COMMUNICATION

ChemComm
www.rsc.org/chemcomm

The probable number of hydrogen-bonded contacts for chemical groups in organic crystal structures†

Lourdes Infantes*[ab] and W. D. Sam Motherwell[a]

[a] *Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, UK CB2 1EZ; Fax: 44 1223 336033; Tel: 44 1223 336408*

[b] *Instituto de Química-Física 'Rocasolano', CSIC, Serrano 119, E-28006 Madrid, Spain. E-mail: xlourdes@iqfr.csic.es; Fax: 34 915 642431; Tel: 34 915 619400*

Received (in Cambridge, UK) 26th February 2004, Accepted 25th March 2004
First published as an Advance Article on the web 23rd April 2004

A new database has been created for chemical groups and their hydrogen-bonded contacts for 41055 organic crystals. The accessible surface of an atom and the ratio between the number of donors and acceptors in a compound have been found to be useful parameters to predict the probable number of contacts for that atom.

Knowledge of hydrogen-bonded interactions is extremely important in the fields of crystal engineering, rational drug design, *etc.* Much work has been done in identifying commonly occurring H-bonding motifs,[1] often called synthons, which could be used to engineer desired crystal packing assemblies.[2] There are many studies of specific chemical groups, and insights can be gained by examination of the patterns of group interactions in the Cambridge Structural Database (CSD),[3] using the ConQuest search program, IsoStar (a library of scatter-plots of group interactions), and Mercury which has a user-friendly interface for non-bonded contact display. However, with current software it is generally difficult or impossible to obtain the average number of hydrogen bonding contacts to a given group. Thus, there would be many applications of a new methodology which could predict for a given molecule what are the most likely interactions of groups in crystal polymorphs, *e.g.* in pharmaceutical materials development. Some factors which are most likely to determine the interaction of groups are their electronic characters, steric accessibilities and the donor/acceptor ratio for the compound as a whole.

Our methodology has been to calculate all contacts between potential hydrogen bonding groups in CSD structures having: OH_{1-3}, or NH_{1-4}, or SH present, no metals present, accuracy R < 0.10, no errors, no disorder, not polymeric, giving 41,055 crystal structures (reference codes). Intermolecular contacts were defined as distances < (sum of van der Waals radii + 0.1 Å), categorised as donor, acceptor, and non-hydrogen-bonds. The data were output from the RPluto[4] program into tables, with atoms assigned an atom code for chemical environment (*e.g.* O in OH, O in C=O), and belonging to a list of 108 common chemical groups (*e.g.* carboxylic COOH, hydroxyl C–OH, oxime C=N–OH, *etc.*). For each donor/acceptor atom a sterically accessible surface,[5] AS, was calculated using a probe of radius 1.2 Å. A relational database, CSDContact, was built using the Microsoft Access software, linked by the CSD reference codes, 117 atom type codes, and 108 group codes. A flexible query language allows selection and counting of occurrences of combinations of properties. For example, "Find all N in secondary amines R–NH–R and count average donor and acceptor contacts"; "Find all O in hydroxyl groups with 1 connection, and get the average value of AS".

What is the average number of donor/acceptor contacts per group? The average counts of donor contacts, ND, and acceptor contacts, NA are given in Table 1 for 21 common groups. We can see that some groups show a very high NA, *e.g.* O of P=O or keto oxygen of COOH, and others almost never accept, *e.g.* N of planar NH_2. Preliminary investigations show that these figures alone do not predict the number of hydrogen bond contacts for a given test molecule, but are highly dependent on the numbers of groups available and ND + NA for the molecule. We would expect different behaviour for one donor, say OH, and 5 acceptor groups, compared to structures with only a single OH.

How does the number of contacts depend on the accessible surface, AS? Each atom type can often exist in several different chemical groups, thus having different electrostatic potentials and steric accessibility. There is a positive correlation between the number of contacts, NC, and the average AS value, < AS >, which is illustrated in a selection of groups in Table 2, averaged over all atoms with the specified NC value.

