Acta Crystallographica Section B
Structural Science
ISSN 0108-7681

Method for the computational comparison of crystal structures

E. L. Willighagen, R. Wehrens, P. Verwer, R. de Gelder and L. M. C. Buydens*

Institute for Molecules and Materials, Radboud University Nijmegen, Toernooiveld 1, NL-6525 ED Nijmegen, The Netherlands

Correspondence e-mail:
l.buydens@science.ru.nl

Received 20 April 2004
Accepted 4 November 2004

A new method for assessing the similarity of crystal structures is described. A similarity measure is important in classification and clustering problems in which the crystal structures are the source of information. Classification is particularly important for the understanding of properties of crystals, while clustering can be used as a data reduction step in polymorph prediction. The method described uses a radial distribution function that combines atomic coordinates with partial atomic charges. The descriptor is validated using experimental data from a classification study of clathrate structures of cephalosporins and data from a polymorph prediction run. In both cases, excellent results were obtained.

1. Introduction

Comparing crystal structures is important in both classification and clustering problems. Classification is important for the understanding of the relation between physical properties and the underlying structure of materials. The specific packing of molecules in a crystal directly influences the physical properties of compounds. As an example, in crystal engineering crystal packings are classified according to intermolecular interactions (Perlstein *et al.*, 1996; Moulton & Zaworotko, 2001; DeGelder *et al.*, 2001; Hollingsworth, 2002; Ilyushin *et al.*, 2002). A second application of the similarity measure is in the clustering stage of *ab initio* crystal structure prediction (Verwer & Leusen, 1998; Lommerse *et al.*, 2000; Motherwell *et al.*, 2002). In this process hundreds or thousands of different hypothetical crystal packings for the same molecule, called polymorphs, are generated. They need to be clustered to obtain representative subsets for which analysis and geometry optimization is feasible.

For the clustering and classification of crystal structures, a properly defined descriptor and a similarity function applied to this descriptor are both required. In the literature several requirements for both the descriptor of crystal structures and the similarity function have been described (Dzyabchenko, 1994; Andrews & Bernstein, 1995; Fábián & Kálmán, 1999). The most obvious requirement for a descriptor–similarity combination is that more dissimilar crystal structures result in larger dissimilarity values. Although this seems trivial, several well known descriptors do not generally satisfy this requirement (Dzyabchenko, 1994; Andrews & Bernstein, 1995; Van Eijk & Kroon, 1997; Fábián & Kálmán, 1999). Many descriptors require a choice of origin or some other setting. Among such descriptors is the combination of unit-cell parameters and fractional coordinates. A descriptor based on reduced unit-cell parameters can vary significantly with only minor lattice distortions (Andrews *et al.*, 1980; Andrews & Bernstein, 1988). While it is in some cases possible to adapt

doi:10.1107/S0108768104028344

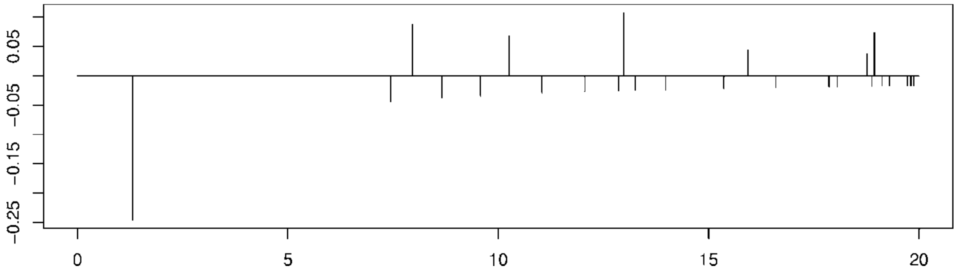


Figure 1
R_eDF for an artificial crystal structure with a positively and a negatively charged atom ($a = 7.97$, $b = 10.26$, $c = 18.77$ Å, $\alpha = \beta = \gamma = 90°$).

the similarity function to deal with such instabilities, we believe that this issue should be addressed by using a proper descriptor.

Recently, powder diffraction patterns have been used to compare the crystal structures of both simulated and experimental structures (Karfunkel *et al.*, 1993; De Gelder *et al.*, 2001). This descriptor does not suffer from the problems mentioned above and has an interpretable physical meaning. A potential disadvantage is that it is not always unique under certain conditions (Karfunkel *et al.*, 1999).

The current article investigates a new direct-space descriptor for comparing crystal structures. It is based on a radial distribution function and includes the electronic properties of the atoms. The descriptor will be described in detail in §2, which will also introduce the dissimilarity measure used to express the dissimilarities between structures using this descriptor.

The validation of the descriptor and the dissimilarity measure is carried out in two ways; first, by comparing the calculated dissimilarity values with empirical values and, secondly, by comparing a clustering created from the calculated dissimilarities with an empirical clustering. Empirical dissimilarity values, however, are not normally known on a continuous scale, but are expressed on a binary scale (identical or not) or are described textually using visual inspection. To our knowledge, there is no data set available from the literature in which the dissimilarities between a set of crystal structures are known on a continuous scale, which is needed for a quantitative validation of the descriptor and its dissimilarity measure. The two data sets for which empirical dissimilarity values and the clustering or classification are obtained are described in §3. These values are used to validate the application of the descriptor and dissimilarity measure.

Experimental details are given in §4 and §5 discusses the calculated dissimilarity values and clusterings for the two data sets.

