Molecular Crystals and Liquid Crystals Science and Technology. Section A. Molecular Crystals and Liquid Crystals

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/gmcl19

Crystal Structure Prediction and the Cambridge Structural Database

W. D. S. Motherwell [a]

[a] Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, CB2 1EZ, U.K.

Version of record first published: 24 Sep 2006

To cite this article: W. D. S. Motherwell (2001): Crystal Structure Prediction and the Cambridge Structural Database, Molecular Crystals and Liquid Crystals Science and Technology. Section A. Molecular Crystals and Liquid Crystals, 356:1, 559-567

To link to this article: http://dx.doi.org/10.1080/10587250108023734

Mol. Cryst. and Liq. Cryst., 2001, Vol. 356, pp.559-567
Reprints available directly from the publisher
Photocopying permitted by license only

Crystal Structure Prediction and the Cambridge Structural Database

W.D.S. MOTHERWELL

Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, U.K.

A computer program has been written to attempt prediction of likely polymorphs of small organic molecules, using a genetic algorithm. The cost function used is based not on energy but the intermolecular atom pair distances as seen in the Cambridge database, for selected similar molecules to the target. Some successful results are described, showing possible applications to molecules not easily treated with empirical energy potentials.

Keywords: crystal structure prediction; database; frequency curves

INTRODUCTION

The traditional method for prediction of the most likely crystal structures for a given molecular compound has been that of lattice energy calculation. The assumption is that the calculated structure with the lowest lattice energy will be the thermodynamically most stable polymorph at all temperatures, and calculations are performed on a set of motionless molecules, ignoring vibrational enthalpy and entropy contribution to the free energy. In recent years it has become apparent that there are many more minima in this calculated energy surface than had been expected, typically 10-100 in the range of 10KJ/mole above the global minimum [1,2]. It is therefore of considerable interest to test if there is any means of using the information latent in the Cambridge Structural Database (CSD) to reduce the number of possible polymorphs in the prediction candidate structures.

Empirical energy atom-pair potentials are usually validated against selected experimental structures and experimental data on thermodynamic properties, well reviewed by Gavezzotti [3]. Recently there have been reports of statistically derived atom-pair potentials derived from surveys of many structures taken from the CSD [4]. A new approach is under development to use the library of interaction scatterplots between chemical fragments, derived from the CSD, known as IsoStar[5]. This paper reports some experiments using direct

comparisons of statistics on interatomic distances between calculated structures and the CSD, without assuming functional forms for energy potentials.

METHODOLOGY

Distance Frequency Curves

The simplest statistic that can be derived from the CSD on intermolecular distances is to count the total number of atom-pair distances for each crystal structure into histogram bins. This histogram if averaged over many structures for a given atom type gives a radial distribution curve, with a standard deviation of the mean value for each histogram bin. Those curves which show a low standard deviation could possibly be used in a predictive manner, because any hypothetical calculated structure that agrees well with these curves in fact fits well with the CSD observations, and thus may be expected to be a possible polymorph.

Some experiments were made with regard to the samples of the CSD which might be used to construct these reference curves. It was found that the curves often have distinct character for each element depending on its immediate chemical environment, i.e. we can identify differences for say carbon in methyl groups, carbon aromatic with one hydrogen, and carbon in carbonyl groups, etc. These are referred to as atomic types, and are automatically assigned by software written within the RPLUTO program [6]. For these curves a normalised contact distance was used as follows: the intermolecular contacts, R_{ij}, are replaced by $R = R_{ij} - R_v$, where R_v is the sum of the van der Waals radii for each atom-pair ij. When a molecule contains several atoms of a given type we average and refer to these curves as C_{methyl}, $O_{carbonyl}$, etc., and count into bins at an interval of 0.2 A. Because of the small number of contacts at low R a simple smoothing function is used to produce an approximate frequency distribution curve, by simple averaging of each bin with its two adjacent neighbours; some examples are shown in Figure 1.

It has been found by visual inspection that these curves generally have a similar shape, and are affected by molecular size. In general there are more unusual interactions in very small molecules near the start of an homologous series, which is visible particularly in the region of the curve near $R = 0$. Ideally one should weight the comparison of curves according to the inverse of the variance of each curve, but in this study all curves are given the same weight.

