Pharmacophoric pattern matching in files of three-dimensional chemical structures: Characterization and use of generalized torsion angle screens

Andrew R. Poirrette and Peter Willett

Department of Information Studies, University of Sheffield, UK

Frank H. Allen

Cambridge Crystallographic Data Centre, University of Cambridge, UK

This paper describes the use of generalized torsion angles for the screening of conformational searches in databases of three-dimensional chemical structures. A generalized torsion angle is defined as the dihedral angle between two vectors, A_1–A_2 and A_3–A_4, in which none, some, or all of the vectors A_1–A_2, A_2–A_3, and A_3–A_4 correspond to formal chemical bonds. The screens consist of a set of four atoms together with an associated angular range, and are identified by a statistical analysis of the frequencies of occurrence of these features in the Cambridge Structural Database. These frequencies are discussed, and the effectiveness of the screens is demonstrated by an extensive series of searches for representative pharamacophoric patterns.

Keywords: conformational pattern, dihedral angle, generalized torsion angle, pharmacophoric pattern, screening system, three-dimensional substructure searching, torsion angle

INTRODUCTION

There is a well-established body of theory and practice associated with the use of database systems for the storage and retrieval of two-dimensional (2D) chemical structures.[1] The last few years have seen the application of these principles to the design of comparable systems for the three-dimensional (3D) structures.[2,3] These systems were originally developed for the specific purpose of carrying out searches for pharmacophoric patterns, but are increasingly being integrated with other tools in computer-aided molecular design programs.[4,5]

Searches in a 3D database system are effected using a two-stage retrieval algorithm, in which an initial screening search is used to eliminate a large fraction of the search file from the detailed geometric search, in which a subgraph-isomorphism algorithm is used to identify those structures that contain the precise query pattern that has been submitted by the user. The geometric search is extremely demanding of computational resources, and there has thus been considerable interest in the development of techniques for the selection of distance-range screens that will ensure a high level of screenout in searches for pharmacophoric patterns that consist of atoms and the associated interatomic distances.[6–9] We are currently engaged in a project to develop comparable techniques for the efficient implementation of angle-based searches, and have recently discussed the use of valence-angle information for screening 3D database searches;[10] here, we report the development of comparable methods for torsion angles. The particular context in which our work has been carried out is the X-ray data in the Cambridge Structural Database produced by the Cambridge Crystallographic Data Centre;[11] however, we believe that the techniques are equally applicable to databases of 3D structures generated by computational procedures such as molecular mechanics or distance geometry.

TORSION ANGLES

Given a set of four atoms A_1, A_2, A_3, and A_4, the torsion angle τ is the angle between the two three-atom planes, $A_1A_2A_3$ and $A_2A_3A_4$. It describes the twist of the vector A_1–A_2 relative to the vector A_3–A_4 when viewed along the vector A_2–A_3.

Torsion angles are defined to lie in the range $-180° < \tau < +180°$. This sign convention, first proposed by Klyne

Address reprint requests to Prof. P. Willett at the Department of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, UK.

Received 14 January 1992; accepted 4 February 1992

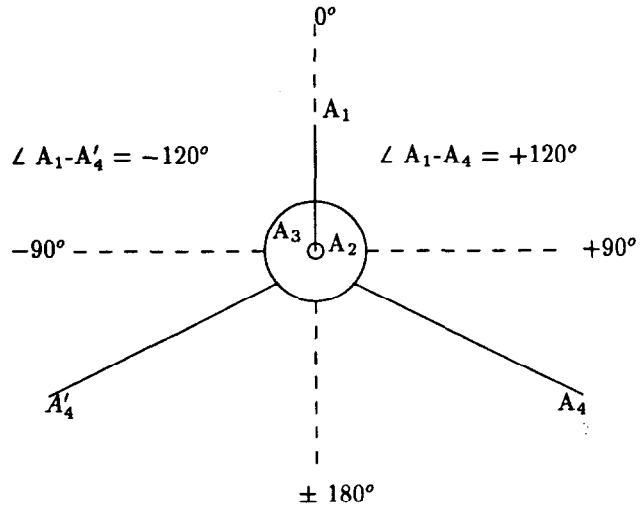


Figure 1. Newman projection for the torsion angle $A_2A_3A_1A_4$, which is obtained by looking along the A_2–A_3 vector of a fragment containing the four atoms A_1, A_2, A_3, and A_4.

and Prelog[12] is best illustrated using a Newman projection,[13] as shown in Figure 1. This projection represents a view along the vector A_2–A_3 of the four-atom fragment mentioned above, and the torsion angle in the figure can be seen to be 120°. This follows from the Klyne-Prelog rule, which states that a clockwise rotation of the first named bond to overlie the third bond is given a positive sign; an anticlockwise rotation, A_1–A_2 to A_3–A_4', is given a negative sign. Throughout this paper, we shall specify a torsion by listing the two central atoms of the torsion first, with these being followed by the atoms attached (whether by a bonded or by a nonbonded interaction) to them. Thus, the angle shown in Figure 1 is defined to be $A_2A_3A_1A_4$. Note that $\tau(A_2A_3A_1A_4)$ is the same as $\tau(A_3A_2A_4A_1)$ in both magnitude and sign.

The torsion angle is very well suited to the screening of 3D substructure searches, since it provides 3D information, unlike an individual valence angle, which is 2D, or an individual bond, which is 1D. 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 2D fragment descriptors.[14]

In this paper, we shall characterize the presence or absence of a bond between two of the four atoms comprising a torsion angle by the letters *B* or *N*, respectively; specifically, we shall discuss four types of generalized torsion angle, these being BBB, NBB, BNN and NNN torsions. The various types of angles will be illustrated using the structure shown in Figure 2:

- A BBB torsion is one in which there are bonds between the four atoms comprising the angle, i.e., the conventional type of torsion angle. Examples include the angles *BCAD, GHEI,* and *EGDH*.
- An NBB torsion is one in which there are bonds between the two outer pairs of atoms forming the four-atom fragment, i.e., A_1–A_2 and A_3–A_4, but in which there is no bond between the central pair of atoms, A_2–A_3; examples include the angles *BDAE, EHDI* and *CADF*. NBB torsions form an important component of the CAVEAT program for the design of enzyme inhibitors.[15]
- A BNN torsion is one in which only the two central atoms of the four-atom fragment are bonded, i.e., A_2–A_3; examples include the angles *BCFE, EGCI,* and *GHDA*.
- An NNN torsion is one that contains no bonded atoms; examples include the angles *BDFG, GIDA* and *EHCA*.

The generalized torsions considered here are analogous to the BB (for bonded-bonded), BN (for bonded-nonbonded) and NN (for nonbonded-nonbonded) valence angles that were investigated in our previous study of generalized valence angles.[10]

It should be noted that two other classes of generalized torsion angle can be generated from a structural fragment such as that shown in Figure 2. These are the NNB torsion angle, where the only bond is that between A_3 and A_4, and the BBN torsion angle, where A_1 is bonded to A_2 and A_2 is bonded to A_3, but A_3 and A_4 are not bonded (the two other possible definitions, NBN and BNB, produce identical angles to those produced by the NNB and BBN definitions). These types of generalized torsion angle were not considered here since they overlap with the types that we have considered. This problem is illustrated by the fragment shown in Figure 2, where examples of BBN torsions include the angles *EFDC* and *ABFD*; however, consideration of these reveals that they duplicate the BNN torsions *FEDC* and *BAFD*, respectively, in that both the BBN and the BNN angles describe the angle between the same pairs of three-atom planes: *EFD* and *FEC* in the first example, and *ABF* and *ABD* in the second example. Similar problems can arise with NNB torsion angles, and we have thus chosen to ignore both the BBN and NNB levels of description.

