J-CAMD 310

Characterising the geometric diversity of functional groups in chemical databases

Susan M. Boyd[a,*], Martin Beverley[a], Leif Norskov[b] and Roderick E. Hubbard[a]

[a]*Department of Chemistry, University of York, Heslington, York YO1 5DD, U.K.*
[b]*Novo-Nordisk, DK-2880 Bagsvaerd, Denmark*

Received 10 April 1995
Accepted 25 July 1995

Keywords: Molecular databases; Molecular diversity; Combinatorial libraries; De novo ligand design

Summary

We have developed a program, HookSpace, which provides a simplistic approach to assessing the diversity of molecular databases. The spatial relationship between pairs of intramolecular functional groups can be analysed in a variety of ways to provide both qualitative and quantitative measures of diversity. Results are described and contrasted for two commercially available databases and a combinatorial library of benzodiazepam derivatives. HookSpace highlights the main differences in molecular content of these data sets.

Introduction

A variety of different approaches to structure-based ligand design have recently been developed [1]. Given the structure of a protein active site, these techniques attempt to suggest molecules that will satisfy the chemical and steric requirements of the active site, possibly providing new lead compounds for pharmaceutical or agrochemical development. There are three distinct categories of approach. The first is to grow or evolve a molecule from a limited number of small templates [2,3]. The second is to search databases of structures from known molecules to find those that will fit into the active site [4,5]. The third approach is to search databases of structural templates that can be linked to particular functional groups to satisfy the chemical and steric complementarity of the binding site [6].

In addition, there has been rapid growth in technologies for both generating and screening combinatorial libraries of molecules [7]. The aim of these techniques is to generate as diverse a set of molecules as possible, and to characterise the range of molecules that have activity. This has led to the search for measures of chemical diversity [8,9] to describe how well a defined set of structures can present to a target binding site the complete range of chemical and physical properties.

In this paper, we describe a simple technique, HookSpace, for calculating a relatively straightforward measure of diversity for a set of molecules. Our main purpose in developing the program was to characterise the diversity in structural databases to be used in the structure-based ligand design program HOOK [6]. For this and other approaches such as CAVEAT [10], an important consideration is the geometric relationship between pairs of functional groups in the molecules of the database used. The HookSpace calculation reflects the distribution of functional groups; the shape of the rest of the molecule is not considered. In addition, the current implementation of the program does not consider flexibility in the molecule. One way to overcome this problem would be to generate a multiconformational database prior to the HookSpace calculation. Nevertheless, within these limitations the program does give a useful indication of the geometric and functional group diversity.

The development of synthetic strategies for production of combinatorial libraries containing many hundreds of thousands of different compounds has generated considerable interest in the concept of chemical diversity. This type of analysis can also be applied to computer-generated representations of molecules, and can thus characterise how spatially diverse the libraries are.

[*] To whom correspondence should be addressed.

Methods

The HookSpace approach has three components. The first is the identification of particular functional groups within molecules in a database. The second is the calculation of a number of geometric properties for each pair of functional groups within each molecule, and the third is the presentation of the results for subsequent analysis. Each of these stages is considered separately below.

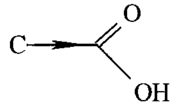
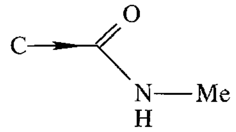
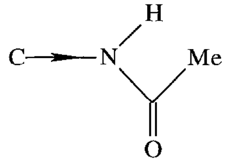
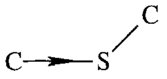
A functional group is user-defined from a template in terms of particular types of atoms, connected by particular types of bonds. Each functional group has a head and tail atom which form the bond (or hook) linking the functional group to a molecule. Functional groups with more than one attachment site (e.g. *N*-methyl acetamide) are considered as two different functional groups.

The results presented here were produced for the functional groups shown in Table 1, although the user could choose to probe any set of functional groups by creating appropriate templates. Indeed, all possible organic functional moieties could be probed using this methodology. Similarly, vectors encoded in the spirit of the CAVEAT [10] approach to intramolecular vector pairs could be characterised in this way.

Results are presented for three databases. The database distributed by the Cambridge Crystallographic Data Centre [11] (referred to here as the Cambridge Structural Database or CSD) contains structural information for all published small-molecule crystal structures. The analyses presented here used version 55 (Spring 1993) of the database, excluding molecules as described in the legend to Table 2. The Available Chemicals Directory [12] (referred to here as the ACD) was provided by MDL (San Leandro, CA) and contains all commercially available molecules, with three-dimensional structures generated using the CONCORD program [13]. The analyses presented here used version 93.1 of the database with some molecules excluded, as described in the legend to Table 2. Finally, a computed database of a combinatorial library of benzodiazepam derivatives (referred to here as benzo) was provided by G. Lauri and P. Bartlett (University of California, Berkely, CA). The structures in this database were generated by in-house fragment-assembly code, and minimised using MM2 [14]. The structures contained within the computational library were based upon established synthetic methodology [15]. All three databases were processed using the Skeleton program [6] to identify functional groups. Only molecules containing two or more functional groups are considered in this analysis.

