Journal of Computer-Aided Molecular Design, 6 (1992) 397–406
ESCOM

J-CAMD 166

Automated site-directed drug design: Searches of the Cambridge Structural Database for bond lengths in molecular fragments to be used for automated structure assembly

P.-L. Chau and P.M. Dean*

Department of Pharmacology, University of Cambridge, Tennis Court Road, Cambridge CB2 1QJ, U.K.

Received 10 March 1992
Accepted 31 May 1992

Key words Drug design; Molecular fragments; Bond lengths; Cambridge Structural Database

SUMMARY

In this paper a database of small frequently occurring molecular fragments is used for the determination of fragment bond lengths from the Cambridge Structural Database. A large number of bond types are described that have not been reported previously.

INTRODUCTION

The goal of automated site-directed drug design is to write an algorithm which takes as input the Cartesian coordinates of the site atoms and generates sensible molecular structures within the prescribed site for consideration as putative ligands. Our approach in this series of papers has been to create a database of small molecular fragments to be used as building blocks for assembly onto molecular graphs which can be generated to span the site. In the previous paper [1], a programme was developed for generating combinatorially all aliphatic chemical fragments containing 3, 4 or 5 atoms of any allowed combination of H, C, N, O, F and Cl. A number of aromatic fragments, fragments containing P and S, and fragments carrying formal charges, were also generated combinatorially. These fragments were searched for using programme packages of the Cambridge Structural Database (CSD), and the frequently occurring ones were retained. In this paper we take the frequently occurring molecular fragments, and statistically ascertain their bond lengths from the CSD. The purpose of this is to provide some geometrical constraints for structure assembly at the stage where the fragments are placed combinatorially on a molecular graph that has been generated in the site.

* To whom correspondence should be addressed.

In this work we aim to produce fragments with transferable bond lengths. At first sight, it seems that we could use general bond lengths from standard tables, e.g. all bonds between any C_{sp3} and another C_{sp3} would be assigned the same length. However, there are problems with this approach. Firstly, there is variation in the bond length of all bond types. The electronic distribution is sensitive to bond length variation. If standard bond lengths were used, the electronic distribution of the fragment would be different. This fact, together with the difference in geometry, means that the electronic properties of poorly standardised fragments would not reflect what they should be in a molecule containing that fragment. The properties would have a low transferability if standard bond lengths were used. Secondly, even if we use bond length tables where the bonds have been sub-classified into different types, we still find that many of the desired bonds in the fragments are missing.

In an important paper on molecular structure, Allen et al. [2] determined the average bond lengths involving the elements H, B, C, N, O, F, Si, P, S, Cl, As, Se, Br, Te and I in organic compounds. They used data from the 1985 version of the CSD. X-ray and neutron diffraction results were used to derive mean bond lengths involving nonhydrogen atoms, while only neutron diffraction data were used to derive mean bond lengths involving hydrogen atoms. In their work, atoms were classified into different types, and the average bond length between two atom types was tabulated. However, many of the bonds that can be generated from combinatorial considerations were not listed. In this work, therefore, an attempt has been made to generate the average bond lengths, for the frequently occurring fragments, from original data using the CSD.

COMPUTING METHODS AND RESULTS

QUEST searches for fragment geometry calculations

CSD searches were performed on the 110 fragments remaining after the initial screening described in the previous paper [1]. However, this time the FDAT files were saved. These files contain numerical structural information which can be processed to give the fragment geometry. More severe constraints were also placed on data retrieval so that only high-quality X-ray structures were included for fragment geometry calculations. Firstly, the CONNSER command, used to search for chemical connectivity, includes an E keyword for the atom properties record for the central atom in the case of acyclic fragments, so that the compounds retrieved contain exactly the number of specified non-hydrogen-bonded atoms. This check prevents some unusual coordination pattern from distorting the geometry of the fragment. Furthermore, other screens were implemented to improve the checking. They are as follows, with the screen number in brackets after the constraint: no entry errors (−34), no disorders (35), no valence errors (−37), organic compounds only (57), chemical and crystallographic connectivities are perfectly matched (85), R-factor < 0.075 (89), entry is error-free at 0.02 Å level (32), atom coordinates field present (153). The last screen was set to ensure that atomic coordinates are available in the entry for fragment geometry calculations.