In this work, we have chosen to use the absolute values of individual torsion angles, i.e., $|\tau|$. Thus, both enantiomorphs of a given 3D molecule will generate identical $|\tau|$ values. This is important since we cannot know, *a priori,* which enantiomorph is represented by a calculated or crystallographically determined coordinate set (unless, of course, the correct absolute stereochemistry has been imposed, or determined, in the X-ray experiment). Certainly, the vast majority of crystal structures contains both enantiomorphs of a given molecule, generated one from another by space group symmetry, and the coordinates recorded in

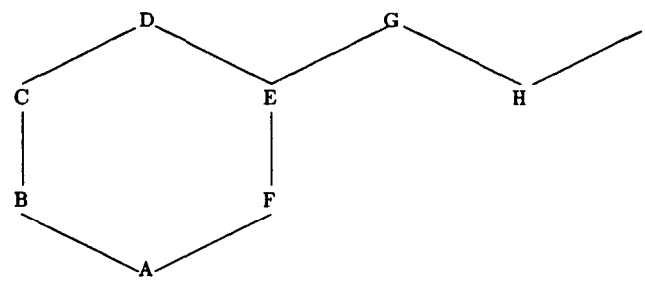


Figure 2. Structural fragment to illustrate the various types of generalized torsion angle that can be generated.

Table 1. Numbers of torsions generated from subsets of the Cambridge Structural Database using various types of generalized torsion angles

Torsion class	Angles generated	Data set size	Angles per compound
BBB	270,058	8000	34
NBB	11,753,328	4000	2938
NBB3	10,784,812	4000	2696
BNN	40,003,911	2000	20,002
NNN	32,287,210	200	161,436

the Cambridge Structural Database will represent only one of these classes.

CHARACTERIZATION OF GENERALIZED TORSION ANGLES

Generation of frequency data

A statistical analysis has been carried out of the frequencies of occurrence of each of the types of generalized torsion angles discussed above. As in our previous study of generalized valence angles,[10] we have generated all occurrences of each of the angular types of interest in subsets of the Cambridge Structural Database, as detailed in Table 1. An inspection of this Table demonstrates clearly the dramatic increase in the number of torsions per structure that accompanies an increase in the number of nonbonded vectors in a torsion definition. Indeed, it was possible to generate the NNN torsions for only 200 structures, and even this required more than 20 CPU hours on an IBM 3083.

The frequency analysis involves representing the various types of torsion angle in a form that is amenable to efficient machine processing. In our experiments, the NBB, BNN and NNN torsions were represented by a canonical string of the form $A_2A_3A_1A_4\ \tau$, where A_1, A_2, A_3, and A_4 are the atomic types, and where τ is the modulus of the torsion angle, i.e., $|\tau|$ as discussed above, rounded to the nearest degree. The canonicalization procedure involved ensuring that $A_2 \leq A_3$, and if these two atoms were both of the same type, that $A_1 \leq A_4$. Canonicalization in this way is made possible by the fact that, as mentioned earlier, $\tau\,(A_2A_3A_1A_4) = \tau\,(A_3A_2A_4A_1)$ in both magnitude and sign. Examples of the canonicalization procedure are presented by Poirrette.[16] The four atoms comprising a torsion were assigned to one of the four atomic classes C, N, O, and Y, these corresponding to carbon, nitrogen, oxygen and anything else, respectively (an analogous classification was used in our study of generalized valence angles[10]). There are a total of 136 unique, canonical sets of four atoms possible using the CNOY classification of atom types, as illustrated in Table 2.

A slightly different characterization of BBB torsion angles was used, based on a detailed consideration of the types of BBB torsion that are known to occur in the Cambridge Structural Database, and of the types of BBB query that might be expected in the context of an operational substructure searching system. The torsions were represented by a string of the form $B\ A_2A_3A_1A_4\ \tau$, where A_1, A_2, A_3, A_4 and τ have the same meaning as before (and where the same canonicalization procedure was used as for the other types of torsion) and where B is the type of the bond linking the two central atoms, A_2 and A_3. The selection of the groups of atoms comprising each torsion was carried out manually, this step involving the use of a more detailed characterization of the various atom types than previously. A total of six types of atom was considered. Three of these, C, N, and O, were the same as with the other types of torsion; however, three types of "other atom" descriptor were used as follows: X represents any atom; Y represents any atom other than C, N, O; and Z is any atom other than C. Only

Table 2. Sets of four atoms used to characterize NBB, BNN, and NNN torsion angles

CCCC	CNOO	COYO	NNCO	NOYC	OOCC	OYOO
CCCN	CNOY	COYY	NNCY	NOYN	OOCN	OYOY
CCCO	CNYC	CYCC	NNNN	NOYO	OOCO	OYYC
CCCY	CNYN	CYCN	NNNO	NOYY	OOCY	OYYN
CCNN	CNYO	CYCO	NNNY	NYCC	OONO	OYYO
CCNO	CNYY	CYCY	NNOO	NYCN	OONN	OYYY
CCNY	COCC	CYNC	NNOY	NYCO	OONY	YYCC
CCOO	COCN	CYNN	NNYY	NYCY	OOOO	YYCN
CCOY	COCO	CYNO	NOCC	NYNC	OOOY	YYCO
CCYY	COCY	CYNY	NOCN	NYNN	OOYY	YYCY
CNCC	CONC	CYOC	NOCO	NYNO	OYCC	YYNN
CNCN	CONN	CYON	NOCY	NYNY	OYCN	YYNO
CNCO	CONO	CYOO	NONC	NYOC	OYCO	YYNY
CNCY	CONY	CYOY	NONN	NYON	OYCY	YYOO
CNNC	COOC	CYYC	NONO	NYOO	OYNC	YYOY
CNNN	COON	CYYN	NONY	NYOY	OYNN	YYYY
CNNO	COOO	CYYO	NOOC	NYYC	OYNO	
CNNY	COOY	CYYY	NOON	NYYN	OYNY	
CNOC	COYC	NNCC	NOOO	NYYO	OYOC	
CNON	COYN	NNCN	NOOY	NYYY	OYON	

nonmetal and nonhydrogen atoms were considered in this analysis.

An inspection of the structures in the Cambridge Structural Database suggested the 20 types of BBB torsion angle that are listed in Table 3. The torsions listed in this table are, for the most part, self explanatory; the initial number indicates the type of bond that exists between atoms A_2 and A_3 and the atomic types are as defined above. In the case of the bonds between the outer atoms forming the fragment, there was no need to define these explicitly in the specification as it was decided to allow any outer bond type for the single-bond torsions and only single outer bonds for the double-bond torsions. Thus, for example, the string 1 CNZX represents a torsion in which the central C and N atoms are linked by a single bond, the C is bonded to any atom other than C by any type of bond and the N is bonded to any atom by any type of bond. The first two torsions listed in Table 3 are an exception to these rules: 0 CCCC represents a torsion in which the central bond is single and all the bonds are acyclic, while 1 CCCC represents the same situation with the sole exception that one or more of the bonds are cyclic. The majority of the torsions in this table involves a central single bond, normally C—C, this reflecting the known flexibility of this very high-frequency moiety. The four torsions in the table that involve a central double bond are needed to screen searches for *cis/trans* stereochemistry and for the small deviations from these two configurations that may occur in, e.g., twisted ethylenes. Aromatic bonds, which are separately identified in the Cambridge Structural Database, were specifically excluded from the central A_2A_3 position. Other multiple bonds are either terminal in character, e.g., C≡N or C=O, or are so rare, e.g., C≡C, that they need not be considered further.

Table 3. Sets of four atom- and bond-types used to characterize BBB torsion angles

0 CCCC	1 CCYY
1 CCCC	1 CNCX
1 CCCN	1 CNZX
1 CCCO	1 COCX
1 CCCY	1 COZX
1 CCNN	1 CYXX
1 CCNO	2 CCCC
1 CCNY	2 CCZX
1 CCOO	2 CCZZ
1 CCOY	2 CNXX

Once the sets of angles had been generated, they were sorted and cumulated to produce the frequency distributions. Examples of the distributions for the different classes can be seen in Figures 3 (BBB), 4 (NBB), 5 (BNN), and 6 (NNN).

