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

Andrew R. Poirrette and Peter Willett

Department of Information Studies, University of Sheffield, Sheffield, UK

Frank H. Allen

Cambridge Crystallographic Data Center, University of Cambridge, Cambridge, UK

This paper describes the use of generalized valence angles for the screening of pharmacophoric pattern searches in databases of three-dimensional chemical structures. A generalized valence angle is defined as the angle between two vectors, AB *and* BC, *which have a common vertex* B, *and in which both vectors correspond to formal chemical bonds; one vector corresponds to a bond and the other to a nonbonded interaction; or both vectors correspond to nonbonded interactions. The screens are identified by a statistical analysis of the frequencies of occurrence of these angle-based features in the Cambridge Structural Database. The occurrence frequencies are discussed and shown to be explicable in terms of small, commonly occurring structural features. The effectiveness of the screens is demonstrated by an extensive series of searches for representative pharmacophoric patterns. The results are compared with those obtained from a similar series of searches using distance-based screens: The latter are found to give a better level of performance, and evidence is presented to suggest that this is due to a high degree of association between the assignments of the angle-based screens.*

Keywords: bond angle, generalized valence angle, pharmacophoric pattern, three-dimensional substructure searching, valence angle

INTRODUCTION

The last few years have seen substantial interest in the development of substructure searching systems for databases of three-dimensional (3D) chemical structures;[1-9] the current status of this rapidly-emerging field has been reviewed by Martin et al.[10] and by Willett.[11] The available systems were originally developed to carry out searches for pharmacophoric patterns, but they are being used increasingly as general tools for rational drug design, e.g., for conformational analysis, structure–activity correlation, and model building.[12,13]

Following the pioneering studies by Gund,[1] searches for pharmacophoric patterns are usually implemented using a two-stage procedure.[3,7] The initial *screening* search is used to eliminate from further consideration large numbers of molecules that cannot possibly contain the query pattern. The screening search involves the comparison of a bit string describing geometrical characteristics of the query pattern with corresponding bit strings for each of the structures in the database that is being searched. Only those few molecules that pass the initial search then undergo the detailed and computationally demanding *geometric* search, which involves representing a query pattern and a database structure as labeled graphs and then checking for the inclusion of the former in the latter by means of a subgraph isomorphism algorithm. The inherent complexity of this stage means that the run-time efficiency of a substructure searching system is crucially dependent on the *screenout,* i.e., on the fraction of the search file that is eliminated by the screening search. This has led to the development of techniques for the selection of screens that will ensure a high level of screenout in searches for pharmacophoric patterns.[2,7,9,14]

The systems for 3D substructure searching that have been reported to date have typically used interatomic distance information as the basis for the screening stage, e.g., the systems developed by Pfizer Central Research,[3] Lederle Laboratories[7] and Chemical Design Limited.[9] Structural characteristics of the query that are describ̈ed in terms of angular information are usually considered only in the final

Address reprint requests to Prof. Willett at the Dept. of Information Studies, University of Sheffield, Western Bank, Sheffield S10 2TN, UK.
Received 4 June 1991; accepted 18 June 1991

geometric searching stage. In this paper, we consider the use of angular information for screening purposes in 3D substructure searching. The particular context in which our work has been carried out is the X-ray crystallographic data in the Cambridge Structural Database (CSD) produced by the Cambridge Crystallographic Data Centre.[15,16] This is the primary source of experimentally determined 3D coordinate data, and CSD offers a wide range of 3D searching facilities, principally through the GSTAT program.[17] However, although effective in operation, these facilities are very slow due to the lack of suitable screening mechanisms. This was acceptable during the early development of CSD, but the increasing size of the database, which now approaches 100 000 compounds, has led to a need to increase the efficiency of 3D searching, and hence to the development of both distance-based and angle-based screens. The latter requirement provided the original basis for the work reported in this paper. However, we believe that the approaches discussed below are also applicable to other 3D database systems, which include sets of coordinates generated by computational procedures, e.g., via force-field calculation[18] or rule-based systems, such as CONCORD.[19]

The paper is organized as follows. The second section discusses ways in which angle-based information may be used in screening, with particular emphasis on the *generalized valence angles,* as defined below, that form the primary focus of the paper. The third section provides a detailed account of the frequency characteristics of generalized valence angles. The fourth section then reports the results of an extensive series of 3D substructure searches that have been carried out using generalized valence angle screens. The section also compares the efficiencies of these sets of screens with sets of interatomic distance screens. The paper concludes with a summary of our main findings.

USE OF ANGULAR INFORMATION

Two types of angle have traditionally been used by chemists in their descriptions of molecular geometry: the *valence angle* and the *torsion angle*. If we consider a contiguously bonded, five-atom fragment $A–B–C–D–E$, then its angular description involves three valence angles ($\angle ABC$, $\angle BCD$, and $\angle CDE$), and two torsion angles ($\angle ABCD$ and $\angle BCDE$). The torsion angles[20] are signed quantities (which are defined to be clockwise positive and anticlockwise negative) that measure, e.g., the degree of twist of atom D relative to atom A when viewed along the bond $B–C$. It is important to emphasize that the accepted definitions of these angular descriptors involve bonded vectors only, as in our example molecular fragment. However, there is no reason, *a priori,* why one should not consider alternative, generalized definitions that involve both bonded and nonbonded vectors. The resulting *generalized valence angles* and *generalized torsion angles* describe geometrical relationships between atoms in 3D space, i.e., they describe subelements of the complete molecular pattern, and hence they can be considered for use as screens in 3D substructure searching. Thus, in our example fragment, there are three distinct types of valence angle that have atom C as the vertex:

- One conventional valence angle, $\angle BCD$, in which both apical atoms are bonded to C.
- Two angles, $\angle ACD$ and $\angle BCE$, in which only one of the apical atoms is bonded to atom C.
- One angle, $\angle ACE$, in which neither of the apical atoms are bonded to atom C.

In this paper, we shall describe these three types of angular feature in terms of the connectedness (or nonconnectedness) of the apical atoms to the vertex atom: we thus have BB (for bonded/bonded), BN (for bonded/nonbonded) and NN (for nonbonded/nonbonded) valence angles, respectively. Thus, for a molecule containing n (nonhydrogen) atoms there are $O(n)$ angles of type BB, $O(n^2)$ angles of type BN, and $O(n^3)$ angles of type NN.

It is, of course, possible to define other generalized angular features in much the same way. Thus, we may consider the more general intervector angles, such as the angle between the vectors AB and DE in our example fragment, and then classify these angles according to the connected or nonconnected nature of the constituent vectors. In the context of generalized torsion angles, the conventional $O(n)$ BBB angles would then be augmented by the $O(n^2)$ BBN and BNB angles, by the $O(n^3)$ BNN and NBN angles, and by the $O(n^4)$ NNN angles. We note that Bartlett et al. have already discussed the use of BNB torsion angles and BN valence angles in the CAVEAT program for the design of enzyme inhibitors.[22]

There are two main approaches to screening. In the first of these, which is the one adopted here and in our previous work on distance-based screens,[2,14] a screen dictionary is created containing a limited number of screens chosen from the potentially vast number of fragments that might be encountered in a large database. Each structure is then represented by a dedicated bit string in which each bit is used to denote the presence or absence of one of the chosen screens; the set of bit strings is usually stored as a bit map so that rapid searching is effected by accessing just those columns of the bit map that correspond to bits that have been set in the query bit string.[3] Alternatively, one can use every distance-based fragment type as a screen, as is done in the Lederle 3DSEARCH system, which currently contains over 13 000 different interatomic distance screens;[7] in this case, the resulting characterizations are normally stored using an inverted list organization.

In the context of angular screens, the availability of the types of angles described above provides a wide range of discriminating keys that could be used to define a query pattern and a database structure for retrieval purposes. However, the sheer number of angles that can be generated from some of the torsion angle definitions could make them unsuitable for screening purposes, using either of the approaches mentioned in the previous paragraph. In the first approach, the fixed and limited size (typically a few hundreds of bits) of the bit string describing an individual molecule means that a very large fraction of the bits is likely to be set, resulting in poor screenout (despite the discriminatory character of the substructural features that are being used to define the geometries of the query pattern and of the database structures). In the second approach, there will be a huge number of inverted lists, with a consequent large storage overhead, and very many of these will need to be intersected at search time, with consequent processing costs. For these reasons, our initial experimental studies of angle-based screening have focused on the three types of generalized valence angles, as described below.

CHARACTERIZATION OF GENERALIZED VALENCE ANGLES

Definition of fragment types

Our studies have used a 5000-structure randomly selected subset of CSD. The elemental types of the atoms in these structures were generalized so that each atom belonged to just one of three classes: C (for carbon), X (for nitrogen or oxygen) and Y (for any nonhydrogen element except for carbon, nitrogen, or oxygen). Thus the three atoms comprising a BB angle define one of the 18 following types of angle: C–C–C, C–C–X, C–C–Y, C–X–C, C–X–X, C–X–Y, C–Y–C, C–Y–X, C–Y–Y, X–C–X, X–C–Y, X–X–X, X–X–Y, X–Y–X, X–Y–Y, Y–C–Y, Y–X–Y, and Y–Y–Y (where the symbol '–' denotes a pair of atoms that are bonded together). Given the occurrence of a fragment P–Q–R, a canonicalization procedure is used to ensure that $P \le R$, so that, e.g., the occurrence of X–C–Y and Y–C–X are posted to the single heading of X–C–Y. In the case of the BN fragments, it is not possible to carry out such a procedure owing to the bonded and nonbonded natures of the two types of interatomic relationship, so that, e.g., C–Y*X is not the same as C*Y–X (where the symbol '*' denotes a pair of atoms that are not bonded together). There are thus 27 possible types of BN angle: C–C*C, C–C*X, C–C*Y, C–X*C, C–X*X, C–X*Y, C–Y*C, C–Y*X, C–Y*Y, X–C*C, X–C*X, X–C*Y, X–X*C, X–X*X, X–X*Y, X–Y*C, X–Y*X, X–Y*Y, Y–C*C, Y–C*X, Y–C*Y, Y–X*C, Y–X*X, Y–X*Y, Y–Y*C, Y–Y*X, and Y–Y*Y. There are 18 possible types of NN angle, after canonicalization, as with the BB angles. However, preliminary tests using NN screen sets showed that they gave very poor levels of screenout. This is a necessary consequence of the fact that $O(n^3)$ NN screens are assigned to each molecule, with the result that the corresponding bit-string characterizations contain a very high percentage of bits set and that, consequently, the great majority of the structures are retrieved in response to most query patterns.[23] Accordingly, a more restricted type of fragment definition was used, which took into account only the fragment types that contained, at most, only a single carbon atom. There are thus 13 possible types of NN angle: C*X*X, C*X*Y, C*Y*X, C*Y*Y, C*C*X, X*C*Y, X*X*X, X*X*Y, X*Y*X, X*Y*Y, Y*C*Y, Y*X*Y, and Y*Y*Y.