Initially, chemical diagrams were obtained for some of these 'hits' using an older version of the CSD system (Mark VIII). It was discovered that many of the fragments are complexed with metal ions, and the bond lengths and bond angles are not characteristic of the same fragments existing in a purely organic, nonmetallic environment. Since most drugs are organic molecules not complexed to metals, metals were excluded from the retrieved compounds by setting MAXA to less

than or equal to 18 (argon). This also increases the accuracy of the atomic coordinates because in X-ray diffraction experiments, heavier nuclei tend to scatter the X-rays more and are hence better localised. In metal complexes the diffraction pattern is dominated by the metals, and the organic moieties are poorly located. Limiting the maximum atomic number of the constituent atoms of the compound to 18 means that the pattern will not be dominated by heavier nuclei in these compounds, and the lighter elements such as carbon, nitrogen or oxygen will be better located.

GSTAT programmes for fragment geometry calculations and statistics

After the QUEST programmes have been executed, the retrieved FDAT subfiles are then used by the GSTAT programmes to calculate the fragment geometry from each hit. The GSTAT query commands include a fragment specification similar to the CONNSER commands, except that in this case crystallographic connectivity is considered instead of chemical connectivity. GSTAT programmes use crystallographic connectivity, and included the following commands to restrict the search to high-quality data only: no disordered structures, avoid error entries, average estimated standard deviation of C-C bonds in range 0.1–1 Å (0.01–0.1 nm), R-factor in the range of 0.002–0.075. (The range was not set from 0 because entries without information on R-factors were set to have an R-factor of 0; a number of such compounds would then be retrieved.)

Again, as in the QUEST commands, the E keyword was included in the atom properties record of all central atoms to exclude unusual coordination patterns in the case of acyclic fragments. Further commands were set to ensure that all atoms in the fragment must be bonded to each other only by the bonds given explicitly in the bond properties record (NOLN), and that no atom within the defined fragment was connected by a cyclic bond to an atom outside the fragment (NOCR). The latter two commands were set to prevent fusion or bridging of the defined fragment. The keyword ALLBOND A or ALLBOND C was also included in the searches of aliphatic or aromatic fragments, respectively.

The assimilation of FDAT subfiles by GSTAT proved to be much more difficult than expected. The reason for this is that different kinds of connectivities are used in QUEST and in GSTAT. GSTAT programmes do not have the concept of chemical bonds: they can only handle different interatomic distances. To ensure that the interatomic distance range specified in the GSTAT bond properties record includes the average bond length, the standard bond length tables were consulted [2]. The interatomic distance range was then set to be the average bond length plus or minus 3 S.D. However, this procedure proved not to be acceptable for all bond types. We decided to estimate the interatomic distance range without using data from Allen et al. [2]. Our stringency criteria were such that the standard deviation for bonds not involving hydrogen should be less than 5% and for bonds involving hydrogen the allowance should be up to 10%. A higher error level was allowed for bonds involving hydrogen because hydrogen nuclei are poorly located in X-ray diffraction studies. Histograms were requested from the GSTAT programme for all bonds to determine if the estimate was acceptable. It was discovered that the listed interatomic distance ranges are not accurate estimates of bond lengths obtained from the CSD: not infrequently one observes that part of the histogram is severed at the boundaries. There were about 340 distinct bond types, but data were only available for about 170 of them. About half of the bond length data required could not be found in the paper of Allen et al. [2]. For example, no bond length statistics were compiled for any bond in pyridazine, 1,3,5-triazine, indole, benzofuran or purine. Examples of common bonds not found in Allen et al. are:

≡C−H	
−N=C<	
>N−N=C<	in aliphatic systems
−O−N=C<	exists in Allen et al. only if O bonded to H
H−N=C<	exists in Allen et al. only if N singly bonded to X
−N=N−N<	
C_{sp3}−(O−O)−	O–O bond
C−C(−C)=N	all bonds
C−C(=N)−N	all bonds
C−C(=O)−N	all bonds
C−C(=N)−H	all bonds.

The listed bond length is either a significant over-estimate or an under-estimate. The reason may be linked to the fact that Allen et al. [2] chose extremely high-quality data, while the criteria in this work are less strict so as to include a larger variety of compounds. A second reason for this discrepancy may be that a slightly larger database than that of Allen et al. was used here for the searches. The third reason for this difference may be that both neutron and X-ray diffraction experimental data were used for determining the lengths of bonds that involve hydrogen atoms, while Allen et al. used neutron diffraction data only.

