Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 17 August 2000
Accepted 16 March 2001

Derivation of a scoring function for crystal structure prediction

Joannis Apostolakis,[a] Detlef Walter Maria Hofmann[b]* and Thomas Lengauer[a,b]

[a] Institute for Algorithms and Scientific Computing, German National Research Center for Information Technology (GMD–SCAI), Schloss, Birlinghoven, 53754 Sankt Augustin, Germany, and [b]Department of Computer Science, University of Bonn, Römerstrasse 164, 53117 Bonn, Germany. Correspondence e-mail: detlef.hofmann@gmd.de

The ever increasing number of experimentally resolved crystal structures supports the possibility of fully empirical crystal structure prediction for small organic molecules. Empirical methods promise to be significantly more efficient than methods that attempt to solve the same problem from first principles. However, the transformation from data to empirical knowledge and further to functional algorithms is not trivial and the usefulness of the result depends strongly on the quantity and the quality of the data. In this work, a simple scoring function is parameterized to discriminate between the correct structure and a set of decoys for a large number of different molecular systems. The method is fully automatic and has the advantage that the complete scoring function is parametrized at once, leading to a self-consistent set of parameters. The obtained scoring function is tested on an independent set of crystal structures taken from the $P1$ and $P\bar{1}$ space groups. With the trained scoring function and *FlexCryst*, a program for small-molecule crystal structure prediction, it is shown that approximately 73% of the 239 tested molecules in space group $P1$ are predicted correctly. For the more complex space group $P\bar{1}$, the success rate is 26%. Comparison with force-field potentials indicates the physical content of the obtained scoring function, a result of direct importance for protein threading where such database-based potentials are being applied.

1. Introduction

The problem of crystal structure prediction is notoriously difficult because, especially for small organic molecules, a large number of distinct structures can be found that show similar physical characteristics, such as energy or density (Gavezzotti, 1998). The actual task of structure prediction can be decomposed into two parts: the creation of a set of plausible solutions that contains at least one conformation which is reasonably close to the correct structure and the ranking of these solutions according to their quality, *i.e.* the probability that they correspond to the correct solution. The two parts are often referred to as sampling and scoring, respectively. Ideally, algorithms should perform a weighted sampling, spending more time searching the parts of conformational space close to the correct solution, thus simultaneously addressing both aspects of the problem. However, this suggests that optimal search algorithms already contain, at least in a rough and approximate manner, some knowledge of the form of the solution. This self-consistency problem can be addressed either by iteration or by use of heuristics, simple rules guiding the search, based on empirical knowledge about the general features of the solutions in similar problems. Certain iterative algorithms, such as simulated annealing, can guarantee convergence, at least in theory, however, this is usually of little practical relevance, since the number of iterations is not strictly limited. Heuristic approaches, when applicable, tend to be very efficient, but can break down for special cases. Efficient heuristic algorithms for sampling the crystal structures of rigid molecules have been developed and tested previously (Hofmann & Lengauer, 1997; Gavezzotti, 1991, 1996).

The present work concentrates on the ranking problem. By making the assumption that the crystal structure has the lowest (free) energy, the energy difference between the experimental crystal structure and any other structure (which we will refer to as a decoy) of the same molecule must be positive. Using this information, one would ideally parametrize a scoring function that yields the correct sign for energy differences. The huge amount of data available on high-resolution experimental structures may be used to obtain such empirical scoring functions. Such procedures have been applied in the field of protein structure prediction with some success (Maiorov & Crippen, 1992; Thomas & Dill, 1999; Zien *et al.*, 2000). Special care has been taken to minimize the amount of prior information used to formulate the scoring function. The scoring function is assumed to be a discrete atomic pair potential with a finite cut-off. No other assumptions concerning its functional form are made.

The quality of the scoring function depends on the method of derivation, the functional form of the class of functions from which the scoring function is chosen, and the quality and quantity of the data. In this work, we describe the derivation of a simple scoring function for the empirical crystal structure prediction of small molecules. The procedure consists of the following three steps: cleaning up of the database, training of the scoring function and validation. Cleaning up the database was found to be important for the quality of the training, since wrong structures often tend to influence learning in parts that are scarcely sampled.

2. Clean up of the database

It is estimated that in some space groups up to 10% of the structures stored in molecular databases may be wrong, *e.g.* a study on the space group *Cc* in the CSD found approximately one tenth of the structures to be incorrect (Marsh, 1997). However, for empirical predictions one needs highly accurate data sets. For our purposes, we used the Cambridge Structural Database System (CSD; Database V5.17, April 1999 Release; Allen & Kennard, 1993). The standard flags for selecting well defined structures were set (insist on coordinates, insist on a perfect match, insist no disorder, insist no polymers, insist R factor $\leq$ 10% and insist on no errors) and the structures were extracted in fdat format. We change in this application from the previously used format mol2 (Hofmann & Lengauer, 1997) to this format to avoid additional problems owing to the conversion routine. In particular, we observed that the conversion fails for nonstandard origins. From the retrieved structures, we removed several structures according to the following criteria:

(i) All structures with a different number of H atoms in the molecular formula than in the coordinate list were removed ($\sim$ 25%). Owing to the $\text{H}\cdots X$ interactions and the general unreliability of H positions, we did not add or correct the positions of the H atoms.

(ii) Only structures of the space groups $P\bar{1}$, $P2_1$, $P2_12_12_1$, $P2_1/c$ and $P1$ were considered. These space groups cover 69.4% of the known crystal structures. The other space groups cannot be handled by our program at present. The structures of the first four space groups were used for training of the scoring function, while space group $P1$ was selected as a test set.