2. The descriptor

To be able to compare crystal structures a descriptor is needed that represents the structure in mathematical form and a dissimilarity measure that expresses the differences between two crystal structures using the descriptor. The resulting dissimilarity values can then be used to cluster or classify the crystal structures by grouping together structures which have a low dissimilarity.

Crystal structures can be uniquely represented by a radial distribution function (RDF) describing the distribution of neighboring atoms around a central atom. Each neighboring atom gives rise to a peak in the function. RDFs are independent of cell choice, and can be physically interpreted. RDFs have been used to describe molecules with the goal of simulating IR spectra (Gasteiger *et al.*, 1996; Hemmer *et al.*, 1999), and have been used in the form of a radial distribution matrix for crystals (Karfunkel *et al.*, 1999). In the latter application each row in the distribution matrix is an RDF describing the interatomic distances for one atom-type pair. As such, the descriptor does not differentiate between, *e.g.* hydroxyl and carbonyl O atoms.

In our approach the RDF is adapted to include more specific information about the atoms. To do so, the RDF is weighted by the electrostatic interactions. To indicate the inclusion of electrostatic information in the descriptor, we will refer to this as the electronic radial distribution function, or R_eDF. The reason for including electrostatics is the assumption that these play a major role in crystal packing (Pauling & Delbrück, 1940; Moulton & Zaworotko, 2001; Desiraju, 1995). By including partial atomic charges, R_eDF focuses on atom groups with large partial charges, in particular functional groups, and differentiates between attractive interactions, between oppositely charged atoms and repulsive interactions.

An atomic R_eDF describes the distribution of coulombic interactions of one atom with the surrounding atoms; the R_eDF for the crystal structure is obtained by summing all the atomic R_eDFs for all N atoms in the asymmetric unit

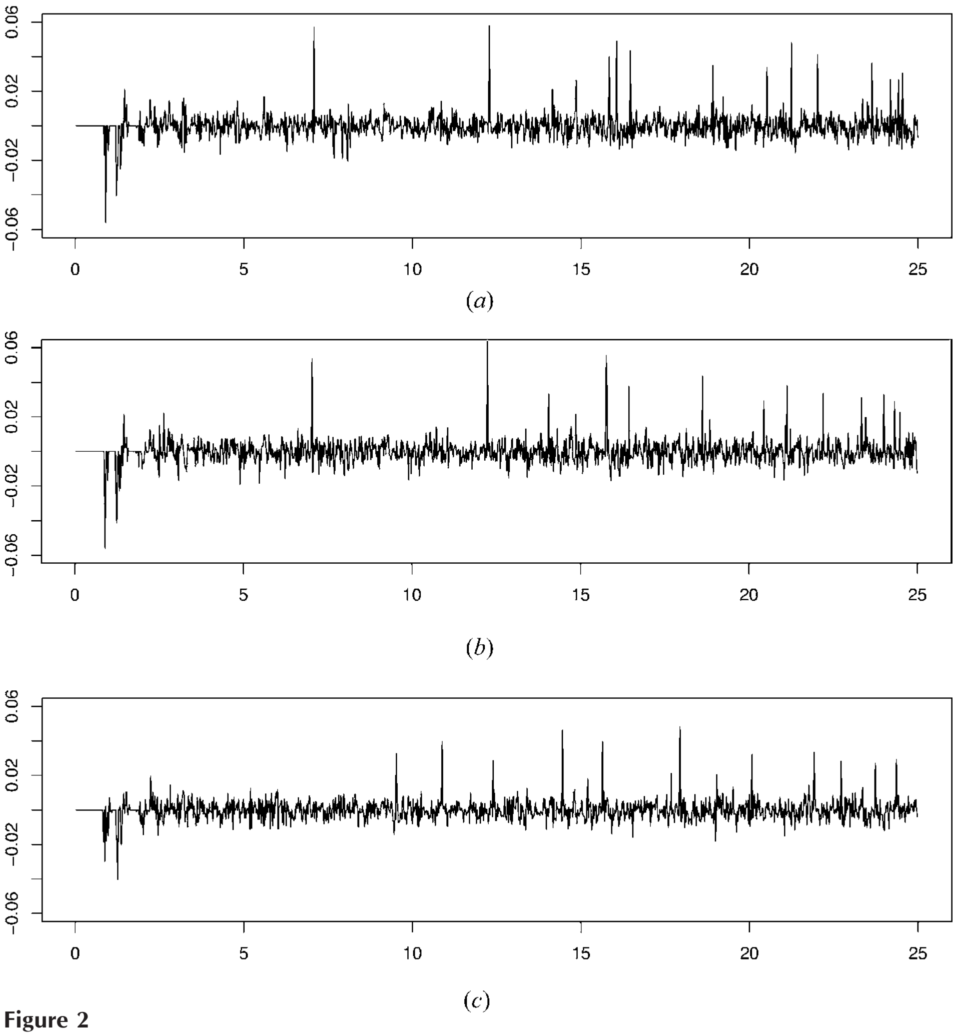
$$\mathrm{R}_e\mathrm{DF}(r) = \sum_{i=1}^{N}\sum_{j=1}^{M}\frac{q_iq_j}{N\cdot r_{i,j}}\,\delta(r - r_{i,j}), \qquad (1)$$

where M is the number of neighboring atoms within a radius r, q_i and q_j are partial atomic charges of the atoms i and j, and δ places the electrostatic interaction at the right distance by its definition $\delta(x) = 1$ if $x = 0$ and $\delta(x) = 0$ if $x \neq 0$. The function is scaled for the number of atoms in the asymmetric unit, N.

