Acta Crystallographica Section B
Structural Science
ISSN 0108-7681

Structure predictions allowing more than one molecule in the asymmetric unit

Bouke P. van Eijck* and Jan Kroon

Department of Crystal and Structural Chemistry, Bijvoet Center for Biomolecular Research, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands

Correspondence e-mail: vaneyck@chem.uu.nl

Received 21 October 1999
Accepted 5 January 2000

The *UPACK* program for crystal structure prediction was extended to allow the possibility of more than one molecule in the asymmetric unit. A search method was developed where the essential parameters (including torsional angles of hydroxyl groups) take random values. Energy minimization and clustering then lead to a list of hypothetical structures. In a test for six hexapyranoses this random search method was found to be approximately equally as efficient as the previously used systematic grid search method for one independent molecule. As a second test, the generation of structures with more than one independent molecule was performed for ethanol. A multitude of possible structures was found, the experimental one (two independent molecules in space group *Pc*) always being present. However, for α-D-mannose (two independent molecules with five unknown hydroxyl torsional parameters each) the number of hypothetical structures was so large that the experimental structure was never encountered. Finally, a crystal structure generation for hydrates of pyranoses and polyalcohols was carried out. When the total number of unknown parameters was less than 20, the experimental structure was encountered. As usual, the empirical energies were not adequate to select that structure from the list of possible ones. Even then, the procedure proved to be valuable in identifying the most probable hydrogen-bonded network in cases where hydrogen positions in the experimental crystal structure were missing.

1. Introduction

In *ab initio* crystal structure prediction the first task is to construct a list of hypothetical structures (Gdanitz, 1997; Verwer & Leusen, 1998). For crystals with one rigid molecule in the asymmetric unit of one of the common space groups, such a list usually contains the experimentally known polymorph(s). Recognizing these, as well as other possible realistic structures, is a difficult problem that will not be the main topic of this paper. Here we shall focus on the problems that arise if more than one molecule in the asymmetric unit is allowed. The obvious consequence is an enormous increase in the number of possible structures.

First a few words about notation. The number of residues in the unit cell is conventionally indicated by Z. The relative multiplicity Z' is defined as $Z' = Z/M$, where M is the multiplicity of the general position. This is not necessarily the same as the number of independent molecules: for instance, $Z' = 1$ if there are two independent molecules in different twofold special positions (Wilson, 1993). Therefore, we shall introduce the symbol Z'' to denote the number of crystallographically nonequivalent molecules. This quantity is important because it gives the number of entities that have to be positioned and oriented independently. For compound structures, *e.g.* hydrates, we have necessarily $Z''>1$ even if there is only one residue in the asymmetric unit ($Z' = 1$).

Some statistical surveys of organic structures occurring in the Cambridge Structural Database (Allen & Kennard, 1993) have been carried out 10 years ago. Although the database has tripled its volume since then, we suppose that the qualitative conclusions are still valid. The occurrence of solvent molecules in crystal structures was estimated at approximately 10% for hydrates and 6% for other solvents (van der Sluis & Kroon, 1989). Another 8% of the structures was found to have $Z'>1$ (Padmaja *et al.*, 1990). The latter occurrence increases to 40% or more for alcohols (Brock & Duncan, 1994; Gavezzotti & Filippini, 1994). Thus, the situation studied in the present paper is by no means exceptional.

In most algorithms for generation of hypothetical crystal structures, one has to decide at the outset how many independent molecules have to be considered. We shall denote this chosen number by G; in principle, the calculations should be repeated for all probable values of G. The distinction between G and Z'' is subtle: if some independent molecules are chemically identical, it is to be expected that a set of structures generated for $G > 1$ will contain some structures with $Z'' < G$ as special cases. Such structures should be described in smaller unit cells or in space groups with higher symmetry.

In principle, all possible crystal structures can be described in space group $P1$ and should be found by a search in that space group for every expected value of G. Pioneering work along this line has been performed by Dzyabchenko (1989), who constructed seven hypothetical benzene structures with $Z'' > 1$. The principle of taking $G > 1$ in space group $P1$ has been used frequently (Gdanitz, 1992; Shoda *et al.*, 1995; Tajima *et al.*, 1995; Gibson & Scheraga, 1995; Williams, 1996) and in this way the crystal structures of several compounds could be predicted. However, they could always be reduced to more complex space groups with $Z'' = 1$. Very recently Gao & Williams (1999) reconsidered this approach, and came to the conclusion that these successes are at least partly due to the high symmetry of the molecules involved.

Recently we have described the possibilities and limitations of *ab initio* crystal structure prediction for carbohydrates (van Eijck & Kroon, 1999). Using a systematic search procedure, lists of possible crystal structures were generated for 32 pyranoses and 24 polyalcohols. These lists were ordered *via* energy in two empirical force fields; they contained hundreds of hypothetical polymorphic structures within an energy window of approximately 25 kJ mol^{-1}. The average energy of the experimental structure with respect to the energetically most favorable one was well below 10 kJ mol^{-1}. Such lists may well serve as a starting point for energy calculations with a more sophisticated force field, like the one that is at present being developed in our laboratory (Mooij, van Duijneveldt *et al.*, 1999).