(iii) Crystal structures with molecules occupying symmetric positions were removed, since the necessary algorithms for these special cases have not been implemented yet.

(iv) Each crystal structure was checked for close contacts. 117 structures with contacts at distances below a threshold of $r_{\min} = 1.3$ Å were removed. The very short distance of 1.3 Å occurs in hydrogen bonds. The hydrogen might be near the center between two O atoms (*e.g.* BAHOXH01, neutron study).

(v) Crystal structures with unusual cell volumes were removed. In a few cases, the assignment of structures to a subgroup is faulty. Such structures can be found by a comparison between an estimated cell volume and the given cell volume (Hofmann, 2001).

In total about three quarters of the database were excluded by the screening flags and the above criteria. Our tests for volume and close contacts indicate 69 structures in error (0.1%) and 87 doubtful structures (0.1%). (The refcodes of these structures are given in the supplementary material[1] and can be found at http://cartan.gmd.de/FlexC/home.html.). This lies far below the mentioned estimation of 10% for space group *Cc*. More and more sophisticated programs will be able to clean up the databases leading to a rapid decrease in the number of incorrect structures and increase the value of the databases.

3. Training of the scoring function

The score of the structure is dependent on the intermolecular contacts in the crystal. Each atom pair of the atom types i and j with a distance r in the kth interval $r_k \leq r < r_{k+1}$ contributes a weight $\epsilon_{i,j,k}$. To minimize computational effort first all contacts $x_{i,j,k}$ are summed up and the frequency is stored sequentially to a vector. This vector $\mathbf{X}$, containing all the intermolecular contacts, is called the structure vector, in analogy to similar applications in computational biology. Finally the frequency of a certain contact $x_{i,j,k}$ is multiplied by its corresponding weight $\epsilon_{i,j,k}$. Thus, the total score of a crystal structure corresponds to the scalar product between the weight vector ϵ and the structure vector $\mathbf{X}$.

$$E = \epsilon\mathbf{X} = \sum_i \sum_j \sum_k x_{i,j,k}\epsilon_{i,j,k}. \tag{1}$$

Distances between 1.3 and 5.5 Å are considered as intermolecular contacts. For the upper distance, we can observe that the weight is almost zero and does not contribute to the total score any more. The lower distance does not occur, because all structures with contacts below 1.3 Å are considered to be wrong. An analysis of the database suggests this threshold for intermolecular hydrogen bonds. The range between 1.3 and 5.5 Å is divided into 30 equidistant intervals k with a width of 0.14 Å.

$$k = \lfloor (r - 1.3\ \text{Å})/30 \rfloor + 1. \tag{2}$$

An example of a structure vector is shown in Table 1. The structure vector starts with two small hydrogen–hydrogen contacts at approximately 1.8 Å and ends with eight oxygen–oxygen contacts near the cut-off at approximately 5.4 Å. This typical organic compound contains only the common elements C, H, N and O.

The scoring function is trained to discriminate the structures found in the database from slightly distorted structures. The distorted structures are termed *decoys* (Maiorov &

[1] Supplementary data for this paper are available from the IUCr electronic archives (Reference: BK0088). Services for accessing these data are described at the back of the journal.

Table 1
Structure vector of BDORLA10.

k	*i*	*j*	Frequency	*k*	*i*	*j*	Frequency
4	1	1	2	30	1	8	24
9	1	1	2	19	6	6	4
11	1	1	2	20	6	6	8
13	1	1	2	21	6	6	4
14	1	1	6	22	6	6	8
15	1	1	6	23	6	6	16
16	1	1	2	24	6	6	6
17	1	1	6	25	6	6	2
18	1	1	4	26	6	6	8
19	1	1	6	27	6	6	10
20	1	1	4	28	6	6	14
21	1	1	8	29	6	6	6
22	1	1	8	30	6	6	8
23	1	1	8	20	6	7	2
24	1	1	4	21	6	7	2
25	1	1	12	22	6	7	4
26	1	1	10	23	6	7	4
27	1	1	8	24	6	7	4
28	1	1	12	25	6	7	4
29	1	1	16	26	6	7	12
30	1	1	10	28	6	7	4
13	1	6	2	29	6	7	2
15	1	6	6	30	6	7	4
16	1	6	12	14	6	8	2
17	1	6	14	15	6	8	4
18	1	6	8	16	6	8	4
19	1	6	12	17	6	8	6
20	1	6	10	18	6	8	14
21	1	6	10	19	6	8	8
22	1	6	10	20	6	8	4
23	1	6	10	21	6	8	10
24	1	6	28	22	6	8	10
25	1	6	12	23	6	8	12
26	1	6	12	24	6	8	26
27	1	6	20	25	6	8	12
28	1	6	30	26	6	8	20
29	1	6	28	27	6	8	10
30	1	6	24	28	6	8	14
18	1	7	4	29	6	8	18
19	1	7	4	30	6	8	12
21	1	7	4	28	7	7	2
23	1	7	2	19	7	8	2
24	1	7	2	20	7	8	4
26	1	7	10	21	7	8	2
27	1	7	4	23	7	8	4
28	1	7	6	24	7	8	4
29	1	7	4	26	7	8	4
6	1	8	2	27	7	8	2
10	1	8	8	28	7	8	6
12	1	8	10	29	7	8	2
13	1	8	4	30	7	8	6
14	1	8	4	11	8	8	2
15	1	8	2	13	8	8	2
16	1	8	2	15	8	8	4
17	1	8	4	16	8	8	4
18	1	8	6	17	8	8	2
19	1	8	10	20	8	8	6
20	1	8	8	21	8	8	6
21	1	8	12	22	8	8	4
22	1	8	4	23	8	8	8
23	1	8	16	24	8	8	4
24	1	8	10	25	8	8	2
25	1	8	24	26	8	8	6
26	1	8	12	28	8	8	2
27	1	8	14	29	8	8	6
28	1	8	18	30	8	8	8
29	1	8	20				