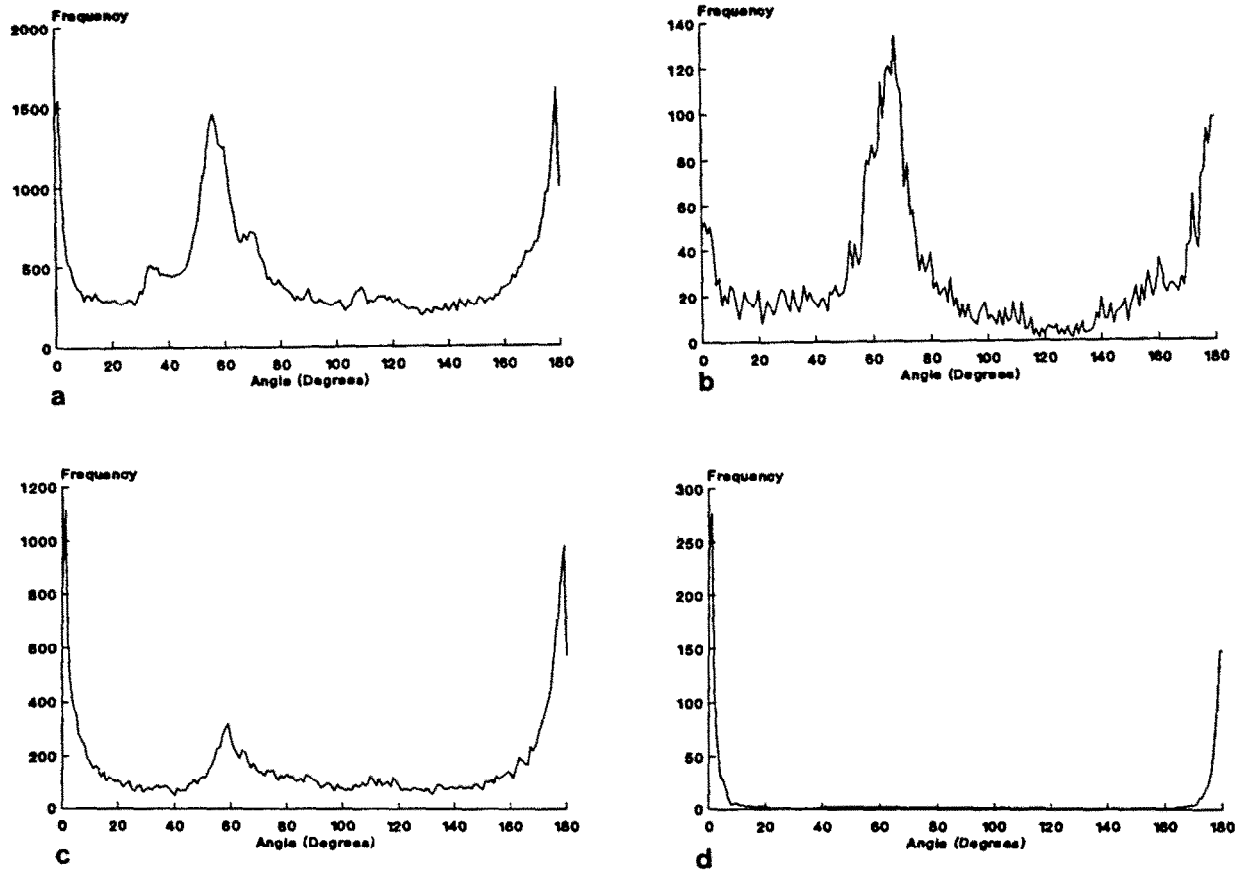


Figure 3. Frequencies of occurrence for BBB torsion angles: (a) 1 CCCC, (b) 1 CCOO, (c) 1 CNCX and (d) 2 CCZZ.

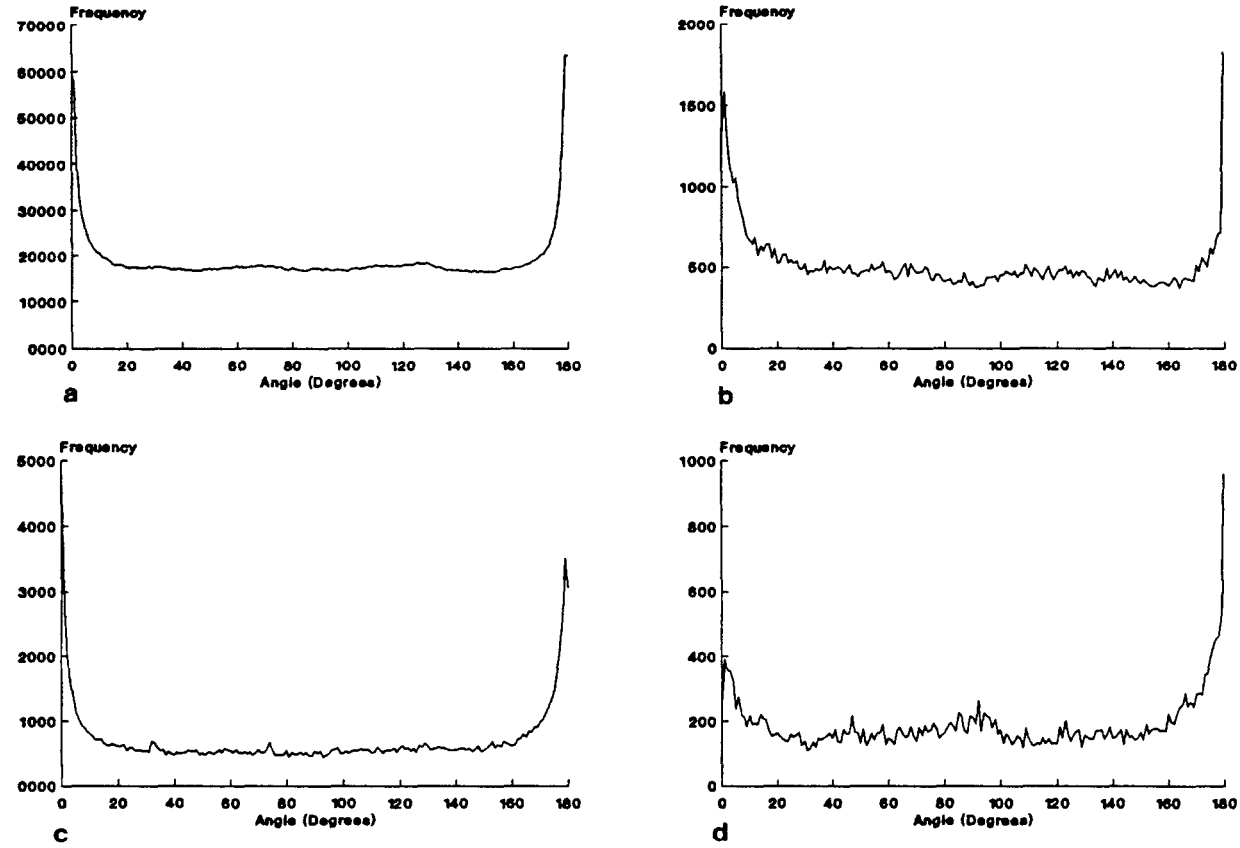


Figure 4. Frequencies of occurrence for NBB torsion angles: (a) CCCC, (b) CCOO, (c) CNNC, and (d) CYNN.