However, the search problem was by no means solved satisfactorily. Occasionally (3 cases out of 56) the experimental structure was not found. More importantly, in selecting the compounds it was necessary to exclude four structures with unusual space groups, six structures with two independent molecules and 11 hydrates (van Eijck & Kroon, 1999). It would be an extremely laborious task to investigate all possible space groups, and the usual limitation to the most populated ones remains necessary in practice. The other two limitations both refer to the case $Z'' > 1$ and are technically almost identical. Up to now our computer package (*UPACK*) could not handle such structures. This limitation has now been removed, as will be described in this paper. We first explain the search strategy and discuss a few test cases to assess its performance. Then we report an attempt to predict the structure of α-D-mannose, where $Z'' = 2$. We finally discuss the possibility of structure prediction for the hydrates that had to be excluded from our previous study.

2. Methods

2.1. Random search

The standard way to generate hypothetical structures with the *UPACK* program package is a systematic grid search (van Eijck & Kroon, 1999). The extension of this method to the case of more than one independent molecule in the asymmetric unit would be quite complex to implement and could easily lead to an unmanageable computational effort. Therefore, that line of approach was abandoned in favor of a random search method.

Random search methods have been used before in crystal structure prediction programs (Verwer & Leusen, 1998). Some use a purely random search, others rely on molecular dynamics sampling or select promising structures on the basis of a Monte Carlo criterion. The commercially available and rather successful *Polymorph Predictor* program package (Molecular Simulations Inc., 1997) belongs to the latter category. We supposed that a straightforward random search may be just as efficient; in any case, it is more simple to implement and to reproduce.

In our procedure the space-group symmetry and the number of independent molecules, G, are chosen in advance. A Cartesian axes system (x, y, z) is used, where the crystallographic a axis coincides with the x axis and the b axis is positioned in the x, y plane. To create a trial structure the components a_x, b_y and c_z are given random values between 4 and 25 Å. For an orthorhombic space group this is sufficient; higher symmetry has not been implemented. In monoclinic space groups the variation of the cell angle β is allowed by assigning a random value (between 0 and a_x) to b_x. In triclinic space groups c_x is similarly chosen between 0 and a_x, and c_y takes a random value between 0 and b_y. So for these three crystal systems the number of lattice parameters to be determined is 3, 4 and 6, respectively.

In each cell constructed in this way the constituent independent molecules are placed with random positions and rotated over random Euler angles ϕ, θ, ψ (van Eijck *et al.*, 1998); to obtain an even sampling it is preferable to assign a random value to $\cos\theta$ rather than to θ itself. This necessitates the determination of an additional six parameters for each independent molecule (less for the first molecule if the space group does not have a fixed origin in one or more directions).

In the alcohols studied in this work the dihedral angles for hydroxyl groups are usually unknown, since they are determined by the hydrogen bonding network of the crystal. Therefore, these dihedral angles can also be set to random values, thus further increasing the number of unknown paramaters. Very often multiple conformations for heavier groups are also possible and then the procedure has to be repeated for every conformation (or, if $G > 1$, for every combination of conformations) that is considered likely to occur as a building block for the crystal. We shall refer to the sum of the number of parameters to be varied for each conformation as the dimensionality of the problem.

In order to consider only structures with a realistic density, a preliminary calculation is performed until 100 structures with reasonable energy are generated. The lowest volume encountered (V_0) is subsequently used in the actual search: the starting value for the parameter c_z is no longer random, but is found from $c_z = V_0/(a_x b_y)$. An apparent disadvantage is that most resulting structures will have an extremely high energy owing to overlapping molecules. However, these can be discarded immediately as soon as any repulsion term exceeds a given value (say, 2000 kJ mol^{-1}). This is detected rapidly and less effort is spent in energy minimization of relatively empty structures.

Accepted structures are subjected to a first energy minimization, where the molecules are already taken to be fully flexible. They are retained if their final energy is less than 40 kJ mol^{-1} with respect to the lowest energy encountered (so far). The resulting preliminary list usually contains equivalent structures, which are not always easily recognized if the unit-cell parameters turn out to be different. Our fast clustering routine (van Eijck & Kroon, 1997) can no longer be applied for $Z'' > 1$ and was replaced by a slower algorithm, which is based on comparison of the interatomic distances between structures (Mooij *et al.*, 1998). This method works only if the energy minimization is well converged, to ensure that the distance lists for equivalent structures are really compatible. An indication of the completeness of the random search procedure can be obtained from the number of equivalent structures: if low-energy structures are found only once, it is advisable to continue the search. This flexibility is an advantage over the grid search method.

The standard procedure is to generate 5000 random structures for each separate conformation. This is followed by clustering, after which a more thorough optimization is carried out for structures within an energy window of 30 kJ mol^{-1}. A second clustering delivers the final list of hypothetical structures. Although the calculations are fairly time consuming, they can be easily carried out on a modern personal computer.