How does the number of contacts depend on donor/acceptor ratio, RDA? The summed values of ND, NA were used from each group atom as an estimate of the effective donor/acceptor ratio, RDA = ΣND/ΣNA for each structure, which is formally different from the ratio of the integer counts of donor or acceptor atoms, *e.g.* a pyramidal amine NH_2 has a maximum of 2 donors and 1 acceptor, but observed ND = 1.24, NA = 0.33. For a given atom with a specified number of contacts NC we calculated the average value of RDA over the structures. Table 2 shows that the number of contacts increases as < RDA > increases.

How does the number of contacts depend on AS and RDA? In the total sample there is a probability F of any group having NC = 0, 1, 2, 3 connections. It can be seen that some groups are much more likely to have no contacts (*e.g.* etheric O, 79%), and some almost always have > 0 contacts (*e.g.* carboxylic OH, 94%). We examined contacts for each group, *e.g.* keto-O, counting the number of occurrences for categories NC = 0, 1, 2, and plotted against binned histogram values for RDA and AS, as in Fig. 1. We see a clear tendency for NC = 0 to occur at lower RDA, and lower AS than for

Table 1 Average number of donor contacts, ND, and acceptor contacts, NA, for some common groups. Nocc is the number of occurrences of this atom in a sample of 41055 crystal structures

Atom Type	ND	NA	Nocc
O of C=O keto	0.00	0.59	8894
O of C=O (COOH)	0.00	0.96	6548
O of C=O (COOR)	0.00	0.38	11330
O of C–O–C	0.00	0.16	38291
O of P=O	0.00	1.12	1211
O of C–OH	0.82	0.47	41546
O of N–OH	1.02	0.16	1448
O of P–OH	1.10	0.33	1061
O of H_2O	2.04	1.10	7444
N of CN	0.00	0.55	2404
N of –N=	0.00	0.44	11801
N of NH_2 (planar)	1.61	0.06	5425
N of NH_2 (pyramidal)	1.24	0.33	983
N of NHR_3^+	0.88	0.00	2095
N of $NH_2R_2^+$	1.81	0.00	2432
N of NH_3R^+	3.58	0.00	3129
N of NH_4^+	5.51	0.00	252
F^-	0.00	3.29	21
Cl^-	0.00	2.68	2535
Br^-	0.00	2.15	939
I^-	0.00	1.14	370

† Electronic supplementary information (ESI) available: details of the calculated properties of atoms and groups in Tables 1 and 2. See http://www.rsc.org/suppdata/cc/b4/b402939a/

DOI: 10.1039/b402939a

NC > 0. There is not such a clear distinction for the cases NC = 1, 2.

Within each category we find a wide spread of RDA and AS values, with no clear limit below which we could say there will be no contacts, showing that the effect of local steric contact geometry is also important. We also find that structures with small numbers of donor and acceptor groups, *e.g.* a subset of 117 molecules with only one hydroxyl OH and one etheric O, depart from the average statistics and, in this case, the etheric O often accepts a contact, F = 68%, not expected from the average for NC > 0 with F = 21%.

As an example of application of the methodology we selected a set of 82 structures where we have one alcoholic OH and two keto-O present. We counted the cases where the OH···O=C contact occurred (80), and whether the O-keto with the higher (50) or lower (17) AS was used. There were 13 cases of an intramolecular OH···O=C contact having no intermolecular connection, and 2 cases of no contact. Thus we have a probability of an intermolecular contact OH···O=C of 82% (67/82), with 61% in favour of the higher AS, 21% for the lower AS, and 18% for no connection.

Table 2 Number of contacts, NC, for some common groups related to average accessible surface, < AS >, and the average donor/acceptor ratio < RDA > over all structures containing the specified group and NC. For each structure RDA = ΣND/ΣNA summed over atoms. Nocc as in Table 1. F is the frequency of occurrence in the sample with values NC = 0, 1, 2, 3