The R_eDF in (1) is a continuous function and is implemented as a discrete function with S intervals of size b, hereafter termed bins

$$\mathrm{R}_e\mathrm{DF}(s) = \sum_{i=1}^{N}\sum_{j=1}^{M}\left\{\frac{q_iq_j}{N\cdot r_{i,j}}D\big[(s+\tfrac{1}{2})b - r_{i,j}\big]\right\}, \qquad (2)$$

where s is the bin index and $s = 0..S$, $r_{i,j}$ is the distance between the two atoms i and j, q_i and q_j are partial atomic charges, and D is

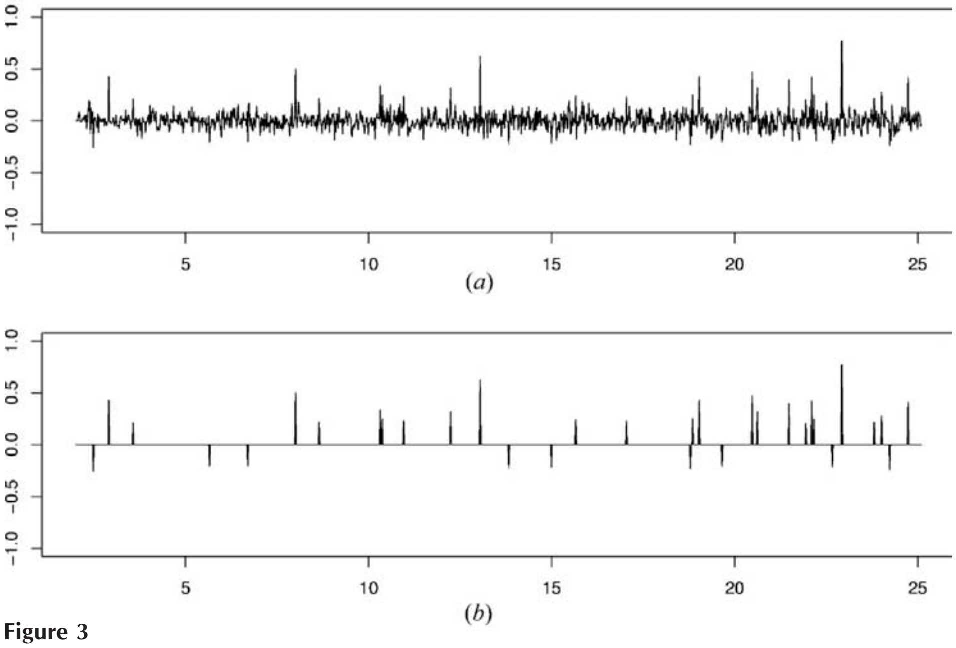


(*a*)

(*b*)

(*c*)

Figure 2
Sample R_eDFs for cephalosporins (*a*) A9, (*b*) A10 from the same class A and (*c*) N19 from a different class N.



(*a*)

(*b*)

Figure 3
R_eDF for one of the simulated estrone structures shown in (*a*), and the effect of cutting away peaks below 20% of the intensity of the highest peak in (*b*).

$$D(x) = \begin{cases} 1 & \text{if } |x| < \tfrac{1}{2}b \\ 0 & \text{if } |x| \geq \tfrac{1}{2}b. \end{cases} \qquad (3)$$

Fig. 1 shows the R_eDF for an artificial crystal with two atoms in the unit cell, a positively and a negatively charged atom ($a = 7.97$, $b = 10.26$, $c = 18.77$ Å, $\alpha = \beta = \gamma = 90^\circ$). The first, negative peak is the interaction between the two atoms at exactly the bonding distance. The other negative peaks are also peaks between two oppositely charged atoms. The overall decrease in intensities is caused by the $1/r$ term in the R_eDF equation. The first positive peak is related to the translation along the a axis, *i.e.* $\pm\mathbf{a}$, and the second peak to the translation along the b axis. The third peak is the translation in the direction $a \pm b$; for this orthogonal structure there are twice as many contributions to this peak as for the first two positive peaks, resulting in the higher intensity.

The R_eDFs of four experimental crystal structures, described in a later section, are given in Figs. 2 and 3. They show a few distinct high intensity peaks and many smaller peaks. The locations of these peaks are specific for the crystal packing: Figs. 2(*a*) and (*b*) show the R_eDFs of two cephalosporin structures from the same class, while (*c*) shows the R_eDF for a different packing.

Fig. 3(*a*) shows the function for a simulated estrone crystal structure; a similar pattern can be observed. Fig. 3(*b*) shows the effect of cutting away peaks with intensities lower than a specific threshold. It was found that the cut-off value must be around 20% of the highest peak. Cutting away the smaller peaks emphasizes the major features in the R_eDF and leads to better discrimination.

Owing to the nature of the R_eDF (Mayo *et al.*, 1990), one can expect positive contributions at those distances which match the translational symmetry in the crystal. However, since such contributions can be canceled out by other, negative contributions they do not

Table 1
Unit-cell parameters of the cephalosporin data set, grouped into seven clusters (A, B, C, D, E, F and N).