Figure 1 shows the essential details of the calculations. For each pair of functional groups within each molecule in the database, the molecule was positioned such that one of the functional groups was orientated along the positive x-axis, with the head atom at the origin. The molecule was then rotated so that the head atom of the second functional group was on the xy-plane, with the head-to-tail vector pointing in the positive z-direction. The position of the second head atom was stored. This procedure was repeated for every pair of intramolecular hooks for every molecule in the database. Each pair of vectors was processed twice, with each head atom being moved to the origin in turn to produce a mapping of head positions in the xy-plane.

TABLE 1
DEFINITION OF THE FUNCTIONAL GROUP TEMPLATES USED TO GENERATE THE PRESENTED RESULTS

Name of group	Description	Structure
ACET	Acetate	
ACAM	*N*-methyl acetamide	
ACA2	*N*-methyl acetamide	
MEOH	Methanol	C—►—OH
MAMM	Methyl ammonium	C—►—NH_3
PHER	Phenyl	C—►—Ph
FLUO	Fluoride	C—►—F
CHLO	Chloride	C—►—Cl
IODI	Iodide	C—►—I
BROM	Bromide	C—►—Br
THIO	Thio ether	
PHOS	Phosphono ether	C—►—O——P

The perception of the functional groups is performed by the program Skeleton [6] as it converts external formatted databases of molecules into a database for use in the HOOK program [6]. The definition of functional groups can, of course, be extended by the user to encompass such functionality as is required.

These data can be analysed in a number of different ways:

(1) The distribution of distances between pairs of head atoms can be plotted as a histogram. For this calculation each vector pair is only counted once.

(2) Tile plots can be produced in which the value plotted at each xy-position represents the number of pairs of functional groups or the number of different combinations of functional groups that occur at this position.

TABLE 2
SUMMARY OF THE FUNCTIONAL GROUPS IN THE DIFFERENT DATABASES

	CSD[a]	ACD[b]	Benzo
No. of molecules	45887	110440	23469
Total no. of groups	99658	438254	127679
(No. of groups / molecule)	(2.2)	(4.0)	
No. of molecules containing > two functional groups	29209	82955	23469
(As % no. of molecules)	(64%)	(75%)	
HookSpace Index	34%	85%	13%

[a] Molecules were selected from the CSD using the following criteria: (1) all molecules should be organic with all hydrogens present, no crystallographic or chemical errors, and no polymeric bonds (either chemical or crystal); (2) no molecules should have carbon-X bonds (where X is a group IA element, a group IIA element, a group IIIA element, a noble gas, a lanthanide, an actinide, Si, P, Pb, As, Bi, Se, Te, Po, At, Zr, Nb, Mo, Tc, Ru, Rh, Pd, Ag, Cd, In, Sn or Sb); (3) no molecules should have carbon-Y-X bonds (where Y is any atom, and X is defined as in (2), but excluding P); and (4) molecules should have at least one carbon atom bonded to another atom.

[b] Molecules were selected from the ACD using the following criteria: (1) molecules contain two or more functional groups and more than four atoms; and (2) molecules do not contain unusual bond lengths (C-C< 1.2 Å, C-H< 1.0 Å, C-X< 1.0 Å) or unusual valence (e.g. pentavalent nitrogen).

These will be referred to as a number tile plot and a chemical tile plot, respectively. Although not presented here, the analysis could also include a measure of how diverse the head-to-tail vector orientations are at each position on the xy-plane.

(3) The frequency of each combination of functional groups can be plotted as a three-dimensional bar chart.

(4) A quantitative measure of spatial diversity of the database is obtained by computing the percentage of non-zero positions in the xy-plane. This value (which is referred to as the HookSpace Index) is obviously dependent on the area and grid spacing chosen for the calculations. The values quoted in this paper are for a grid of 20 × 20 Å.

For each of these calculations, it is possible to limit the analysis to a selected set of functional groups to study the distribution of a particular functionality in the database. The SQUID program [16] was used to perform all data processing and to produce plots.

Results

The three databases contain structures generated by quite different methods. The following results give an indication of how the HookSpace approach can highlight differences and similarities in the distribution of functional groups within each data set.