2.2. Force fields

In this study the *UNITAT* force field (van Eijck & Kroon, 1999) was used. This is a force field where the groups CH, CH_2 and CH_3 (but not OH) are replaced by united 'atoms'. It was derived from *GROMOS*87 (van Gunsteren & Berendsen, 1987) by making a few adaptations to improve the modeling of carbohydrate structures. Its simplicity is advantageous to reduce the computational effort involved in structure generation and initial energy minimization. Afterwards, the resulting list of structures may be used as a starting point for continued energy minimization with a more sophisticated force field. The assumption is that there is a one-to-one correspondence between energy minima in the various force fields. This is not always the case and occasionally one loses a few structures that would have been found if the structure generation had been carried out directly in the other force field. We shall report some results obtained with continued energy minimization in the all-atom *OPLS* force field (Jorgensen *et al.*, 1996; Damm *et al.*, 1997).

These force fields had to be extended to be able to model crystal water. In *UNITAT* the *SPC/E* model (Berendsen *et al.*, 1987) was used, as appropriate for a descendent of *GROMOS*, with intramolecular force constants 4644 kJ mol^{-1} Å^{-2} for H—O stretch and 385 kJ mol^{-1} rad^{-2} for H—O—H bending (van Gunsteren *et al.*, 1996). In *OPLS* the *TIP3P* potential (Jorgensen *et al.*, 1983) was used, with intramolecular terms from Dang & Petitt (1987).

The performance of a force field must be examined in two ways. Firstly, experimentally known structures of related compounds should be reasonably well maintained upon energy minimization. In that case there will be no problem to establish which of the hypothetical structures correspond to an experimentally observed polymorph: the list of generated structures can be searched for the energy and geometrical details that correspond to the energy-minimized experimental structure. On the other hand, if observed geometries are badly deformed upon energy minimization, the whole exercise is rather pointless. Even if the deformed structure is present in the list (and we have encountered one case, propane, where it even was the one with lowest energy), the situation could hardly be classified as a successful structure prediction.

Secondly, the energy of the experimental structure should not be much higher than the energy of the hypothetical structure with lowest energy (the global minimum). This energy difference (denoted by ΔE) can be estimated after structure generation in a few relevant space groups, after which one may hope that no new structures with a significantly lower energy will be encountered. The ranking of the experimental structure with respect to the one with lowest energy (denoted by R) is more difficult to assess: unless the ranking is fairly low, it will generally deteriorate even further with every additional set of structures that is generated.

3. Test for six hexapyranoses

Our efforts have traditionally been directed towards the study of carbohydrate crystal structures. In our first attempt at crystal structure prediction we studied six hexapyranoses with the *GROMOS* force field (van Eijck *et al.*, 1995). In four cases the experimental structure had a satisfactorily low energy, but in the other two the energy was up to 24 kJ mol^{-1} higher than

Table 1
Comparison of grid search and random search for six hexapyranoses.

Rankings R and energy differences ΔE (kJ mol^{-1}) refer to a search in the experimental space group ($P2_12_12_1$) for three conformations. In the random search the experimental structure was found N(exp) times.

	*ADGALA*03	*GLUCSA*10	*ADTALO*10	*COKBIN*	*BDGLOS*01	*GLUCSE*01
R(grid)	6	9	10	23	51	35
R(random)	7	9	10	24	52	33
ΔE (grid)	5.6	5.5	2.5	5.2	13.0	10.2
ΔE (random)	5.6	5.5	2.4	4.7	13.1	10.2
N(exp)	9	4	13	12	4	8

Table 2
Numbers of ethanol structures in a few space groups.

G is the number of independent molecules allowed in the search. Only structures within 5 kJ mol^{-1} from the global minimum have been counted.

G	$P1$	$P\overline{1}$	Pc	$P2_1$	$P2_1/c$
1	0	0	3	3	30
2	29	249	299	275	
4	5668				

for the most favorable hypothetical structure. Later we found that a force field change as slight as replacing *GROMOS* by *UNITAT* could reduce this upper limit to 13 kJ mol^{-1} (van Eijck & Kroon, 1999).

In the present study we repeated the latter calculation using the random search method, in order to compare its performance with the previous results from the grid search. For each compound the three possible conformations of the exocyclic CH_2OH group were taken as building blocks to create structures in the common experimental space group $P2_12_12_1$. For each conformation there are five unknown hydroxyl dihedral angles, so the dimension of the problem is 14. The *UNITAT* force field was used with default settings for the random search.

The rankings and energy differences of each experimental structure with respect to the global minimum are given in Table 1. The structures are identified by their Refcodes in the Cambridge Structural Database (Allen & Kennard, 1993). For comparison, the corresponding grid search entries from Table 3 of our previous paper (van Eijck & Kroon, 1999) are also given; the correspondence is excellent except for *COKBIN*, where apparently the structure with lowest energy was missed in the random search. This is surprising, since most low-energy structures were encountered more than once; for instance, the experimental structure was found 12 times (Table 1). The total computing time (on a Silicon Graphics Origin 200 equipped with a 180 MHz R10000 processor) was 55 h for the grid search and 67 h for the random search. All in all, the grid search method appears to be slightly more effective for this problem.