Atom Type	Group Type	NC	< AS >	< RDA >	F (%)	Nocc
O-keto O=C	R_2C=O	0	0.284	0.568	52	4448
		1	0.337	0.754	39	3396
		2	0.355	0.890	8	650
		3	0.375	0.952	1	95
O-acid O=C	COOH	0	0.347	0.506	26	1673
		1	0.357	0.675	56	3655
		2	0.380	0.899	14	922
		3	0.385	0.903	3	191
O-ester C=O	COOR	0	0.267	0.524	66	7521
		1	0.319	0.692	29	3333
		2	0.320	0.737	4	425
O-ether	C–O–C	0	0.096	0.758	79	19738
		1	0.11	0.983	20	4987
		2	0.125	1.032	1	281
O of P=O	R_3P=O	0	0.206	0.339	24	98
		1	0.274	0.455	63	259
		2	0.257	0.693	12	51
		3	0.282	1.243	1	6
O-alcohol OH	R_3C–OH	0	0.129	0.856	18	1239
		1	0.140	0.850	57	3885
		2	0.171	1.224	23	1534
		3	0.178	1.085	2	104
O-phenolic OH	C–OH	0	0.204	0.870	30	2371
		1	0.244	0.893	48	3771
		2	0.272	1.270	19	1531
		3	0.272	1.273	2	165
O-acid –OH	COOH	0	0.262	0.533	6	393
		1	0.232	0.645	74	4772
		2	0.245	0.791	17	1071
		3	0.247	0.886	3	213
N-cyano	CN	0	0.479	0.565	57	1333
		1	0.490	0.898	34	799
		2	0.489	1.077	8	180
		3	0.506	1.232	1	24
N of –N=	–N=	0	0.086	0.672	60	4689
		1	0.163	0.952	37	2861
		2	0.178	1.022	2	189

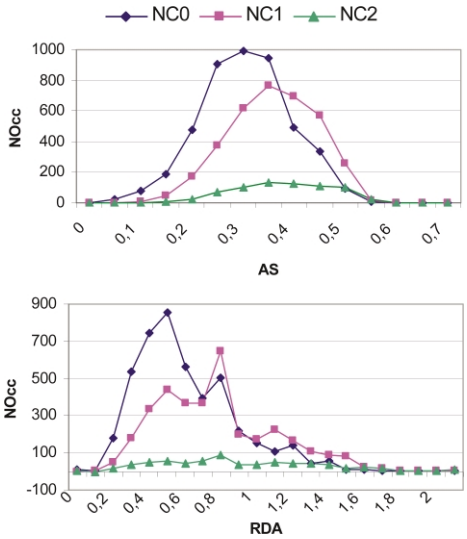

Fig. 1 Diagram for the keto-O group showing the number of occurrences, Nocc, observed for the average accessible surface, < AS >, and the average donor/acceptor ratio, < RDA >. Separate graphs are given for the number of contacts NC = 0, 1 and 2.

Knowledge of the most probable interactions in a crystal is valuable information to assist prediction of the packing of organic molecules or to select the most likely from a set of possible polymorphic forms.[6] We are using the database to investigate the effects of the relative geometric location of groups, and to study in detail the effect of RDA and AS on the number of contacts, and also to identify preferred group–group interactions. The question of prediction of the number of contacts for an atom in a given molecule can only be answered with a certain probability, as there is complex dependency on the number of available polar groups and the hydrophobic crystal packing of molecules. However, we suggest that a database approach will eventually provide such probabilities with increasing confidence.

Notes and references

1 F. H. Allen, W. D. S. Motherwell, P. R. Raithby, G. P. Shields and R. Taylor, *New J. Chem.*, 1999, **23**, 25.
2 G. R. Desiraju, *Angew. Chem., Int. Ed. Engl.*, 1995, **34**, 2311.
3 F. H. Allen, *Acta Crystallogr., Sect. B*, 2002, **58**, 380; F. H. Allen and W. D. S. Motherwell, *Acta Crystallogr., Sect. B*, 2002, **58**, 407.
4 *RPLUTO, A program for crystal structure visualisation*, Cambridge Crystallographic Data Centre, Cambridge, 2002 (http://www.ccdc.cam.ac.uk).
5 L. Infantes and S. Motherwell, *Struct. Chem.*, 2004, **15**, 173.
6 J. D. Dunitz, *Chem. Commun.*, 2003, 545.