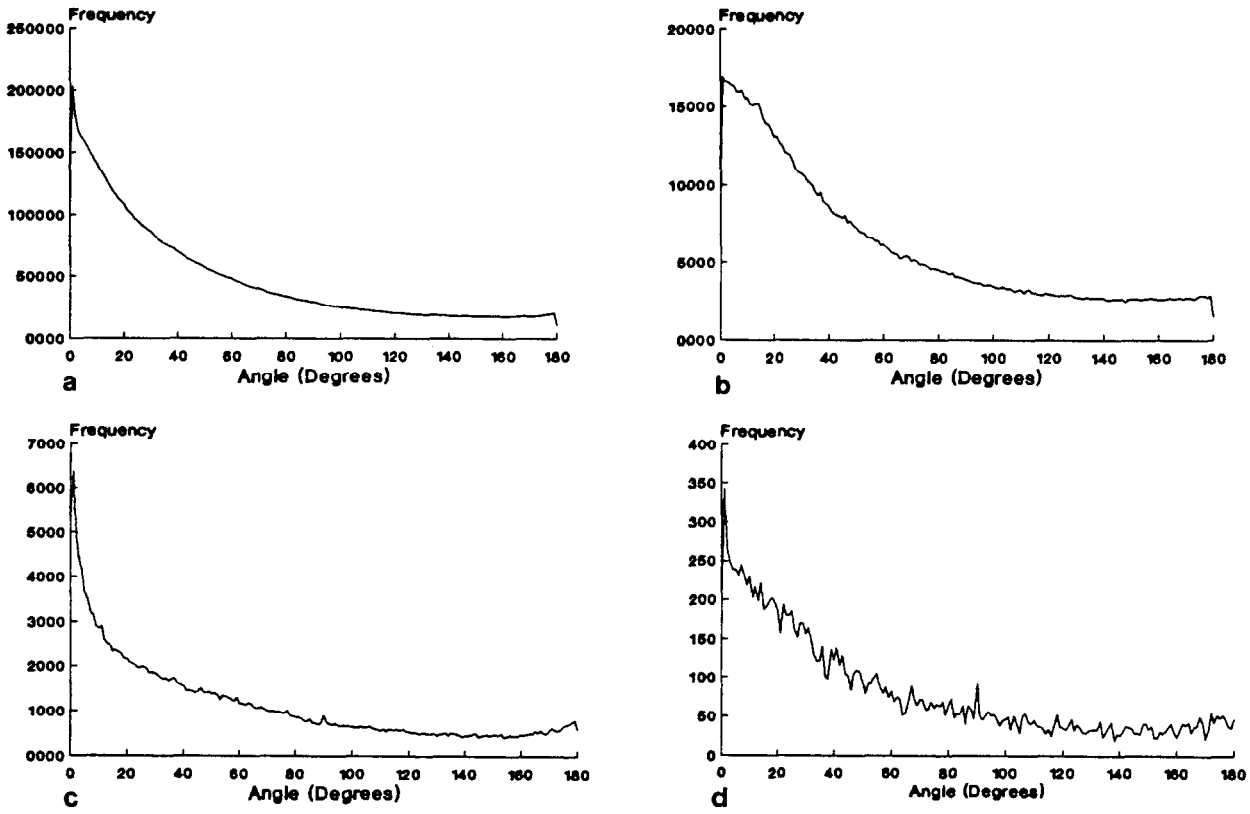


Figure 5. Frequencies of occurrence for BNN torsion angles: (a) CCCC, (b) CCOO, (c) CNNC, and (d) CYNN.

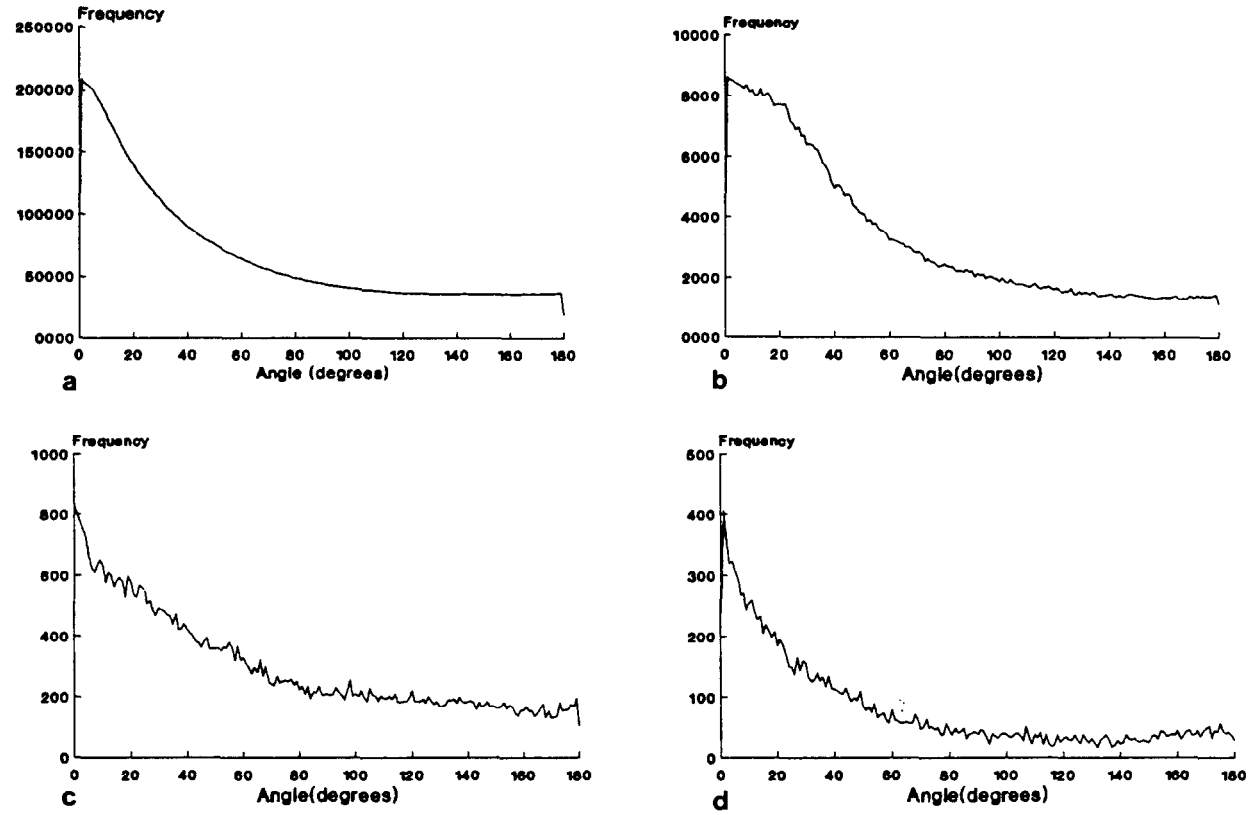


Figure 6. Frequencies of occurrence for NNN torsion angles: (a) CCCC, (b) CCOO, (c) CNNC, and (d) CYNN.

BBB torsion angles

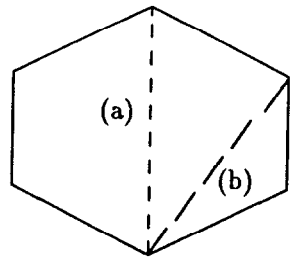
Figure 3 shows the frequency distributions for the BBB fragments, 1 CCCC, 1 CCOO, 1 CNCY and 2 CCZZ: of these, the first three contain a central single bond, and the last a central double bond. The shapes seen are typical of those exhibited by all of the fragments in the class. For the single-bond fragments, three peaks at around 0°, 60°, and 180° are shown in all cases. These peaks are as expected, representing the typical synperiplanar (0°), synclinal (60°), and antiperiplanar (180°) conformational arrangements about C—C single bonds. The presence of many puckered small and medium ring systems is revealed in the broad synclinal peak at 40°–80°. Anticlinal conformations (in side chains) are rare, the "peak" at 100°–130° in Figure 3a being scarcely distinguishable from the general background. The fourth distribution in Figure 3 is typical of torsion angles about a central double bond, with the huge peaks at 0° and 180°, reflecting *cis* and *trans* relationships, respectively. The *cis* double bonds predominate due to their presence in unsaturated rings.