Crippen, 1992). A function that, for a set of crystal structures and corresponding decoys, yields lower scores for the crystal structures than for the corresponding decoys is termed *consistent* with this set. The assumption that such a function does indeed exist is based on the hypothesis that the crystal structures observed in the experiments attain the global minima of free energy. This hypothesis is widespread in chemistry, even if in some cases metastable structures can crystallize; a phenomenon referred to as polymorphism. If the scoring function closely approximates the free energy function for every system (besides the polymorphic) it will be able to discriminate between crystal structures and decoys. Therefore, the terms energy and score can be interchanged. The method is limited by the approximation of the atom-pair potential used here and in common force fields. Although a consistent function for all crystal structures may exist, in principle, it may be impossible to model with a pair potential.

For the generation of decoys, the cell vector **a** of the experimental structure is elongated or shortened by a random value up to ±1 Å. For each structure, 15 decoys are produced. The number of decoys per structure has been kept low, because further decoys are relatively similar and the information contained in the different decoys overlaps. The low number of decoys allows for the use of more crystal structures, which, as already mentioned, have a higher information content than decoys.

The aim of the training is to find a weight vector ϵ for which the following condition is met for every molecule in the training set.

$$E_n = \epsilon\mathbf{X}_n \leq E_{\rm in} = \epsilon\mathbf{X}_n, \qquad (3)$$

where $\mathbf{X}_n$ and $\mathbf{X}_{\rm in}$ are the conformations corresponding to the crystal structure and a decoy, respectively. We can also write:

$$0 \leq \epsilon\mathbf{X}_{\rm in} - \epsilon\mathbf{X}_n = \epsilon(\mathbf{X}_{\rm in} - \mathbf{X}_n). \qquad (4)$$

Therefore, the decoy conformations are represented relative to the corresponding crystal structure. In this formulation, the problem can be solved by any of a number of standard techniques. In this case, a simple descent procedure for minimizing the perceptron function J

$$J(\epsilon) = \sum_{\rm in} \max\big(0, \epsilon(\mathbf{X}_n - \mathbf{X}_{\rm in})\big) \qquad (5)$$

was used. J can be understood as a measure of the errors made by the scoring function defined by the parameter vector ϵ. The structure vectors are normalized to avoid the different data entering the function J with different weights.

Simple descent schemes are known to converge in a finite number of steps for consistent data, *i.e.* if a parameter vector ϵ exists that satisfies all inequalities required by the data (Duda & Hart, 1973). For inconsistent data, our method does not converge to a particular solution. However, after a number of iterations, solutions are obtained that are accurate enough for our purposes. In practice, the data are always inconsistent. This is because a low ratio of variables (parameter vector components) to the number of data is used to avoid extreme overfitting. Therefore, the procedure is discontinued after a constant number of iterations (1000) and the parameter vector yielding the lowest number of errors is kept as the solution vector. In order to obtain a complete potential, we have assigned a more or less arbitrary value of 10 to all bins that were not trained. This value is much higher than the weights obtained by training and corresponds to an infinite term which effectively screens out all conformation sampling distances that were not in the data set of crystal structures and were therefore not trained for. After the initial training, the program *FlexCryst* was used for structure prediction of the compounds in space group $P\bar{1}$. In a second step, crystal structures of space group $P\bar{1}$ were generated and scored with *FlexCryst*. All candidate structures with a lower score than the corresponding experimental structure (approximately 20 000) were collected and used again with the original training set for learning. Unlike the initial decoys, these structures differ significantly from the experimental structures. This type of decoy generation is significantly more time consuming, however, it leads to structures that are harder to learn than the naive decoys. The scores obtained after this round are presented and discussed shortly.

4. Structure generation

For validation of the scoring function, the atom-pair potentials obtained were applied to the crystal structure prediction of organic molecules. It is well known that crystal structures are highly sensitive to intermolecular potentials. Sometimes thousands of quite different local minima can fall within a narrow energy range (40 kJ mol^{-1}), as has been shown for monosaccharides (van Eijk *et al.*, 1995). This fact is reflected by the phenomenon of polymorphism (Threlfall, 1995) in nature, where many different crystal structures can often be found for one molecule.