Atomic coordinates of the fragments were output in Cartesian coordinates with reference to the inertial axes of the fragment. An averaging command was also included so that the first fragment was taken as reference, and the 8 possible orientations of the inertial axes of succeeding fragments were examined to see which of them gave the closest fit to the reference. The least-squares method was chosen, and the sums of squares of the distances between corresponding atoms were minimised. Finally, all the fragments were superposed to form a giant fragment, and the geometry of this giant fragment was output in Cartesian coordinates for later use in molecular orbital calculations.

Fragment geometry calculations and statistics

Combined QUEST and GSTAT programmes were used to calculate the average geometry of the 110 fragments that survived the initial two screenings. In order that the results obtained would be statistically significant, at least 10 observations from different compounds were required. In practice, a much larger number of observations was used, because of the need to ascertain the frequency distribution. However, this could not always be achieved. Many fragments gave so few hits that it was impossible to determine if the distribution was normal from the histogram. One fragment, in fact, achieved no hits at all when these normal stringency criteria were included. Therefore, in some runs the selection criteria were relaxed, e.g., a higher R-factor or error level was permitted, or for some runs, there may not be any discrimination on R-factor or error level at all.

On the other hand, one fragment, benzene, occurred so frequently that the number of FDAT subfiles retrieved exceeded the store. The selection criteria, therefore, were tightened to limit the number of 'hits' achieved. The quality of the entries in this case was improved.

Another complication with 3 of the fragments C–O–C, C–O–H and $(C)_2$–N–H, was that the bond lengths between the central atom and one or two of the peripheral atoms showed a bimodal distribution when the statistical calculations were performed. Using the special Q or T keyword

TABLE 1
BOND LENGTHS FOR THE 3-ATOM FRAGMENTS[a]

Fragment letter	Bond 1 (Å)	Bond 2 (Å)
3.a	1.185 ± 0.015 (65)	1.464 ± 0.013 (65)
3.b	1.142 ± 0.013 (557)	1.433 ± 0.023 (557)
3.c	1.175 ± 0.014 (64)	0.962 ± 0.091 (64)
3.d	1.607 ± 0.033 (109)	1.149 ± 0.022 (108)
3.e	1.267 ± 0.023 (44)	1.468 ± 0.016 (44)
3.f	1.239 ± 0.023 (47)	1 460 ± 0.035 (47)
3.g	1.286 ± 0.015 (196)	1.379 ± 0.032 (196)
3.h	1.279 ± 0.012 (99)	1.400 ± 0.024 (99)
3.ı	1.293 ± 0.023 (11)	0.882 ± 0.076 (11)
3.j	1.293 ± 0.061 (30)	1.303 ± 0.070 (30)
3.kl	1.343 ± 0.036 (12)	1.364 ± 0.032 (12)
3.k2	1.332 ± 0.020 (1740)	1.452 ± 0 021 (1738)
3.k3	1.420 ± 0.019 (137)	1.422 ± 0 017 (137)
3.l	1.420 ± 0.037 (71)	1.414 ± 0.023 (71)
3.m	1.370 ± 0.022 (33)	1.455 ± 0.017 (33)
3.nl	1.307 ± 0.021 (726)	0.948 ± 0 122 (726)
3.n2	1.422 ± 0.021 (1101)	0.881 ± 0.116 (1101)
3.o	1.400 ± 0.017 (162)	0.930 ± 0.108 (162)
3.p	1.787 ± 0.035 (181)	1.791 ± 0.033 (181)
3.q	1 450 ± 0.023 (62)	1.586 ± 0.029 (62)
3.r	1.814 ± 0.033 (146)	2.030 ± 0.021 (146)

[a]The structures of the molecular fragments, together with the bond key and fragment letter code, are given in Fig 1 of [1].

to determine the hybridization state of the peripheral atom, it was discovered that each peak was associated with a hybridization state, with the *sp2* hybridization state showing a shorter bond length than the *sp3* state. These fragments were therefore sub-classified and treated as different fragments in the search, statistical analysis and subsequent hydrogen addition and quantum chemistry calculations. In the end, 113 distinct fragments were obtained. To avoid repetition in drawing the molecular fragment structures, the reader is referred to Figs. 1–8 of the previous paper [1] where the fragment and bond keys are given. Bond-length data derived in this paper are expressed in Tables 1–8 and the order corresponds exactly to the order in the figures drawn previously [1]. In these tables the mean bond lengths, together with their standard deviation and number of observations in parentheses, are given for each bond of all the molecular fragments.