NBB torsion angles

Figure 4 shows the frequency distributions for the NBB fragments CCCC, CCOO, CNNC, and CYNN. All of the distributions show distinct peaks at around 0° and 180° with a flat area between these two peaks; the shape of the distributions is typical of all the fragments generated in the NBB class. The peaks at 0° and 180° are due, in large part, to the very frequent occurrence of phenyl rings in the Cambridge Structural Database. The contribution of these rings is illustrated in Figure 7, which shows the idealized NBB angles that are generated from the two types of marked vector, (a) and (b), in a phenyl ring, and which demonstrates the very large numbers of NBB angles at 0° and 180° that occur in this substructural moiety. NBB angles close to 0° or 180° are also generated by puckered ring systems, e.g., cyclopentane and cyclohexane.

In addition to the normal NBB angles, we have also investigated restricted types of NBB torsion, which are referred to subsequently as an NBB3 torsion and an NBB26 torsion. These are generated in exactly the same way as a NBB torsion, subject to the sole constraint that the nonbonded central atoms, A_2 and A_3, must be separated by a distance of at least 3 Å (for a NBB3 torsion) or 2.6 Å (for a NBB26 torsion). The resultant angles are similar to those used in the CAVEAT program,[15] which employs NBB torsions in which the central atoms are separated by distances between 3 Å and 9 Å. The distance constraint was introduced here in an attempt to exclude the many intra-ring torsions, which provide the main contributions to the large peaks at 0° and 180° in the normal NBB distributions. The vector (a) in Figure 7 is around 2.8 Å in length and thus all NBB torsions about this vector will be eliminated by the 3-Å constraint; analogously, the vector (b) is around 2.4 Å in length and thus all NBB torsions about this vector will be eliminated by the 2.6-Å constraint.

Figure 8 shows the overall frequency distributions for the



τ Taken About Vector	Number Of Such Vectors In A Phenyl Ring	Total Number Of NBB Angles	
		0°	180°
(a)	3	6	6
(b)	6	12	12
Total		18	18

Figure 7. Examples of idealized NBB torsion angles resulting from the two types of nonbonded vector, (a) and (b), in the phenyl ring.

NBB, NBB26 and NBB3 classes, i.e., when summed over all of the 136 possible combinations of atomic types. The effect of the distance constraints is to reduce substantially the peaks at 0° and 180°, but to leave the remainder of the distribution almost completely unchanged. The greater degree of equifrequency evident in the NBB26 and NBB3 distributions suggests that these types of torsion might produce more effective screen sets than the unconstrained NBB distributions.

BNN torsion angles

Figure 5 shows the frequency distributions for the four BNN torsions CCCC, CCOO, CNNC and CYNN. The shapes are all similar and typical of the distributions obtained for this type of torsion, with a large peak around 0° that falls away to level-off by about 90°.

This shape is explicable by simple statistical considerations, as illustrated by the molecule in Figure 9. The figure contains 20 of the BNN torsions around the bonded vector 8–9, these angles being calculated using the SYBYL molecular-modeling package. The angles have been chosen so that 10 of them involve atoms on the same side of this vector and 10 of them involve atoms on the opposite sides; in what follows, we refer to these as *same angles* and *opposite angles*. It will be seen that the same angles are very much lower, with a mean of 8°, than the opposite angles, with a mean of 168°. Entirely comparable results are obtained with different vectors and different molecules.

Given a bond E within a molecule, let there be N atoms on one side of E and M atoms on the other side of E. There are thus a total of

$$\frac{N(N-1)+M(M-1)}{2}$$

unique same torsion angles, and

$$N \times M$$

unique opposite torsion angles. Let these totals be T_S and T_O, respectively. Let C be the total number of atoms in the molecule, excluding the two atoms that form the chosen bond, E, so that

$$C = M + N$$

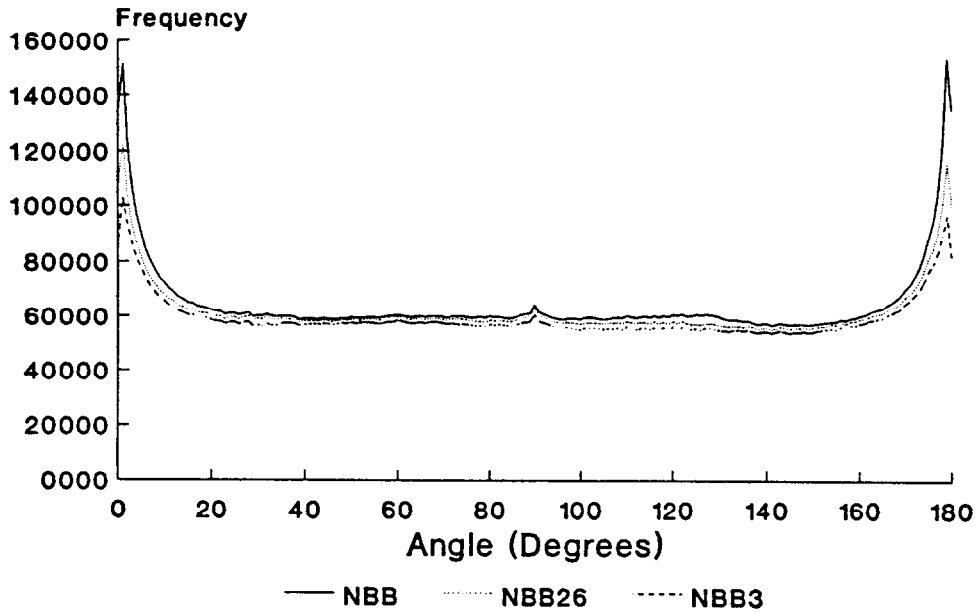


Figure 8. Frequency of occurrence for NBB, NBB26, and NBB3 torsion angles.

and thus

$$T_O = N(C - N)$$

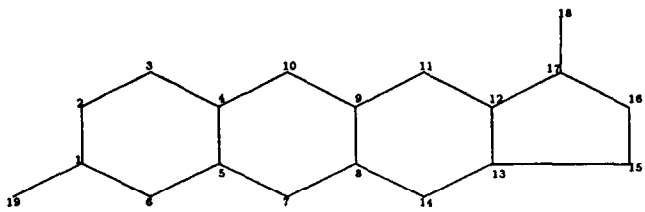
Differentiating with respect to N, and setting the result to zero, the maximum value for T_O is obtained when

$$N = \frac{C}{2} = M$$

The minimum value for T_O is obtained when N (or M) is zero, i.e., when the chosen bond is at one "end" of the molecule. By a similar argument, the maximum value for T_S is obtained when N (or M) is zero, and the minimum value is obtained when

$$N = \frac{C}{2} = M$$

Thus, both T_S and T_O will depend on the location of E within the molecule, with T_S increasing (and T_O decreasing) the further that E is away from the center of the molecule. The analysis in Figure 9 shows that same angles are close to 0°, and thus the further that E lies from the center, the more it will contribute to the peak near 0°. In any molecule, there are many more atoms and bonds near the edge of the molecule than there are atoms and bonds near the center; accordingly, the number of same angles will tend to be far greater than the number of opposite angles, resulting in the peak near 0° that is observed in all of the BNN frequency distributions.



Torsion Angles About The Vector 8-9 (Sign Removed)			
Opposite Angles		Same Angles	
Atoms	Angle	Atoms	Angle
19-11	153	19-5	5
2-13	178	1-10	25
4-12	179	2-4	1
6-13	173	1-6	10
5-16	179	3-1	8
19-17	174	18-11	21
3-11	139	16-13	2
19-18	175	17-12	10
10-14	178	17-18	2
7-16	152	2-5	1
Mean	168	Mean	8

Figure 9. Examples of BNN torsion angles generated about the bonded vector 8–9 in the molecule above. Ten of the torsion angles were generated by taking two atoms on the same side of the vector and ten were generated by taking two atoms on opposite sides. The angles are rounded to the nearest degree and were calculated using the SYBYL molecular modeling package.

