is fixed for a given target structure, and thus LA produces a ranking in decreasing order of the number of atom-triplets common to the target and to a dataset structure. The measure thus takes no account of the number of triplets in each of the dataset structures, and there is some evidence to suggest that such unnormalized coefficients are of less use than coefficients that contain an appropriate normalization factor in the denominator (such as the Dice and Tanimoto coefficients).[36]

We had expected that the effectiveness of the Bemis-Kuntz measures would increase as more and more atoms were included in the subsets that were considered (so that the BK-4 results would be better than those for BK-3, which would, in their turn, be better than those for BK-2). However, the figures in Tables 1–3 show no such increase as the more-specific fragments are used for the generation of the frequency distributions. The two torsion-angle measures perform well, with one or other of them giving the best performance for five of the datasets (A, D, E, F and G); however, many of the differences in the mean numbers of actives are very small (occurring only in the second decimal place).

The Kendall coefficient of concordance[37] was used to test the null hypothesis, H_o, that there is no significant difference in the performances of the various similarity measure when they were applied to the eight QSAR datasets. This test measures the extent to which k rankings of the same set of N objects are in agreement with each other. In this case, $N = 7$, the number of different similarity measures (BNB, NBN, BK-2, BK-3, BK-4, LT and LA), and $k = 8$, the number of rankings of these similarity measure, i.e., the number of datasets. The similarity measures were ranked by assigning rank 1 to the measure that gave the highest mean number of actives in the top-ranked molecules, rank 2 to the measure that gave the next-highest mean number of actives, etc.

The Kendall test statistic, W, was computed using the mean numbers of actives for the 7 similarity measures that are listed in Tables 1–3. The W values when the top-5, top-10 and top-20 molecules were considered are 0.327, 0.261 and 0.162, respectively. The significance of these values may be established using the χ^2 test, since

$$\chi^2 \approx k(N-1)W$$

with $N - 1$ degrees of freedom. Here, $N = 7$ and there are thus 6 degrees of freedom. Reference to the critical values for χ^2 shows that the top-5 value for W is significant at the 0.01 level of statistical significance and the top-10 value at the 0.05 level. For these results, then, it is possible to reject the null hypothesis and hence to accept that there is a significant difference in the performances of the various similarity measures. Given such a significant difference, the measures may be compared using their overall mean ranks.[37] This suggests the following relative levels of effectiveness

BNB > BK-2 > NBN > LT > BK-4 > BK-3 > LA

for the top-5 searches, and

NBN > BNB > BK-2 > BK-3 > LT > BK-4 > LA

for the top-10 searches. It can be seen that these two rankings are in broad accord with each other.

A referee noted that the LA method was far inferior to the remainder and suggested that the Kendall analysis should be repeated, after the exclusion of the LA results. In this case, where N is 6, it is possible to reject the null hypothesis only for the top-10 values, where the calculated values for W are 0.186, 0.337 and 0.107, with no significant differences being found for either the top-5 or the top-20 values.

A series of *consensus searches* was carried out. A consensus search is one that yields a ranking in which the rank for the I-th compound in a dataset is the mean of the ranks that are observed for that I-th compound in each of the seven individual searches (BNB, NBN, BK-2, BK-3, BK-4, LT and LA). These seven similarity measures use different types of information: if they are complementary in nature, at least to some extent, then better results might be expected than those obtained from the use of an individual measure. In fact, the consensus searches were not found to be significantly better than the best of the individual searches.[35]

5.2. Carbohydrate Data Set. The results of the searches for the four carbohydrate classes are detailed in Table 4, the main body of which lists the values obtained for the median-rank measure defined in Section 4.

Table 4 shows that the NBN and BK-4 measures give the best results. Both of these measures involve descriptors derived from four-atom groupings. It is surprising, then, that the BNB measure, which is also based on quadruplets yields the worst results of all in this table, whereas it gave a consistently high level of performance for the QSAR datasets. It would appear that the single nonbonded vector and two bonded vectors used in this measure do not probe the molecular shape very effectively: only the nonbonded vector provides a description of molecular size, since the two bonds are regarded as being of constant length and are omitted in this description. Further, the other two parameters in the BNB descriptors are both angular and measure intervector and torsional relationships of short-range (bonded) interactions. Both of these problems are essentially rectified in the NBN approach, which incorporates two long-range (nonbonded) vectors and their torsional relationships. It is possible that an extension into the NNN area (i.e., the use of torsion angles in which none of the four atoms are connected[19]) may be even more successful, and this is currently under investigation.