Table 2 contains a summary of the functional group characteristics of the three databases analysed in this study. Not surprisingly, the functional groups of the benzo database do not cover as much HookSpace as those of the other two databases. It is interesting to see that the proportion of the molecules in the databases containing more than two functional groups is only a little higher for the ACD than for the CSD, even though there are proportionately a lot more functional groups per molecule. This can be understood from Fig. 2, which compares the number of functional groups per molecule in the different databases. The distribution in the benzo database is a consequence of its construction, but the major difference in the average number of groups per molecule in the ACD (4.0) compared to that in the CSD (2.2) is seen to arise from the higher number of molecules in every category in the ACD that have more than two functional groups.

Table 3 compares the relative occurrence of each class of functional group in each database. Again, the distribution of groups in the benzo database is determined by the

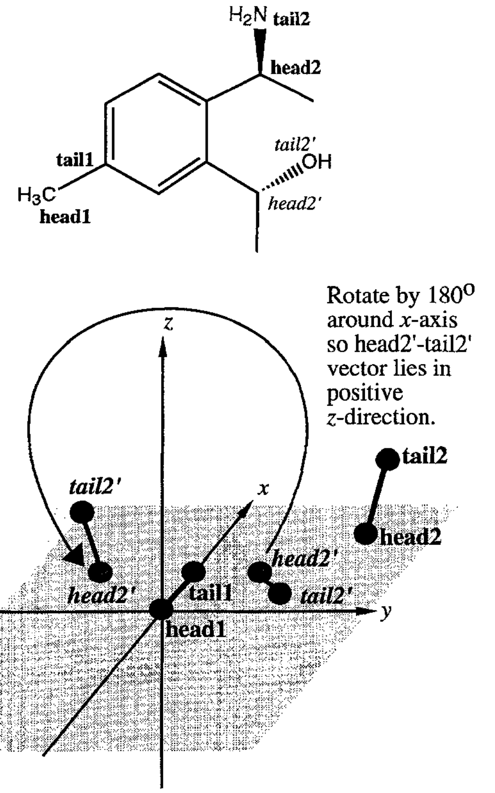


Fig. 1. Schematic illustration of the calculation of functional group geometries.

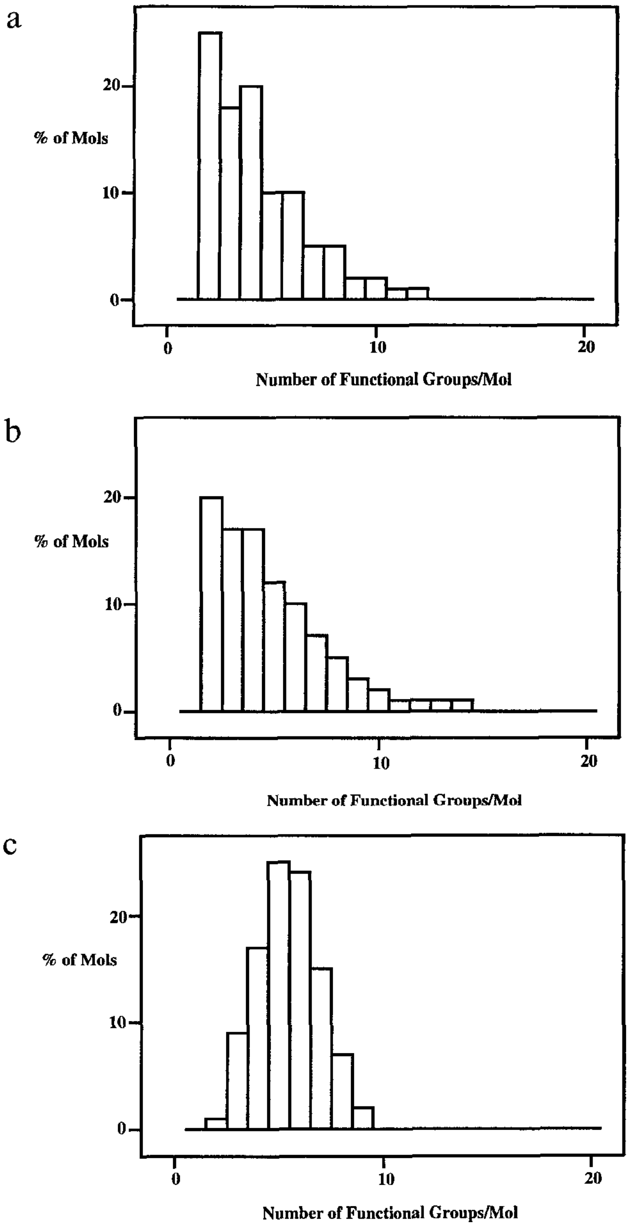


Fig. 2. Plots of the percentage of molecules having a particular number of functional groups per molecule for the (a) CSD, (b) ACD and (c) benzo databases.

common skeletal framework of the molecules. The distributions for the ACD and CSD are noticeably different. For the ACD, MAMM has the highest number of groups, whereas for the CSD it is PHER. Interestingly, there are substantially higher percentages of FLUO and CHLO functional groups in the ACD than in the CSD, but about the same proportion of BROM and IODO in each.