This explanation would suggest that there should be no change in the overall shape of the distribution if distance constraints are applied, as was done in the case of the NBB distributions. This was, indeed, found to be the case when constraints of 2.6 Å and 3.0 Å were applied to the distances for the two pairs of nonbonded atoms, i.e., A_1A_2 and A_3A_4, the only change resulting from the application of these constraints being a reduction in the total number of all types of torsion angle.

NNN torsion angles

Figure 6 shows the frequency distributions for the four NNN torsions CCCC, CCOO, CNNC, and CYNN. Once again, the shapes are very similar to each other, and are typical of all NNN torsions. The shapes of the distributions are similar to those for the BNN torsions, and are explicable in exactly the same way (except that in this case, the central vector is nonbonded, rather than being bonded as with the BNN torsions).

USE OF GENERALIZED TORSION-ANGLE SCREENS

In this section, we discuss the experiments that were carried out to evaluate the effectiveness of generalized torsion angles for screening 3D substructure searches.

Generation and assignment of screen sets

The data in Table 1 demonstrate clearly the very large numbers of BNN and NNN torsions in a molecule. These numbers imply that correspondingly large numbers of screens would be assigned to compounds and queries if these types of torsion were used for screening purposes, and we have accordingly restricted our attention to the less frequently occurring BBB, NBB, and NBB3 torsions.

The procedure followed for screen-set generation and screen assignment was precisely the same as in our previous study.[10] For the chosen types of generalized torsion, screen sets of sizes 128, 256, and 512 members were generated using the screen-set generation algorithm described by Cringean *et al.*[9] The screens that are generated consist of sets of atoms with associated angle ranges, these being chosen so that each screen occurs approximately equifrequently in a file of structures. Thus, the screens are of the form $A_2A_3A_1A_4A_{2'}A_{3'}A_{1'}A_{4'}$ τ τ': this descriptor represents all of the possible torsions between the set of atoms $A_2A_3A_1A_4$ with an associated angle τ and the set of atoms $A_{2'}A_{3'}A_{1'}A_{4'}$ with an associated angle τ' (note that $A_2A_3A_1A_4$ and $A_{2'}A_{3'}A_{1'}A_{4'}$ may be identical in the case of frequently occurring sets of four atoms). Part of the 128-member, BBB screen set is shown in Figure 10.

The experiments used a 5000-compound subset of the Cambridge Structural Database. Screens were assigned to

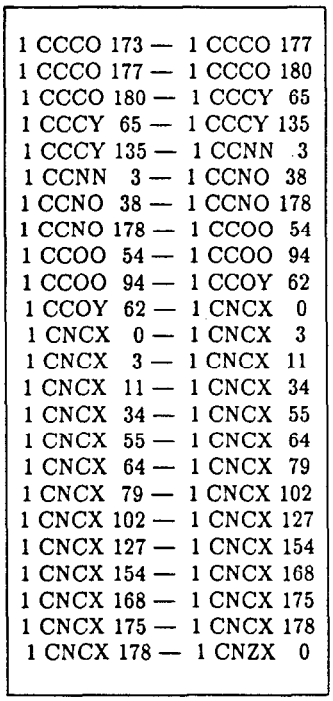


Figure 10. Screens 67–90 of the 128-member BBB screen set.

Table 4. Density of screen assignment for screen sets based on the three types of torsion angle

Torsion type	Screen set size		
	128	256	512
BBB	0.14	0.08	0.05
NBB	0.52	0.45	0.39
NBB3	0.49	0.43	0.37

all of the compounds in this subset using each of the three types of screen set. Table 4 details the proportion of the bits that were set when each of the screen sets was assigned to this file. The density of screen assignment clearly reflects the number of angles generated per compound, the values for the BBB assignments being much lower than for the two sets of NBB assignments.

Generation of queries

Sets of 3D query patterns were obtained by randomly selecting two, three, or four torsion angles (of whichever type was required) from structures in a 5000-molecule subfile of the Cambridge Structural Database. In all, there were 300 automatically generated patterns for each size of pattern. Previous studies in this department[17,18] have used sets of published pharmacophoric patterns, these being based on interatomic distances. To test the efficiency of the BBB screens generated in the present work, we have manually specified typical torsional constraints that have been employed in the location of conformational archetypes in the Cambridge Structural Database. The 19 query patterns defined for the BBB searches are shown in Table 5.

Each of the query patterns, whether generated manually or automatically, was searched within tolerance values ± 0, i.e., an exact-match search, 5°, 10°, and 20°. Because of the torsional phase change at ± 180°, and also because of our use of $|\tau|$ values throughout this work, care was needed in the generation of the τ-ranges when the specified torsion angle was close to 0° or to 180°. Consider a $|\tau|$ value of 176°, for which a ± 10° tolerance yields a range of 166°–186° or, taking account of the phase change at ± 180°, a Klyne-Prelog range of 166° to – 174°. The use of $|\tau|$ values implies a range of 166° to 174°, which is obviously incorrect since it excludes the original query angle of 176°. In fact, our use of $|\tau|$, restricting all values to the enantiomorph-insensitive 0°–180° scale, implies a *reversal of direction* in passing through 180° rather than a reversal of sign. The true range for our example of 176 ± 10° is hence 166°–180°. These considerations apply to all values of $|\tau|$ in Table 5 that lie within $t°$ of 0° or 180°, where $t°$ is the absolute tolerance that is employed in the search (and the same considerations also apply to all of the automatically generated query patterns). This special treatment is due to the circular nature of the variable τ, which invalidates normal linear arithmetic approaches to the generation of ranges, statistical descriptions, etc.[19,20]

Screening performance

The main function of a screen set is to reduce the number of structures that must undergo the time-consuming geometric search. The efficiency of a screen set is thus generally measured in terms of the number of structures that is eliminated by the screen search. Assume that a database contains N structures and that only n of these match the query pattern at the screen level; then the *screenout* is defined to be

$$\frac{N-n}{N}$$

so that a large (small) value for the screenout corresponds to an efficient (inefficient) substructure search. In our experiments, N was 5000, corresponding to the subset of the Cambridge Structural Database mentioned above.

The median screenout results for the automatically generated BBB queries are listed in Table 6. As expected, the screenout increases in line with an increase in screen-set size and with a decrease in the tolerance; however, it will be seen that high levels of screenout are achieved even with the smallest screen sets and with the largest tolerances. Similar conclusions apply to the results for the (automatically generated) NBB and NBB3 queries, which are listed in Table 7 and Table 8, respectively.

The median screenout results for the 19 manually-generated BBB queries are listed in Table 9, while the screenout results for the individual queries are demonstrated by the figures in Table 10. It will be seen that individual screenouts of 1.00 are recorded in a number of instances and the screenouts often remain high even with a tolerance of ± 20° as is illustrated, for example, by the mixed pattern, where none of the screenouts is less than 0.98. Those queries that

Table 5. Manually defined BBB query patterns

Pattern	Typical of	Torsions specified
Chair1	6-membered ring	1 CCCC 60
Substituted-chair	6-membered ring	1 CCCC 60, 1 CCCC 180
Boat1	6-membered ring	1 CCCC 0, 1 CCCC 60
Twist-boat	6-membered ring	1 CCCC 33, 1 CCCC 70
Envelope	6-membered ring	1 CCCC 0, 1 CCCC 30, 1 CCCC 60
Half-chair1	6-membered ring	1 CCCC 0, 1 CCCC 25, 1 CCCC 55, 1 CCCC 75
Chair2	6-membered ring + Heteroatom	1 CCCC 60, 1 CCCN 60, 1 CNCX 60
Chair3	6-membered ring + Heteroatom	1 CCCC 60, 1 CCCO 60 1 COCX 60
Half-chair2	6-membered ring + Heteroatom	1 CCCC 55, 1 CCCC 75, 1 CCCO 55, 1 COCX 20, 1 COCX 0, 1 CCCO 20
Twist-chair1	7-membered ring	1 CCCC 40, 1 CCCC 54 1 CCCC 72, 1 CCCC 88
Twist-chair2	7-membered ring + Heteroatom	1 CNCX 40, 1 CCCN 88, 1 CCCC 54, 1 CCCC 72
Chair4	7-membered ring	1 CCCC 0, 1 CCCC 64, 1 CCCC 66, 1 CCCC 84
Crown	8-membered ring	1 CCCC 88
Boat2	8-membered ring	1 CCCC 0, 1 CCCC 74
Extended	Acyclic	0 CCCC 180
Gauche-twist	Acyclic	0 CCCC 60, 0 CCCC 180
Mixed	Mixed cyclic/acyclic	0 CCCC 60, 0 CCCC 180, 1 CCCC 60, 1 CCCC 180
Double1	Double-bond	2 CCCC 15, 2 CCCC 160
Double2	Double-bond	2 CCCC 0, 2 CCCC 180

are most strongly affected by the tolerance tend to be those that contain only one or two angles, e.g., the chair and substituted-chair patterns.