Cluster	*a*	*b*	*c*	*α*	*β*	*γ*
A	23.47	7.12	14.93	90.0	108.27	90.00
	23.42	6.97	15.00	90.0	110.41	90.00
	23.46	7.12	14.89	90.0	108.57	90.00
	23.41	7.11	14.81	90.0	108.15	90.00
	23.39	7.20	14.76	90.0	108.58	90.00
	23.02	7.15	14.55	90.0	104.64	90.00
	23.40	7.06	14.92	90.0	109.80	90.00
	23.43	7.11	14.88	90.0	108.19	90.00
	23.49	7.08	14.85	90.0	108.95	90.00
	23.45	7.03	14.84	90.0	110.55	90.00
B	7.11	21.72	30.96	90.0	90.00	90.00
	7.00	20.99	30.69	90.0	90.00	90.00
	7.11	21.86	32.31	90.0	90.00	90.00
	7.09	21.27	31.00	90.0	90.00	90.00
C	14.92	7.38	20.50	90.0	105.77	90.00
D	23.56	7.13	18.69	90.0	109.38	90.00
E	7.07	10.70	14.23	87.15	79.00	89.74
F	15.40	7.30	23.57	90.00	99.35	90.00
N	10.87	9.51	12.39	90.00	98.70	90.00
	10.91	9.41	12.20	90.00	98.53	90.00

always show up in the R$_e$DF. Moreover, peaks not related to translational symmetry are especially interesting, because they provide information additional to periodicity.

Fig. 4 shows the R$_e$DF for cephalosporin structure A1 (top) and the locations of peaks caused by the translational symmetry. Clearly, a significant number of peaks are not caused by translational symmetry and contain additional structural information. Each peak consists of many contributing atom pairs resulting in a netto positive (repulsive) or negative (attractive) peak in the function.

Dissimilarities between crystal structures are represented by the difference between the two corresponding R$_e$DFs. For this, a weighted cross correlation (WCC) is used (De Gelder *et al.*, 2001), which is applied to the high intensity peaks of the R$_e$DF.

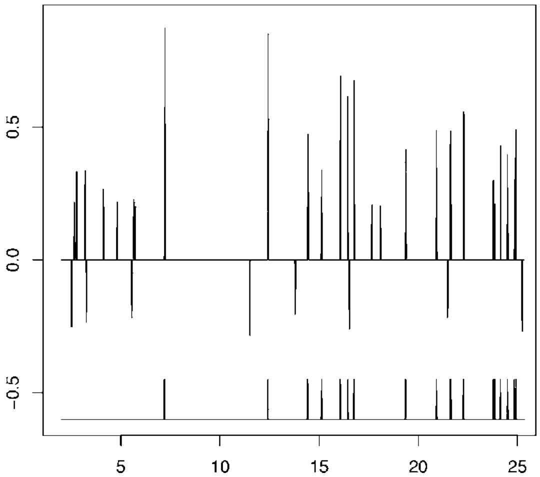


Figure 4
This figure shows that the nature of the peaks in the R$_e$DFs is not only describing the translation symmetry of the crystal structure: the top function is the R$_e$DF of cephalosporin A1 after applying the peak selection. The bottom black line shows the locations originating from translation symmetry.

Table 2
Visual criteria used to classify the 1128 dissimilarities between the 48 crystals.

The qualifier ++ represents *almost identical*, + *similar*, − *dissimilar*. See text for a more specified definition.

Dissimilarity class	Number of pairs	Unit-cell parameters	Placement in cell	Orientation in cell
Identical	8	++	++	++
Similar	21	+	+	+
Dissimilar	1099	−	−	−

3. Data

Two data sets are used in this article to show the application of the descriptor. The first data set contains the experimental crystal structures of the inclusion complexes of cephalosporins. These 20 structures are classified into seven classes, but there is no information about the similarity between structures other than belonging or not belonging to the same class. To our knowledge, there is no data set available from the literature in which the dissimilarities between all crystal structures are known on a continuous scale, which would be ideal to validate the proposed descriptor and its dissimilarity measure. The second data set contains simulated polymorphs of estrone, for which detailed information is available about the dissimilarities between the structures, as explained below. The 48 structures in this data set are classified into 25 classes based on visual inspection, as described below.

3.1. Cephalosporin data set

The cephalosporin data set consists of 20 clathrate structures of cephalosporins (Kemperman *et al.*, 2000; De Gelder *et al.*, 2001). The 20 compounds were classified into seven isomorphic classes based on their crystal form: A, B, C, D, E, F and N. Class A has ten structures, all in the $C2$ space group. Class B has four structures in the $P2_12_12_1$ space group. Classes C, D, E and F all have one structure, and have space groups $P2_1$, $C2$, $P1$, and $P2_1$ respectively. Class N has two structures which both have the $P2_1$ space group. A brief overview of the unit-cell parameters of this data set is given in Table 1. Further details on these structures can be found in Kemperman *et al.* (2000) and De Gelder *et al.* (2001).

For a set of 20 crystal structures, there are 190 unique pairs of structures $[\frac{1}{2} * n * (n-1) = \frac{1}{2} * 20 * 19]$. The dissimilarity associated with each pair is unknown. However, it is known whether the pair is a within-cluster or a between-cluster pair, *i.e.* the dissimilarity of a pair of structures from the same class is marked as within-cluster, and for a pair of structures that do not belong to the same structure class it is marked as between-cluster.

Table 3
An overview of the estrone dataset showing the lengths *a*, *b* and *c* of the orthogonal unit-cell axes of the 48 structures, and the direction (*a*, *b* or *c* direction) and form of the hydrogen-bond chain (linear or zigzag).