The histograms in Fig. 3 show the distribution of distances between pairs of functional groups within a molecule for a database. The most striking difference between the distributions is that the ACD contains molecules in which the functional groups are separated by a much larger distance than is seen in the CSD. This is mainly an artefact of the method used for generating the three-dimensional structures for the ACD. The longest distance of 69.4 Å is for a large peptide (a parathyroid hormone), which is generated in an extended linear conformation. In contrast, the largest molecule in the CSD is *bis* (3-beta-acetoxy-D:C-friedo-oleano-8-en-29-oic) anhydride [17], a molecule with two steroid-like skeletons linked by an anhydride group, with a distance between the functional groups of 32.1 Å.

A striking similarity between all three plots is the very small number of functional groups separated by between 1.8 and 2.0 Å, which is the region between one and two bond lengths where there are few molecules. The smallest interfunctional group distance in the ACD is determined by the selection criteria applied in preparation of the database (see Methods and Table 2). For the CSD, the smallest separation distance between the two groups is found in *n*-decaheptanoic acid [18], where the distance between the methanol head atom and the adjacent acetic acid head atom is 0.95 Å.

Figure 4 shows the central results of the HookSpace analyses, which are tile plots of the distribution of the pairs of functional groups. The plots are not necessarily symmetrical about the x-axis, as the second functionality vector is rotated to point towards the positive z-direction. Only certain molecules containing coplanar functional groups will possess the required symmetry to produce symmetrical entries on the plot.

The left-hand side of each pair is the number tile plot, in which the density at each xy-position represents how many pairs of functional groups have a particular geometry relative to each other. The distinctive features of these plots are that most of the orientations for head–head vectors that are separated by 6 Å or less appear to be satisfied by the database, with the exception of the inner ring of points, which represent less than a bond

TABLE 3
PERCENTAGE DISTRIBUTION OF THE DIFFERENT TYPES OF FUNCTIONAL GROUPS IN THE THREE DATABASES

Type	Database		
	CSD	ACD	Benzo
PHER	29.4	22.8	49.4
ACET	12.4	7.0	5.5
MAMM	20.4	28.1	0.0
MEOH	13.4	8.8	0.0
ACAM	6.8	6.1	13.5
ACA2	6.4	5.8	13.5
THIO	0.3	0.2	0.0
PHOS	1.2	0.9	0.0
CHLO	4.8	8.9	12.0
FLUO	2.5	9.5	6.0
BROM	2.1	1.6	0.0
IODO	0.3	0.3	0.0