DISCUSSION

The results obtained here complement the molecular fragments database, derived from a search of all frequently occurring fragments, with a complete description of the bond lengths in the fragments. These fragments are most probably stable and easily synthesizable. The bond lengths of the fragments are determined from X-ray and neutron diffraction data. Previous work on mean bond lengths [2] used compounds involving up to 15 elements, but data for many of the common bonds were found to be missing. This work concentrates on the bonds of only the fragments, but

TABLE 2
BOND LENGTHS FOR THE 4-ATOM FRAGMENTS CONTAINING ONE DOUBLE BOND[a]

Fragment letter	Bond 1 (Å)	Bond 2 (Å)	Bond 3 (Å)
4.2.a	1.342 ± 0.032 (223)	1.477 ± 0.035 (224)	1 484 ± 0.037 (224)
4.2.b	1.287 ± 0.022 (132)	1.490 ± 0.028 (132)	1.495 ± 0.023 (132)
4.2.c	1.220 ± 0.032 (297)	1.491 ± 0.042 (297)	1.493 ± 0.038 (297)
4.2.d	1.356 ± 0.028 (72)	1.501 ± 0.034 (72)	1.381 ± 0.040 (72)
4.2.e	1.345 ± 0.023 (52)	1.485 ± 0.032 (52)	1.352 ± 0.044 (52)
4.2.f	1.326 ± 0.030 (786)	1.470 ± 0.034 (786)	1.004 ± 0.080 (784)
4.2.g	1.306 ± 0.023 (34)	1.501 ± 0.019 (34)	1.331 ± 0.034 (34)
4.2.h	1.228 ± 0.016 (584)	1.515 ± 0.023 (584)	1.338 ± 0.019 (584)
4.2.i	1.301 ± 0.040 (8)	1.495 ± 0.026 (8)	1.318 ± 0.060 (8)
4.2.j	1.279 ± 0.019 (53)	1.462 ± 0.028 (53)	0.994 ± 0.054 (53)
4.2.k	1.205 ± 0.025 (2019)	1.501 ± 0.023 (2019)	1.321 ± 0 037 (2019)
4.2.l	1.212 ± 0.024 (33)	1.452 ± 0.020 (33)	1.035 ± 0.068 (33)
4.2.m	1.421 ± 0.029 (12)	1.339 ± 0.013 (12)	1.341 ± 0.016 (12)
4.2.n	1.367 ± 0.031 (44)	1.348 ± 0.040 (44)	0.983 ± 0.099 (44)
4.2 o	1.331 ± 0.019 (14)	1.366 ± 0.032 (14)	1.033 ± 0.092 (14)
4.2.p	1.310 ± 0.028 (127)	1.014 ± 0.078 (127)	1.013 ± 0.080 (127)
4.2.q	1.332 ± 0.022 (138)	1.339 ± 0.027 (137)	1.331 ± 0.025 (137)
4.2.r	1.242 ± 0.015 (149)	1.354 ± 0.023 (149)	1 348 ± 0.022 (149)
4.2.s	1.300 ± 0.020 (8)	1.325 ± 0.035 (8)	0.995 ± 0.087 (8)
4.2.t	1.209 ± 0.012 (128)	1.343 ± 0.016 (128)	1.351 ± 0.020 (128)
4.2.u	1.233 ± 0.022 (49)	1.331 ± 0.016 (49)	1.015 ± 0.066 (48)
4.2.v	1.493 ± 0.038 (14)	1.326 ± 0.023 (14)	1.727 ± 0.015 (14)
4.2.w	1.345 ± 0.024 (81)	1.335 ± 0.021 (81)	1.691 ± 0.017 (81)

[a]The structures of the molecular fragments, together with the bond key and fragment letter code, are given in Fig. 2 of [1].