Cluster	*a*	*b*	*c*	Direction of hydrogen-bond chain	Form of chain
A	7.063	11.530	19.481	*c*	Zigzagged
	7.971	10.262	18.772	*c*	Zigzagged
	8.427	10.958	17.286	*c*	Zigzagged
B	7.742	9.110	23.163	*c*	Linear
	7.658	9.188	22.419	*c*	Linear
	7.691	8.865	23.262	*c*	Linear
	7.706	8.910	24.038	*c*	Linear
C	6.457	12.421	19.679	*c*	Zigzagged
	6.678	13.305	18.966	*c*	Zigzagged
D	5.946	12.940	20.499	*c*	Zigzagged
	6.332	13.037	19.066	*c*	Zigzagged
E	8.687	10.067	18.082	*b*	Linear
	9.381	9.432	18.147	*b*	Linear
F	8.742	13.276	13.617	*b*	Linear
	9.649	12.309	13.279	*c*	Linear
G	7.456	14.441	15.324	*b*	Zigzagged
	8.281	13.521	15.177	*c*	Zigzagged
	9.025	11.533	15.931	*c*	Zigzagged
H	8.507	10.087	18.943	*b*	Linear
	9.331	9.410	17.887	*b*	Linear
I	6.903	9.589	23.719	*c*	Zigzagged
J	7.980	10.539	18.293	*b*	Linear
K	9.868	12.127	13.063	*c*	Linear
L	7.969	10.597	18.254	*b*	Linear
	7.969	10.597	18.255	*b*	Linear
M	7.968	13.259	14.687	*b*	Linear
	7.968	13.259	14.688	*b*	Linear
N	7.581	10.387	19.439	*b*	Linear
	7.581	10.387	19.439	*b*	linear
O	9.306	9.445	18.111	*b*	Linear
	9.306	9.445	18.111	*b*	Linear
P	7.733	9.526	21.196	*b*	Zigzagged
	7.733	9.526	21.196	*b*	Zigzagged
Q	7.500	12.300	17.088	*c*	Linear
	7.500	12.300	17.088	*c*	Linear
R	8.560	13.268	14.186	*c*	Linear
	8.560	13.268	14.186	*c*	Linear
S	7.829	13.975	15.743	*b*	Zigzagged
	7.829	13.975	15.743	*b*	Zigzagged
T	7.135	10.876	20.431	*c*	Zigzagged
	7.442	10.043	22.177	*c*	Zigzagged
U	9.183	13.104	13.198	*c*	Linear
	9.750	12.673	13.044	*c*	Linear
V	7.235	11.743	19.066	*c*	Zigzagged
	7.293	10.763	20.544	*c*	Zigzagged
W	7.772	9.123	23.078	*c*	Zigzagged
X	7.302	13.266	16.788	*b*	Linear
Y	9.228	13.127	13.254	*b*	Linear

3.2. Estrone data set

The second data set consists of 48 simulated crystal structures of the estrone steroid, which has three known naturally occurring polymorphs (CSD refcodes: ESTRON10, ESTRON11 and ESTRON12; Busetta *et al.*, 1973). Two thousand polymorphic structures were generated using the Polymorph Predictor module in *Cerius*^{2} (Verwer & Leusen, 1998; Molecular Simulations Inc., 1997). The method used by this program consists of a generation step where random crystal structures are generated. After the removal of duplicates, the energies of the remaining 1278 structures were minimized using a force field. For this data set, the estrone molecule was kept rigid and the $P2_12_12_1$ space group symmetry was imposed during the initial generation. The energy minimization was carried out with the DREIDING-2.21 force field using Ewald summation to calculate the van der Waals and Coulomb interactions. Electrostatic potential (ESP)-derived atomic charges for estrone were calculated using *GAUSSIAN*94 (Frisch *et al.*, 2001) with the HF/6-31G* basis set.

From the 1278 structures, a set of 48 structures were selected in the low-energy region which represent the crystal structures that might be found in nature. The densities of these simulated structures are in the range 1.043–1.173 g cm^{-3}, while the experimental structures have densities around 1.2 g cm^{-3}. It is common for predicted crystal structures to have different densities, due to the force field used. The energies are in the region of 21.06 kJ mol^{-1}.

To classify the crystal structures, the 1128 pairwise comparisons between the 48 estrone structures ($\frac{1}{2} * 48 * 47$) were manually grouped into three dissimilarity classes by visual inspection. Classification of the pairwise dissimilarities was carried out by trying to overlap the crystal structures. However, an attempt has been made to quantify the differences in terms of packing parameters. These properties were taken into account during the clustering: cell parameters, placement in the cell and orientation in the cell (see Table 2). The cell parameters were compared and show big differences (for −), small differences (for +) or hardly any differences (for ++). The placement in the cell is compared visually: ++ indicates that the four molecules in the unit cell can be placed on top of each other perfectly within 0.01 Å, + indicates that they fit well and − means that they cannot be aligned simultaneously. Similarly, for the rotations around the various axes, ++ indicates that the molecules in the two structures have an identical orientation, + indicates a rotation up to *ca* 10°. Larger rotations do not occur in the data sets, as the actual molecular packing becomes different. The number of dissimilarity classes is chosen to reflect the number of visually distinguishable dissimilarity types in the above analysis.

The first dissimilarity class is called *identical*, as the structures are *visually* identical. The second class is called *similar* and consists of pairs of crystal structures that show small displacements or small rotations of the molecules in the unit cell, but the location of the molecules in the cell and the cell parameters itself are similar. The third class is called *dissimilar* and consists of all dissimilarities not classified in the other two classes. No further distinction between dissimilarities can be made in this class. Note that the first two classes have far fewer structure pairs than the *dissimilar* class, which reflects the diversity of the data set.