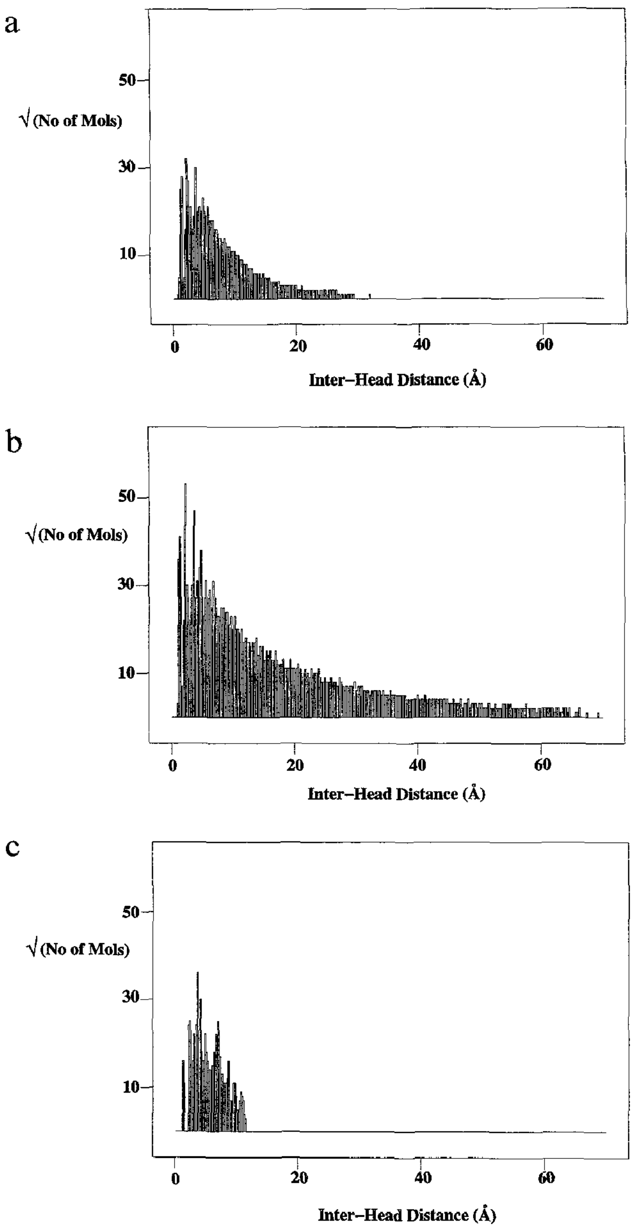


Fig. 3. Histograms of the head–head distance between functional group pairs in the (a) CSD, (b) ACD and (c) benzo databases. For clarity, the square root of the number of times a distance occurs is plotted to allow lower frequency values to be seen in the distribution.

length. The more populated bands lie in concentric semicircular formations around the origin, becoming less dense along the region of the x-axis, and less distinct as the distance of the band from the origin increases. The spacing between these bands corresponds well to inter-head separation by multiples of a typical carbon–carbon bond length for points close to the origin. Not surprisingly, the pattern becomes less pronounced with increasing distance from the origin, as the number of possible modes of atom connection increases. The sparsity of points along the x-axis may be explained by the rarity of cases where two hooks in a molecule have a linear relationship (there are, of course, examples such as *para*-disubstituted benzenes, where the functionality vectors will be linearly related, thus giving rise to points along the x-axis).

There are two striking differences between the number tile plots of the ACD and the CSD. Firstly, the extended conformations and thus larger dimensions of the ACD molecules result in a more even distribution of functional group distances in the xy-plane. Secondly, the CSD distribution shows much more pronounced sets of concentric circles, due to the greater accuracy of the crystallographic bond lengths and angles compared with the ACD structures. There is a distinctive, localised cluster of points on the ACD and CSD plots lying on the y-axis close to the origin. In both cases, these correspond to molecules in which the head–tail vector for group 1 is overlaid with the tail–head vector of group 2, the vector heads thus being separated by one C-C bond length. In the CSD, the most populated point on the plot, with an occupancy of 1.46×10^4, is located at position (0.0,1.4), and corresponds to a distance of 1.4 Å (one C-C bond length separation). In the ACD, the most populated point lies at (1.0,2.0) (occupancy 5.8×10^4), corresponding to a distance of 2.4 Å, i.e., the distance found between the two α-carbon atoms of a secondary or tertiary amine. The high frequency of occurrence of this distance for the ACD is characteristic of the peptide content of the database.

The right-hand side plots in Fig. 4 show tile plots for each database in which the density at each xy-position represents how many different types of functional group pairs have a particular geometry relative to each other. The overall shape of the plot is the same as that of the number plot. However, the pattern of shading of the tile plot positions is different, showing that the distribution of the different types of functional groups is more even than the actual number of functional groups found. This is because the more pronounced concentric bands seen in the number tile plot are produced by repeated occurrences of the same types of functional groups in the same orientations relative to each other.

Figure 5 depicts two three-dimensional histograms for each database, the bottom figure showing the number of occurrences of the various combinations of functional groups found in the database and the top figure showing the HookSpace Index for each of these combinations. This HookSpace Index is calculated by counting how many of the xy-positions on the number tile plot are occupied by just that particular pair of functional groups. The most striking difference between the two databases is the large number of MAMM–MAMM pairs in the ACD. For the CSD, the most common pair of functional groups comprises two benzene rings. In general, there are small differences in the frequency of different combinations of functional groups in the different databases. For example, there is a proportionately higher number of FLUO–FLUO combinations in the ACD than in the CSD.