Generalized valence angles were generated for each of the 5000 CSD structures. The occurrences of the angles were cumulated and plotted as frequency distributions for each of the 18 BB, 27 NB, and 13 NN fragment types. Some typical distributions are illustrated in Figures 1, 2, and 3 for the BB, BN and NN classes, respectively. These distributions are discussed in the remainder of this section.

BB angles

The BB distributions of Figure 1a–h, representing conventional valence angles, show a wide variation in shape. However, these variations are readily interpreted in terms of the functional groups, ring systems, and metal coordination geometries that are chemically common and therefore are well represented in any random subset of structures derived from CSD. Thus, the C–C–C valence angle distribution of Figure 1a shows two dominant peaks: a sharp maximum at 120°, arising from benzenoid and ethylenic systems having C sp^2 atoms as the vertex of the valence angle, and a broader shoulder at ~ 105–112° corresponding to the normal (and more flexible) tetrahedral angle at C sp^3. The only other

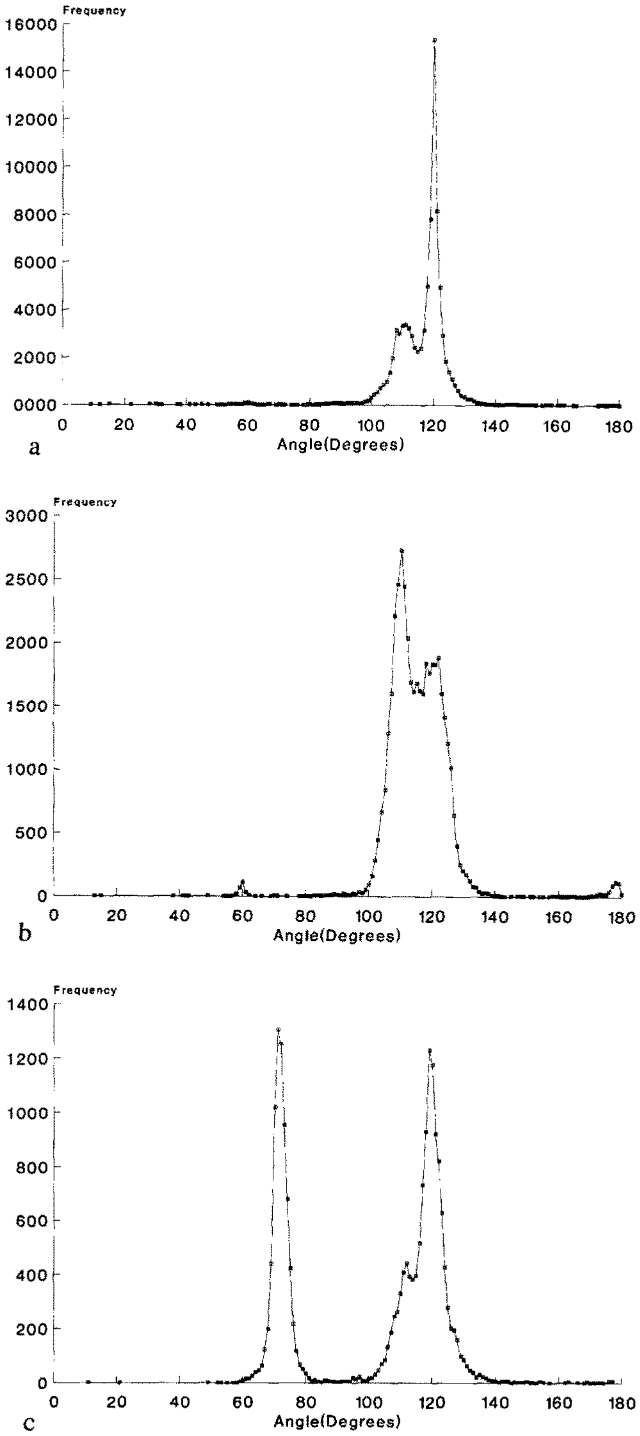


Figure 1. Frequencies of occurrence for BB fragments: a, C–C–C; b, C–C–X; c, C–C–Y; d, C–Y–C; e, X–C–Y; f, Y–Y–Y; g, X–Y–Y; and h, Y–X–Y.

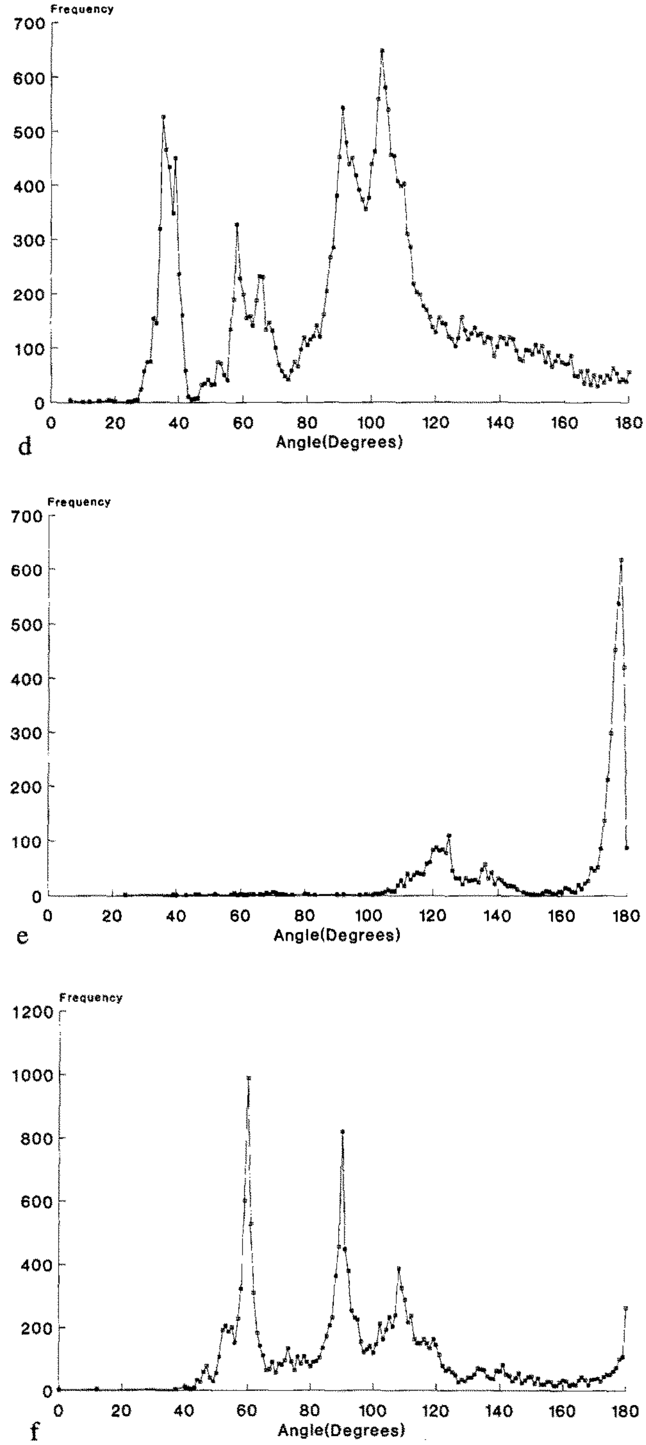


Fig. 1 cont'd

areas in Figure 1a that show any density at all are at 60° (cyclopropane rings), 90° (cyclobutane rings), and 180° (acetylenic systems). The dominance of benzenoid systems, which occur in some 47% of CSD entries and which generate a minimum of six independent angles in each case, is such that the smaller peaks are virtually invisible in Figure 1a. These points are exemplified better in Figure 1b, which shows the distribution of C–C–X valence angles. Here the

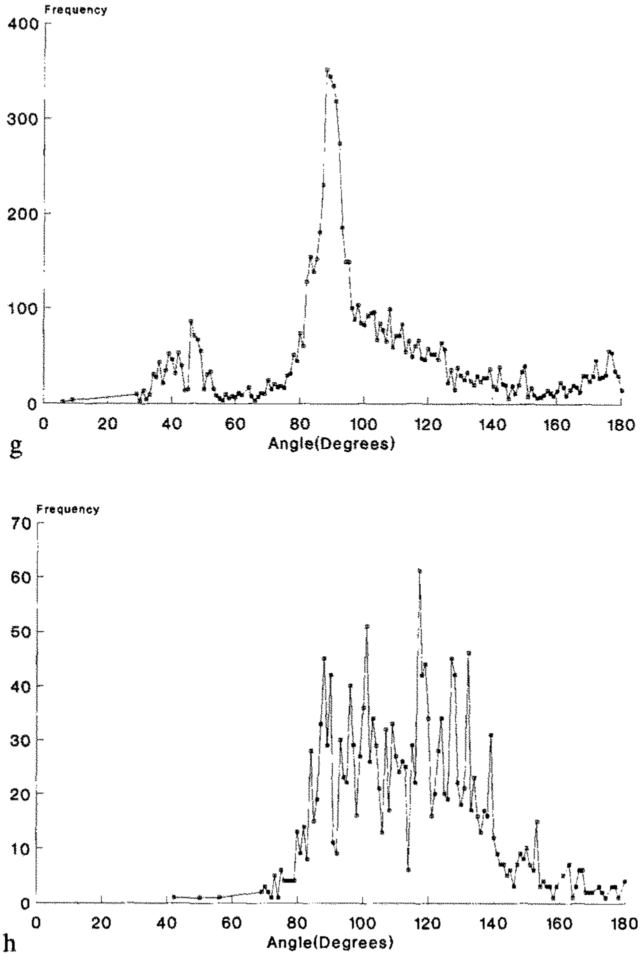


Fig. 1 cont'd

tetrahedral peak is larger than that due to the C sp^2 vertices, since a restricted number of the benzene rings are directly substituted by X atoms (i.e., N or O), and the valence-angle multiplicity is reduced from six (for C–C–C angles in the ring) to two (for each C–C–X involving an N or O substituent). Further, small peaks at ~60° (from epoxides and aziridine) and at 180° (from substituted acetylenes and cyano groups) are now clearly visible.