4. Test for ethanol

4.1. Structure generation

Ethanol is a very suitable test molecule because it crystallizes with two independent molecules in the space group Pc. Moreover, the two molecules differ in conformation: one hydroxyl group is *trans*, the other is *gauche* (Jönsson, 1976). A high-pressure modification is also known, with one independent molecule in space group $P2_1/c$ (Allan & Clark, 1999). Most calculations are relatively fast and it is possible to verify the completeness of the search algorithm as well as the ability to find the experimental structure.

Structures with $G = 1$ were generated in space groups $P2_1/c$, $P\overline{1}$, $P2_12_12_1$, $P2_1$, $Pbca$, $C2/c$, $Pna2_1$, Cc, $Pca2_1$, $C2$, $P1$, $Pbcn$ and Pc. These are the most populated space groups for structures with $Z' = Z'' = 1$ (Mooij, van Eijck *et al.*, 1999). The random search as well as the grid search were applied. There is a clear distinction between the group of structures with lowest energy (all falling within a range of ~ 5 kJ mol^{-1}) and other, less favorable ones. In the former category we find hydrogen bonds (O$\cdots$O distances less than 2.80 Å and O$\cdots$H—O angles close to 180°), which are absent in the other structures. The settings of the grid search led to 137 hydrogen-bonded structures and only 21 others. Of the first group, the random search missed three structures, but found 28 new ones. Moreover, about 750 structures without hydrogen bonds were generated. Evidently, the standard settings of the random search allow better sampling, at the cost of producing many uninteresting solutions.

Structures with $G = 2$ were generated in space groups $P1$, $P\overline{1}$, Pc and $P2_1$. In each of these space groups 10 000 structures were generated, of which only a limited number of hydrogen-bonded structures remain at the end of the procedure (Table 2). The total number of structures in $P1$ with $G = 2$ was 134. In principle, this set should contain all structures from space groups $P\overline{1}$, Pc and $P2_1$ with $G = 1$ (55 in total). This was not the case: 26 were missing, three of which belonged to the class of hydrogen-bonded structures. The explanation is that crystal structures may be unstable if space-group symmetry is no longer enforced (Mooij *et al.*, 1998). After expansion to $P1$ and further energy minimization only five structures remained missing, all with relative energies of at least 14 kJ mol^{-1}.

In space group $P2_1/c$ with $G = 1$ we found 30 hydrogen-bonded structures, of which 26 were stable upon removal of the symmetry constraints. All of the latter should be present in space groups $P\overline{1}$, Pc and $P2_1$ with $G = 2$. It was found that there were only four, one and two missing structures, respectively.

This experience suggests that the structure generations with $G = 2$ are reasonably complete. This is confirmed by the observation that it is rare to find a hydrogen-bonded structure only once in the random search.

As an ultimate test, 20 000 structures were generated in space group $P1$ with $G = 4$. All previously found structures should be recovered, but this was hardly the case: for instance, only six out of the 26 expected $P2_1/c$ structures were present. This is not surprising since most structures are found only once, indicating that a continued search would produce many

Table 3
Hydrates of carbohydrates and polyalcohols.

Substance	Class	Space group	N(conf)	N(water)	Remarks
BAXNAP	p	$C2$	2	1	
DMGALP	p	$P2_12_12_1$	3	1	x
GEHDAY	i	$P1$	1	1	
*GLUCMH*11	p	$P2_1$	1	1	
MANHEP	p	$P2_12_12_1$	2	1	x
*MGALPY*01	p	$P2_12_12_1$	2	1	n
MYTOLD	i	$P2_1/a$	0	2	x
*RHAMAH*12	p	$P2_1$	0	1	n
SUNGUD	a	$C2/c$	6	1	
JIFDEH	p	$P4_12_12$	1	1/2	
MBDGPH	p	$P4_12_12$	2	1/2	

Class: p = pyranose, a = polyalcohol, i = inositol. N(conf) is the number of torsional degrees of freedom in a molecule, not including hydroxyl groups. N(water) is the number of water molecules in the asymmetric unit. Remarks: n = neutron diffraction; x = some hydrogen coordinates unknown.

more new structures. The dimension of the problem is 34 and, although ethanol is a very simple molecule, the computational demands are large.

4.2. Experimental structures

An altogether different question is whether or not the experimental structures are present in the list of hypothetical structures. Here not only the structure generation is involved, but the quality of the force field plays a major role.

It turned out that the experimental low-temperature structure (Pc, $Z'' = 2$) was indeed found, in Pc with $G = 2$ as well as in $P1$ with $G = 4$. The energy difference with the global minimum is $\Delta E = 1.1$ kJ mol^{-1} in the *UNITAT* force field. The ranking is difficult to assess. For instance, if only $G = 1$ structures in the 13 studied space groups were considered, we would find $R = 23$. However, if we include the five calculations with $G > 1$, there are already 104 structures with lower energy than the experimental one.

The calculations were continued with the *OPLS* force field. Here the energy difference was found to be ΔE = 1.6 kJ mol^{-1}. The two rankings, as defined above, are 13 and 32, respectively. We have seen earlier (van Eijck & Kroon, 1999) that this force field produces larger energy differences than *UNITAT*.