TABLE 3
BOND LENGTHS FOR THE 4-ATOM FRAGMENTS CONTAINING ONLY SINGLE BONDS[a]

Fragment letter	Bond 1 (Å)	Bond 2 (Å)	Bond 3 (Å)
4.1.a	1.464 ± 0.020 (59)	1.464 ± 0.018 (58)	1.462 ± 0.020 (59)
4.1.b	1.387 ± 0.031 (27)	1.450 ± 0.012 (27)	1.456 ± 0.012 (27)
4.1.c	1.399 ± 0.020 (14)	1.391 ± 0.051 (14)	1.422 ± 0.055 (14)
4.1.d	0.880 ± 0.096 (30)	1.372 ± 0.019 (30)	1.392 ± 0.022 (30)
4.1.e	0.900 ± 0.093 (405)	1.453 ± 0.019 (405)	1.333 ± 0.017 (404)
4.1.f	0.929 ± 0.115 (21)	1.466 ± 0.015 (21)	1.473 ± 0.015 (21)
4.1.g	0.927 ± 0.088 (179)	1.394 ± 0.020 (177)	1.347 ± 0.019 (175)
4.1.h	0.918 ± 0.087 (391)	0.920 ± 0.086 (390)	1.327 ± 0.018 (382)
4.1.i	0.935 ± 0.089 (75)	0.928 ± 0.106 (75)	1.422 ± 0.020 (75)
4.1.j	1.468 ± 0.018 (14)	1.615 ± 0.008 (14)	1.021 ± 0.046 (14)

[a]The structures of the molecular fragments, together with the bond key and fragment letter code, are given in Fig. 3 of [1].

TABLE 4
BOND LENGTHS FOR THE 5-ATOM FRAGMENTS[a]

Fragment letter	Bond 1 (Å)	Bond 2 (Å)	Bond 3 (Å)	Bond 4 (Å)
5.a	1.535 ± 0.021 (105)	1.529 ± 0.014 (105)	1.530 ± 0.017 (105)	1.533 ± 0.015 (105)
5.b	1.530 ± 0.019 (186)	1.525 ± 0.016 (186)	1.528 ± 0.015 (186)	1.484 ± 0.025 (186)
5.c	1.519 ± 0.020 (152)	1.520 ± 0.017 (152)	1.518 ± 0.017 (152)	1.457 ± 0.024 (152)
5.d	1.528 ± 0.016 (139)	1.522 ± 0.014 (139)	1.525 ± 0.014 (139)	1.000 ± 0.075 (139)
5.e	1.519 ± 0.013 (4)	1.525 ± 0.017 (4)	1.523 ± 0.018 (4)	1.521 ± 0.062 (4)
5.f	1.551 ± 0.032 (5)	1.534 ± 0.033 (5)	1 471 ± 0.028 (5)	1.388 ± 0.037 (5)
5.g	1.526 ± 0.017 (329)	1.525 ± 0.017 (329)	1.477 ± 0.025 (329)	0.999 ± 0.065 (329)
5.h	1.551 ± 0.007 (8)	1.553 ± 0.030 (8)	1.385 ± 0.018 (8)	1.385 ± 0.023 (8)
5.ı	1.526 ± 0.017 (283)	1.523 ± 0.017 (283)	1.430 ± 0.018 (283)	1.006 ± 0.068 (282)
5.j	1.518 ± 0.021 (1145)	1.515 ± 0.021 (1145)	1.010 ± 0.071 (1144)	1.011 ± 0.071 (1142)
5.k	1.534 ± 0.021 (5)	1.456 ± 0.011 (5)	1.452 ± 0.009 (5)	1.058 ± 0.073 (5)
5.l	1.510 ± 0.019 (681)	1.479 ± 0.026 (681)	1.008 ± 0.065 (678)	1.002 ± 0.070 (681)
5.m	1.528 ± 0 022 (12)	1.407 ± 0.019 (12)	1.409 ± 0.020 (12)	1.015 ± 0.090 (12)
5.n	1.498 ± 0 025 (677)	1.438 ± 0.024 (677)	1.011 ± 0.072 (676)	1.009 ± 0.067 (677)
5.o	1.513 ± 0.000 (2658)	0.983 ± 0.066 (2763)	0.983 ± 0.068 (2763)	0.984 ± 0.070 (2767)
5.p	1 528 ± 0.028 (102)	1.321 ± 0.018 (102)	1.322 ± 0.016 (102)	1.321 ± 0.017 (102)
5.q	1.453 ± 0.009 (10)	1.451 ± 0.010 (10)	0.943 ± 0.027 (10)	1.019 ± 0.054 (10)
5.r	1.462 ± 0.022 (1130)	0.988 ± 0.075 (1127)	0.986 ± 0.079 (1130)	0.990 ± 0 079 (1129)
5.s	1.422 ± 0.026 (11)	1.383 ± 0.019 (11)	0.983 ± 0.093 (11)	0.992 ± 0.058 (11)
5.t	1.433 ± 0.021 (1186)	0.985 ± 0.071 (1183)	0.989 ± 0.067 (1185)	0.990 ± 0.070 (1184)
5.u	1.516 ± 0.020 (6)	1.501 ± 0.052 (6)	1.506 ± 0.046 (6)	1 851 ± 0.053 (6)
5.v	1.521 ± 0.011 (12)	1.523 ± 0.013 (12)	1.000 ± 0.087 (12)	1.802 ± 0.010 (12)
5.w	1.540 ± 0.030 (68)	1.768 ± 0.018 (68)	1.768 ± 0.028 (68)	1.767 ± 0.020 (67)
5.x	1.498 ± 0.020 (93)	1.015 ± 0.071 (93)	1.015 ± 0.092 (94)	1.781 ± 0.023 (94)
5.y	1.801 ± 0.020 (211)	1.513 ± 0.016 (210)	0.982 ± 0.075 (211)	0.979 ± 0.070 (210)
5.z	1.780 ± 0.026 (284)	0.977 ± 0.087 (284)	0.964 ± 0 078 (284)	0.972 ± 0.078 (284)