Based on the visually determined dissimilarities, identical and similar crystal structures were grouped, leading to 25 true classes, labeled A to Y. Table 3 shows the members of each class. The diversity of the unit-cell axes between the structures is apparent from this table. The similarity within classes is mostly clear, for example in class A.

An additional analysis has been carried out to quantify the similarity of the structures within the classes: for all structures the hydrogen-bonding pattern was determined as described by two variables. As estrone has only one hydrogen-bond donor and only one acceptor, the bonding pattern can only exist in the form of chains. Thus, the axis along which the chain is directed is given, as well as the form of the chain: linear or zigzag. In all cases the structure pairs with *identical* and *similar* similarity values show an identical scheme of hydrogen-bond chains. The hydrogen-bonding patterns are given in Table 3 and support the clustering found by visual analysis of the structures.

4. Experimental

For both data sets the R$_e$DF was used with a bin size of 0.02 Å and in a domain of [2,25] Å. The bin size was chosen such that the high intensity peaks were clearly visible. Below 2 Å there is mostly intramolecular information, which does not describe crystal packing and is therefore not included in the chosen domain. The distance up to which the R$_e$DF is calculated, 25 Å, is found to be the smallest distance containing enough informative peaks and is used for both data sets. When calculating the dissimilarities between the R$_e$DFs with the WCC measure, a triangle is used of 0.6 Å, which is about half a bond length. Much larger and much smaller values showed worse clustering results.

The descriptor is validated for both data sets, by grouping all dissimilarities calculated with the descriptor into the dissimilarity classes, as defined earlier. The median, minimal and maximal dissimilarity values for the classes can be compared and ideally show distinct classes. The larger the overlap between two dissimilarity classes, the worse the descriptor. The better the trend in the calculated dissimilarity values, the better the descriptor.

In addition to this, the calculated dissimilarities are used to cluster the crystal structures into a dendrogram using hierarchical average-linkage clustering. The dendrogram can be cut at a height yielding a certain number of clusters. Cutting at a small height will give many clusters, while cutting at a large height will give only a few clusters. The height at which the dendrogram is cut is chosen to give the number of clusters that matches the number of classes defined for that data set.

Finally, the simulated estrone structures are matched against the experimentally determined ESTRON10 structure to find the structure with the same packing. This is done by calculating the R$_e$DF for the experimental and simulated structures and calculating the dissimilarity between ESTRON10 and all of the simulated structures. The structure with the smallest dissimilarity to ESTRON10 is identified to have the same packing.

The simulated structures are not matched against the ESTRON11 polymorph which also has $P2_12_12_1$ symmetry, because the hydroxyl group in ESTRON11 points in a different direction to that in the simulated structure, leading to a different packing. Neither were they matched against ESTRON12 which has a different space-group symmetry.

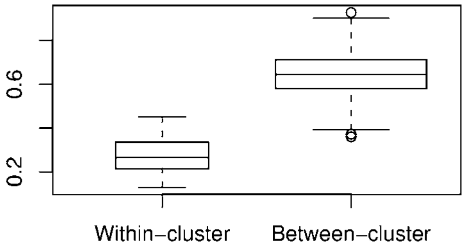


Figure 5
Box plot for dissimilarities between the two defined dissimilarity classes (within-cluster and between-cluster) calculated for the cephalosporin structures with the R$_e$DF.

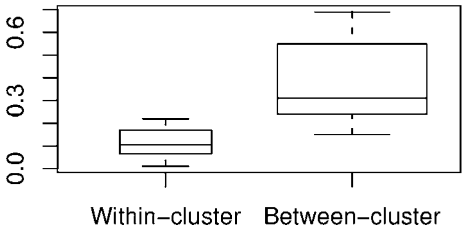


Figure 6
Box plot for dissimilarities between the two defined dissimilarity classes (within-cluster and between-cluster) calculated for the cephalosporin structures using powder diffraction data (see De Gelder *et al.*, 2001).

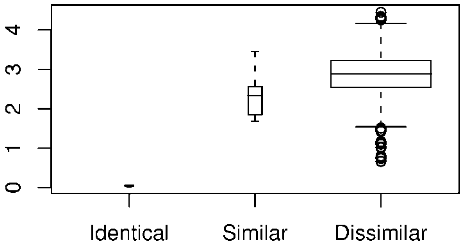


Figure 7
Box plot for dissimilarities between the estrone crystal structures grouped by the three dissimilarity classes as defined in Table 2, calculated with the R$_e$DF. The widths of the boxes are proportional to the number of objects in that class. The circles in this plot indicate dissimilarities that fall outside the fourth quantile of the distribution.

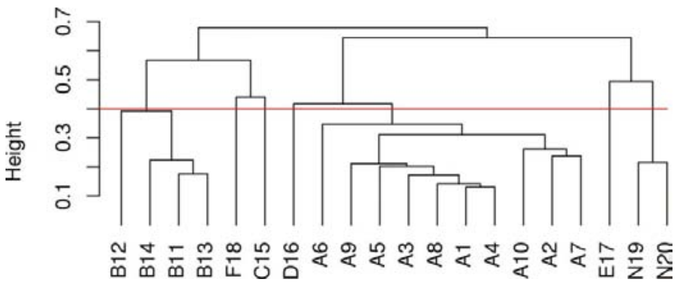


Figure 8
Dendrogram for the cephalosporin data set calculated with the optimized descriptor for the 20 structures with average linkages. The seven structure classes that are compared with the known classes (A, B, C, D, E, F, N) were determined by cutting the dendrogram at a height of 0.4.