Recently, we developed a sophisticated *ab initio* intermolecular potential which has also been applied to ethanol (Mooij, van Eijck *et al.*, 1999). Here the energy difference was ΔE = 0.4 kJ mol^{-1} and a ranking $R = 4$ was found for 500 $G = 2$ structures, which were the best ones out of 1200 generated in 12 space groups with MSI's *Polymorph Predictor* (Molecular Simulations Inc., 1997; Verwer & Leusen, 1998) using the *DREIDING*2.21 force field and a special technique to treat the conformational freedom of the hydroxyl group.

In the *UNITAT* force field the global minimum occurred in space groups $P2_1$ ($G = 2$) and $P1$ ($G = 4$). Both structures were found to be equivalent and could be reduced to an elegant structure in space group $P4_3$ (or $P4_1$) with $Z'' = 1$. In the *OPLS* force field the lowest energy was found only once, for a $P\overline{1}$ structure with two molecules in the asymmetric unit. The three best structures in the *ab initio* potential all had $Z'' = 1$. Thus, we still cannot give an explanation for the observation that ethanol crystallizes with two independent molecules in the asymmetric unit.

Energy minimization of the high-pressure structure ($P2_1/c$, $Z'' = 1$) at zero pressure led, of course, to a considerably deformed structure. Interestingly, this structure was found to be even lower in energy than the low-temperature one: ΔE = 1.0 kJ mol^{-1} in *UNITAT* and 0.8 kJ mol^{-1} in *OPLS* as well as in the *ab initio* force field. It was present in the four relevant lists for $G \leq 2$. Contrary to expectations, this structure has no particularly small volume, and no improvement in ranking was obtained by calculating the list of enthalpy values at 3 GPa.

5. Calculations for α-D-mannose

Encouraged by these results, the structure prediction of α-D-mannopyranose was attempted. This substance crystallizes in space group $P2_12_12_1$ with two nonequivalent molecules (Longchambon *et al.*, 1976). It should be noted that the published structure must be in error; an acceptable geometry can be obtained if we assume that all hydrogen y-coordinates of the second molecule must take the opposite sign. After this correction the structure was well maintained upon energy minimization in the *UNITAT* force field. In the *OPLS* force field a few dihedral angles are rather far off, so at best only a deformed structure can be predicted.

The exocyclic dihedral angle O6—C6—C5—O5 is 67° for the first molecule and −64° for the second. This observation emphasizes that it should not be taken for granted that one molecular building block is sufficient. Rather, all possible combinations of the two dihedral angles should be taken as starting structures: three if only these two *gauche* forms are assumed, six if also the *trans* conformation is allowed. The latter conformation is usually supposed to be unfavourable owing to the parallel bonds C6—O6 and C4—O4 (Jeffrey, 1990).

However, it soon became clear that even with the correct pair of building blocks the experimental crystal structure was not going to be found. After generating over 60 000 structures in the experimental space group $P2_12_12_1$, the energy of the best structure was still 0.4 kJ mol^{-1} higher than that of the experimental one. Moreover, only few duplicate structures were present, which is an indication that the parameter space (which has 25 dimensions, since each molecule has five unknown hydroxyl dihedral angles) was not at all sampled adequately. An order-of-magnitude estimate can be made in the following way. In three groups of 20 000 structures the numbers of equivalent structures were 11, 21 and 18. The average is 17, obviously with a large uncertainty. Let the total number of feasible structures be N and let us assume (incorrectly) that each of them can be found in the random search with the same probability, $1/N$. In 20 000 structures there are 2×10^8 pairs, for each of which there is a probability of $1/N$ that the pair consists of two equivalent structures. Neglecting the probability of finding more than two equivalent structures, we find $2 \times 10^8/N = 17$. So it is seen that roughly N = 10 million trial random structures would be needed to find one

Table 4
Proposed fractional coordinates for missing or unreliable H atoms.

The hydrogen coordinates have been obtained by adding the O—H vectors calculated in the OPLS force field to the published oxygen coordinates.

Substance	Carrier	*x*	*y*	*z*
DMGALP	O1	0.287	0.443	0.374
	O3	0.747	0.492	0.217
	or	0.705	0.504	−0.131
	O6	0.235	0.681	0.033
	O7(H_2O)	0.291	0.297	0.035
		0.400	0.323	0.168
MANHEP	O1	0.132	0.229	0.299
	O2	−0.003	0.509	0.236
	O3	−0.020	0.800	0.434
	O4	0.199	0.948	0.293
	O5	0.199	1.228	0.750
	O6	0.230	0.843	0.961
	O8(H_2O)	0.182	0.531	−0.105
		0.160	0.383	0.027
MYTOLD	O1	0.230	0.500	0.110
	O2	0.115	0.425	−0.458
	O3	0.423	0.251	−0.458
	O4	0.308	0.166	−0.098
	O5	0.658	0.268	0.481
	O6	0.359	0.432	0.452
	O7(H_2O)	0.409	0.137	0.276
		0.386	0.046	0.237
	O8(H_2O)	0.150	0.079	−0.435
		0.319	0.080	−0.416

particular solution. To solve this problem computers have to become much faster or, preferably, a better algorithm should be found.