[a]The structures of the molecular fragments, together with the bond key and fragment letter code, are given in Fig 4 of [1].

TABLE 5
BOND LENGTHS FOR CHARGED ATOM FRAGMENTS[a]

Fragment letter	Bond 1/5 (Å)	Bond 2/6 (Å)	Bond 3 (Å)	Bond 4 (Å)
Q a	1.496 ± 0.012 (14)	1.492 ± 0.014 (14)	1.443 ± 0.010 (14)	1.595 ± 0.014 (14)
	1.575 ± 0.011 (14)	0.945 ± 0.049 (14)		
Q.b	1.500 ± 0.011 (14)	1.497 ± 0.009 (14)	1.827 ± 0.025 (14)	1.570 ± 0.018 (14)
	0.974 ± 0 067 (14)			
Q.c	1.450 ± 0.016 (93)	1.445 ± 0.015 (93)	1.445 ± 0.017 (93)	1.782 ± 0.024 (93)
Q d	1.251 ± 0.018 (325)	1.249 ± 0.018 (324)	1.528 ± 0.021 (325)	
Q.e	0.943 ± 0.086 (282)	0.951 ± 0.086 (282)	0.942 ± 0.084 (282)	1.487 ± 0.014 (282)
Q.f	1.493 ± 0.014 (69)	0.952 ± 0.092 (69)	1.493 ± 0.016 (69)	0.946 ± 0.085 (69)
Q.g	1.500 ± 0.013 (75)	1.495 ± 0.014 (75)	1.498 ± 0.018 (75)	0.962 ± 0.082 (75)
Q.h	1 320 ± 0.013 (106)	0.902 ± 0.070 (105)	0.905 ± 0.067 (106)	

[a]The structures of the molecular fragments, together with the bond key and fragment letter code, are given in Fig. 5 of [1].

TABLE 6
BOND LENGTHS FOR 5-MEMBERED RINGS[a]

Fragment letter	Bond 1/5 (Å)	Bond 2 (Å)	Bond 3 (Å)	Bond 4 (Å)
R5.a	1.371 ± 0.014 (49)	1.386 ± 0.017 (49)	1 400 ± 0.017 (49)	1.376 ± 0.020 (49)
	1.367 ± 0.015 (49)			
R5.b	1.364 ± 0.021 (61)	1.322 ± 0.015 (61)	1.401 ± 0.018 (61)	1.371 ± 0.018 (61)
	1.357 ± 0.017 (61)			
R5 c	1.352 ± 0.017 (56)	1.312 ± 0.012 (56)	1.375 ± 0.012 (56)	1.361 ± 0.017 (56)
	1.370 ± 0.011 (56)			
R5.d	1.372 ± 0.017 (30)	1.315 ± 0.012 (30)	1.355 ± 0.010 (30)	1.327 ± 0.013 (30)
	1.337 ± 0.011 (30)			
R5.e	1.360 ± 0.009 (12)	1.306 ± 0.015 (12)	1.396 ± 0.011 (12)	1.306 ± 0.013 (12)
	1.368 ± 0.009 (12)			
R5.f	1.357 ± 0.012 (25)	1.290 ± 0.013 (25)	1.360 ± 0.011 (25)	1.319 ± 0.012 (25)
	1.340 ± 0.014 (25)			
R5.g	1.367 ± 0.018 (76)	1 337 ± 0.022 (76)	1.420 ± 0.020 (76)	1.336 ± 0.023 (76)
	1.369 ± 0.017 (76)			
R5.h	1.416 ± 0.017 (33)	1.313 ± 0.017 (33)	1.422 ± 0.018 (33)	1.347 ± 0.019 (33)
	1.354 ± 0.015 (33)			