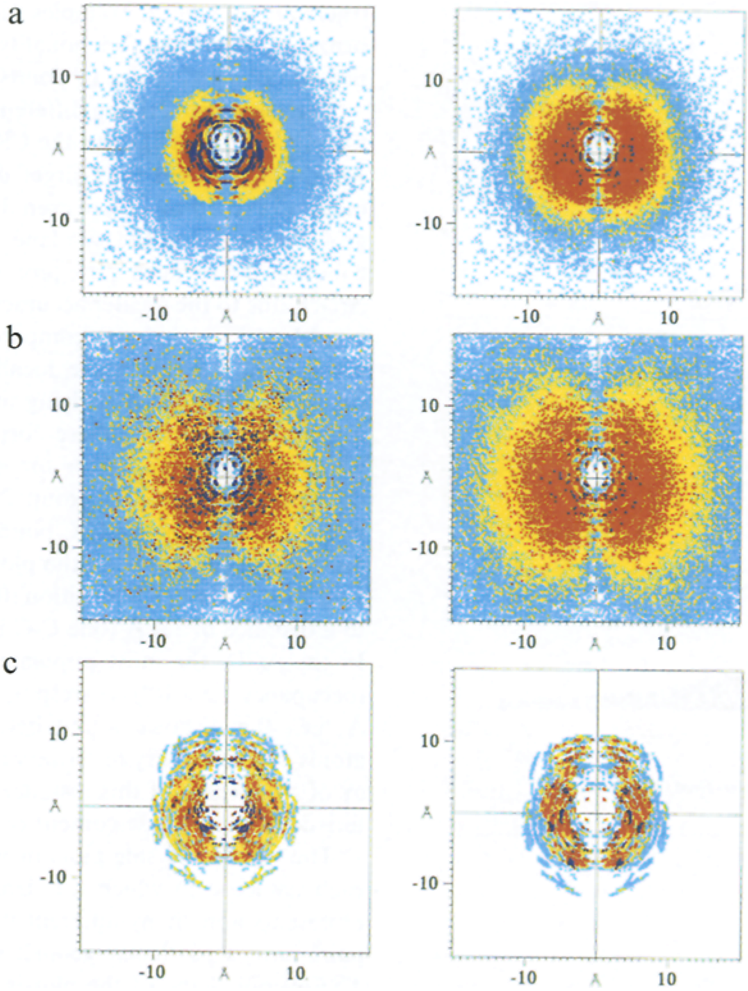


Fig. 4. HookSpace results for the (a) CSD, (b) ACD and (c) benzo databases with the tiles representing 0.2 Å in the x- and y-direction. The plot on the left is the number tile plot in which the shading of each xy position is related to how many pairs of functional group have that particular geometry. The plot on the right is the functional group tile plot, in which the shading of each xy-position is related to how many different types of functional group pairs have that particular geometry (the maximum is 12, see Table 1). The plots are contoured to best display the functional group distribution for each database. For the CSD, the number tile plot is contoured at 1, 25, 75 and 150, and the functional group tile plot is contoured on 1, 4, 8 and 12 occurrences. The ACD number tile plot is contoured on 1, 25, 100 and 500, with the functional group plot contoured on 1, 4, 8 and 12. The benzo database number plot is contoured on 1, 10, 75 and 500, and the functional group plot is contoured on 1, 2, 3 and 5. The contours are represented by the colours cyan, yellow, red and blue, respectively.

The differences between the databases, due in part to the different methods of structure generation employed, are highlighted by the number and HookSpace Index distributions of the functional groups. For the CSD, the HookSpace Index for each combination of functional groups is approximately similar to the relative number of pairs of groups seen. In contrast, for the ACD, the HookSpace Index of the MAMM–MAMM combination dominates the plot. This is due to the large number of peptides in the database that are in extended conformations. This difference between the distributions of pairs of functional groups between CSD and ACD can be seen particularly in the plots for two PHER groups and for two MAMM groups from each database, shown in Fig. 6. The difference emphasises the more extended nature of the ACD molecules compared to those in the CSD.

Discussion

The HookSpace method provides a simple description of functional group distribution in a structural database by analysing the number of different types of functional groups that are at particular distances and angles from each other. This basic information can be analysed in a number of different ways to characterise a database of molecules. Perhaps the most useful are the tile plots, showing graphically and concisely the distribution of functional groups in databases. This is of some importance in assessing the value of different databases when using approaches to structure-based ligand design such as HOOK [6] and CAVEAT [10]. In addition, although there are limitations in the definition of functional groups and in the inability of the current program to consider molecular