The LT measure does slightly better than BK-3, which uses a comparable level of description, and is again often better than the LA measure. It is noticeable that here we have the expected progression for the Bemis-Kuntz measures, since performance improves considerably as we move from atom-pairs (BK-2) through triplets (BK-3) to the quadruplets of BK-4.

5.3. Use of Atom-Type Information. For comparison with the results in Tables 1–4 discussed thus far, the last column of each table contains the results obtained with the *Individual Distances*, or ID, measure of Pepperell and Willett.[24] We have noted that 2-D similarity-searching systems use the Tanimoto coefficient, as defined in Section 2.2. The ID measure simply involves replacing the occurrences of 2-D substructural fragments in this expression with interatomic distances. Let P and Q be two atoms in a dataset molecule, D, separated by a distance of $D\_DIST(P,Q)$, and let R and S be two atoms in the target molecule, T, separated by a distance $T\_DIST(R,S)$. These two interatomic distances are defined as being common to T and D if: $D\_DIST(P,Q)$ and $T\_DIST(R,S)$ are the same to within a user-defined tolerance (0.5 Å in our experiments) and if the atomic types of the pairs of atoms P and Q and R and S are equivalent. *COMMON* in the Tanimoto expression is then the number of such common distances, while $N(T)$ and $N(D)$ are the numbers of distances in T and D, respectively. This measure provides a natural level of comparison since it uses not only distance information, as in BK-2, but also includes atom-type information, which is specifically excluded by the measures studied thus far. The columns headed ID in Tables 1–4 give the results obtained when this measure was used with the eight QSAR datasets and the four carbohydrate classes.

It will be seen that the ID measure gives the best level of performance with four of the QSAR datasets (B, E, F and G) and with the first and the third of the carbohydrate classes; in this latter case, it also does well with the other two classes, with all of the median positions for this measure being only slightly inferior to the best possible results. The classification of distances for the hexopyranose sugars into C≈C, O≈O and C≈O types automatically enhances the dimensionality of the simple distance descriptors: this enhancement is clearly of benefit in this dataset, where the constituent molecules have almost equal numbers of carbons and oxygens.

We thus conclude that, hardly surprisingly, the inclusion of atom-type information often provides a more effective structural characterisation for shape-based similarity searching than a measure that does not take such information into account. That said, many of the differences here are very small, and at least some of the purely-geometric measures do provide a fair level of effectiveness.

6. CONCLUSIONS

In this paper, we have compared several measures that may be used for similarity searching in databases of 3-D chemical structures. The measures use a range of distance and angular information to encode the shape of a 3-D molecule, without consideration of chemical attributes such as elemental type or physico-chemical properties.

If we are to obtain descriptors that are devoid of atom-type information (as we believe we must for a general description of molecular shape), then we can only conclude that a treatment of four-atom units represents our best chance of success (as is exemplified by the rankings of the various similarity measures that have been discussed in Section 5.1). An important point is that any four-atom unit can be treated as a general, if distorted, tetrahedron. Except in very specialised cases, this unit is naturally imbued with three-dimensionality, a fact that is not true of measures based on pairs or triplets of atoms. What is important, then, is the choice of tetrahedra from the $N(N-1)(N-2)(N-3)/24$ that are available in a molecule containing N atoms, and the choice of suitable geometric descriptors for each such tetrahedron for inclusion in a similarity measure. Both of these points are now being actively pursued, and we hope to report the results of further experiments (using both the best of the measures discussed here and other types of NBN and NNN fragments) in the near future.

Finally, we note that a practical similarity-searching system must be both *effective*, i.e., result in the retrieval of molecules that are judged by the users of the system to be structurally related to the target molecules that they have submitted, and *efficient*, i.e., enable the identification of these related structures with the minimum use of computer resources. In this paper, we have considered only the effectiveness of the various measures when they were applied to several small datasets; considerations of efficiency will arise when the most promising of the measures are applied to datasets of non-trivial size in a later stage of the project.