The distribution of C–C–Y (where Y is any atom other than C, N, or O) valence angles is shown in Figure 1c. This again shows a doublet in the range 100–125°, corresponding to angles having C sp^3 and C sp^2 as the vertex atoms, as might be expected from the significant proportion of halogen substituents present in CSD entries. However, Figure 1c also shows a sharp peak in the 70–75° area, arising from C–C–M (where M corresponds to a metal atom) fragments in π-cyclopentadienyl (metallocene) structures. Each metal-ring π-interaction generates 10 independent angles of this type, a multiplicity that accounts for the peak height. The metallocenes are also responsible for the low-angle peaks, at 30–40° and 60–70°, which are present in the C–Y–C distribution of Figure 1d. Both result from C–M–C π-bonded fragments. The lower peak arises from fragments in which the two carbons have a bonding 1, 2 relationship in the cyclopentadienyl ring, while the 60–70° peak is due to fragments where the two carbons have a 1, 3-transannular relationship. The remaining broad 90–110° peak in Figure 1d arises from the various arrangements of the very common terminal carbonyl and cyano ligands around metal centers. Indeed, it is the linearity of these *M*–C=O and *M*–C=N systems that account for the dominant 180° peak in the *X*–C–*Y* distribution of Figure 1e.

All of the remaining 13 BB distributions are susceptible to analogous interpretation based upon chemical knowledge and some prior knowledge of the structure types that dominate CSD. In general, those distributions that involve only atoms of classes C and *X* (i.e., N, O) show just a few discrete peaks that correspond to the limited range of directed valencies available to these atoms. However, those distributions that involve atoms of type *Y*, a group which is dominated by the halogens and, particularly, by the metallic elements, show a much greater variety of valence angle values. This is as expected from the wide range of coordination numbers exhibited by these elements, together with the relatively low energies that are required to deform coordination geometries from their ideal configurations. This variability is well illustrated in the distributions of *Y*–*Y*–*Y*, *X*–*Y*–*Y*, and *Y*–*X*–*Y* valence angles shown in Figures 1f, 1g, and 1h, respectively. The *Y*–*Y*–*Y* distribution shows peaks at 60° (from triangulo-trimetal arrangements), 90° (from square planar and octahedral coordinations), 110° (tetrahedral, with a small trigonal shoulder at 120°), and 180° (from octahedral coordinations). The *X*–*Y*–*Y* distribution of Figure 1g is dominated by square-planar coordinations, and also shows the wide variability of angles at metal centers, while the *Y*–*X*–*Y* distribution of Figure 1h arises from N, O-bridging of two metal centers, the variability being due to the different ligand environments of the *X* atom and to the very wide range of metals encompassed by the *Y* atoms.

BN angles

Four of the BN distributions are illustrated in Figure 2. They are fully representative of the complete set of 27 distributions, all of which show a similar shape. This consists of an underlying curve with a broad maximum in the 15–150° range. The distributions are usually slightly skewed, rising steeply from 0°, but declining more steadily towards 180°. In many cases, sharp discrete maxima rise from this general curve, almost always at angles that are less than 100°. Major peaks at 0° occur in only two of the distributions, C–*X***Y* and the C–*Y***X* distribution illustrated in Figure 2d and discussed below. Large peaks at 180° are not observed.

By comparison with the BB valence angle distributions of Figure 1, the BN distributions summarize pattern de-

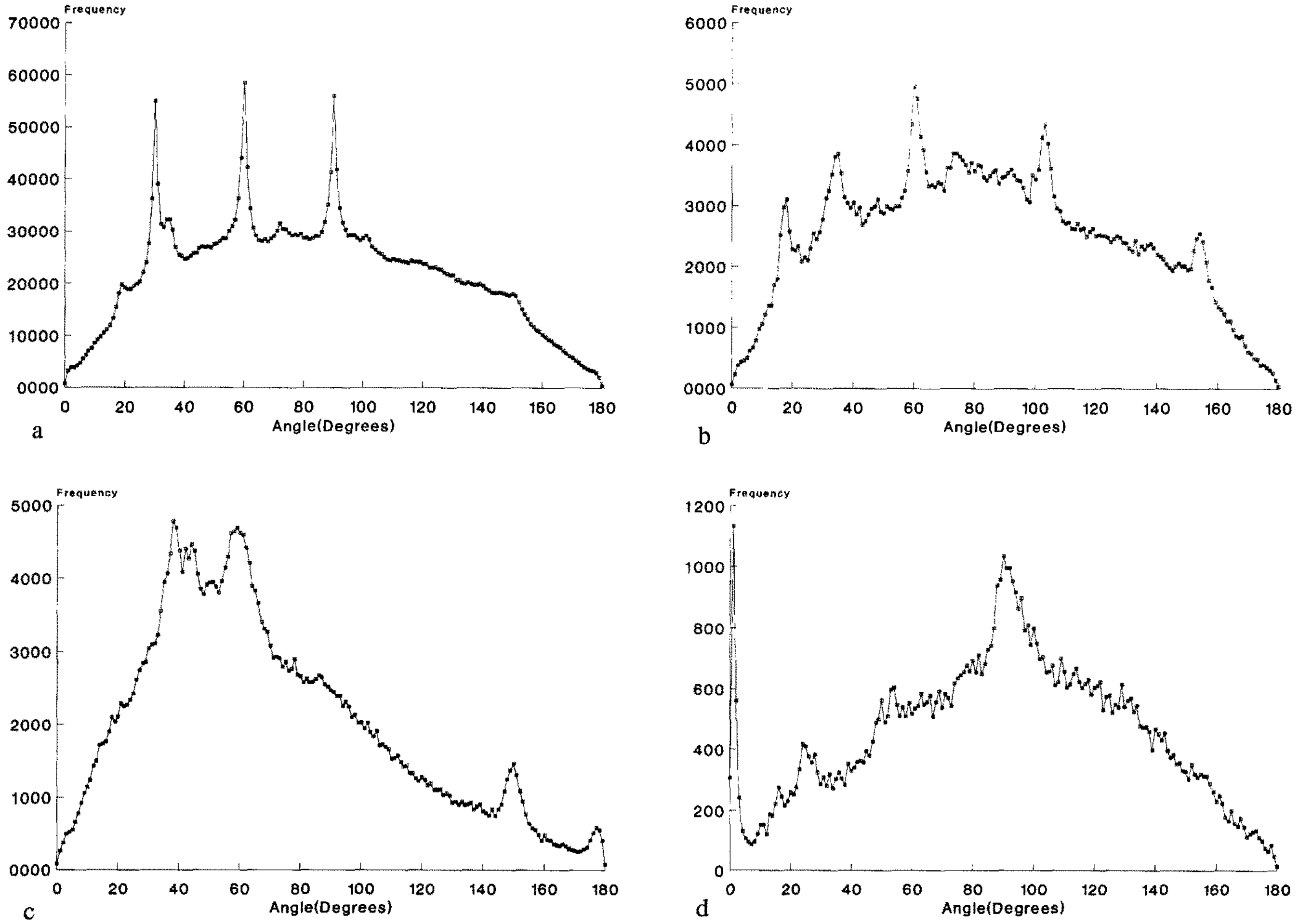


*Figure 2. Frequencies of occurrence for BN fragments: a, C–C*C; b, C–C*Y; c, Y–C*C; and d, C–Y*X.*