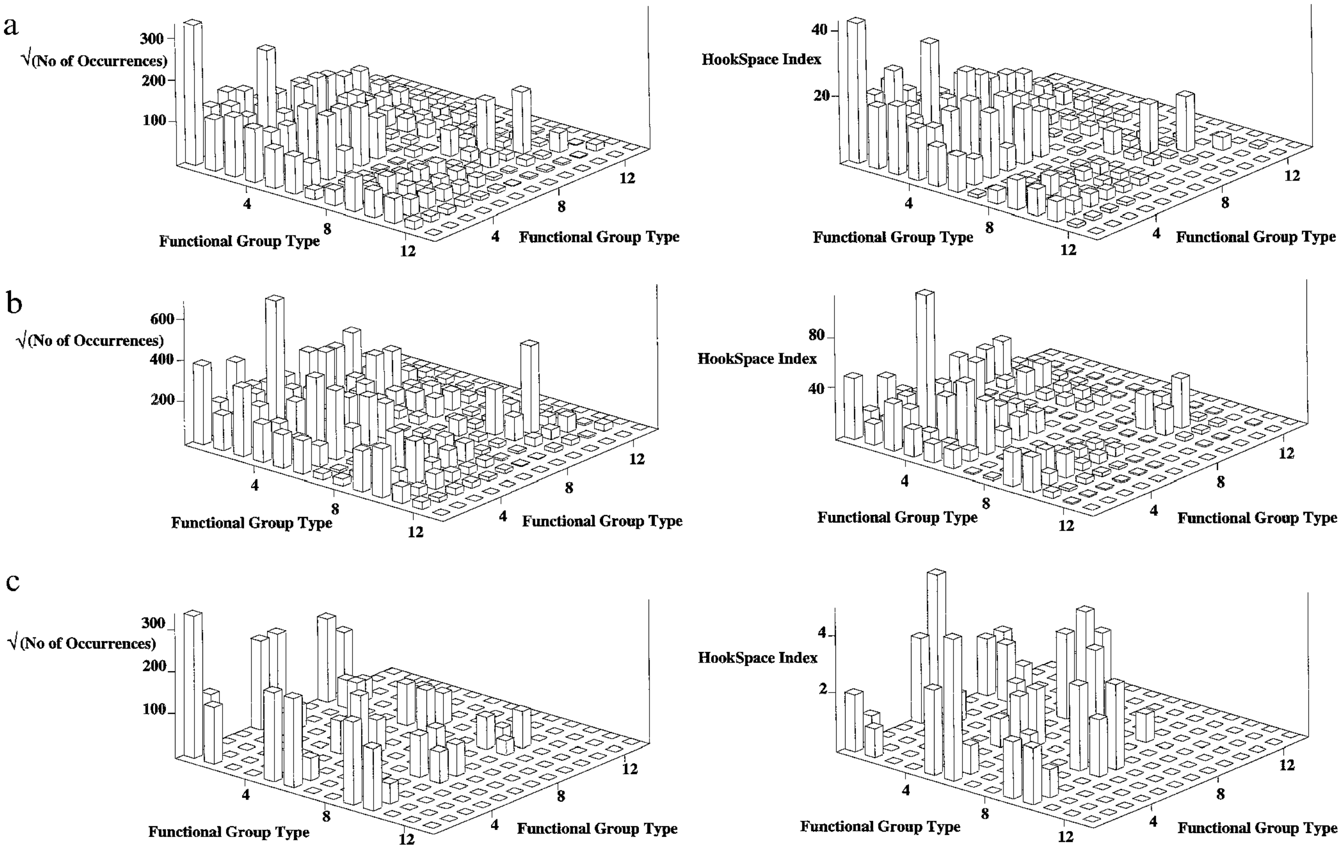


Fig. 5. 3D histograms showing the number of occurrences and the HookSpace Index for each type of functional group pair for the (a) CSD, (b) ACD and (c) benzo databases. To aid visualisation of the results, the square root of the number of occurrences of each functional pair is plotted on the histogram. The histogram columns correspond to the following functional groups: 1 = PHER; 2 = ACET; 3 = MAMM; 4 = MEOH; 5 = ACAM; 6 = ACA2; 7 = THIO; 8 = PHOS; 9 = CHLO; 10 = FLUO; 11 = BROM; and 12 = IODI.

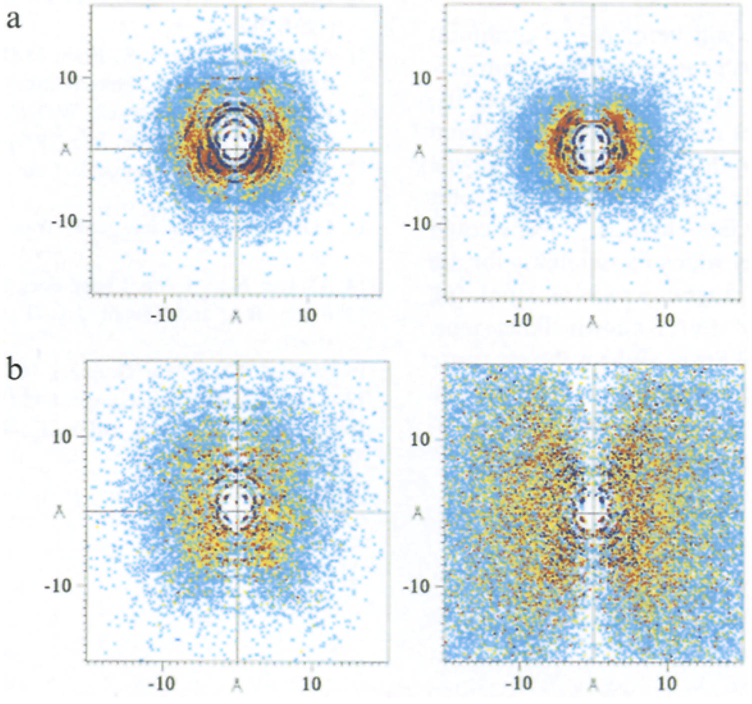


Fig. 6. (a) Functional group number tile plots for the CSD, for pairs of PHER (left) and MAMM (right) functional groups. Both plots are contoured on 1, 5, 10 and 25. (b) The same plots for the ACD, for pairs of PHER (left) and MAMM (right) functional groups. Both plots are contoured on 1, 5, 20 and 100.

flexibility, the HookSpace method can be considered as characterising one aspect or measure of structural diversity.