Most present methods (Willock *et al.*, 1995; Holden *et al.*, 1993; Gavezzotti, 1996; Hofmann & Lengauer, 1997) assume that the conformation of the molecule is known or restricted to a few alternatives. During sampling, the conformation of the molecule remains fixed. This assumption is justified for pigments, which are rigid owing to large π systems, and for steroids owing to the high connectivity of the atoms. The methods of crystal structure prediction generate a list of possible candidate structures. These structures are ranked according to their score. A prediction is successful if the top-ranking structure is nearly identical to the experimentally determined structure. This is not always the case owing to problems in structure generation, inaccuracies of the scoring function and polymorphism. During structure generation, the variable space is incompletely scanned and generated structures can deviate somewhat from the minimum (especially in the case of discrete modeling). Inaccuracies of the scoring function are due to scarcity of data or the neglect of three-body interactions. In these cases, the experimental structure is not ranked best, but should be found among the top-ranking structures. Finally, polymorphic structures are metastable and do not correspond to the global (free) energy minimum. For the generation of crystal structures *FlexCryst* (Hofmann & Lengauer, 1997, 1998, 1999) was used. This algorithm follows the nuclei concept (Gavezzotti, 1991). In the simple case of space group $P1$, we start with a molecule, calculate energetically favorable chains, extend these chains to planes and finally search for proper three-dimensional structures. In general, all crystal structures can be generated in four steps or less, by successive addition of symmetry elements. Algorithms from the docking program *FlexX* (Rarey, Kramer *et al.*, 1996; Rarey, Wefing & Lengauer, 1999; Kramer *et al.*, 1999) are applied to automatically find putative symmetry operations. For the calculation of proper aggregates, we search for preferred positions of single atoms. The complete interaction between two molecules is given by the sum of the interactions of their atoms.

$$E_{IJ} = \sum_{i}^{n_I} \sum_{j}^{n_J} p(i, j, r_{ij}). \tag{6}$$

To find the preferred interaction points, the molecule is embedded into a grid of 0.5 Å^3. (This grid size results from a trade-off between the accuracy of the results and the computational efficiency.) At each grid point, a hypothetical C atom is placed and its interaction with the molecule is calculated. The best interaction points are retained for further processing.

In the next step, symmetry operations are determined that map one of these interaction points p to an atom c. Each symmetry operation is defined by a rotation matrix $\mathbf{W}$ and a translation vector $\mathbf{w}$. In general our condition can be written as:

$$\begin{pmatrix} W_{xx} & W_{xy} & W_{xz} \\ W_{yx} & W_{yy} & W_{yz} \\ W_{zx} & W_{zy} & W_{zz} \end{pmatrix} \begin{pmatrix} p_x \\ p_y \\ p_z \end{pmatrix} + \begin{pmatrix} w_x \\ w_y \\ w_z \end{pmatrix} = \begin{pmatrix} c_x \\ c_y \\ c_z \end{pmatrix}. \tag{7}$$

The rotation matrix $\mathbf{W}$ involves three variables, the three Euler angles, while the translation vector $\mathbf{w}$ involves three additional variables. The variables of the rotation matrix are scanned in steps of 5° and the translation vector is determined by the solution of the equation. In the special case of pure translation and inversion, the procedure simplifies. For a translation, the rotation matrix is equal to the unit matrix $\mathbf{1}$ and for an inversion the rotation matrix is $-\mathbf{1}$. For each of the determined symmetry operations, the energy of the corresponding dimer is calculated and the energetically favored symmetry operations are retained.

The combination of several of these symmetry operations defines a crystal structure of a certain space group. To generate a structure of $P1$, three translations are combined, for $P\bar{1}$ four inversions, for $P2_1$ one screw axis and two translations, for $P2_1/c$ one screw axis and two inversions, and for $P2_12_12_1$ two screw axes. The volume of the generated structures is calculated first. Only if the volume of the structure deviates from the estimated value by less than 25% is the structure retained for further processing. The retained structures are ranked and sorted according to the scoring function described above.

5. Results

The scoring function was trained on structures from the space groups $P\bar{1}$, $P2_1$, $P2_1/c$ and $P2_12_12_1$. The trained scoring function was tested on the space group $P1$. The generation of structures in space group $P1$ is simpler and faster than for other space groups. Thus, problems that may relate to crystal structure generation are avoided. The molecule conformation was extracted from the database and used as input to *FlexCryst*. The generated structures were ranked and compared to the experimental crystal structure. As a similarity measure, we used the largest distance between the unit-cell edges and the nearest grid point of the experimental structure.

In Fig. 1, the measure of similarity is illustrated for the simplified case of the plane group $p1$. The experimental structure is defined by the two vectors b_1 and b_2. These two vectors define the grid (circles). The calculated structure is defined by the unit-cell vectors b'_1 and b'_2. The similarity measure corresponds to the maximum of the distances r_1 and r_2 of the unit-cell vectors to the nearest grid point $s = \max(r_1, r_2)$. The origin $\mathbf{o}$ can be chosen arbitrarily in this space group, but must be taken into account for other space groups, *e.g.* for space group $P\bar{1}$ the origin must be an inversion center. For application of this measure of similarity, the molecular conformations have to be superimposed. In *FlexCryst*, the molecule remains fixed in space and the cell is constructed around the molecule, thus the molecules are superimposed, by default.

This similarity index is very fast and simple to calculate, but not as universally used as others (Lommerse *et al.*, 2000; Gavezzotti & Filippini, 1995) based on root-mean-square deviations. It is restricted to one specific conformation of a molecule in a particular space group. Other similarity measures are also able to recognize similar crystal structures if they are described in different space groups, for example, a small distorted crystal described in a subgroup with the undistorted crystal structure. The similarity index s and the root mean-square deviation r.m.s., considering only the molecule centers of the six closest molecules around the central molecule, is related by $s \simeq 1.32 \pm 0.17$r.m.s. in space group $P1$. In space group $P\bar{1}$, the relation is $s \simeq 2^{1/3}1.32$r.m.s. $= 1.64$r.m.s.. The factor $2^{1/3}$ takes into account that the unit cell contains two molecules. Taking the maximum deviation avoids the averaging effect inherent to the calculation of the r.m.s.