We noted in our previous analysis of generalized valence angles that, all other things being equal, searches specifying large tolerances were best effected using screen sets that contained many different sets of atoms with fairly wide angular ranges (rather than using screen sets that contained a smaller number of sets of atoms with narrower, more precise angular ranges).[10] A similar pattern of behavior is observed here as is evidenced by comparing the BBB screen sets, which contain only 20 sets of atoms, with the NBB and

Table 6. Median screenout, averaged over 300 searches of 5000 structures taken from the Cambridge Structural Database, for automatically generated patterns of BBB torsion angles

Number of BBB torsion angles	Screen set size	Tolerance in degrees			
		0	5	10	20
2	128	0.94	0.87	0.84	0.80
	256	0.97	0.90	0.86	0.82
	512	0.99	0.92	0.88	0.84
3	128	0.97	0.92	0.89	0.86
	256	0.99	0.94	0.91	0.87
	512	1.00	0.96	0.93	0.89
4	128	0.98	0.94	0.92	0.88
	256	1.00	0.96	0.94	0.91
	512	1.00	0.97	0.95	0.92

Table 7. Median screenout, averaged over 300 searches of 5000 structures taken from the Cambridge Structural Database, for automatically generated patterns of NBB torsion angles

Number of NBB torsion angles	Screen set size	Tolerance in degrees			
		0	5	10	20
2	128	0.64	0.62	0.62	0.59
	256	0.68	0.67	0.64	0.63
	512	0.73	0.68	0.66	0.64
3	128	0.67	0.65	0.65	0.63
	256	0.74	0.70	0.68	0.67
	512	0.78	0.74	0.70	0.70
4	128	0.68	0.67	0.66	0.64
	256	0.74	0.72	0.70	0.69
	512	0.81	0.77	0.75	0.72

Table 8. Median screenout, averaged over 300 searches of 5000 structures taken from the Cambridge Structural Database, for automatically generated patterns of NBB3 torsion angles

Number of NBB3 torsion angles	Screen set size	Tolerance in degrees			
		0	5	10	20
2	128	0.65	0.64	0.63	0.63
	256	0.70	0.67	0.66	0.64
	512	0.75	0.70	0.68	0.66
3	128	0.68	0.67	0.66	0.65
	256	0.76	0.71	0.70	0.69
	512	0.80	0.75	0.73	0.71
4	128	0.71	0.69	0.68	0.67
	256	0.78	0.72	0.71	0.70
	512	0.82	0.79	0.77	0.74

NBB3 screen sets, which both contain 136 sets of atoms (as detailed in Tables 2 and 3, respectively); the BBB searches show a greater fall in screenout than do the NBB and NBB3 searches in moving from a tolerance of $\pm 0°$ to one of $\pm 20°$.

The valence-angle study also included an extensive analysis of the equifrequency of assignment and the degree of screen association for BB, BN, and NN screen sets,[10] since theoretical and experimental studies of 2D screening systems have demonstrated that the efficiency of substructure searching is maximized if the screens have approximately equal frequencies of occurrence and if there are no positive statistical associations between these occurrences. It was shown in that study that the screen-set selection algorithm resulted in the identification of screens that exhibited a very high level of equifrequency of assignment when used for screening a database of structures; and that there were strong positive associations between these assignments, the degree of association being inversely related to the number of bonds in the generalized valence angle (so that the associations were greatest for the NN screen sets and smallest for the BB screen sets). Poirrette[16] details the results of an analogous study of the equifrequency of assignment and the interscreen associations for the generalized torsion angle screens considered in this paper. The findings of this analysis are entirely comparable to those that we have reported previously for generalized valence angle screens. Specifically, all of the screen sets exhibited a very high level of equifrequency of assignment, but the strength of the positive interscreen associations increased in line with the number of nonbonded interactions in the torsion definition. In the case of the generalized valence angles, this latter factor, coupled with the higher bit-string densities that are observed when nonbonded valence angles are considered, was shown to be responsible for the observation that the BB screen sets gave better screenout than the BN screen sets, which, in their turn, gave better screenout than the NN screen sets.[10]

Table 9. Median screenout, averaged over 19 searches of 5000 structures taken from the Cambridge Structural Database, for manually generated patterns of BBB torsion angles

Screen set size	Tolerance in degrees			
	0	5	10	20
128	0.98	0.94	0.92	0.88
256	1.00	0.96	0.94	0.91
512	1.00	0.97	0.95	0.92

Table 11 lists the mean interscreen associations for the BBB, NBB and NBB3 screen sets; when considered in conjunction with the bit-string densities in Table 4, it is hardly surprising that the BBB screenouts in Table 6 are superior to the NBB and NBB3 screenouts in Tables 7 and 8. It will be seen that the NBB3 associations are marginally greater than the NBB associations, despite the fact that the NBB3 screenouts are never less than, and usually superior to, the NBB screenouts; for these two types of screen sets, then, it would seem that the lower bit-string density (as shown by Table 4) is of greater importance than the screen associations in determining the precise level of screenout that is obtained. The fact that the NBB3 screen sets gave a generally superior level of performance than the NBB screen sets provides quantitative support for the suggestion of Bartlett *et al.* that a distance threshold should be used when generating NBB torsion angles.[15]

CONCLUSIONS

In this paper, we have discussed the use of generalized torsion angles for the implementation of angle-based screening in 3D substructure searching systems. It is possible to generate many different types of generalized torsion angle; however, some of these definitions overlap, and we have thus restricted our attention to the BBB, NBB, BNN, and NNN torsion angles.

The frequency distributions for BBB torsions are explicable in simple chemical terms; similar comments apply to the NBB torsions, which are dominated by the peaks at 0° and 180° that result from very commonly occurring ring systems, including phenyl rings. The peaks in the NBB distributions can be substantially reduced by adding a distance constraint when angles are being generated; the resulting distributions are much flatter than the basic NBB distributions. The distributions for the BNN and NNN torsions are characterized by a single broad peak near 0° that then falls rapidly away at larger angles, and we have shown that this behavior can be explained by purely topological considerations. BNN and NNN torsion angles occur very frequently in molecules, which makes them unsuitable for screening purposes with fixed-length bit strings of the sort considered here, and our substructure searching experiments have thus used only screen sets that are based on BBB and NBB torsions. These screen sets have been shown to give high levels of screenout with both automatically generated and manually generated query patterns.