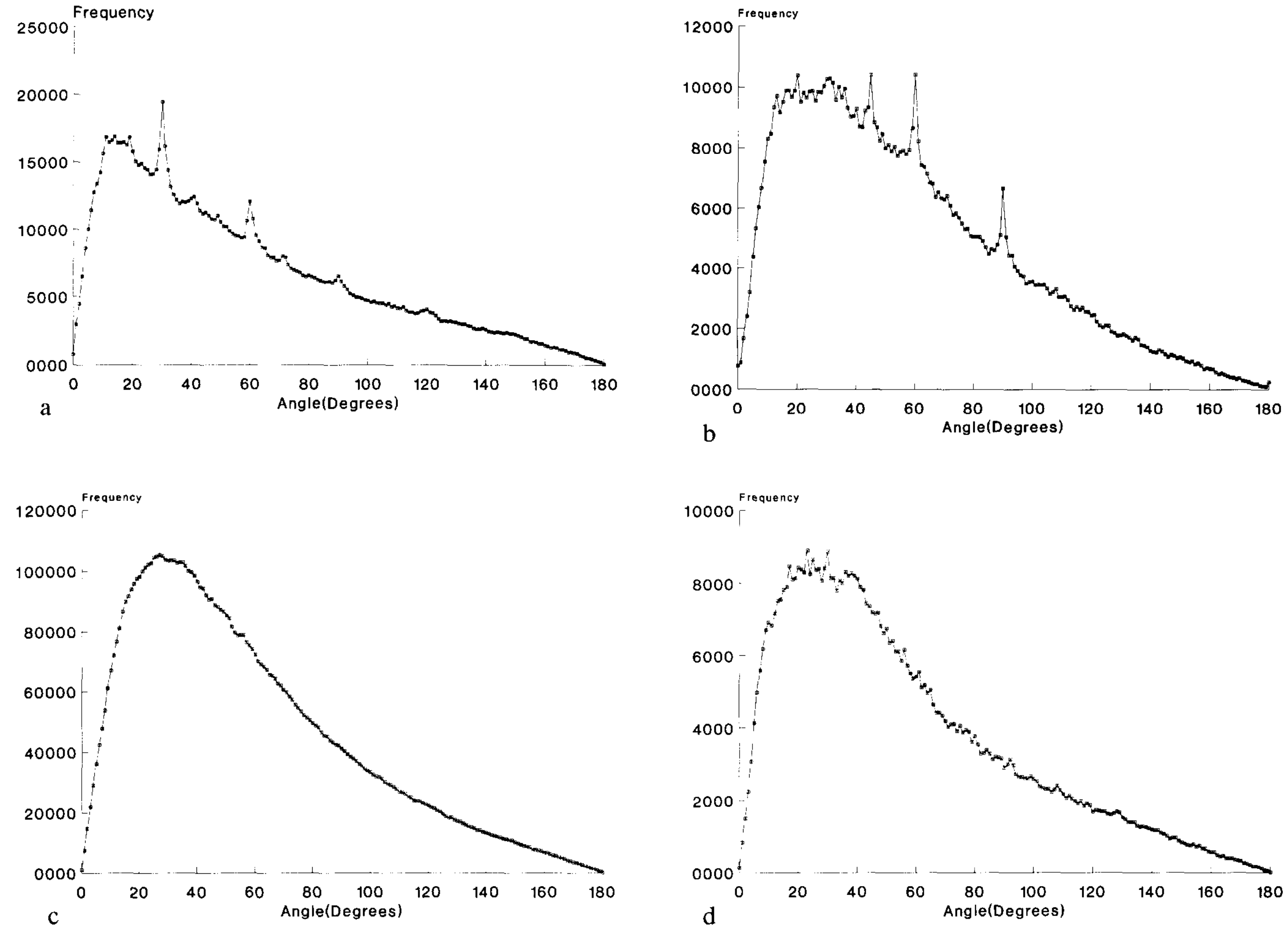
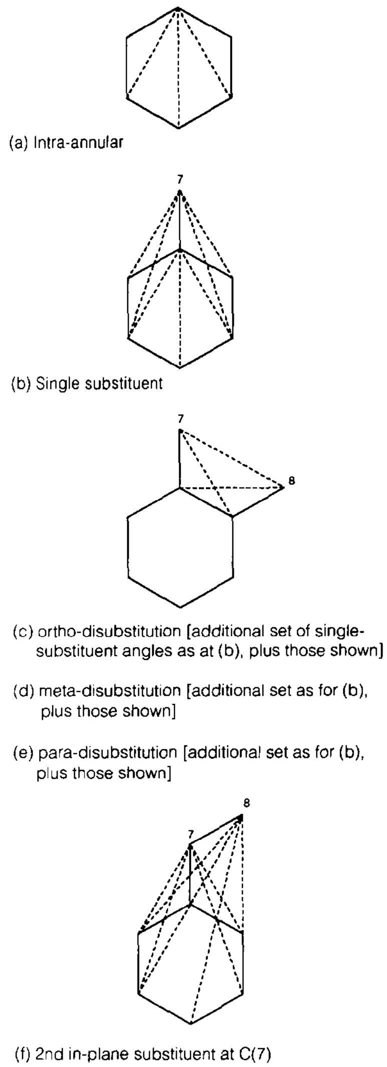


Figure 3. Frequencies of occurrence for NN fragments: a, C*C*C; b, Y*Y*Y; c, X*C*X; and d, Y*X*Y.

scriptors that are unfamiliar to structural chemists (and similar comments apply to the NN distributions in Figure 3). They are, however, readily interpretable in terms of the predominant 3D structural motifs identified in the discussion of the BB distributions. Thus, we already know that the BN distribution for C–C*C fragments, as shown in Figure 2a, is dominated by benzenoid aromatics. In Figure 4, we summarize the idealized BN angles that are generated by a benzene ring itself, and by various types of simple in-plane substituents. In Figure 5(a), we show how the BN frequency table for a specific benzenoid compound can be built from the "components" of Figure 4. The sharp maxima at 30°, 60°, and 90° in the full distribution of Figure 2a are already apparent in the distribution for this single simple molecule. These peaks are primarily due to the twelve-fold multiplicity of the intra-annular BN angles, coupled with the common occurrence of these angle values in the substituent interactions. Further, the small shoulders at 15° and 150°, predicted by Figure 5, are also clearly visible in Figure 2a. We have also calculated BN distributions for a number of simple aliphatic and alicyclic molecules. In hexamethylcyclohexanes, for example, we obtain a distribution that is very similar to that from the planar analogue hexamethylbenzene, except that the major peaks are more diffuse (30–36°, 54–60°, and 88–93°), and intermediate values are more evenly represented up to 110°. Of the 204 BN angles in each distribution, 48 exceed 110° in the planar molecule, while only 28 exceed this value in the substituted chair-form cyclohexyl-compound.

The C–C*Y curve of Figure 2b exhibits five major peaks at 15, 30°, 60°, 105°, and 150°. Given the occurrence of large numbers of halogeno-substituted phenyl rings in CSD, inspection of Figure 4b (with atom 7 taken as the halogen) shows that these are the only C–C*halogen angles possible and that each occurs with twofold multiplicity. The remaining angles of 150° (multiplicity 2) and of 180° (multiplicity 1) in Figure 4b are Y–C*C if atom 7 is a halogen; the small peaks that correspond to these values can be seen quite clearly in the full Y–C*C distribution of Figure 2c. (Similarly, the final angle of 0° from Figure 4b is C–Y*C if atom 7 is a halogen and a small peak at this position can be observed in the full distribution, which is not illustrated here).

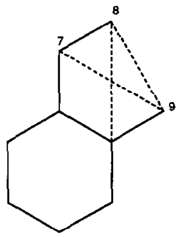
Apart from the halogen–C*C angles noted above, the Y–C*C distribution of Figure 2c has dominant peaks at 35–45° and at ~60°. The latter arises from metal–C*C angles in π-cyclopentadienyl complexes, angles in which the nonbonded carbons have a 1, 3-intraannular relationship.



	Angles	Multiplicity
(a) Intra-annular	BN	BN (NN)
	30	12 (12)
	60	12 (6)
	90	12
(b) Single substituent	0	1
	15	4
	30	4
	60	2
	105	2
	150	4
	180	1
(c) ortho-disubstitution [additional set of single-substituent angles as at (b), plus those shown]	60	2
	90	2
(d) meta-disubstitution [additional set as for (b), plus those shown]	30	2
	135	2
(e) para-disubstitution [additional set as for (b), plus those shown]	0	2
	180	2
(f) 2nd in-plane substituent at C(7)	0	1
	15	2
	30	4
	45	2
	60	3
	75	1
	90	3
	105	1
	120	2
	135	2
	150	2
	180	1

Figure 4. Incremental build-up of BN valence angles as a planar benzene ring is progressively substituted by in-plane substituents: a, intra-annular angles; b, single substituent angles; c, ortho-disubstituent angles additional to the set of single-substituent angles shown in Figure 4b, d, meta-disubstituent angles additional to the set of single-substituent angles shown in Figure 4b; e, para-disubstituent angles additional to the set of single substituent angles shown in Figure 4b; and f, the effect of a second in-plane substituent on atom 7

These angles have eightfold multiplicity in the complete unit, and are the "complement" of the C–*M*–C BB valence angles of Figure 1d. Thus, a pair of angles is generated by the approximately equilateral triangle formed by two *M*–C π-bonds and the C*C nonbonded interaction. The low angle 35–45° peak in Figure 2c is related in a similar complementary manner to the broad 90–110° BB valence angle peak in Figure 1d, which has been ascribed to O, N═C–*M*–C═O, N units. A triangle, now roughly isosceles, is again formed by the C–*M* bonds and the C*C nonbonded vector. Indeed, it is the *M*–C═N, O unit that also dominates the C–*Y***X* BN distribution of Figure 2d. Here, the large peak at 0° is obviously due to the intraligand C–*M***X* angles, while the dominant peak at 90° is from the same source, but with the carbon and the *X* (i.e., N or O) atom in neighboring ligands.



(a) Frequencies of occurrence of idealized BN angles.

Angle	Ring	C7	C9	ortho	C8	C8-C9	Total
0		1	1		1		3
15		4	4		2		10
30	12	4	4		4		24
45					2		2
60	12	2	2	2	3	2	23
75					1		1
90	12			2	3	2	19
105		2	2		1		5
120					2		2
135					2		2
150		4	4		2		10
165							0
180		1	1		1		3

(b) Frequencies of occurrence of idealized NN angles.

Angle	Ring	C7	C9	ortho	C8	C8-C9	Total
0		1	1		3		5
15		6	6		12		24
30	12	5	5	2	8	7	39
45		4	4	6	4	1	19
60	6	3	3	2	2	4	20
75		2	2	2	1	2	9
90		2	2	2	2	1	9
105					1	1	2
120		2	2		2	1	7
135							
150					2		2
165							
180					1		1

Figure 5. Frequencies of occurrence of BN (a) and NN (b) fragments for the illustrated simple, disubstituted benzene compound. These frequencies may be determined by inspection of the appropriate parts of Figure 4, with additional occurrences arising from the C8–C9 interaction that is indicated in the compound by dashed lines

NN angles

All of the 18 NN distributions have a shape that is even more regular than those of the BN distributions discussed above. This shape, of which Figures 3a–d are typical, is a positively skewed distribution with a mode in the 20–30° range. In the vast majority of cases the curve is smooth, but for the C*C*C and *Y***Y***Y* fragments shown in Figures 3a and 3b, respectively, sharp and discrete peaks rise from this curve. For the C*C*C case, the shape of the underlying curve, and also the positions of the maxima, can be understood in terms of an NN analysis of simple benzenoid compounds that is entirely analogous to the analysis of Figure 4. Figure 5b summarizes the results of this NN analysis for the illustrated planar disubstituted phenyl ring. The distribution of the 137 NN angles in this single unit is entirely representative of the full C*C*C distribution shown in Figure 3a. Thus, Figure 5b predicts a clear maximum at 30°, with a second (ca. half-height) peak at 60°. The initial peak at ~15°, and the small 'blips' at 45°, 75°, 90°, and 120° in Figure 3a, are also consistent with the results of Figure 5b. As with the corresponding BN distribution for C–C*C, as shown in Figure 2a, it is the high multiplicity of the intraannular angles that gives rise to the very strong 30° and 60° peaks. Further, we note that the distributions due to the individual substituents, listed in columns 2, 3 and 5 of Figure 5b, illustrate the underlying skewed shape of the full distributions of Figure 3. The sharp peaks in the $Y*Y*Y$ NN distribution of Figure 3b are entirely consistent with the various metal coordination geometries displayed by Figure 1f. Thus, NN angles of 45° and 90° will predominate in square-planar and octahedral arrangements, while NN angles of 60° will arise from tetrahedral arrangements. The two other NN distributions of Figures 3c and 3d are unremarkable, but are included here to reinforce the overall shape similarity in this series.