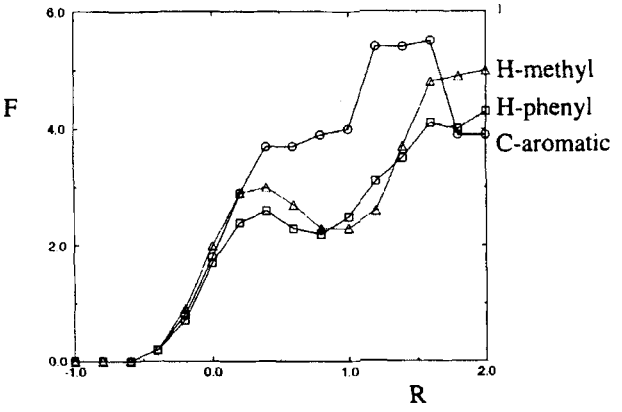

FIGURE 1. Typical distance frequency curves for hydrogen and carbon

The program cost function, Q, is calculated for each relevant atom type to produce curves F_{calc} for the hypothetical structure, which are compared with F_{ref} for the reference curves, where $Q = \Sigma (F_{calc} - F_{ref})^2$, summed over all atoms in the molecule. The range of bins for comparison is typically R_v-1.0 to R_v+1.0 A.

These curves contain information on the 3D packing environment of each atom type, in a one-dimensional isotropic representation. It has been found that the standard deviation of curves for atom types are generally lower if the CSD reference set of molecules is selected to those molecules of approximately isomeric formulae to the target molecule, with the same set of chemical functional groups. This is conveniently achieved by using the molecular empirical formula search features in the CSD search program, Quest, together with the 2D-similarity feature.

One of the features of this cost-function is that it can be used to deal with atom types which are generally considered difficult to treat by the energy potential method, such as bromine, iodine or metal complexes. Provided there are sufficient similar molecules in the CSD one can use this method, and indeed it has been successful with a bromine compound mentioned below, and a metal pi-complex (CSD code ALYLRH)

Penalty functions

We may also recognise any persistent H-bonding motifs, or other intermolecular contacts, in the CSD reference set. These may be expressed in terms of a penalty function, P which can be set to an arbitrary high value when a particular motif expected because of high occurrence in the reference set, does not appear in the calculated structure. In this case we simply add the value with a weighting value, w, thus, $Q_{total} = Q + wP$. At present this selection of persistent motifs is not automated, but derived by visual inspection of the reference set using RPLUTO visualisation. The penalty function may be codified for computation by considering a given number expected contact distances within a given spherical radius around the atom type, not using any directional information, (EXPSPH command), or by expecting actual specified contacts between named atom types, with specified geometry defined by polar coordinates (EXPCON command). Generally in the studies presented here the approach has been to use the simplest non-directional description, to leave the maximum number of possible geometric motifs accessible to the GA process.

A penalty function is always applied to prevent too close approach of atoms, which is related to the typical repulsive term in empirical atom pair potentials, but much softer in effect, increasing only according to $w(R_{ij}-R_{csd})^2$, where R_{csd} is the closest approach found in the CSD reference set for element pair ij, using the sum of the van der Waals radii as the default setting.

Program details

A computer program, RANCEL, has been developed which uses a genetic algorithm [7] to search for the global minimum in a cost function Q. The molecule is treated as a rigid body, one molecule per asymmetric unit, in a set of predefined space groups most commonly found in the CSD. The control parameters have been found by experiment to give reasonable results with a fixed population size of 100, started randomly within a cell of up to 30 A axial lengths, cycled through 100 generations, with a crossover rate of 0.8 and mutation rate of 0.05. For each space group the run is repeated 20 times using different random number seeds. A 20-atom molecule, with 20 seeds, requires about 10 min CPU per space group on a Sun Ultra 10, 333MHz)

Naphthalene derivatives

A set of 60 naphthalene derivatives were selected for initial tests of the methodology. The first point to be verified was that if one uses the frequency distribution curves derived from a given CSD structure then the program usually finds a well defined global minimum in Q close to the CSD structure, with cell dimensions matching within 5 - 10%. Although this may seem a trivial result, it does confirm that there is sufficient 3D information included in these isotropic curves to define the structures. It is also found that these structures are close packed with reasonable density and calculated energy, using empirical potentials.

