Probabilities of formation of bimolecular cyclic hydrogen-bonded motifs in organic crystal structures: a systematic database analysis

Frank H. Allen,*[a] Paul R. Raithby,*[b] Gregory P. Shields[a,b] and Robin Taylor[a]

[a] *Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, UK CB2 1EZ*
[b] *Department of Chemistry, Lensfield Road, Cambridge, UK CB2 1EW*

A database methodology has been developed to locate all possible occurrences of bimolecular cyclic hydrogen-bonded motifs in the Cambridge Structural Database and to calculate their probabilities of formation; the top 24 motifs involving D–H···A (D, A = N or O) are identified.

The design process in crystal engineering depends crucially on the high probabilities of formation of a limited number of strong intermolecular interactions.[1,2] These non-covalent motifs, termed bimolecular couplings[3] or supramolecular synthons[2] by analogy with molecular synthesis, often involve strong hydrogen bonds. Despite their successes[1,2] crystal structure prediction and retrosynthesis depend heavily on evidence gleaned from visual surveys of limited subsets of existing crystal structures.[4] To extend the list of potential H–bonded synthons, we must identify and categorise the diversity of motifs that occurs, *e.g.* in the > 170 000 crystal structures recorded in the Cambridge Structural Database (CSD).[5] This data driven approach establishes the topologies, chemical constitutions and probabilities of formation of all H-bonded motifs in an objective manner.

Here we report probabilities of formation for cyclic bimolecular motifs that incorporate D–H···A (D, A = N or O) bonds. Our procedure identifies the important structure-determining H-bonded motifs that link pairs of molecules in crystal structures, *i.e.* we are interested in the immediate non-covalently bonded environment(s) of the molecule(s) that comprise the crystal chemical unit (CCU)‡. Once identified, we envisage that extension of a structure using each of these primary links can be visualised in a manner that mimics crystal growth, and that we can then generate and describe complete networks for structure classification and comparison purposes.[6]

The CSD System program Quest3D,[5] which can locate general non-bonded contacts according to pre-defined geometrical criteria, has been modified to investigate H–bonded motifs. For each target molecule in the CCU, the modifications are (a) identify potential donors (D–H) and acceptors (A), (b) locate bonds D–H···A involving the target molecule according to preset geometrical constraints, (c) establish whether pairs of H-bonds between the target and a neighbouring molecule form a cyclic motif, classify other H-bonds as isolated, (d) establish motif ring sizes (using shortest intramolecular bond paths), their symmetry (both topological and crystallographic), and their chemical constitutions and (e) provide interactive graphical display of each cyclic motif.

Despite the generality of our modifications to Quest3D, the initial analysis specifically addresses cyclic systems: it provides probability statistics for bimolecular cyclic motifs involving N–H or O–H donors and N or O acceptors, *i.e.* motifs that incorporate the strongest H-bonds and, hence, can be regarded as structure determining. In our automated procedure, all D–H bond lengths were normalised to mean values from neutron diffraction,[7] and H···A distance limits were established from histograms generated using the October 1996 release of the CSD: 2.30 Å in N–H···N, 2.25 Å in N–H···O, and 2.20 Å in O–H···N and O–H···O. Additionally, the D–H···A angle was required to be > 90°. CSD entries were accepted for analysis if they were classified as error-free, organic, and had $R < 0.10$.

The size, symmetries and chemical constitution of each cyclic motif were combined into a chemical topology record and these were sorted to yield raw occurrence values (N_{occ}). While these figures are interesting, their interpretation is complicated by the fact that they depend on the number of times that the various donor and acceptor groups occur in the CSD. For example, the carboxylic acid ring motif might have a high N_{occ} value simply because carboxylic acids are common in the CSD. To correct for this effect, we determined a probability of occurrence (expressed as a percentage), Prob = 100 N_{occ}/N_{poss}, for each of the 75 most common motifs. Here, N_{poss} is the number of times that the motif could possibly occur, given: (a) the number of structures in the CSD that contain the required functional group(s), and (b) the number of times that the groups(s) occur in the CCU of each structure. The automatic generation of N_{poss} is complicated, *inter alia*, by the need to consider: (a) the presence or absence of topological symmetry, (b) which overlapping fragments can be used simultaneously in forming motifs, (c) the three-dimensional geometry of certain fragments, (d) the effects of bifurcation and other multi-centre bonds and (e) the sharing of H-bonds between two rings. We have generated a set of logical rules§ which take account of these considerations; the effects of (c) are not straightforward to assess, and no account has been taken of competition for donor-