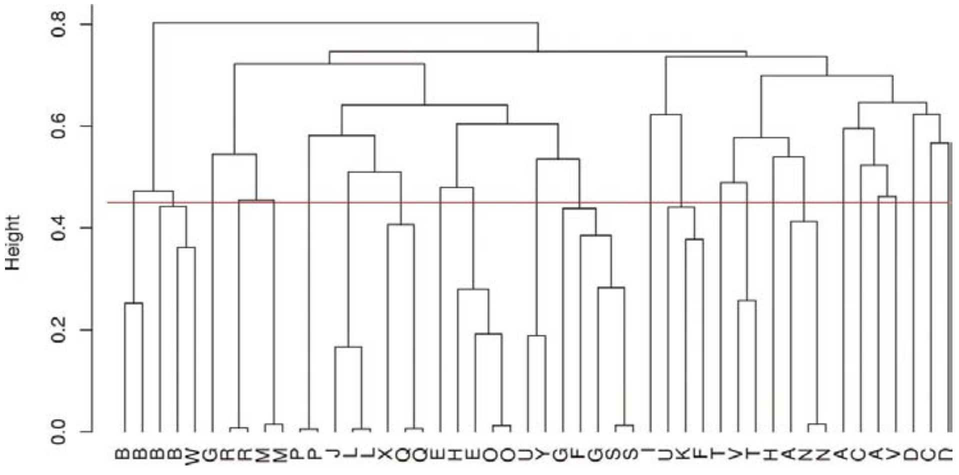


Figure 9
Dendrogram of the 48 structures in the estrone data set clustered with the average linkage using the dissimilarity values calculated with the optimized descriptor. The 25 clusters that are compared with the validation set were determined by cutting the dendrogram at a height of 0.44 (horizontal line). Object labels are taken from Table 3.

Neither experimental structure has a corresponding structure in the simulated data set.

The calculation of R_eDF descriptions for crystal structures and dissimilarity measures is implemented in C++. The clustering of structures based on their dissimilarity matrix is carried out in *R* (Gentleman & Ihaka, 1996) with the average linkage method. Calculations were performed on both Solaris and GNU/Linux systems.

5. Results

5.1. Dissimilarity classes

The descriptor is validated by calculating the dissimilarity values between all pairs of crystal structures. The dissimilarity values calculated for the cephalosporin data set are shown as box plots in Fig. 5, where the within-cluster and between-cluster groupings are based on the known classification. As desired, the two medians show a rise going from the *within-cluster* class to the *between-cluster* class. There is, however, a slight overlap between the two dissimilarity classes. The calculated dissimilarities on the basis of powder diffraction patterns (De Gelder *et al.*, 2001) are shown in Fig. 6 and show the same increase for the median and overlap, although the separation of the classes is better with the R_eDF descriptor.

The results for the estrone data set are plotted as box plots in Fig. 7. The calculated dissimilarities are an order of magnitude larger than those for the cephalosporin set. This is caused by the higher intensities of the peaks in the estrone R_eDFs. The medians in the plot show a gradual rise going from the *identical* class to the *dissimilar* class. This is what one would expect, but the figure shows that the two most dissimilar classes are not fully separated. The *identical* class is completely separated from the other two dissimilarity classes.

5.2. Dendrograms and partitionings

The dendrogram determined for the cephalosporin data set with the new descriptor using average linkage is given in Fig. 8. Given a properly chosen height, it predicts the true classes without errors. Partitioning the dendrogram into seven clusters was done by cutting the tree at a height of 0.4 (horizontal line).

The use of the R_eDF descriptor for the experimental data set was compared with the dendrogram determined on the basis of powder diffraction patterns (see Fig. 5*d* in De Gelder *et al.*, 2001). The latter shows a clustering which is essentially correct, but the dendrogram based on the R_eDF gives a better discrimination of the separate groups.

The dendrogram for the 48 crystal structures of estrone was calculated with the average linkage from the R_eDF-generated dissimilarities and is given in Fig. 9. The dendrogram shows that the crystal structures which are known to have a dissimilarity in the *identical* class (clusters L–S) are correctly grouped together. The structures from cluster B, with dissimilarities in the *similar* class are grouped together, but cluster A, also with dissimilarities in the *similar* class, is scattered over the right hand side of the dendrogram. This reflects the fact that the dissimilarities for the two dissimilarity classes have an overlap (see Fig. 7). A partitioning with 25 clusters is generated from the dendrogram by cutting at a height of 0.45 (horizontal line).

5.3. Matching ESTRON10

In the case of the simulated estrone structure, it is interesting to know if the method is able to tell which simulated

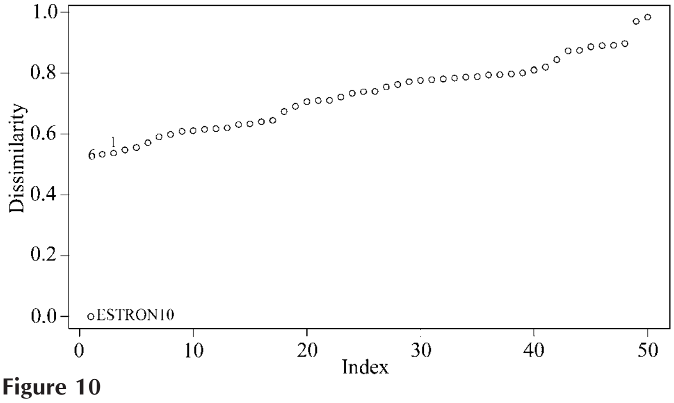


Figure 10
Dissimilarities between the experimentally found polymorph (ESTRON10) and all simulated estrone structures (1–48). Structures 6 and 1 have the same packing as the ESTRON10, and are identified with the new descriptor.

structure matches an experimental structure. This has been done for ESTRON10 and the results are given in Fig. 10. The R_eDF for ESTRON10 is calculated in the same way as for the simulated structures, and the dissimilarity measure is able to identify structures 6 and 1 having the same packing. Structures 6 and 1 both belong to cluster A with a dissimilarity between them in the *similar* class.