One may wonder how such a complex structure is actually realised during crystal formation. Obviously there exists a mechanism that is much more subtle than our brute-force search method. To compare the energetics with hypothetical $Z'' = 1$ structures, a standard grid search was carried out for three molecular conformations in space group $P2_12_12_1$. Here the best *UNITAT* structure was 0.7 kJ mol^{-1} higher in energy than the experimental one. It is tempting, but premature, to conclude that the structure with two independent molecules really has an energetic advantage. In the *OPLS* force field structures were encountered with energies up to 12.7 kJ mol^{-1} lower than for the (deformed) experimental one.

6. Calculations for hydrates

6.1. Modeling of crystal structures

In our previous study (van Eijck & Kroon, 1999) 11 hydrates had to be excluded. They are listed in Table 3. Of course, these structures exhibit complex hydrogen-bonded networks where the crystal water plays an essential role. A complication in their modeling is that H atoms are not always reliably determined by X-ray diffraction. In any case, three experimental structure determinations are incomplete: there are H atoms missing that are necessary to define the dihedral angles of the hydroxyl groups and the orientations of the water molecules. With the present version of *UPACK* it is possible to start the search with the known parameters and to try various possibilities for the missing ones. In this way a very limited number of possible structures was generated which are compatible with the observed unit cell data as well as with the experimental carbon and oxygen positions.

Table 5
Geometry shifts upon energy minimization of experimental structures for hydrates of carbohydrates and polyalcohols.

Each entry represents the root-mean-square value of the largest shifts encountered for each compound listed in Table 3. Unreasonably positioned non-hydroxyl H atoms have been disregarded. Missing hydrogen coordinates have been added according to Table 4.

	UNITAT	*OPLS*
Bond lengths without H (Å)	0.032	0.041
Bond lengths with H (Å)	0.24	0.22
Angles without H (°)	3.9	3.7
Angles with H (°)	11.1	11.5
Improper dihedral angles (°)	2.3	–
Torsions without H (°)	12.2	9.4
Torsions with H (°)	19.1	19.8
O···O distances < 3.5 Å	0.21	0.22
Cell axes (%)	4.4	3.9
Cell angles (°)	1.6	1.2
Molecular centers of gravity (Å)	0.14	0.17
Molecular Euler angles without H_2O (°)	4.8	5.4
Molecular Euler angles with H_2O (°)	9.8	9.6

The results are given in Table 4. For *DMGALP* (Dea *et al.*, 1974) two structures are acceptable, with an energy difference of 1 kJ mol^{-1} both in the *UNITAT* and the *OPLS* force fields. They differ only in the position of the H atom carried by O3. In the less favourable structure this atom is shared between two acceptors in an almost planar three-center hydrogen bond arrangement. One of the acceptors (O4) is not otherwise participating in hydrogen bonding. The next candidate structure, with a different orientation of the water molecule, has a relative energy of 5 kJ mol^{-1}. The unusual eclipsed conformations of the two methoxy groups are difficult to model: they are reproduced within 30° in *UNITAT* and within 15° in *OPLS*. The results for *MANHEP* (Taga & Osaki, 1969) are straightforward: the proposed structure has the lowest energy in both force fields and there is no alternative. For *MYTOLD* (Lomer *et al.*, 1963) there are two structures within 0.1 kJ mol^{-1} in the *UNITAT* force field, but one of them has by far (5 kJ mol^{-1}) the lowest energy in the *OPLS* force field. Here quite a few published hydrogen positions are unrealistic and all hydrogen coordinates were replaced by calculated ones.

To assess the performance of the two force fields, the shifts of all relevant geometry parameters upon energy minimization of the experimental crystal structures were calculated. The geometry parameters were classified in 13 types and for every type the largest shift was determined for each structure. The root-mean-square values of these maxima, averaged over the 11 hydrates, are reported in Table 5. Except for neutron diffraction studies, parameters without H are more trustworthy than those with H. The H atoms in the methoxy groups of *BAXNAP* and *GEHDAY* were disregarded because unreasonable valence angles occurred.

Table 6
Crystal structure generation for hydrates of carbohydrates and polyalcohols.

N(hydroxyl) is the number of rotatable hydroxyl groups. D is the dimension of the search problem. N(exp) is the number of times the experimental structure was found; numbers between parentheses refer to an extended calculation of 50 000 trials. Only the experimental conformation was used. See text for the investigated space groups.

Substance	Space group	N(hydroxyl)	D	N(exp)	ΔE	R	ΔE	R
DMGALP	$P2_12_12_1$	3	18	6	9.2	23	17.6	32
BAXNAP	$C2$	4	19	1 (3)	4.3	6	0.0	1
*MGALPY*01	$P2_12_12_1$	4	19	4	16.4	142	1.6	2
*RHAMAH*12	$P2_1$	4	19	4	2.7	10	5.1	3
GEHDAY	$P1$	5	20	42	1.3	3	6.3	8
*GLUCMH*11	$P2_1$	5	20	0 (10)	9.4	58	8.5	23
MANHEP	$P2_12_12_1$	6	21	0 (3)	3.8	9	0.0	1
SUNGUD	$C2/c$	7	23	0	13.4		22.1	
MYTOLD	$P2_1/a$	6	28	0	11.9		5.9	
Average					8.0			7.5

6.2. Structure generation

In any successful structure generation, the experimentally observed polymorph(s) should obviously be present in the list of hypothetical structures. These structures are built from free molecules that can usually occur in several plausible conformations. To limit the already considerable amount of computation to be performed, only the experimentally observed conformations were used; in a real structure prediction the work involved would be multiplied by a factor in the order of 3^N, where N is the number of conformational degrees of freedom given as N(conf) in Table 3. Therefore, if the intramolecular energy differences between the various conformations are not well modeled, this exercise will not show it.