For our validation, two structures were considered similar if the measure of similarity $\max(r_1, r_2, r_3, o)$ was below 1.8 Å. This limit was chosen to be comparable to the application of statistically derived potentials (Hofmann & Lengauer, 1997) and is required by the mesh used there. In the discrete approach of structure generation, cell vectors must fall on mesh points. In a mesh of 1.0 Å, two neighboring mesh points have the distance in the diagonal of $(3 \times 1^2)^{1/2} \simeq 1.8$ Å. The finer mesh of 0.5 Å, applied in this work, suggests a limit of $(3 \times 0.5^2)^{1/2} \simeq 0.86$ Å. Most of the similar structures fulfil this requirement also, but for the sake of comparison with the earlier study the larger boundary is applied.

The validation procedure was applied to the space groups $P1$ and $P\bar{1}$. $P1$ is the most simple group for the problem of structure generation, because this group has only nine degrees of freedom (three angles, the length of the unit-cell vectors and the orientation angles of the molecule in the cell). Therefore, this group allows a fast test of the scoring function to a large number of molecules. The generation of structures in this space group is nearly always successfull (85.5%) and the scoring function can be tested best. In a second test we validated the algorithm for space group $P\bar{1}$. $P\bar{1}$ has 12 degrees of freedom (in addition, the origin of the unit cell has to be determined), which is the highest number of degrees of freedom among all the space groups. For this space group, the problems of structure generation and scoring function interfere. In this sense, the first problem refers to the test of the scoring function and the second test on $P\bar{1}$ the problem of crystal structure prediction.

All 239 structures of space group $P1$ in the database fulfilling the screening conditions has been collected. In 174 cases (72.8%), the structure ranked best was similar to the experimental structure and the prediction was considered successful. The more stringent cut-off of 0.86 Å reveals 144 cases (60.3%) as correct. The experimental structure was found to have a lower energy than all generated structures in 138 cases (57.7%). Fig. 2 shows the overall distribution of the similar structures. In 205 of 239 cases (85.8%), a structure that

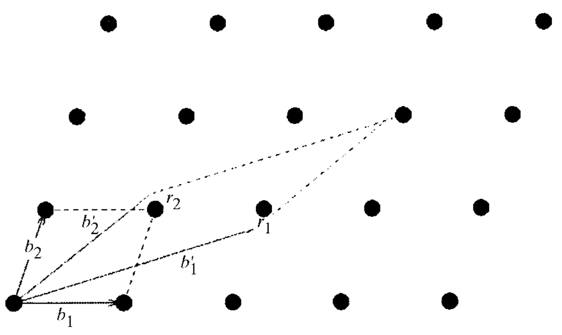


Figure 1
Similarity index for the plane group $p1$.

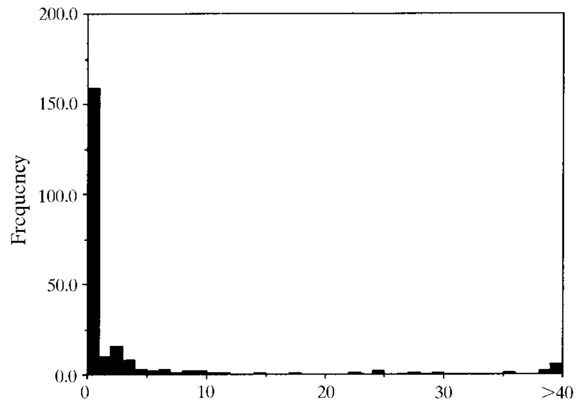


Figure 2
Histogram of the rank of structures similar to the experimental structure in space group $P1$.

is similar to the experiment is among the possible candidates. In three cases, the ranks are above 40 and the prediction cannot be considered successful. The results compare favorably with results published earlier (Hofmann & Lengauer, 1997), which were based on statistical potentials. There the success rate was 52.7% for a set of structures of space group $P1$ restricted to the most common atoms in organic chemistry.

In the second test, we applied the sampling algorithm and the scoring function to 54 structures of the space group $P\bar{1}$, not contained in the training set. Applying the smaller threshold of 0.86 Å structure prediction succeeded in five cases (9%) and a similar structure was ranked first. With application of the former threshold of 1.8 Å, the prediction was successful in 14 cases (25.9%). This compares favorably with previously published success rates of 13.6% achieved by statistically derived potentials. Among the generated structures, a similar structure was found in 23 cases (42.6%). Further improvement of the structure generation can be achieved by retaining more nuclei during the successive construction of crystal structures, but the success rate competes with the calculation time of this step. The scoring was correct in 35 cases (64.8%), in the sense that the experimental structure was assigned the lowest energy and is comparable to the corresponding value for $P1$. The distribution of the ranks is shown in Fig. 3. The calculation times are longer (9 min per molecule) and the success rate is lower for the space group $P\bar{1}$ because for $P\bar{1}$ 12 free variables have to be determined, three more than for the space group $P1$. For $P1$, the choice of the origin is unrestricted; in $P\bar{1}$, the origin has to coincide with an inversion center.

The newly derived potentials contain all interactions found in the database. This includes even such rare atoms as Np or Dy (of course, these potentials are poor due to the scarcity of the data) and allows for the handling of compounds containing these rare elements. Assuming the validity of pairwise additive interactions, the inclusion of compounds with these rare elements does not affect or improve the potentials for the other atoms.