The large dissimilarity between the simulated structures and ESTRON10 is due to the fact that the set of simulated structures is the result of a molecular mechanics optimization. Force-field artifacts lead to longer unit-cell axes than experimentally found; therefore, the *y* scales of Fig. 10 and Fig. 7 are not directly comparable. The important thing here is that the order of dissimilarities is correct. It also makes comparing the dissimilarities of ESTRON10 *versus* 6 and 1 with the dissimilarity of ESTRON10 *versus* the third most ESTRON10-like compound less intuitive; the small differences in those three values do not necessarily indicate that the third structure has almost the same packing as ESTRON10 as structure 6 does.

6. Conclusions

This article presents a new computational method to compare crystal structures. It is conceptually easy and contains only a few parameters to tune; within broad ranges, the exact values of these parameters have little influence on the results. The method is, therefore, very general. It correctly shows increasing dissimilarity values when going from identical crystal structures to similar, and finally to dissimilar structures. It is difficult to order dissimilar structures in a meaningful way and, therefore, the main use of the descriptor is twofold: to gather similar structures from a large set and to recognize the most similar structure from a set of candidate structures. Both have numerous and important applications.

This research was sponsored by the NWO, the Netherlands Organization for Scientific Research.

References

Andrews, L. C. & Bernstein, H. J. (1988). *Acta Cryst.* A**44**, 1009–1018.

Andrews, L. C. & Bernstein, H. J. (1995). *Acta Cryst.* A**51**, 413–416.

Andrews, L. C., Bernstein, H. J. & Pelletier, G. A. (1980). *Acta Cryst.* A**36**, 248–252.

Busetta, B., Courseille, C. & Hospital, M. (1973). *Acta Cryst.* B**29**, 298–313.

De Gelder, R., Wehrens, R. & Hageman, J. A. (2001). *J. Comput. Chem.* **22**, 273–389.

Desiraju, G. R. (1995). *Angew. Chem. Int. Ed.* **34**, 2311–2327.

Dzyabchenko, A. V. (1994). *Acta Cryst.* B**50**, 414–425.

Fábián, L. & Kálmán, A. (1999). *Acta Cryst.* B**55**, 1099–1108.

Frisch, M. J. *et al.* (2001). *GAUSSIAN*. Pittsburg, PA: Gaussian, Inc.

Gasteiger, J., Sadowski, J., Schuur, J., Selzer, P., Steinhauer, L. & Steinhauer, V. (1996). *J. Chem. Inf. Comput. Sci.* **36**, 1030–1037.

Gentleman, R. & Ihaka, R. (1996). *J. Comput. Graph. Stat.* **5**, 299–314.

Hemmer, M. C., Steinhauer, V. & Gasteiger, J. (1999). *Vib. Spectrosc.* **19**, 151–164.

Hollingsworth, M. D. (2002). *Science*, **295**, 2410–2413.

Ilyushin, G., Blatov, N. & Zakutin, Y. (2002). *Acta Cryst.* B**58**, 948–964.

Karfunkel, H., Wilts, H., Hao, Z., Iqbal, A., Mizuguchi, J. & Wu, Z. (1999). *Acta Cryst.* B**55**, 1075–1089.

Karfunkel, H. R., Rohde, B., Leusen, F. J. J., Gdanitz, R. J. & Rihs, G. (1993). *J. Comput. Chem.* **14**, 1125–1135.

Kemperman, G. J., De Gelder, R., Dommerholt, F. J., Raemakers-Franken, P. C., Klunder, A. J. H. & Zwanenburg, B. (2000). *J. Chem. Soc. Perkin Trans. 2*, **7**, 1425–1429.

Lommerse, J. P. M., Motherwell, W. D. S., Ammon, H. L., Dunitz, J. D., Gavezzotti, A., 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.

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

Molecular Simulations Inc. (1997). *Cerius*2 User Guide, ch. 7. San Diego: Molecular Simulations Inc.

Motherwell, W. D. S., Ammon, H. L., Dunitz, J. D., Dzyabchenko, A., Erk, P., Gavezzotti, A., Hofmann, D. W. M., Leusen, F. J. J., Lommerse, J. P. M., Mooij, W. T. M., Price, S. L., Scherega, H., Schweizer, B., Schmidt, M. U., Van Eijck, B. P., Verwer, P. & Williams, D. E. (2002). *Acta Cryst.* B**58**, 647–661.

Moulton, B. & Zaworotko, M. J. (2001). *Chem. Rev.* **101**, 1629–1658.

Pauling, L. & Delbrück, M. (1940). *Science*, **92**, 77–79.

Perlstein, J., Steppe, K., Vaday, S. & Ndip, E. M. N. (1996). *J. Am. Chem. Soc.* **118**, 8433–8443.

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

Verwer, P. & Leusen, F. J. J. (1998). *Computer Simulation to Predict Possible Crystal Polymorphs*, Vol. 12 of *Reviews in Computational Chemistry*, ch. 7. New York: Wiley-VCH.