ACKNOWLEDGMENT

We thank the Cambridge Crystallographic Data Centre, the Science and Engineering Research Council and Tripos Associates for funding, and the referees for comments on an earlier draft of this paper. The Krebs Institute for Biomolecular Research is a designated centre for the Science and Engineering Research Council Molecular Recognition Initiative.

REFERENCES AND NOTES

(1) Willett, P. *Three-Dimensional Chemical Structure Handling*; Research Studies Press: Taunton, U.K., 1991.
(2) Martin, Y. C. 3D Database Searching in Drug Design. *J. Med. Chem.* **1992**, *35*, 2145–2154.
(3) Willett, P. A Review of Three-Dimensional Chemical Structure Retrieval Systems. *J. Chemom.* **1992**, *6*, 289–305.
(4) Artymiuk, P. J.; Bath, P. A.; Grindley, H. M.; Pepperrell, C. A.; Poirrette, A. R.; Rice, D. W.; Thorner, D. A.; Wild, D. J.; Willett, P.; Allen, F. H.; Taylor, R. Similarity Searching in Databases of Three-Dimensional Molecules and Macromolecules. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 617–630.
(5) Johnson, M. A., Maggiora, G. M., Eds. *Concepts and Applications of Molecular Similarity*; Wiley: New York, 1990.
(6) Willett, P. Similarity Searching in Databases of Three-Dimensional Chemical Structures. In *Information Systems and Data Analysis. Studies in Classification, Data Analysis and Knowledge Organization*; Bock, H. H., Lenski, W., Richter, M. M., Eds.; Springer: Heidelberg, 1993; Vol. 4, in press.
(7) CONCORD Users Manual. Tripos Associates, St. Louis, MO, 1988, and University of Texas at Austin.
(8) Allen, F. H.; Davies, J. E.; Galloy, J. J.; Johnson, O.; Kennard, O.; Macrae, C.; Mitchell, E. M.; Mitchell, G. F.; Smith, J. M.; Watson, D. G. The Development of Versions 3 and 4 of the Cambridge Structural Database System. *J. Chem. Inf. Comput. Sci.* **1991**, *31*, 187–204.
(9) Willett, P. *Similarity and Clustering in Chemical Information Systems*; Research Studies Press: Letchworth, U.K., 1987.
(10) Barnard, J. M.; Downs, G. M. Clustering of Chemical Structures on the Basis of Two-Dimensional Similarity Measures. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 644–649.
(11) Pepperrell, C. A.; Taylor, R.; Willett, P. Implementation and Use of an Atom Mapping Procedure for Similarity Searching in Databases of 3-D Chemical Structures. *Tetrahedron Comput. Methodol.* **1990**, *3*, 575–593.
(12) van Geerestein, V.; Perry, N. C.; Grootenhuis, P. D. J.; Haasnoot, C. A. G. 3D Database Searching on the Basis of Ligand Shape Using the SPERM Prototype Method. *Tetrahedron Comput. Methodol.* **1990**, *3*, 595–613.
(13) Gund, P. Three-Dimensional Pharmacophoric Pattern Searching. *Prog. Mol. Subcell. Biol.* **1977**, *5*, 117–143.
(14) Jakes, S. E.; Willett, P. Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures: Selection of Inter-Atomic Distance Screens. *J. Mol. Graphics* **1986**, *4*, 12–20.
(15) Bemis, G. W.; Kuntz, I. D. A Fast and Efficient Method for 2D and 3D Molecular Shape Description. *J. Comput.-Aid. Mol. Des.* **1992**, *6*, 607–628.
(16) Nilakantan, R.; Bauman, N.; Venkataraghavan, R. New Method for Rapid Characterization of Molecular Shapes: Applications in Drug Design. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 79–85.
(17) Nilakantan, R.; Bauman, N.; Dixon, J. S.; Venkataraghavan, R. Topological Torsion: A New Molecular Descriptor for SAR Applications. Comparison with Other Descriptors. *J. Chem. Inf. Comput. Sci.* **1987**, *27*, 82–85.
(18) Weiner, S. J.; Kollman, P. A.; Case, D. A.; Singh, U. C.; Ghio, C.; Alagona, G.; Profeta, S.; Weiner, P. A New Force Field for Molecular Mechanical Simulation of Nucleic Acids and Proteins. *J. Am. Chem. Soc.* **1984**, *106*, 765–784.