[a]The structures of the molecular fragments, together with the bond key and fragment letter code, are given in Fig. 6 of [1].

is comprehensive and has generated the mean bond length values for many entries missing from the work of Allen et al. [2]. The bond lengths of all these fragments show small standard deviations (always $< 10\%$ of the mean bond length for bonds involving hydrogen, and $< 5\%$ for bonds not involving hydrogen), meaning that the bond lengths of these fragments do not change appreciably when their environment is changed. This transferability of bond length properties is very important; in structure assemblies using these fragments, we can assume that the bond lengths of fragments are constant without incurring large error margins.

TABLE 7
BOND LENGTHS FOR 6-MEMBERED AROMATIC RINGS[a]

Fragment letter	Bond 1/5 (Å)	Bond 2/6 (Å)	Bond 3 (Å)	Bond 4 (Å)
R6.a	1.388 ± 0.010 (2373)	1.384 ± 0.010 (2373)	1.376 ± 0.012 (2373)	1.376 ± 0.012 (2373)
	1.384 ± 0.010 (2373)	1.387 ± 0.011 (2373)		
R6 b	1.337 ± 0.013 (154)	1 388 ± 0.015 (155)	1.379 ± 0.012 (155)	1.381 ± 0.012 (155)
	1.377 ± 0.012 (155)	1 337 ± 0.011 (154)		
R6.c	1.349 ± 0.019 (10)	1.333 ± 0.014 (10)	1.416 ± 0.020 (10)	1.358 ± 0.016 (10)
	1.403 ± 0.014 (10)	1.328 ± 0.023 (10)		
R6.d	1.333 ± 0.015 (72)	1.334 ± 0.016 (72)	1.338 ± 0.016 (72)	1.389 ± 0.023 (72)
	1.379 ± 0.021 (72)	1.341 ± 0.013 (72)		
R6.e	1.334 ± 0.012 (34)	1.390 ± 0.021 (34)	1.334 ± 0.013 (34)	1.332 ± 0.012 (34)
	1.387 ± 0.022 (34)	1.332 ± 0.015 (34)		
R6.f	1.331 ± 0.016 (22)	1.338 ± 0.017 (22)	1.332 ± 0.017 (22)	1.333 ± 0.017 (22)
	1.340 ± 0.017 (22)	1.333 ± 0.015 (22)		

[a]The structures of the molecular fragments, together with the bond key and fragment letter code, are given in Fig. 7 of [1]

TABLE 8
BOND LENGTHS FOR FUSED RINGS[a]