There are two main applications for this type of analysis. The first is to appreciate the diversity within a database of molecules, be it a database of existing molecules or a database generated to reflect a combinatorial chemistry programme. Secondly, the diversity could be used as a property to guide the development of combinatorial libraries of molecules which span the available geometric and functional space that could be available within a target binding site.

As mentioned above, a major limitation of the technique as currently implemented is that it uses a rigid description of the molecules, which is a serious deficiency for databases in which the structures for the molecules were generated automatically. This is a particular issue which could explain many of the differences between the ACD and CSD. The CSD represents real crystal structures, whereas the structures for the ACD were generated using the CONCORD program, which generates molecules in rather extended conformations. This problem could be overcome by introducing some conformational search technique into the HookSpace programs, or by pre-generation of acceptable conformations in the Skeleton database before using HookSpace. Both of these methods could be rather expensive in computer time, suggesting that the major value for HookSpace is in looking at rigid molecules.

Conclusions

The program HookSpace will be of use to computational and medicinal chemists in screening both commercial and in-house databases of molecules to sample the suitability of these databases for use in targeted ligand design. Usc of the program in this way could also provide tools to build a tailor-made database for a particular query. For example, appropriate chemical screens could be applied to a database, after which its suitability for use with rational ligand design programs such as Hook [6], against a particular receptor site of known dimensions, could be re-assessed by HookSpace. Where the structure of the receptor site is not known, a data set could be similarly identified where the spatial arrangement of functional groups satisfies a defined pharmacophore.

The program can be obtained on request from the authors.

Acknowledgements

This work was supported in part by a Strategic Research Fund grant from Zeneca and by funding from the SERC. The authors acknowledge useful discussions with Bob Bywater, John Bradshaw and Tony Wilkinson. We are particularly grateful to George Lauri and Paul Barlett for providing the benzodiazepam database.

References

1 a. Verlinde, C.L.M.J. and Hol, W.G.J., Structure, 2 (1994) 577.
b. Kuntz, I.D., Meng, E.C. and Shoichet, B.K., Acc. Chem. Res., 27 (1994) 117.

2 Moon, J.B. and Howe, W.J., Protein Struct. Funct. Genet., 11 (1991) 314.

3 Jones, G., Willett, P. and Glen, R.C., J. Mol. Biol., 245 (1995) 43.

4 DesJarlais, R.L., Sheridan, R.P., Seibel, G.L., Dixon, J.S., Kuntz, I.D. and Venkataraghavan, R., J. Med. Chem., 28 (1988) 849.

5 Lawrence, M.C. and Davis, P.C., Protein Struct. Funct. Genet., 12 (1992) 31.

6 Eisen, M.B., Wiley, D.C., Karplus, M. and Hubbard, R.E., Protein Struct. Funct. Genet., 19 (1994) 199.

7 Alper, J., Science, 264 (1994) 1399.

8 a. Keanan, J.K., Tsai, D.E. and Keene, J.D., Trends Biochem. Sci., 19 (1994) 57.
b. Dunbar Jr., J.B., Abs. Papers Am. Chem. Soc., 208 (1994) 124.

9 Martin, E.J., Blaney, J.M., Siani, M.A., Spellmeyer, D.C., Wong, A.K. and Moos, W.H., J. Med. Chem., 38 (1995) 1431.

10 Lauri, G. and Bartlett, P.A., J. Comput.-Aided Mol. Design, 8 (1994) 51.

11 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.W., Rodgers, J.R. and Watson, D.G., Acta Crystallogr., B35 (1979) 2331.

12 ACD-3D, v. 93.1, available from Molecular Design Ltd, San Leandro, CA.

13 CONCORD, available from Tripos Associates Inc., St. Louis, MO.

14 Allinger, N.L., J. Am. Chem. Soc., 99 (1977) 8127.

15 Bunin, B.A. and Ellman, J.A., J. Am. Chem. Soc., 114 (1992) 10997.

16 Oldfield, T.J., J. Mol. Graphics, 10 (1992) 247.

17 Nakai, H., Acta Crystallogr., C45 (1989) 1465.

18 Larson, K. and Von Sydow, E., Acta Crystallogr. Scand., 20 (1966) 1465.