Table 1 Probability-ordered statistics for the 24 motifs in the CSD having $N_{occ} > 25$ and Prob > 20%

Motif No.	N_{occ}	N_{poss}	Prob (%)	Symm[a] (%)
1	93	96	97	—
2	199	218	91	—
3	36	40	90	—
4	36	44	82	—
5	62	82	76	—
6	206	354	58	—
7	158	290	55	—
8	79	154	51	—
9	86	192	45	—
10	38	92	41	68
11	45	114	40	—
12	47	120	39	—
13	44	118	37	—
14	204	556	37	75
15	58	159	37	62
16	39	111	35	64
17	29	83	35	79
18	847	2541	33	65
19	99	306	32	—
20	93	341	27	76
21	172	660	26	64
22	50	206	24	—
23	876	3687	24	62
24	84	404	21	64

[a] Motif possess crystallographic symmetry; other motifs are topologically (hence crystallographically) asymmetric.

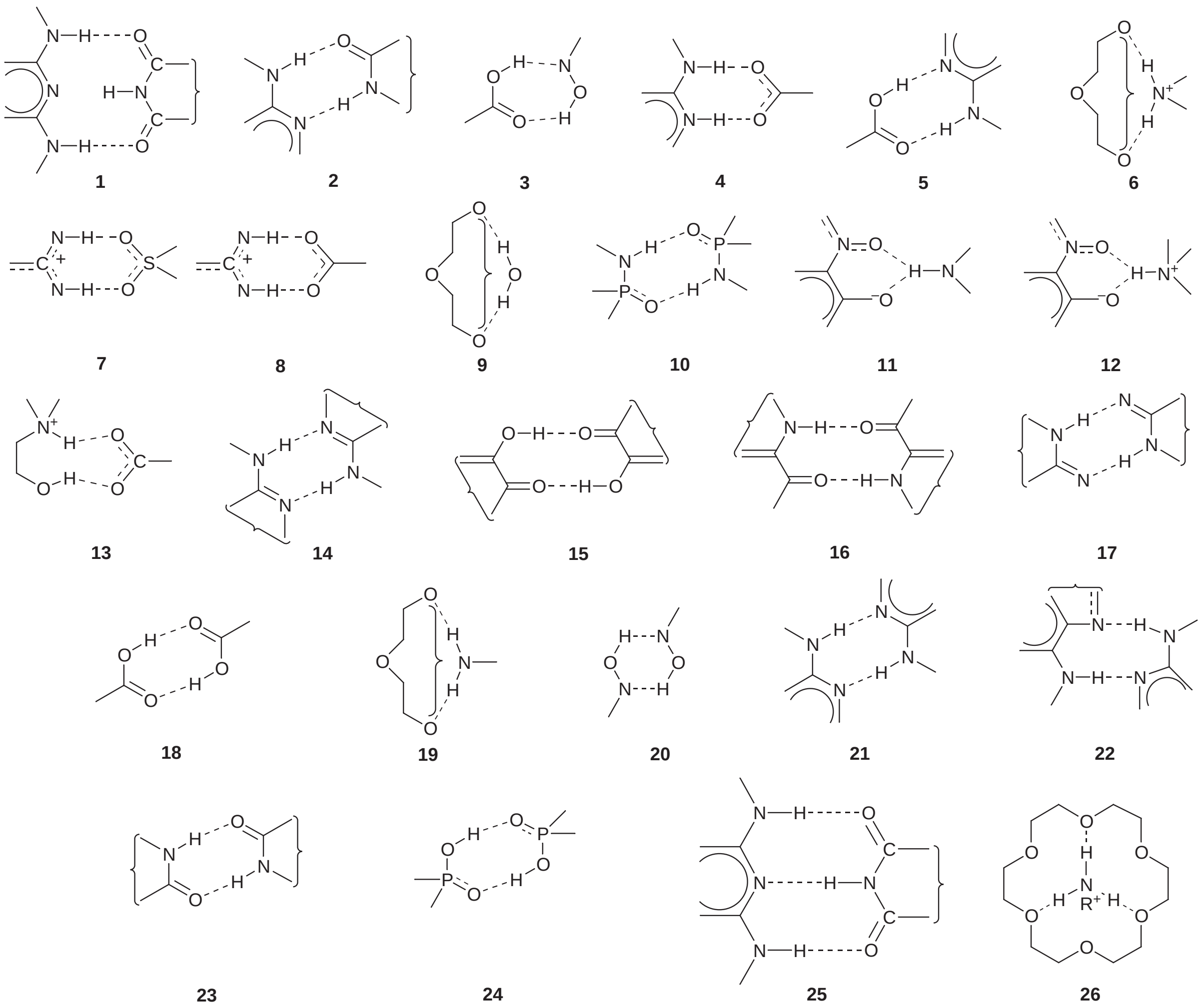

Fig. 1 Probability-ordered H-bonded motifs **1**–**24** and recognition motifs **25** and **26** (see text)

H between different acceptors or *vice versa*. Although the rules are capable of further refinement, principally for rarer situations, we believe that the relative probabilities for the most common motifs presented in Table 1 are reliable.