Other methods of parametrization, *e.g.* exact *ab initio* calculations, may be impractical for time reasons. Moreover, conventional methods of parametrization fit only a few parameters at a time and are very difficult to automate. Statistically derived potentials (Sippl, 1993) suffer from a mixture of intermolecular and intramolecular correlations (Mitchell *et al.*, 1999). To separate these two influences, the bridge function must be known, which connects the atom-pair distribution function with the potential (Henderson & Sokolowski, 1996). In general, this function is not available and the separation of these two influences has to be made based on approximations (Bahar & Jernigan, 1997).

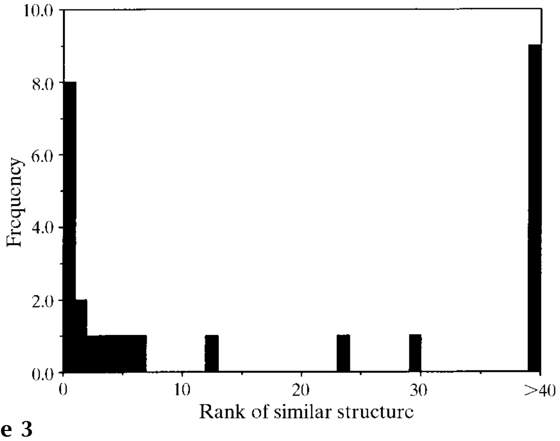


Figure 3
Histogram of the rank of structures similar to the experimental structure in space group $P\bar{1}$.

Table 2
Correct ranking by different methods.

Potential	$P\bar{1}$		$P1$	
	26 structures		169 structures	
Trained	18	69%	148	88%
Dreiding	25	96%	159	94%

6. Comparison with a force field

We compared the new potentials and the Dreiding force field (Mayo *et al.*, 1990) with respect to their performance in identifying experimental structures or structures resembling these. For comparison, we retained only structures where *SYBYL* was able to assign charges (169 structures in $P1$ and 26 in $P\bar{1}$). For each of these structures, we generated crystal structures with *FlexCryst* and selected the best 20 according to the score. To this set we added the experimental structure and assigned charges with *SYBYL* according the Gasteiger–Hückel method. For comparison we checked if the energy function scores a structure out of the 21 candidates first, which is similar to the experimental structure. The new score performs somewhat worse than the Dreiding force field in identifying reasonable structures. The trained scoring function ranked a structure best that is similar to the experimental one in 148 out of 169 cases (88%) in $P1$ and 18 out of 26 for $P\bar{1}$ (69%). The force field succeeded 159 (94%) and 25 (96%) times, respectively, in the same task (Table 2). This test somewhat favors the force field, because the generated candidates are selected according to their score from the trained scoring function. These structures are the likeliest candidates lower in the score than the experiment. The force field, instead, may correctly assign the order of these structures, but it cannot be excluded that other structures will be lower in energy than the selected structures and/or the experiment.

An analysis reveals that structures scored correctly by the force field, though incorrectly by the trained scoring function, often show ions, highly charged atoms or hydrogen bonds. This suggests that the present scoring function resembles mainly the dispersion and exchange repulsion energy, in common force fields often summed to the Lennard–Jones or Buckingham potential. This suggests that the present score can be improved by further differentiation of atom types to include information about the partial charges of the atoms and their local environment.

The new score correlates with the dispersion and exchange repulsion energy

$$E_{\rm vdW}[{\rm kJ}] \simeq 210 E_{\rm trained}. \tag{8}$$

The correlation coefficient between the derived score and the dispersion–repulsion energy is 0.83 (Fig. 4). Structures with positive energy or positive score were not used for the regression. For some structures, the exchange repulsion energy surmounts the dispersion energy and the sum of both becomes positive if a contact in the experimental structure is very short. The reason can be an unusual contact or a misplaced atom. The same can be observed with the new score. It becomes infinite (a high positive number, in practice) if the structure contains a contact which is not observed in the training set. The correlation with the sum of the total energy is poor. A few ionic crystals perturb the correlation and the new scoring function obviously does not describe ionic structures correctly.