The idea of using the curves from one chemical isomer to predict the structure of other isomers was tested on a set of three dinitronaphthalenes. The curves from 1,5-dinitronaphthalene (DNNAPH, $P2_1/a$) were used for runs on the 1,8-dinitro compound (DNTNAP, $P2_12_12_1$). This turned out to be quite successful, even though the chemical environment of the nitro-groups is very different, (Figure 2.)

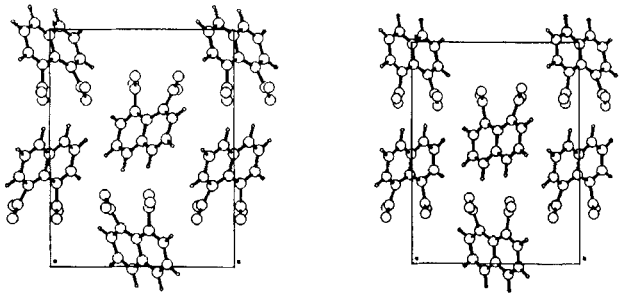

FIGURE 2. Comparison of the DNTNAP structure found by GA (left), cell 11.53 14.85 5.58, with CSD entry (right) 11.37 14.97 5.39.

It is interesting to note that the short contacts C-H...O are reproduced, with no special penalty function except that R_{csd} CH...O was set at 2.54 as found in DNNAPH. The Q value is not as sensitive as was hoped, and it is possible to find other polymorphs with similar Q values. There is

scope for modifying the Q function with a weighting scheme related to the standard deviations of the reference curves, which is under investigation.

Some naphthalene derivatives were found not to give lowest Q values unless certain other expected contact information was coded as penalty functions. It was found that 1,5-dibromo-naphthalene (DBRNAP), could not be found with lowest Q, unless we add expectation of the frequently observed C-H...π interactions in aromatic molecular packing. This was encoded with a sphere command, EXPSPH, expecting 2 H-atoms within 3.40 A of each carbon. Further work is in hand for automatic creation of expected contacts by scanning the reference set of CSD structures.

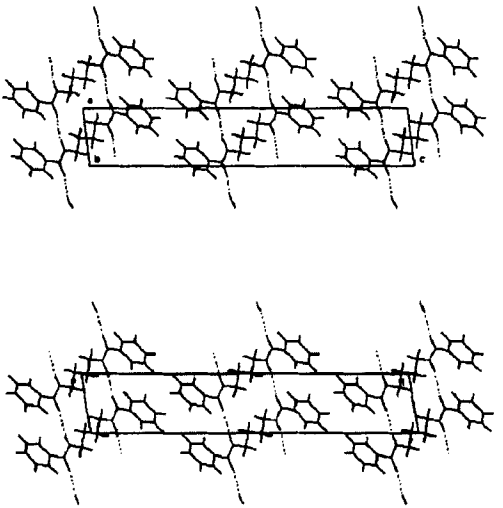

FIGURE 3. Result for the diamide MEBZAM, showing RANCEL lowest Q structure (top) and CSD entry (bottom). Space group is $P2_1/c$, Z=2, and hydrogen bonds are shown forming ladders. An origin shift (1/2,0,0) is visible, and permitted in this space group.

Diamide examples

As part of a study on packing patterns of diamides a CSD reference set of about 20 diamide structures was selected, in which there was no other H-bonding group besides the amide. To test RANCEL on particular members the distance frequency curves excluded the contribution of that member. Examination of close contacts showed that we can expect H-bonds NH...O < 2.20 seen in all 20 structures, and also C-H...O bonds < 2.80. The relevant R_{csd} values were set, and also expectation values, for sphere O 3.10 expect 4 H, sphere H(N) 2.20 expect 1 O. These simple conditions were found sufficient to give structures with the lowest Q values within the observed spacegroup, with the correct packing motifs (Figure 3). Tests on several other amides where the extended H-bonding motifs are different e.g. DEBHEX, showed prediction of the correct packing, even with these non-directional penalty functions

2-cyano-4-hydroxythiophene

This compound was one of the unpublished structures used in a blind test structure prediction workshop held at CCDC in May 1999 [8]. The RANCEL program was not successful in this case, although it predicted the correct H-bonding scheme. Post-analysis of the experimental structure (now published [9]) has been interesting in revealing the critical penalty functions which need to be included to produce a lowest-Q correct structure, see Figure 4.