EVALUATION OF PERFORMANCE

Experimental details

Screen sets were generated from the BB, BN, and NN fragment occurrence data. These screen sets were assigned to each of the query patterns and to each of the 5000 CSD structures, and the resulting characterizations were then evaluated in an extended series of 3D substructure searches.

Generation of screen sets. There has been considerable interest in the development of methodologies for the selection of sets of screens that will provide high screenout in substructure searches (in both 2D and 3D). This work has demonstrated the importance of *equifrequency*, i.e., the idea that fragments chosen as screens should occur approximately equifrequently in the database that is to be searched.[2] This concept underlies the screens used in the present investigation, which were all generated by means of the screen set selection algorithm described by Cringean et al.[14]

In essence, the screen set selection algorithm involves the following three steps:

(1) Assume that some class of fragment descriptor has been selected as the basis for the screening system that is to be implemented: in the context of the present paper, this will be the class of BB, BN, or NN valence angles.

(2) Generate all occurrences of this fragment class in a sample of chemical structures. Sort the occurrences into increasing alphanumeric order and cumulate them to give a *fragment dictionary*, in which each particular type of the chosen fragment class is stored with its frequency of occurrence.

(3) Assume that a screen set is to be created that contains S screens. Partition the fragment dictionary into S partitions, each of which contains approximately the same number of fragment occurrences. Each of the resulting partitions then corresponds to one of the screens that are available for assignment to database structures or to query substructures.

The screen set selection algorithm used here provides an effective way of implementing the partitioning step, Step 3. In summary, the algorithm first makes a rough division of the dictionary into approximately the required number of partitions and then refines the lower and upper bonds of these to give the final required number of partitions: the procedure is described in detail by Cringean et al.[14] The range of fragment values in each partition resulting from the use of this algorithm then defines one of the screens in the screen set. Thus a typical BB screen might be of the form C–C–C 93 103, denoting a valence angle containing three carbon atoms bonded together at angles between 93° and 103° (in our work, all angles have been rounded to the nearest degree). A database structure or query substructure is encoded by generating each fragment in turn and searching the dictionary to identify the partition in which the fragment occurs. For example, the screen mentioned previously would be assigned to any molecule that contained three carbons having a BB valence angle V such that $93 \leq V \leq 103$. Once the appropriate screen has been identified, the corresponding bit is set in the bit string that is used for the screening stage of the substructure search.

This screen set selection procedure was used to create screen sets containing 128, 256, or 512 screens. The screen sets were used to produce a bit-string representation of each of the 5000 CSD molecules; these bit strings were then stored as a bit map for rapid query processing.

We have noted previously that distance-based screening mechanisms are widely used for 3D substructure searching systems. For comparison purposes, we have generated sets of 128, 256, and 512 interatomic distance screens, using the same algorithm as for the generation of the angle screen sets[14] and with comparable atomic classes, i.e., C, X, and Y.

Query patterns. A subfile containing 100 structures was created from the file of 5000 structures. A query pattern was generated by randomly selecting two, three, or four valence angles from each of the structures in this subfile, so that there were 100 patterns for each size. Each pattern was searched for with tolerances of ±0° (i.e., exactly as specified), 1°, 2°, 3°, 4°, or 5°. In fact, three different sets of 100 structures were created; the results presented here correspond to just one of these three sets of queries. Analogous distance-based searches were carried out with tolerances of ±0.0, 0.1 and 0.5 Å on each of the interatomic distances in the query patterns.

A previous study at Sheffield used a set of published pharmacophoric patterns that consisted of nonbonded atoms and the associated interatomic distances.[3] These were converted into angle-based patterns and searched for, using the NN screen sets (because of the nonbonded nature of the atoms in these patterns). The distances in the published patterns are normally specified as distance ranges: When these were converted to angles, it was found that many of the resulting angular ranges were very wide. For example, the antileukemic pharmacophoric pattern[24] contains three atoms and the interatomic distance ranges 8.62 ± 0.5 Å, 7.08 ± 0.56 Å, and 3.35 ± 0.65 Å; these distance ranges result in a query pattern that contained the NN valence-angle ranges 126 ± 45°, 39 ± 33°, and 16 ± 14°, i.e., 81–171°, 6–72° and 2–30°. Searches were carried out using these ranges as calculated, and also taking the midpoint in the range and then searching with tolerances of ±0° and ±5°; thus, the antileukemic pattern searched for using the latter tolerance figure contained the NN valence angle ranges 121–131°, 34–44°, and 11–21°.

Measurement of 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: the most widely used measure of performance is the *screenout*. 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.

Results and Discussion

Equifrequency of screen assignment. The screen set generation algorithm used in this work is intended to identify angular or distance ranges that occur approximately equifrequently in the database that is to be screened. The degree of equifrequency is commonly measured by the *relative entropy*, H_R, which is defined as follows. Assume that a screen set contains S screens and that the Ith such screen has been assigned to a database of structures $FREQ(I)$ times ($1 \leq I \leq S$). Let $TOTAL\_FREQ$ be defined by

$$TOTAL\_FREQ = \sum_{I=1}^{S} FREQ(I)$$

then the relative entropy H_R is calculated from

$$H_R = \frac{-1}{log_2 S} \sum_{I=1}^{S} G(FREQ(I))$$

where $G(F(I)) = 0$ if $FREQ(I) = 0$ and

$$\frac{FREQ(\mathrm{I})}{TOTAL\_FREQ}\, log_2 \left(\frac{FREQ(I)}{TOTAL\_FREQ}\right)$$

otherwise

If the frequencies are exactly equal, so that $FREQ(I) = FREQ(J)$ for all I and J, then H_R is unity; in general however, H_R will be rather less than this upperbound value, the difference from unity increasing as the $FREQ(I)$ values become more disparate. Thus H_R provides a simple and direct measure of the effectiveness of a screen set generation algorithm in selecting equifrequently occurring fragments. Table 1 lists the relative entropies of assignment for each of the CXY screen sets, the numbers of fragment occurrences that acted as the input to the screen set generation algorithm, and the H_R values for this original occurrence data. It will be seen that all of the screen sets exhibit a high degree of relative entropy, and that these values are much greater than those for the original data. The table also contains the analogous results for the distance-based screen sets.

Degree of screen association. Equifrequency of screen assignment is a necessary, but not sufficient, condition for a high level of screenout to be obtained in a wide range of substructure searches. Account also must be taken of the degree of association between the screens once they have been assigned.[25,26] If two fragments I and J occur in proportions $P(I)$ and $P(J)$ of the compounds in a database, then a search that specifies both of these screens will eliminate a proportion $P(I) \times P(J)$ of the compounds if, and only if, the assignments of the two screens are statistically independent of each other. If the assignments are not independent, the screenout will be less than or greater than expected, depending upon whether there is a positive or a negative association, respectively. The association of 2D screens has been investigated by Adamson et al.,[25] who described a method for investigating the extent of the associations in a set of screen assignments. The geometric relationships that exist between many of the screens that are assigned to each molecule in the work reported here suggested that strong associations might be present that could affect the screenout obtainable in 3D substructure searches; we have accordingly adopted the methods of Adamson et al. to determine whether this is, in fact, the case.

Table 1. Frequencies of occurrence and relative entropies for angle-based and distance-based screen sets

Screen set	Number of fragments	Relative entropy			
		Initial	128	256	512
BB	300 119	0.79	0.98	0.95	0.92
BN	10 881 746	0.89	1.00	0.99	0.99
NN	54 995 776	0.92	1.00	1.00	0.99
Distances	3 237 428	0.87	1.00	1.00	1.00

Consider a database of structures that has been screened using a set of S screens. An $S \times S$ matrix P is constructed, in which the IJth element $P(I, J)$ contains the proportion of the database that has been assigned both the Ith and the Jth screens. The association $V(I, J)$ between these two screens is then given by

$$V(I, J) = \frac{P(I, J) - P(I) \times P(J)}{\sqrt{(P(I) - P(I)^2)(P(J) - P(J)^2)}}$$

$V(I, J)$ is equal to $+1.00$ (-1.00) if all (none) of the structures that contain the Ith screen contain the Jth screen; more generally, $V(I, J)$ will lie between these two limiting values.

Figures 6–8 show the relative frequency distributions for the degrees of association obtained with the three types of screen set, with each figure showing the distributions for the 128-, 256-, and 512-member screen sets. An inspection of these figures shows that for a given size of screen set, i.e., 128, 256, or 512, the magnitude of the associations increases as one moves from BB to BN and then to NN. Thus, while some of the BB associations are negative, the preponderance of positive-valued BN and NN associations suggests that these screen sets will give relatively low screenout (even leaving aside the problem of the bit-string packing densities discussed below). Both the BN and NN distributions reveal a fair amount of structure within this general trend. For example, the NN distribution shows fairly well-defined peaks at associations in the ranges 0.2–0.4 and 0.4–0.6, while the BN distribution shows a number of peaks at around 0.0, 0.3–0.4, and 0.5–0.6.