(19) Poirrette, A. R.; Willett, P.; Allen, F. H. Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures: Characterisation and Use of Generalised Torsion Angles. *J. Mol. Graphics* **1993**, *11*, 2–14.
(20) Poirrette, A. R.; Willett, P.; Allen, F. H. Pharmacophoric Pattern Matching in Files of Three-Dimensional Chemical Structures: Characterisation and Use of Generalised Valence Angles. *J. Mol. Graphics* **1991**, *9*, 203–217.
(21) Bartlett, P. A.; Shea, G. T.; Telfer, S. J.; Waterman, S. CAVEAT: A Program to Facilitate the Structure-Derived Design of Biologically Active Molecules. In *Molecular Recognition: Chemical and Biochemical Problems*; Roberts, S. M., Ed.; Royal Society of Chemistry: Cambridge, U.K., 1990; pp 182–196.
(22) Fisanick, W.; Cross, K. P.; Rusinko, A. Similarity Searching on CAS Registry Substances. 1. Global Molecular Property and Generic Atom Triangle Geometric Searching. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 664–674.
(23) Fisanick, W.; Cross, K. P.; Forman, J. C.; Rusinko, A. Experimental System for Similarity and 3D Substructure Searching of CAS Registry Substances. 1. 3D Substructure Searching. *J. Chem. Inf. Comput. Sci.* **1993**, *33*, 548–559.
(24) Pepperrell, C. A.; Willett, P. Techniques for the Calculation of Three-Dimensional Structural Similarity Using Inter-Atomic Distances. *J. Comput.-Aid. Mol. Des.* **1991**, *5*, 455–474.
(25) Cringean, J. K.; Pepperrell, C. A.; Poirrette, A. R.; Willett, P. Selection of Screens for Three-Dimensional Substructure Searching. *Tetrahedron Comput. Methodol.* **1990**, *3*, 37–46.
(26) Henry, D. R.; Jurs, P. C.; Denny, W. A. Structure–Antitumor Activity Relationships of 9-Anilinoacridines Using Pattern Recognition. *J. Med. Chem.* **1982**, *25*, 899–908.
(27) Yuta, K.; Jurs, P. C. Computer-Assisted Structure-Activity Studies of Chemical Carcinogens. Aromatic Amines. *J. Med. Chem.* **1981**, *24*, 241–251.
(28) Stuper, A. J.; Jurs, P. C. Structure-Activity Studies of Barbiturates Using Pattern-Recognition Techniques. *J. Pharm. Sci.* **1978**, *67*, 745–751.
(29) Chastrette, M.; Zakarya, D.; Elmouaffek, A. Structure–Odor Relations of the Nitrobenzene Musk Family. *Eur. J. Med. Chem.* **1986**, *21*, 505–510.
(30) Walsh, D. B.; Claxton, L. D. Computer-Assisted Structure-Activity Relationships for Nitrogenous Cyclic Compounds Tested in Salmonella Assays for Mutagenicity. *Mutat. Res.* **1987**, *182*, 55–64.
(31) Rose, S. L.; Jurs, P. C. Computer-Assisted Studies of Structure–Activity Relationships of N-Nitroso Compounds Using Pattern Recognition. *J. Med. Chem.* **1982**, *25*, 769–776.
(32) Jurs, P. C.; Chou, J. T.; Yuan, M. Computer-Assisted Structure–Activity Studies of Chemical Carcinogens. A Heterogeneous Data Set. *J. Med. Chem.* **1979**, *22*, 476–483.
(33) Stouch, T. R.; Jurs, P. C. Computer-Aided Studies of the Structure–Activity Relationships between the Structure of Some Steroids and Their Anti-Inflammatory Activity. *J. Med. Chem.* **1986**, *29*, 2125–2136.
(34) Allen, F. H.; Fortier, S. Stereochemical and Configurational Classification of the Hexopyranose Sugars Using Numerical Clustering Methods. *Acta Crystallogr., Sect. B*, submitted for publication.
(35) Bath, P. A. Ph.D. Thesis. Manuscript in preparation.
(36) Van Rijsbergen, C. J. *Information Retrieval*; Butterworth: London, 1979.
(37) Siegel, S.; Castellan, N. J. *Non-Parametric Statistics for the Behavioural Sciences*; McGraw-Hill: London, 1988.