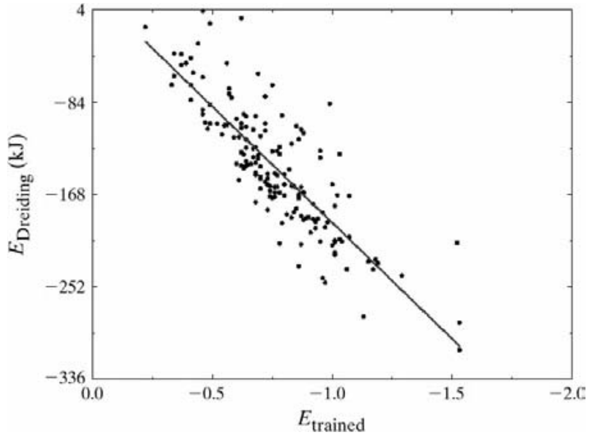
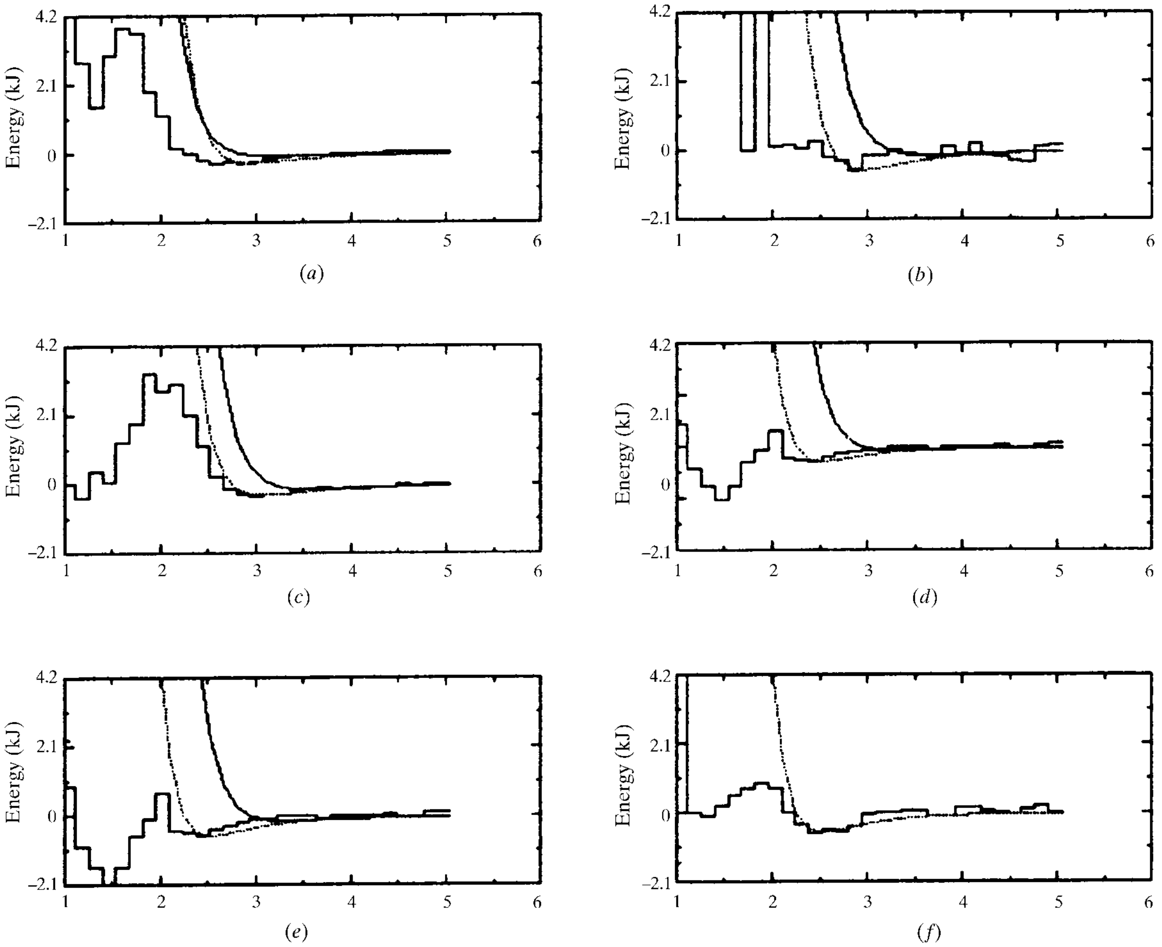


Figure 4
Comparison of score *versus* crystal energy.

The dependence of the score on distance can be understood as an empirical potential and as such can be compared after calibration with the above factor to the force-field pair potentials. In the following, we discuss the interactions of hydrogen with the first elements in detail. These potentials reflect the features of the trained potentials in dependence of the quality and quantity of data.

The curve for H$\cdots B$ shows the approximate form of a Lennard–Jones potential (Fig. 5*b*). The energy is large for small distances, reaches a minimum and then approaches zero for larger distances, however, the curve is noisy. This behavior



(*a*) (*b*) (*c*) (*d*) (*e*) (*f*)

Figure 5
Interaction potential for hydrogen with the common elements of the first row.

is observed for interactions, which occur rarely in the database. The fitted Lennard–Jones potential (dotted line) resembles the force field (dashed line), but the trained potential curve indicates a deeper well depth and a slightly shorter van der Waals radius.

The potentials of H with H (Fig. 5*a*), C (Fig. 5*c*), and F (Fig. 5*f*) qualitatively resemble van der Waals potentials, except for short distances; *e.g.* the H$\cdots$C potential shows a local minimum at 1.7 Å, *i.e.* at a distance smaller than the van der Waals minimum at 3.4 Å. Such local minima are caused by inaccurate structures with contacts at the corresponding distances. Some examples are the structures FEBGUO, PUDGOK and KUHKIH. The structure FEBGUO is not excluded in the screening, but a detailed analysis indicates one intermolecular C$\cdots$H distance of 1.47 Å. The structure is wrongly assigned to $P12_1/a1$ rather than $P12_1/c1$. Similar errors occur for KUHKIH and PUDGOK ($P12_1/c1$ rather than $P12_1/n1$). Such artefacts in the potentials can be avoided by further screening methods.

The interactions of H with N (Fig. 5*d*) and O (Fig. 5*e*) are particularly interesting. Besides the van der Waals minimum there is a second minimum at a shorter distance. This minimum corresponds to the hydrogen bond and is deeper than the van der Waals minimum, in agreement with the well known strength of hydrogen bonds. H atoms capable of forming hydrogen bonds have additional terms in the force field and our trained potentials strongly support this. The average potential, which mixes interactions such as C—H$\cdots$O and O—H$\cdots$O, can be separated by defining additional atom types, which take into account the chemical environment of the atom. As long as the number of defined atom types remains low, it can improve the scoring function. As the number of atom types becomes too large, the quality of the scoring function will deteriorate due to overfitting. At present, the atom types correspond to the elements. The different interactions in the above example are not contained in the O$\cdots$H potential, they are treated implicitly through the O$\cdots$O and C$\cdots$O potentials. The hydrogen bond is thus described by the sum of the O$\cdots$H and O$\cdots$O potentials.