The probability-ordered Fig. 1 and Table 1 include the 24 motifs that have $N_{\mathrm{occ}} > 25$ and Prob > 20%. Motifs **1** and **2** are part of the triply H-bonded motif **25** (**1** is the envelope ring), having some analogies with nucleotide recognition in DNA, and a potent supramolecular synthon.[2,8] The other very high probability motifs, **3**–**5**, all occur in relatively few individual structures (12, 17 and 20 respectively), but **5** is noted by Desiraju.[2] Motif **6** represents a portion of an 18-crown-6 ether–ammonium complex. Here, three N-H donors from $[RNH_3]^+$ normally lie above the ring and the complete pattern, **26**, is best regarded as the recognition motif. Analogous complexes of 18-crown-6 with water (**9**) or RNH_2 (**19**) have lower but still appreciable probabilities.

The low probabilities of formation of the carboxylic acid **18** and cyclic amide **23** motifs might be considered surprising. However, they reflect chemical environments that contain many competing acceptors and donors, together with steric influences, particularly in the amide case. If the analysis is restricted to mono- and di-carboxylic acids without other functional groups, then Prob for motifs **18** and **23** rises to 95.5 and 84.8% respectively. The cyclic amides of motif **23** have their H-donors constrained to be *cis* to the =O acceptor. In flexible acyclic amides, which are usually *anti*, the overall probability of forming a cyclic motif is only 8.4%. However, if the analysis is restricted to structures having only primary or secondary amide functions, then the probability rises to 44.1% for the cyclic case and to 15.7% for the acyclic case. Using these statistics we are now analysing why particular motifs do not occur with greater frequency, and what alternative interactions are possible.

Notes and References

† E-mail: allen@ccdc.cam.ac.uk

‡ The crystal chemical unit (CCU) contains the complete unique molecule(s) that comprise the crystal structure. Where molecular and crystallographic symmetry coincide the fractional molecules are expanded by symmetry to form the CCU prior to the investigation of H-bonded motifs.

§ A brief description of these rules is provided as Supplementary material which is available on the World Wide Web: http://www.rsc.org/suppdata/cc/1998/1043.

1 C. B. Aakeröy and K. R. Seddon, *Chem. Soc. Rev.*, 1993, **22**, 397.

2 G. R. Desiraju, *Angew. Chem., Int. Ed. Engl.*, 1995, **34**, 2311.

3 W. Jones, V. R. Pedireddi, A. P. Chorlton and R. Docherty, *Chem. Commun.*, 1996, 997.

4 G. A. Jeffrey and S. Takagi, *Acc. Chem. Res.*, 1978, **11**, 264; L. Leiserowitz and A. T. Hagler, *Proc. R. Soc. London, Ser. A*, 1983, **338**, 133.

5 F. H. Allen and O. Kennard, *Chem. Des. Autom. News*, 1993, **8**, 1; 31.

6 M. C. Etter, *Acc. Chem. Res.*, 1990, **23**, 120; J. Bernstein, R. E. Davis, L. Shimoni and N.-L. Chang, *Angew. Chem., Int. Ed. Engl.*, 1995, **34**, 1555.

7 F. H. Allen, O. Kennard, D. G. Watson, L. Brammer, A. G. Orpen and R. Taylor, *J. Chem. Soc., Perkin Trans. 2*, 1987, S1.

8 W. Saenger, *Principles of Nucleic Acid Structure*, Springer-Verlag, New York, 1984, ch. 6; J.-M. Lehn, M. Mascal, A. DeCian and J. Fischer, *J. Chem. Soc., Chem. Commun.*, 1990, 479; G. M. Whitesides, E. E. Simanek, J. P. Mathais, C. T. Seto, D. N. Chin, M. Mammen and D. M. Gordon, *Acc. Chem. Res.*, 1995, **28**, 37.

Received in Cambridge, UK, 19th February 1998; 8/01424H