Given a molecule containing X atoms, there are order $O(X)$ BBB and order $O(X^2)$ NBB torsion angles. There are also order $O(X)$ BB valence angles and order $O(X^2)$ BN

Table 10. Individual screenout results for 19 searches of 5000 structures taken from the Cambridge Structural Database for manually generated patterns of BBB torsion angles

	Screen set size											
	128				256				512			
Pattern	0	5	10	20	0	5	10	20	0	5	10	20
Chair1	0.91	0.69	0.61	0.55	0.91	0.69	0.63	0.55	0.91	0.69	0.62	0.56
Substituted-chair	0.95	0.82	0.74	0.58	0.98	0.81	0.75	0.68	0.98	0.82	0.75	0.69
Boat	0.99	0.82	0.74	0.66	0.99	0.82	0.76	0.67	0.99	0.83	0.76	0.68
Twist-boat	0.93	0.82	0.75	0.63	0.98	0.84	0.77	0.64	0.91	0.85	0.78	0.64
Envelope	0.99	0.87	0.80	0.68	1.00	0.89	0.82	0.69	1.00	0.90	0.83	0.71
Half-chair1	1.00	0.89	0.81	0.69	1.00	0.91	0.83	0.70	1.00	0.92	0.84	0.71
Chair2	0.99	0.94	0.82	0.89	0.99	0.96	0.93	0.91	1.00	0.97	0.94	0.91
Chair3	0.99	0.95	0.88	0.83	1.00	0.95	0.91	0.87	1.00	0.96	0.94	0.88
Half-chair2	1.00	0.97	0.96	0.90	1.00	0.98	0.95	0.90	1.00	0.99	0.97	0.91
Twist-chair1	0.98	0.89	0.81	0.69	1.00	0.90	0.83	0.69	1.00	0.92	0.84	0.70
Twist-chair2	0.99	0.97	0.96	0.91	1.00	0.99	0.98	0.93	1.00	0.99	0.98	0.95
Chair4	0.99	0.87	0.80	0.70	1.00	0.89	0.82	0.70	1.00	0.90	0.83	0.71
Crown	0.86	0.78	0.70	0.60	0.90	0.78	0.72	0.62	0.97	0.81	0.73	0.62
Boat2	0.98	0.83	0.74	0.66	0.98	0.82	0.77	0.66	0.99	0.84	0.76	0.67
Extended	0.91	0.91	0.91	0.91	0.94	0.94	0.90	0.90	0.96	0.93	0.91	0.91
Gauche-twist	0.95	0.95	0.95	0.95	0.97	0.97	0.96	0.93	0.99	0.97	0.95	0.95
Mixed	1.00	0.99	0.99	0.98	1.00	0.99	0.99	0.98	1.00	0.99	0.99	0.98
Double1	0.84	0.84	0.84	0.82	0.94	0.94	0.90	0.84	0.98	0.98	0.94	0.85
Double2	0.95	0.84	0.82	0.82	0.98	0.85	0.85	0.84	0.99	0.88	0.86	0.85

Table 11. The mean association between the assignment frequencies for screen sets based on the three types of torsion angle

Torsion type	Screen set size		
	128	256	512
BBB	0.08	0.05	0.04
NBB	0.16	0.17	0.18
NBB3	0.19	0.18	0.19

valence angles,[10] and it is thus interesting to note that a comparison of Tables 6 and 7 with our previous screenout results for BB and BN valence angles[10] shows that the two sets of screens give almost identical levels of performance, with neither of them being obviously superior to the other. We thus believe that both valence angles and torsion angles are appropriate for use in operational systems for 3D substructure searching, and facilities for the screening of conformational searches based on BBB torsion angles will be provided in a forthcoming version of the Cambridge Structural Database system; we are now investigating the use of angle-based definitions for 3D similarity searching.

ACKNOWLEDGMENTS

We thank the British Library Research and Development Department for funding this study under grant number RDD/G/872, and Tripos Associates for provision of the SYBYL molecular-modeling package.

REFERENCES

1 Ash, J.E., Warr, W.A., and Willett, P. (Eds.) *Chemical Structure Systems* Ellis Horwood, Chichester, 1991
2 Martin, Y.C., Bures, M.G., and Willett, P. Searching databases of three-dimensional structures. In: *Reviews in Computational Chemistry* (K.B. Lipkowitz and D.B. Boyd, Eds.) VCH, New York, 1990, pp. 213–263
3 Willett, P. *Three-Dimensional Chemical Structure Handling* Research Studies Press, Taunton, 1991
4 Martin, Y.C. Computer design of potentially bioactive molecules by geometric searching with ALADDIN. *Tetrahedron Computer Methodology* 1990, **3,** 15–25
5 Bures, M.G., Black-Schaefer, C., and Gardner, G. The discovery of novel auxin transport inhibitors by molecular modelling and three-dimensional pattern analysis. *J. Computer-Aided Molecular Design* 1991, **5,** 323–334
6 Jakes, S.E. and Willett, P. Pharmacophoric pattern matching in files of 3-D chemical structures: selection of inter-atomic distance screens. *J. Mol. Graphics* 1986, **4,** 12–20
7 Sheridan, R.P., Nilakantan, R., Rusinko, A., Bauman, N., Haraki, K.S., and Venkataraghavan, R. 3DSEARCH: a system for three-dimensional substructure searching. *J. Chemical Information and Computer Sciences* 1989, **29,** 255–260
8 Murrall, N.W. and Davies, E.K. Conformational free-

dom in 3-D databases. 1. Techniques. *J. Chemical Information and Computer Sciences* 1990, **30,** 312–316
9 Cringean, J.K., Pepperrell, C.A., Poirrette, A.R., and Willett, P. Selection of screens for three-dimensional substructure searching. *Tetrahedron Computer Methodology* 1990, **3,** 37–46
10 Poirrette, A.R., Willett, P., and Allen, F.H. Pharmacophoric pattern matching in files of 3-D chemical structures: characterization and use of generalized valence angle screens. *J. Mol. Graphics* 1991, **9,** 203–217
11 Allen, F.H., Davies, J.E., Galloy, J.J., Johnson, O., Kennard, O., Macrae, C.F., Mitchell, E.M., Mitchell, G.F., Smith, J.M., and Watson, D.G. The development of Versions 3 and 4 of the Cambridge Structural Database system. *J. Chemical Information and Computer Sciences* 1991, **31,** 187–204
12 Klyne, W. and Prelog, V. Description of the steric relationships across single bonds. *Experientia* 1960, **16,** 521–523
13 Newman, M.S. A notation for the study of certain stereochemical problems. *J. Chemical Education* 1955, **32,** 344–347
14 Nilakantan, R., Bauman, N., Dixon, J.S., and Venkataraghavan, R. Topological torsion: a new molecular descriptor for SAR applications. Comparison with other descriptors. *J. Chemical Information and Computer Sciences* 1987, **27,** 82–85
15 Bartlett, P.A., Shea, G.T., Telfer, S.J., and Waterman, S. CAVEAT: a program to facilitate the structure-derived design of biologically active molecules. In: *Molecular Recognition: Chemical and Biochemical Problems* (S.M. Roberts, Ed.) Cambridge, Royal Society of Chemistry, 1990, pp. 182–196
16 Poirrette, A.R. *The Use of Angle-Based Fragments for Screening Databases of Three-Dimensional Chemical Structures,* PhD thesis, University of Sheffield, 1992
17 Jakes, S.E., Watts, N.J., Willett, P., Bawden, D., and Fisher, J.D. Pharmacophoric pattern matching in files of three-dimensional chemical structures: evaluation of search performance. *J. Molecular Graphics* 1987, **5,** 41–48
18 Clark, D.E., Willett, P., and Kenny, P.W. Pharmacophoric pattern matching in files of three-dimensional chemical structures: use of bounded distances for the representation and searching of conformationally-flexible molecules. *J. Mol. Graphics* 1992, **10,** 194–204
19 Upton, G.J.G. and Fingleton, B. *Spatial Data Analysis by Example. Volume 2. Categorical and Directional Data* Chichester, John Wiley, 1989
20 Allen, F.H. and Johnson, O. Automated conformational analysis from crystallographic data. 4. Statistical descriptors for a distribution of torsion angles. *Acta Crystallographica* 1991, **B47,** 62–67