The results demonstrate that in a real structure prediction problem one can perform a fast screening of the possible structures with *Flexcryst* and the new score, and postprocess the best structures with a force-field energy for even higher accuracy.

7. Conclusions

In this work, we determined a scoring function for use in structure prediction by using the structural information in the CSD database and information on incorrect structures obtained with the program *Flexcryst*. For this purpose, we cleaned up the database, applied a learning machine to a training set and validated the derived potentials on a test set of molecules not included in the training set.

We have used a simple method from machine learning to obtain an empirical scoring function for small-molecule crystal structure prediction. Unlike previous work (Maiorov & Crippen, 1992; Thomas & Dill, 1999), the assumptions made about the functional form of the scoring function and its relation to the data were minimal. The scoring function corresponds to a discrete pair potential with a cut-off at 5.4 Å. When the statistics are sufficient, the obtained potentials are smooth and show significant similarity to Lennard–Jones potentials. This is of particular interest because this information is not included in the learning process.

In contrast to other parametrizations, all parameters are optimized simultaneously. This avoids the problem of interfering parameters of step-by-step procedures, which require several iterations to become consistent (Foloppe & Mackerell, 2000). The scores obtained in this work can be simply improved by further screening and the use of new experimentally resolved structures. Given enough memory to store all the structures, the parameterization itself is very fast (a few hours on a single processor of a Sun Enterprise 4000, 250 MHz) and automatic.

Tests of the new scoring function on an independent set of structures showed improved predictive power over statistical potentials derived with the inverse Boltzmann method (success rate of 72% for the new method over 53% for inverse Boltzmann methods for structures from *P*1; Sippl, 1993; Hofmann & Lengauer, 1997). A simple comparison with a commercially available force field has indicated that our approach for obtaining a *de novo* energy function in an automated way is very promising.

D. W. M. Hofmann and T. Lengauer would like to thank S. Motherwell for the support in all questions concerning the database, M. Schmidt and P. Erk for discussions on the experimental and industrial aspects of crystal structure prediction, and the Deutsche Forschungsgemeinschaft (project No. LE 491/12–1) for financial support.

References

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

Bahar, I. & Jernigan, R. L. (1997). *J. Mol. Biol.* **266**, 195–214.

Duda, R. O. & Hart, P. E. (1973). *Pattern Classification and Scene Analysis.* New York: John Wiley and Sons.

Eijck, B. van, Mooij, W, & Kroon, K. (1995). *Acta Cryst.* B**51**, 99.

Foloppe, N. & Mackerell, A. D. (2000). *J. Comput. Chem.* **21**, 86–104.

Gavezzotti, A. (1991). *J. Am. Chem. Soc.* **113**, 4622–4629.

Gavezzotti, A. (1996). *Acta Cryst.* B**52**, 201–208.

Gavezzotti, A. (1998). *Cryst. Rev.* **7**, 5–121.

Gavezzotti, A. & Filippini, G. (1995). *J. Am. Chem. Soc.* **117**, 12299–12305.

Henderson, D. & Sokolowski, S. (1996). *J. Chem. Phys.* **104**, 2971–2975.

Hofmann, D. (2001). Submitted.

Hofmann, D. & Lengauer, T. (1997). *Acta Cryst.* A**53**, 225–235.

Hofmann, D. & Lengauer, T. (1998). *J. Mol. Mod.* **4**, 132–144.

Hofmann, D. & Lengauer, T. (1999). *J. Mol. Chem. (Theochem.)* **474**, 13–23.

Holden, J. R., Du, Z. Y. & Ammon, H. L. (1993). *J. Comput. Chem.* **14**, 422–437.

Kramer, B., Rarey, M. & Lengauer, T. (1999). *Proteins*, **37**, 228–241.

Lommerse, J. P. M., Motherwell, W. D. S., Ammon, H. L., Dunitz, J. D., Gavezzotti, G., Hofmann, D. W. M., Leusen, F. J. J., Mooij, W. T. M., Price, S. L., Schweizer, B., Schmidt, M. U., van Eijck, B. P., Verwer, P. & Williams, D. E. (2000). *Acta Cryst.* B**56**, 697–714.

Maiorov, V. N. & Crippen, G. M. (1992). *J. Mol. Biol.* **227**, 876–888.

Marsh, R. E. (1997). *Acta Cryst.* B**53**, 317–322.

Mayo, S., Olafson, B. D. & Goddard, W. A. III (1990). *J. Chem. Phys.* **94**, 8897–8909.

Mitchell, J. B. O., Laskowski, R. A., Alex, A. & Thornton, J. M. (1999). *J. Comput. Chem.* **20**, 1165–1176.

Rarey, M., Kramer, B., Lengauer, T. & Klebe, G. (1996). *J. Mol. Biol.* **261**, 470–489.

Rarey, M., Wefing, S. & Lengauer, T. (1996). *J. Comput.-Aided Mol. Des.* **10**, 41–54.

Sippl, M. (1993). *J. Comput.-Aided Mol. Des.* **7**, 473–501.

Thomas, P. & Dill, K. (1999). *Proc. Natl Acad. Sci. USA*, **93**, 11628–11633.

Threlfall, T. (1995). *Analyst*, **120**, 2435–2460.

Willock, D., Price, S., Leslie, M. & Catlow, C. (1995). *J. Comput. Chem.* **16**, 628–647.

Zien, A., Zimmer, R. & Lengauer, T. (2000). *J. Comput. Biol.* **7**, 483–501.