Two of the compounds (*JIFDEH* and *MBDGPH*) crystallize in a tetragonal space group and cannot be handled (yet) owing to program limitations. Of the remaining nine compounds seven are optically pure. They were studied in the four most common enantiomorphous space groups: $P1$, $P2_1$, $P2_12_12_1$ and $C2$. Two compounds (*MYTOLD* and *SUNGUD*) consist of achiral molecules. They were studied in the same space groups, augmented with only the experimental one; the computational effort for considering all common space groups would be excessive. Therefore, no significant ranking can be given here.

The structure generation was carried out with the default number of 5000 trial structures for each space group. The results are given in Table 6, ordered with respect to the dimensionality of the search problem. In the first five cases the experimental structure was present in the list of hypothetical structures, so the structure generation can be said to fulfill its primary requirement. In the last four cases the experimental structure was not found. Evidently the practical limit is somewhere around 20 dimensions; the success for *GEHDAY* is due to its space group $P1$, where usually only very few low-energy structures can be constructed. A marginal case is *BAXNAP*, where the experimental structure was found only once. Statistically, this means just as little as the failure of finding the experimental structure for *GLUCMH* and *MANHEP*. To improve the statistics 50 000 additional trial structures were generated for these three compounds, with a corresponding increase in the rate of success (Table 6). For the two achiral compounds the situation is next to hopeless, especially for *MYTOLD* where two water molecules must be placed.

The performance of the force fields with respect to calculating a low energy for the experimental structure is worse than for the anhydrous pyranoses, where average ΔE values of 4.1 and 6.6 kJ mol^{-1} were found for *UNITAT* and *OPLS*, respectively (van Eijck & Kroon, 1999). Some improvement of the parameters for water in a crystalline environment might be possible, but that is outside the scope of the present work.

7. Discussion

The random search technique is comparable in efficiency to the grid search, and its implementation and use is much more straightforward. Generally the number of equivalent structures decreases with increasing energy. So the deepest minima in the potential energy landscape are not only deeper but also broader than others, and they attract more starting structures. This must be the explanation that the random search is often successful, even in cases where it is by no means complete. One could speculate that a broad minimum might provide a kinetic advantage in the crystallization process, but we found no outstandingly large numbers of equivalent solutions for the experimental structure.

Success or failure of the search depends to a large extent on the number of unknown parameters. For problems with a dimensionality of less than about 20, the experimental structure was generally present in the list generated with the simple *UNITAT* force field. However, the rankings were by no means good enough to allow a reliable *ab initio* structure prediction. Continuation of the energy minimization with the *OPLS* force field yielded only marginally better results.

Once again it is seen that, at present, crystal structure prediction can only be useful in practice if auxiliary experimental data are available. In this work we encountered three cases where only the hydrogen positions in the published crystal structure were unknown. Here it was possible to generate a complete list of acceptable hydrogen-bonded networks and to select the most probable one with good confidence. Generally, considerably less experimental information is available. If an indexed powder diffraction diagram is available, it is at present often possible to solve the structure without any recourse to energy calculations (Harris & Tremayne, 1996; Louër, 1998; Harris, 1999). Several publications have appeared where it was claimed that a crystal structure could have been determined by using unindexed powder diffraction data to discriminate between various structures proposed by force-field calculations (Karfunkel *et al.*, 1996; Schmidt & Dinnebier, 1999; Filippini *et al.*, 1999). Obviously, for applications such as this a force field should be targeted towards obtaining reliable lattice parameters and atomic coordinates rather than accurate energy differences.

In the absence of diffraction data, the number of non-equivalent molecules in a unit cell could be determined from solid-state NMR measurements. Whether or not solvent molecules are incorporated in a crystal may be determined beforehand, but crystal structure predictions could also be used to estimate the energy and density of unknown hydrates. Interestingly, to this end the actual crystal structure is not relevant since these properties are almost the same for all probable structures, and all one needs is a reliable force field.

We are grateful to Wijnand Mooij for providing details of his results on ethanol and for numerous helpful discussions.

References

Allan, D. R. & Clark, S. J. (1999). *Phys. Rev. B*, **60**, 6328–6334.

Allen, F. H. & Kennard, O. (1993). *Chem. Des. Autom. News*, **8**, 31–37.

Berendsen, H. J. C., Grigera, J. R. & Straatsma, T. P. (1987). *J. Phys. Chem.* **91**, 6269–6271.

Brock, C. P. & Duncan, L. L. (1994). *Chem. Mater.* **6**, 1307–1312.

Damm, W., Frontera, A., Tirado-Rives, J. & Jorgensen, W. L. (1997). *J. Comput. Chem.* **18**, 1955–1970.

Dang, L. X. & Petitt, B. M. (1987). *J. Phys. Chem.* **91**, 3349–3354.