This behavior was investigated by selecting an individual

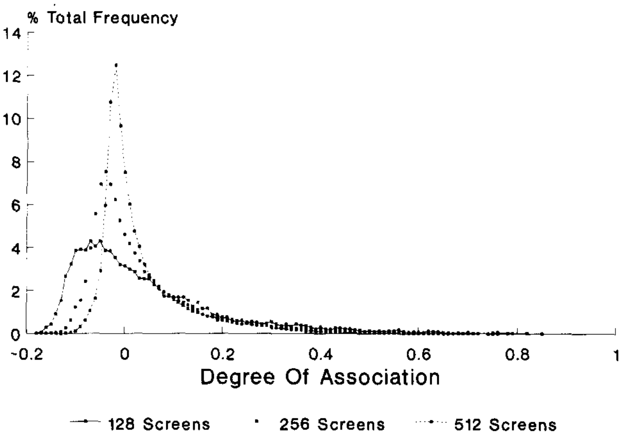


Figure 6. Distribution of screen associations for BB screen sets

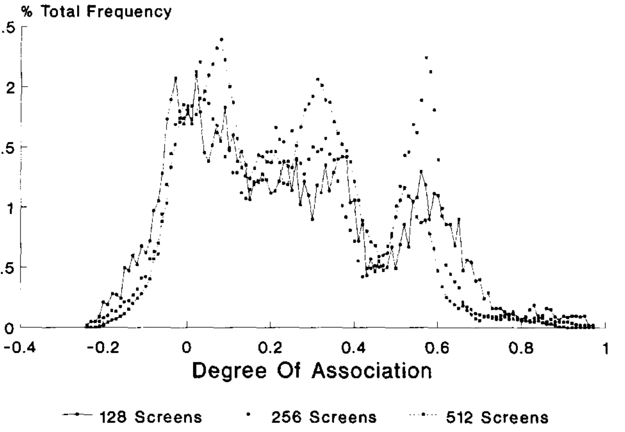


Figure 7. Distribution of screen associations for NB screen sets

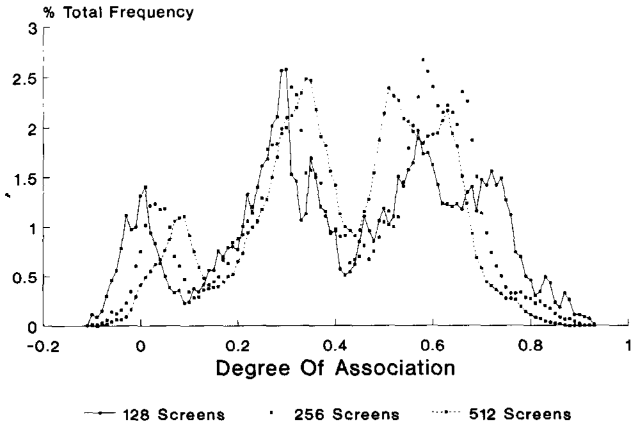


Figure 8. Distribution of screen associations for NN screen sets

screen X and determining the association of X with each of the other screens in its screen set. Examples of this behavior, using each of the three types of screen set, are shown in Figures 9a, 10a and 11a: comparable results are obtained if other screens are chosen. An inspection of these distributions shows clearly defined peaks and troughs, i.e., the chosen screen X has high associations with some groups of screens, and low associations with other groups. In fact, the strongest associations occur with fragments of the same class as X so that, e.g., if X is a C–C*C screen, then the peaks occur in that part of the distribution that corresponds to the C–C*C screens. Moreover, the locations of these peaks and troughs correspond to what we shall refer to as *boundary points*: places in the screen set where the type of screen changes from one fragment class to another. For example, boundary points exist where the screens change from C–C*C to C–C*X or from X–Y*X to X–Y*Y.

The boundary points for each screen set have been overlayed onto the association distributions as shown in Figures 9b, 10b, and 11b: it will be seen that there is an extremely close fit between the locations of the boundary

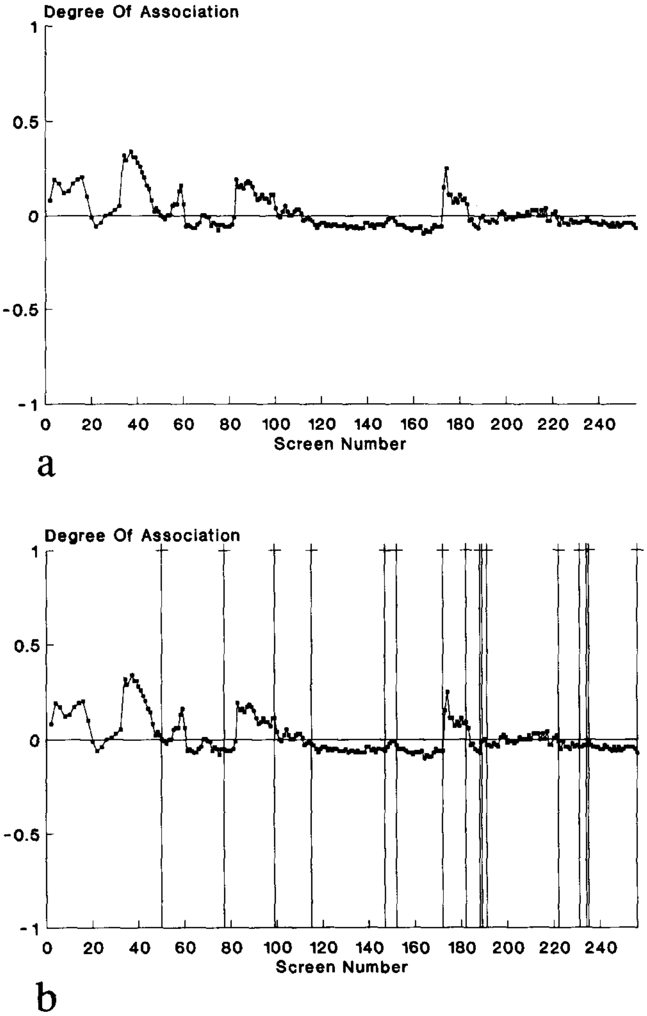


Figure 9. a, Association of screen 36 in the 256-member BB screen set with the other screens in this set; and b, as (a) but with the boundary points superimposed

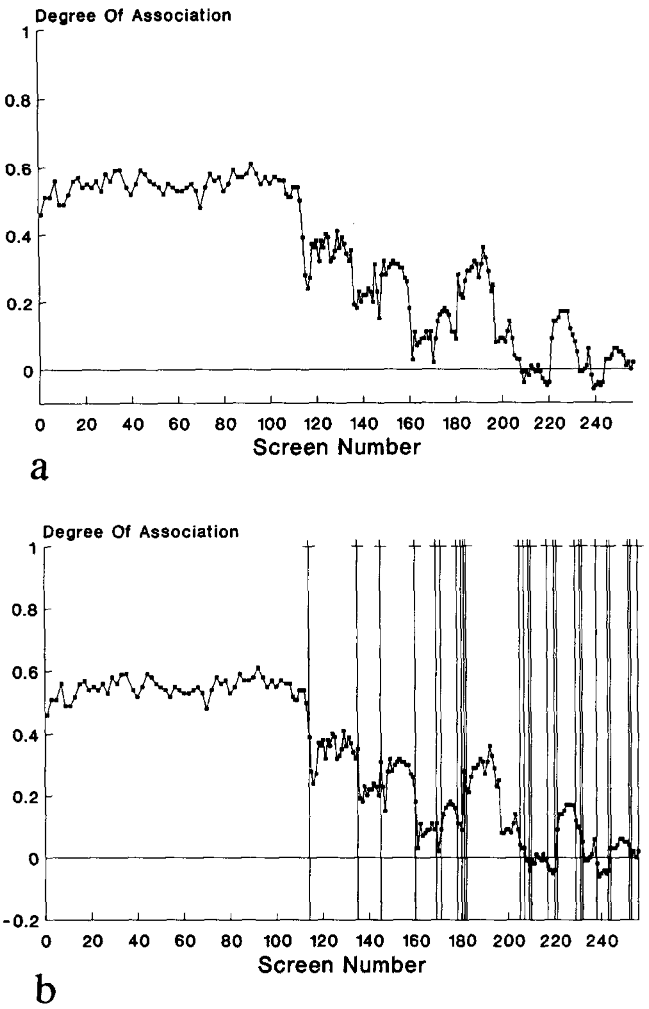


Figure 10. a, Association of screen 36 in the 256-member BN screen set with the other screens in this set; and b, as (a) but with the boundary points superimposed

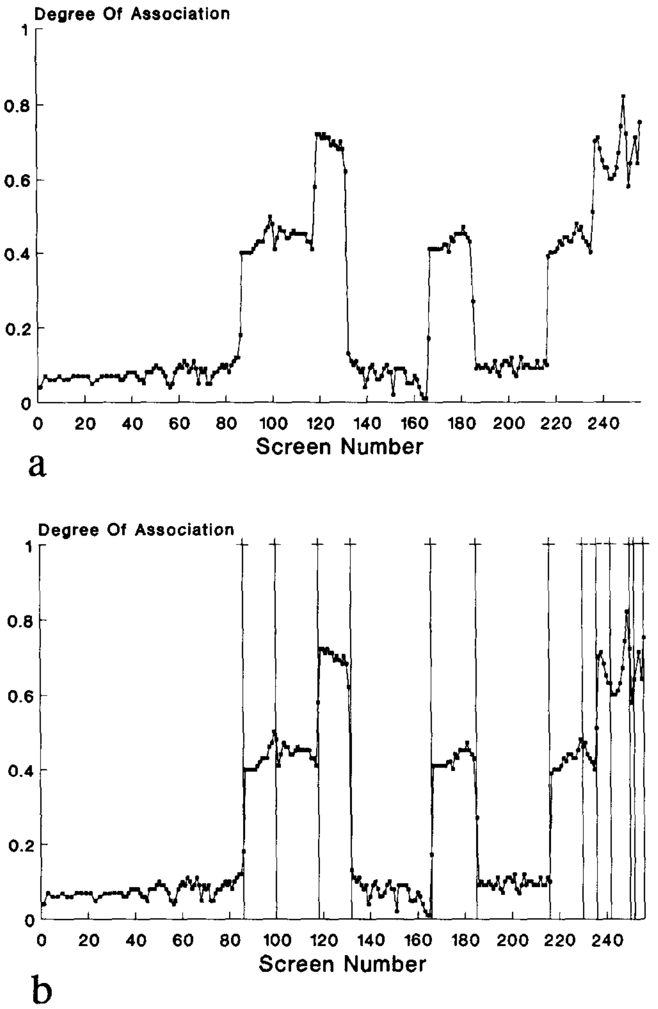


Figure 11. a, Association of screen 253 in the 256-member NN screen set with the other screens in this set; and b, as (a) but with the boundary points superimposed

points and of the peaks and troughs. For example, Figure 10 shows the association plot for screen 36 in the 256-member BN screen set, which corresponds to the range C–C*C 55 56. The plot shows a high degree of association, around 0.5, with the screens prior to the first boundary point, all of which are also C–C*C screens. Figure 11 presents similar data for screen 253 in the 256-member NN screen set, which corresponds to the range *X***Y***X* 73—*Y***Y***Y* 13. The first boundary point is at screen 85 and corresponds to the boundary between the many C**X***X* screens and the other types of screen; screen 253 is not a C**X***X* screen and the associations are correspondingly low. The screens between 85 and the next boundary point, at screen 99, are all C**X***Y* screens and the associations are much higher, around 0.4. The full set of boundary points and the corresponding fragment types at these points are listed in Table 2.