Fragment letter	Bond 1/5/9 (Å)	Bond 2/6/10 (Å)	Bond 3/7/11 (Å)	Bond 4/8 (Å)
Rf.a	1.372 ± 0.010 (59)	1.395 ± 0.013 (59)	1.379 ± 0.010 (59)	1.400 ± 0.009 (59)
	1.435 ± 0.011 (59)	1.364 ± 0.014 (59)	1.375 ± 0.017 (59)	1 376 ± 0.011 (59)
	1.393 ± 0.008 (59)	1.409 ± 0.008 (59)		
Rf.b	1.390 ± 0.018 (67)	1.380 ± 0.016 (67)	1.388 ± 0.014 (67)	1.379 ± 0.014 (67)
	1.483 ± 0.015 (67)	1.400 ± 0.012 (67)	1.400 ± 0.015 (67)	1.485 ± 0.013 (67)
	1.382 ± 0.013 (67)	1.382 ± 0.012 (67)		
Rf.c	1.381 ± 0.010 (20)	1.403 ± 0.009 (20)	1.381 ± 0.010 (20)	1 390 ± 0.008 (20)
	1.393 ± 0.012 (20)	1.322 ± 0.026 (20)	1.359 ± 0.025 (20)	1 379 ± 0.012 (20)
	1.391 ± 0.011 (20)	1.395 ± 0.011 (20)		
Rf.d	1.332 ± 0.011 (54)	1.340 ± 0.010 (54)	1.349 ± 0.012 (54)	1 405 ± 0.011 (54)
	1.387 ± 0.009 (54)	1.312 ± 0.009 (54)	1.371 ± 0.014 (54)	1.375 ± 0.010 (54)
	1.343 ± 0 007 (54)	1.383 ± 0.009 (54)		
Rf.e	1.401 ± 0.018 (29)	1.396 ± 0.018 (29)	1.393 ± 0.010 (29)	1.391 ± 0.017 (29)
	1 487 ± 0.028 (29)	1.411 ± 0.039 (29)	1.404 ± 0.044 (29)	1.482 ± 0.024 (29)
	1.385 ± 0.014 (29)	1.381 ± 0.014 (29)		
Rf.f	1.372 ± 0.012 (142)	1.411 ± 0.013 (142)	1.359 ± 0.013 (142)	1.417 ± 0.012 (141)
	1.418 ± 0.010 (142)	1.360 ± 0.013 (141)	1.403 ± 0.011 (142)	1.367 ± 0.010 (142)
	1.419 ± 0.011 (142)	1.426 ± 0.013 (142)	1.423 ± 0.011 (142)	
Rf g	1.324 ± 0.020 (94)	1.409 ± 0.018 (94)	1.361 ± 0.022 (94)	1.413 ± 0.016 (94)
	1.414 ± 0.016 (94)	1.367 ± 0.019 (94)	1.408 ± 0 018 (94)	1.370 ± 0.019 (94)
	1 420 ± 0.016 (93)	1.367 ± 0.015 (94)	1.413 ± 0.015 (94)	
Rf.h	1.308 ± 0.013 (12)	1.435 ± 0.023 (12)	1.316 ± 0.015 (12)	1.357 ± 0.013 (12)
	1.415 ± 0.018 (12)	1.362 ± 0.013 (12)	1.408 ± 0.008 (12)	1.375 ± 0.018 (12)
	1.404 ± 0.013 (12)	1.364 ± 0.013 (12)	1.411 ± 0.019 (12)	

[a]The structures of the molecular fragments, together with the bond key and fragment letter code, are given in Fig. 8 of [1]

In this work, we have also calculated the Cartesian coordinates of the average fragments. This can be used by the structure assembly programme, currently under development, to build up a molecule from fragments. We have neglected bond angles because the bond length is more useful in the preliminary assessment of fragment assembly by placement on molecular graphs with standard bond angles. Whether a fragment can be located in a certain part of a binding site can easily be tested by comparing the size of the site and the bond lengths of the fragment. In later stages, the Cartesian coordinates can be used.

The results from this work should be useful in drug design. All the bonds surveyed are found in commonly occurring fragments; this information forms a large database of judiciously chosen high-quality data. In the next paper we extend the investigation to a study of the electronic distribution in all the molecular fragments [3]. We would then have a large knowledge base of molecular fragments, including their constituent atoms, Cartesian coordinates, bond lengths and electronic distributions, for site-directed drug design.

ACKNOWLEDGEMENTS

We would like to acknowledge Dr. Frank Allen of the Cambridge Structural Database for his

constant patience and advice whilst this work was in progress. We are indebted to the Croucher Foundation (P.-L.C.) and the Wellcome Trust for personal financial support through the Principal Research Fellowship scheme (P.M.D.). Part of the work was carried out in the Cambridge Centre for Molecular Recognition supported by the SERC.

REFERENCES

1 Chau, P.-L. and Dean, P.M., J. Comput.-Aided Mol. Design, 6 (1992) 385.
2 Allen, F.H., Kennard, O., Watson, D.G., Brammer, L., Orpen, A.G. and Taylor, R., J. Chem. Soc. Perkins Trans. II, (1987) S1.
3 Chau, P.L. and Dean, P.M., J. Comput.-Aided Mol. Design, 6 (1992) 407.