The unsuccessful calculations included only the expected OH...N condition to be less than 2.00, on the basis of examination of CSD molecules which contained just OH and CN as donor/acceptors e.g. PECBII. The best Q-value structure showed the correct H-bonded chain, but wrongly packed. This is an example of a molecule which has very few quasi-isomers in the CSD, and also few examples of small molecules containing only the pairs of H-bonding groups, CN and OH

However, examination of the IsoStar database for interaction between O and thiophene S showed several examples of short contacts. Also the IsoStar scatterplot of CN and O shows a significant number of contacts O...C perpendicular to the CN bond around 3.50. We noted also short CH...O contacts might be expected from the acidic H. So by re-running RANCEL with the expected contact conditions of OH...N 2.00, S...O 3.40, O...CN 3.50 and CH...O 2.80, it was found that the correct structure did appear amongst the lowest Q values.

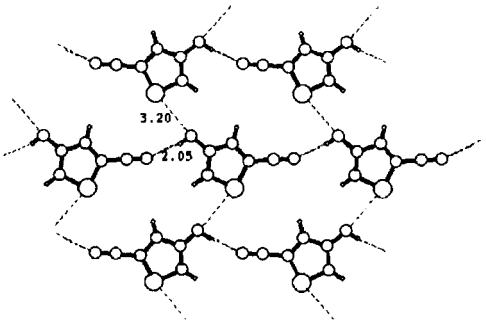

FIGURE 4. The structure in P21/c consists of near-planar sheets parallel to the (1 0 -1) plane with OH...N hydrogen bonds of 2.05 A and close contacts S...O of 3.20.

This illustrates, of course, the fundamental weakness of the method when there are insufficient examples of similar molecules with their functional groups in the required close proximity, which obviously give rise to considerable electrostatic effects in this case.

CONCLUSIONS

These preliminary investigations have shown that intermolecular distance frequency curves contain much packing information, and in some cases can give reasonable predictions of polymorphs, without recourse to energy calculation. Because the frequency curves have some dependence on the chemical element ratios in the molecules, and also the molecular size, it has been found that most success occurs when using the curves from a set of selected CSD molecules that are approximate isomers of the target. Non-directional information on interactions is often sufficient to allow prediction, and it is hoped that inclusion of directional information might improve the cost function. The present Q function is not sufficiently sensitive to give reliable prediction but does tend to converge on a small set of possible polymorphs; it may well be useful in combination with energy calculation in reducing the set of likely polymorphs. As the CSD grows in size the problem of limited sets of reference molecules should decrease.

References

[1] S.L. Price and K.S. Wibley, *J. Phys. Chem. A* **101**, 2198–2206 (1997).

[2] B.P. van Eijck, A.L. Spek, W.T.M. Mooij & J. Kroon, *Acta Cryst.* **B54,** 291–299 (1998).
[3] A. Gavezzotti, *Crystallography Reviews,* **7,** 5–121 (1998).
[4] D.W.M. Hofmann and T. Lengauer, *Acta Cryst.,* **A53,** 225–235 (1998).
[5] I.J. Bruno et al., *J. Comput.-Aided Molec. Des.,* **11,** 525–537 (1997).
[6] RPLUTO program, down-loadable from CCDC at *www.ccdc.ac.uk*(1999).
[7] R. Storn and K. Price, *J. Global Optimization,* **11,** 341–359 (1997).
[8] J.P.M. Lommerse et al., *Acta Cryst., B,* submitted, (1999).
[9] A.J. Blake et al., *Acta Cryst.,* **B55,** in press, (1999).