The rationale for this behavior is the multiple occurrences in the BN and, to a still greater extent, in the NN screen assignments that result from the way in which fragments are generated and screens assigned. Thus, once one particular type of screen has been assigned to a structure, large numbers of others of the same type will also be assigned

Table 2. Fragment types and boundary points for the association plot from screen 253 in the 256-member NN screen set. The association plot is shown in Figure 11.

Fragment type	Boundary point
C**X***X*	85
C**X***Y*	99
C**Y***X*	117
C**Y***Y*	130
*X**C**X*	165
*X**C**Y*	184
*X***X***X*	215
*X***X***Y*	229
*X***Y***X*	235
*X***Y***Y*	241
*Y**C**Y*	249
*Y***X***Y*	252
*Y***Y***Y*	256

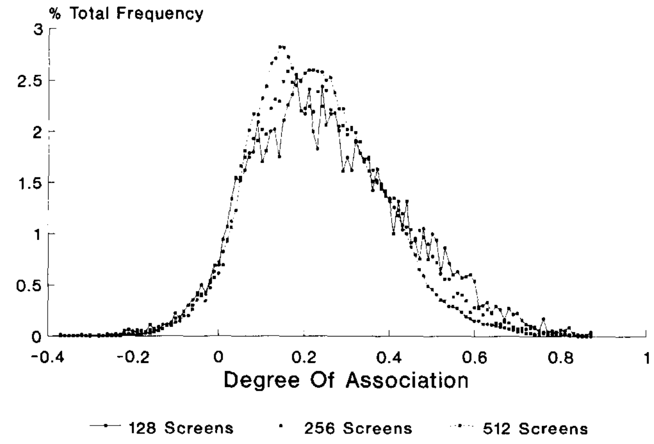


Figure 12. Distribution of screen associations for distance screen sets

to that molecule. This behavior is discussed above and is exemplified by the analyses in Figure 5.

For comparison with Figures 6–8, Figure 12 shows the corresponding distributions for the three distance-based screen sets. It will be seen that these distributions are far less skewed towards the positive end of the scale of associations, with a nontrivial percentage of the associations being less than zero. On this criterion, therefore the distance-based screens are expected to give better levels of screenout than comparable sets of angle-based screens.

Screenout performance

Angle-based screens. We now consider the efficiency of the screen sets when they are used for 3D substructure searching. The main experimental results are presented in Tables 3, 4, and 5, which list the median screenout when averaged over the set of 100 queries using the BB, BN, and NN screen sets, respectively.

The median figures listed here hide the great variation in screenout that is obtained with different query patterns, with some queries that consisted exclusively of carbons giving screenouts as low as 0.20. For all three types of fragment, the screenout rises with an increase in the number of angles in the query, and falls with an increase in the angular tolerance that is applied to the query pattern: these findings are entirely in line with expectation. Moreover, the screenout at a given tolerance decreases as one moves from BB to BN, and thence to NN. This arises from the increasing associations discussed above and from the densities of assignment in the bit strings that are used to represent each of the molecules in the database; these densities are listed in Table 6.

Increasing the tolerance of a search often appears to have less effect on the screenout of the BN searches than on the BB and NN searches. For example, if we consider a 512-member screen set with two angles per query, the median screenout drops from 0.74 to 0.69 (for BN) as the tolerance is increased from 0° to 5°; the decreases for the corresponding BB and NN screen sets are 0.99 to 0.88 and 0.65 to 0.54, respectively. The reason for this behavior appears to lie in the numbers of fragment types that are available in the three types of screenset: specifically, as noted above, there are 18, 27, and 13 possible combinations of three atoms classes in the BB, BN, and NN screen sets, respectively. Thus, when a screen set of a particular size is generated, each BN screen will, on average, represent a larger fraction of the distance range for a particular three-atom class than will the corresponding BB and NN screens. Since the average BB or NN range is smaller than the average BN range, a query pattern with wide tolerances will result in more bits being set in the query bit string when BB or NN screens are used than when BN screens are used. The effect will become more pronounced the greater the tolerance, and this is, indeed, what is observed in practice.

Comparison of angle-based and distance-based screens. The figures in Tables 3–5 demonstrate that valence angle screens can eliminate large numbers of molecules from the geometric searching stage of a 3D substructure search, even when only a few angles are specified in a query. For comparison, the results of the searches based on distance screen-

Table 3. Median screenout, averaged over 100 searches of 5000 structures taken from the Cambridge Structural Database, for patterns of BB valence angles

Number of BB valence angles	Screen set size	Tolerance (degrees)					
		0	1	2	3	4	5
	128	0.95	0.91	0.90	0.88	0.86	0.82
2	256	0.98	0.95	0.93	0.91	0.88	0.87
	512	0.99	0.96	0.93	0.92	0.90	0.88
	128	0.98	0.95	0.93	0.92	0.90	0.89
3	256	0.99	0.97	0.95	0.94	0.93	0.91
	512	1.00	0.98	0.96	0.95	0.94	0.92
	128	0.99	0.96	0.94	0.92	0.92	0.91
4	256	1.00	0.98	0.97	0.96	0.94	0.94
	512	1.00	0.99	0.97	0.96	0.94	0.94

Table 4. Median screenout, averaged over 100 searches of 5000 structures taken from the Cambridge Structural Database, for patterns of BN valence angles

Number of BN valence angles	Screen set size	Tolerance (degrees)					
		0	1	2	3	4	5
	128	0.64	0.64	0.64	0.63	0.63	0.63
2	256	0.66	0.65	0.65	0.65	0.65	0.63
	512	0.74	0.74	0.73	0.70	0.70	0.69
	128	0.65	0.64	0.64	0.64	0.64	0.64
3	256	0.68	0.68	0.67	0.67	0.67	0.66
	512	0.80	0.78	0.77	0.75	0.74	0.72
	128	0.67	0.66	0.66	0.65	0.65	0.65
4	256	0.72	0.70	0.70	0.70	0.68	0.68
	512	0.82	0.80	0.79	0.78	0.77	0.76

Table 5. Median screenout, averaged over 100 searches of 5000 structures taken from the Cambridge Structural Database, for patterns of NN valence angles

Number of NN valence angles	Screen set size	Tolerance (degrees)					
		0	1	2	3	4	5
	128	0.52	0.52	0.51	0.51	0.50	0.49
2	256	0.58	0.56	0.55	0.53	0.52	0.51
	512	0.65	0.61	0.58	0.56	0.55	0.54
	128	0.56	0.55	0.54	0.54	0.53	0.51
3	256	0.62	0.61	0.58	0.58	0.56	0.54
	512	0.69	0.66	0.62	0.60	0.59	0.57
	128	0.58	0.57	0.56	0.56	0.54	0.54
4	256	0.64	0.62	0.60	0.58	0.58	0.56
	512	0.71	0.67	0.63	0.62	0.60	0.58

ing are listed in Table 7. The bit-string densities of the distance-based screens are most similar to those for the NN angle screens (see Table 6); a comparison of the figures when exact query patterns are used, i.e., queries with zero tolerances, suggests that the distance screens give a better level of screenout (although this is due, in part at least, to the lower densities and associations for the distance-based screens). It is not really possible to compare the searches with nonzero tolerances: while tolerances of ±5° and ±0.5 Å are both entirely reasonable search parameters, it is not obvious that they correspond to searches of equal specificity.

In considering the results, it should be noted that the patterns were generated by drawing atoms at random (subject only to the constraint that three, two, or none of the atoms were bonded together in the BB, BN, and NN queries, respectively), so that many of the patterns consisted of just carbon atoms. This is in contrast to many of the published studies of distance-based 3D substructure searching systems, which have focused on searches for pharmacophoric patterns that contain a large proportion of heteroatoms and which have resulted in much higher levels of screenout than those obtained here.[3] In addition, the actual numbers of screens used here are much less than in operational systems: for example, the MACCS-3D system uses over 2000 screens[13] while Lederle Laboratories' 3DSEARCH system uses over 13 000 series.[7] The final point of note is that both BB and BN valence angles contain topological information that could be used to increase screenout above the levels shown in Tables 3 and 4, given a conventional 2D screening system in addition to the 3D screens considered here.

Table 6. Bit-string densities

Screen set size	BB	BN	NN	Distance
128	0.16	0.67	0.62	0.53
256	0.10	0.61	0.58	0.44
512	0.06	0.51	0.50	0.33

Table 7. Median screenout, averaged over 100 searches of 5000 structures taken from the Cambridge Structural Database, for patterns containing interatomic distances

Number of distances	Screen set size	Tolerance (Å)		
		0.0	0.1	0.5
	128	0.62	0.54	0.48
2	256	0.74	0.62	0.51
	512	0.86	0.70	0.57
	128	0.68	0.59	0.49
3	256	0.81	0.67	0.56
	512	0.92	0.75	0.59
	128	0.74	0.64	0.56
4	256	0.88	0.72	0.60
	512	0.95	0.80	0.62

The first of the factors referred to in the previous paragraph is evidenced by considering the results that were obtained when the set of eight published pharmacophoric patterns[3] mentioned above were searched using the valence-angle screens. The patterns all consist of interatomic distances (or distance ranges) and these distances were converted into angular ranges, as described. The results of the searches are listed in Table 8, which also contains the results obtained when the original distance-based patterns were searched using distance screens. The upper part of this table shows that the angle-based searches give poor levels of screenout with very many more structures needing to undergo the geometric search than when distance-based screens are used. These low levels of screenout are caused, in large part, by the high proportion of oxygen and nitrogen atoms in the various pharmacophoric patterns: in the CXY screen sets, both of these elemental types are subsumed under the common atomic descriptor *X*. Further sets of angle-based and distance-based screens were created in which four types of atomic descriptor were employed, these being, C, N, O, and Y, i.e., Y covers all elemental types other than CNO. The screenouts obtained when these screen sets were used are listed in the lower half of Table 8, where it will be seen that there is a substantial improvement in the performance of the angle-based screens relative to the distance-based screens.