Dea, I. C. M., Murray-Rust, P. & Scott, W. E. (1974). *J. Chem. Soc. Perkin Trans. II*, pp. 105–108.

Dzyabchenko, A. V. (1989). *Sov. Phys. Crystallogr.* **34**, 131–133.

Eijck, B. P. van & Kroon, J. (1997). *J. Comput. Chem.* **18**, 1036–1042.

Eijck, B. P. van & Kroon, J. (1999). *J. Comput. Chem.* **20**, 799–812.

Eijck, B. P. van, Mooij, W. T. M. & Kroon, J. (1995). *Acta Cryst.* B**51**, 99–103.

Eijck, B. P. van, Spek, A. L., Mooij, W. T. M. & Kroon, J. (1998). *Acta Cryst.* B**54**, 291–299.

Filippini, G., Gavezzotti, A. & Novoa, J. J. (1999). *Acta Cryst.* B**55**, 543–553.

Gao, D. & Williams, D. E. (1999). *Acta Cryst.* A**55**, 621–627.

Gavezzotti, A. & Filippini, G. (1994). *J. Phys. Chem.* **98**, 4831–4837.

Gdanitz, R. J. (1992). *Chem. Phys. Lett.* **190**, 391–396.

Gdanitz, R. J. (1997). *Theoretical Aspects and Computer Modeling of the Molecular Solid State*, edited by A. Gavezzotti, pp. 185–201. Chicester: John Wiley and Sons.

Gibson, K. D. & Scheraga, H. A. (1995). *J. Phys. Chem.* **99**, 3765–3773.

Gunsteren, W. F. van & Berendsen, H. J. C. (1987). *GROMOS*. University of Groningen, The Netherlands.

Gunsteren, W. F. van, Billeter, S. R., Eizing, A. A., Hünenberger, P. H., Krüger, P., Mark, A. E., Scott, W. R. P. & Tironi, I. G. (1996). *Biomolecular Simulation: The GROMOS96 Manual and User Guide.* Hochschulverlag, Zürich.

Harris, K. D. M. (1999). *J. Chin. Chem. Soc.* **46**, 23–34.

Harris, K. D. M. & Tremayne, M. (1996). *Chem. Mater.* **8**, 2554–2570.

Jeffrey, G. A. (1990). *Acta Cryst.* B**46**, 89–103.

Jönsson, P.-G. (1976). *Acta Cryst.* B**32**, 232–235.

Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. (1983). *J. Chem. Phys.* **79**, 926–935.

Jorgensen, W. L., Maxwell, D. S. & Tirado-Rives, J. (1996). *J. Am. Chem. Soc.* **118**, 11225–11236.

Karfunkel, H. R., Wu, Z. J., Burkhard, A., Rihs, G., Sinnreich, D., Buerger, H. M. & Stanek, J. (1996). *Acta Cryst.* B**52**, 555–561.

Lomer, T. R., Miller, A. & Beevers, C. A. (1963). *Acta Cryst.* **16**, 264–268.

Longchambon, F., Avenel, D. & Neuman, A. (1976). *Acta Cryst.* B**32**, 1822–1826.

Louër, D. (1998). *Acta Cryst.* A**54**, 922–933.

Molecular Simulations Inc. (1997). *Polymorph Predictor: Cerius2 User's Guide.* Molecular Simulations Inc., San Diego, California, USA.

Mooij, W. T. M., van Duijneveldt, F. B., van Duijneveldt-van de Rijdt, J. G. C. M. & van Eijck, B. P. (1999). *J. Phys. Chem. A*, **103**, 9872–9882.

Mooij, W. T. M., van Eijck, B. P. & Kroon, J. (1999). *J. Phys. Chem. A*, **103**, 9883–9890.

Mooij, W. T. M., van Eijck, B. P., Price, S. L., Verwer, P. & Kroon, J. (1998). *J. Comput. Chem.* **19**, 459–474.

Padmaja, N., Ramakumar, S. & Viswamitra, M. A. (1990). *Acta Cryst.* A**46**, 725–730.

Schmidt, M. U. & Dinnebier, R. E. (1999). *J. Appl. Cryst.* **32**, 178–186.

Shoda, T., Yamahara, K., Okazaki, K. & Williams, D. E. (1995). *J. Mol. Struct. (Theochem.)* **333**, 267–274.

Sluis, P. van der & Kroon, J. (1989). *J. Cryst. Growth*, **97**, 645–656.

Taga, T. & Osaki, K. (1969). *Tetrahedron Lett.* **51**, 4433–4434.

Tajima, N., Tanaka, T., Arikawa, T., Sakurai, T., Teramae, S. & Hirano, T. (1995). *Bull. Chem. Soc. Jpn*, **68**, 519–527.

Verwer, P. & Leusen, F. J. J. (1998). *Reviews in Computational Chemistry*, edited by K. B. Lipkowitz and D. B. Boyd, Vol. 12, pp. 327–365. New York: Wiley-VCH.

Williams, D. E. (1996). *Acta Cryst.* A**52**, 326–328.

Wilson, A. J. C. (1993). *Acta Cryst.* A**49**, 795–806.