The improvement in performance with the CNOY screen sets is due to the fact that a much smaller number of screens is assigned when the more discriminating atomic descriptors are used, as is evidenced by the figures in Table 9. This table lists the numbers of bits that were set in the query bit strings for each of the eight pharmacophoric patterns when the query searched was the full angular range as calculated from the original distance range. It will be seen that the use of the CNOY screen sets results in far fewer screens being assigned to the query patterns, with the concomitant increase in screenout that is evident from Table 8. It is possible that further improvements in performance could be obtained if more sophisticated atomic descriptors were to be employed, e.g., the five-level characterizations used in the 3DSEARCH system.[7] The relative merits of using a small number of

Table 8. Median screenout for searches using eight published pharmacophoric patterns with *CXY* and CNO*Y* screen sets. Three figures are listed for each of the angle-based searches: the first of these represents the searches that were carried out with an angular range calculated from the distance range in the original distance-based pattern, and the second and third represent the searches that were carried out for the midpoint of this range with tolerances of ±0° and ±5°, respectively

Atom types	Screen set size	NN angle screens			Distance screens
	128	0.29	0.49	0.35	0.55
CXY	256	0.29	0.54	0.39	0.62
	512	0.30	0.68	0.41	0.72
	128	0.58	0.62	0.62	0.69
CNO*Y*	256	0.63	0.68	0.66	0.81
	512	0.66	0.74	0.70	0.86

atomic descriptors with the associated short distance ranges, as against a larger number of atomic descriptors with the associated large distance ranges, is discussed in detail by Poirrette.[23]

The final set of results, in Table 10, details the median screenouts that were obtained in searches for the eight pharmacophoric patterns that used both angle-based and distance-based screens. The entry in the *IJ*th element of the main body of the table is the median screenout for a search that used both the *I*th distance-based screen set and the *J*th angle-based screen set. Thus, as an example, a median screenout of 0.74 was obtained in searches that used two CNOY screen sets, one of which contained 128 distance ranges and the other of which contained 512 angle ranges. An inspection of these figures shows little benefit in using such combined representations, when compared with a distance-based screen set containing an equal number of screens.

Table 9. Numbers of NN query screens for angle-based pharmacophoric pattern searches using *CXY* and CNO*Y* screen sets, and with tolerances calculated from the original distance-based pharmacophoric patterns

Pattern number	*CXY* screen sets			CNO*Y* screen sets		
	128	256	512	128	256	512
1	21	42	73	8	12	17
2	19	31	80	6	7	11
3	6	10	16	4	5	4
4	13	23	58	8	14	26
5	19	40	66	5	7	13
6	7	11	24	4	5	5
7	16	28	71	4	4	7
8	2	3	7	2	2	2

Table 10. Median screenout for searches using eight published pharmacophoric patterns with combination of angle-based and distance-based *CXY* and CNO*Y* screen sets

Screen set type	Distance screen set size	Angle screen set size		
		128	256	512
	128	0.58	0.58	0.58
CXY	256	0.65	0.65	0.65
	512	0.73	0.73	0.73
	128	0.71	0.71	0.74
CNO*Y*	256	0.83	0.83	0.83
	512	0.87	0.88	0.88

CONCLUSIONS

Current systems for screening 3D substructure searches are generally based on the use of interatomic distances, where each screen represents some pair of atoms together with an associated distance range. In this paper, we have evaluated the use of analogous screens based on generalized valence angles, so that the screens consist of a set of three atoms together with an associated angular range. In addition to conventional valence angles, where the central atom at the apex of the angle is bonded to both of the other atoms, we have also considered valence angles where this central atom is bonded only to one or to neither of the other two atoms.

An analysis of the occurrence frequencies of these three types of angle shows that the distributions are dominated by angles that are generated by small (often planar) structural motifs: a phenyl ring and its direct substituents, the metallocenes, various classical metal coordination spheres, simple functional groups like carboxy and amido, and simple ligands like carbonyl and cyanide.

Sets of BB, BN, and NN screens have been generated using the fragment occurrence data. The generalized valence-angle screens would seem to be less discriminating than interatomic distance screens when comparable searches are carried out. Evidence is presented to suggest that this is related to the bit-string densities and to the degree of association between the screens in the larger angle-based screen sets. The associations are particularly noticeable in the BN and NN screen sets, and reflect the dominance of the simple, frequently occurring structural motifs that have been mentioned previously. Many of these are topological in character, and thus the screenout figures reported here could be improved by the provision of 2D screening facilities. Similar comments apply to the BB screen sets, where each screen consists of a sequence of three bonded atoms. Even so, the levels of screenout that have been obtained here would be sufficient to bring about substantial reductions in the computational requirements of geometric searching for angular patterns. We are currently investigating the use of conventional and generalised intervector and torsion angles as bases for screen generation: This work will be reported shortly.

ACKNOWLEDGEMENTS

We thank the British Library Research and Development Department for funding, and Catherine A. Pepperrell for assistance with the distance-based searches.

REFERENCES

1 Gund, P. Three-dimensional pharmacophoric pattern searching. *Progress in Molecular and Sub-Cellular Biology* 1977, **5**, 117–143

2 Jakes, S.E. and Willett, P. Pharmacophoric pattern matching in files of 3D chemical structures: selection of interatomic distance screens. *J. Mol. Graphics* 1986, **4**, 12–20

3 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. Mol. Graphics* 1987, **5**, 41–48

4 Brint, A.T. and Willett, P. Pharmacophoric pattern matching in files of three-dimensional chemical structures: comparison of geometric searching algorithms. *J. Mol. Graphics* 1987, **5**, 49–56

5 Martin, Y.C., Danaher, E.B., May, C.S., and Weininger, D. MENTHOR, a database system for the storage and retrieval of three-dimensional molecular structures and associated data searchable by substructural, biologic, physical or geometric properties. *J. Comp. Aided Mol. Design* 1988, **2**, 15–29

6 Van Drie, J.H., Weininger, D., and Martin, Y.C. ALADDIN: an integrated tool for computer-assisted molecular design and pharmacophore recognition from geometric, steric, and substructure searching of three-dimensional molecular structures. *J. Comp. Aided Mol. Design* 1989, **3**, 225–251

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. Chem. Inf. Comp. Sci.* 1989, **29**, 255–260

8 Moock, T.E., Christie, B., and Henry, D. MACCS-3D: a new database system for three-dimensional molecular models. In: *Chemical Information Systems: Beyond the Structure Diagram.* (D. Bawden and E.M. Mitchell, eds.) Ellis Horwood, Chichester, 1990, pp. 42–49

9 Murrall, N.W. and Davies, E.K. Conformational freedom in 3D databases. 1. Techniques. *J. Chem. Inf. Comp. Sci.* 1990, **30**, 312–316

10 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

11 Willett, P. *Three-Dimensional Chemical Structure Handling.* Research Studies Press, Taunton, 1991

12 Martin, Y.C. Computer design of potentially bioactive molecules by geometric searching with ALADDIN. *Tetrahedron Comp. Meth.* 1990, **3**, 15–25

13 Christie, B.D., Henry, D.R., Guner, O.F., and Moock, T.E. MACCS-3D: a tool for three-dimensional drug design. *Proceedings of the 14th International On-line Information Meeting* (D. Raitt, ed.) Learned Information, London, 1990, 137–161

14 Cringean, J.K., Pepperrell, C.A., Poirrette, A.R., and Willett, P. Selection of screens for three-dimensional substructure searching. *Tetrahedron Comp. Meth.* 1990, **3**, 37–46

15 Allen, F.H., Bellard, S., Brice, M.D., Cartwright, B.A., Doubleday, A., Higgs, H., Hummelink T., Hummelink-Peters, B.G., Kennard, O., Motherwell, W.D.S., Rogers, J.R., and Watson, D.G. The Cambridge Crystallographic Data Centre: computer-based search, retrieval, analysis and display of information. *Acta Cryst.* 1979, **B35**, 2331–2339

16 Allen, F.H., Kennard, O., and Taylor, R. Systematic analysis of structural data as a research technique in organic chemistry. *Acc. Chem. Res.* 1983, **16**, 146–161

17 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. Chem. Inf. Comp. Sci.* 1991, **31**, 187–204.

18 Burkert, U. and Allinger, N.L. *Molecular Mechanics.* American Chemical Society, Washington DC, 1982

19 Rusinko, A., Sheridan, R.P., Nilakantan, R., Haraki, K.S., Bauman, N., and Venkataraghavan, R. Using CONCORD to construct a large database of three-dimensional coordinates from connection tables. *J. Chem. Inf. Comp. Sci.* 1989, **29**, 252–255

20 Klyne, W. and Prelog, V. Description of the steric relationships across single bonds. *Experientia* 1960, **16**, 521–523

21 Gannon, M.T. and Willett, P. Sampling considerations in the selection of fragment screens for chemical substructure search systems. *J. Chem. Inf. Comp. Sci.* 1979, **19**, 251–253

22 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

23 Poirrette, A.R., PhD thesis, University of Sheffield, in preparation.

24 Zee-Cheng, K.Y. and Cheng, C.C. Common receptor-complement feature among some antileukemic compounds. *J. Pharm. Sci.* 1970, **59**, 1630–1634

25 Adamson, G.W., Lambourne, D.R., and Lynch, M.F. Analysis of common structural characteristics of chemical compounds in a large computer-based file. Part III. Statistical association of fragment incidence. *J. Chem. Soc., Perkin I* 1972, 2428–2433

26 Hodes, L. Selection of descriptors according to discrimination and redundancy. Application to chemical substructure searching. *J. Chem. Inf. Comp. Sci.* 1976, **16**, 88–93