

COMPACK: a program for identifying crystal structure similarity using distances

James Alexander Chisholm^{a,b*} and Sam Motherwell^a^aCambridge Crystallographic Data Centre, UK, and ^bPfizer Institute for Pharmaceutical Materials Science, UK. Correspondence e-mail: chisholm@ccdc.cam.ac.uk

A method is presented for comparing crystal structures to identify similarity in molecular packing environments. The relative position and orientation of molecules is captured using interatomic distances, which provide a representation of structure that avoids the use of space-group and cell information. The method can be used to determine whether two crystal structures are the same to within specified tolerances and can also provide a measure of similarity for structures that do not match exactly, but have structural features in common. Example applications are presented that include the identification of an experimentally observed crystal structure from a list of predicted structures and the process of clustering a list of predicted structures to remove duplicates. Examples are also presented to demonstrate partial matching. Such searches were performed to analyse the results of the third blind test for crystal structure prediction, to identify the frequency of occurrence of a characteristic layer and a characteristic hydrogen-bonded chain.

© 2005 International Union of Crystallography
Printed in Great Britain – all rights reserved

1. Introduction

Similarity between crystal structures can be identified at various levels. For example, two crystal structures may have similar crystal symmetry, similar unit-cell parameters and may show chemical similarity such as the presence of common synthons. In this work, we present a method of comparing crystal structures to identify similarity in molecular packing. Structures are matched if molecular structures, represented by atom types and atom connectivity, match exactly and if molecular packings, represented using interatomic distances, can be matched to within specified tolerances.

The task of identifying structural similarity is necessary when deciding whether a newly determined structure is the same as an existing structure. Also, in the field of crystal structure prediction (CSP), thousands of hypothetical structures are generated and then optimized by lattice energy. Many structures converge to the same structure during the optimization step and it is highly beneficial to remove duplicates to produce a list of unique structures. Comparing structures by visual inspection can be slow and prone to human bias and so we have developed this practical computational method for identifying identical structures.

Computational methods do exist for comparing crystal structures, such as the computer program *CRYCOM* (Dzyabchenko, 1994). In that program the crystal structure is represented using unit-cell, space-group and fractional coordinate data. Structures with the same space group are matched if unit cell and atomic coordinate data can be matched to within specified tolerances. As choices often exist for the crystal axes, origin and asymmetric unit, all possible equivalent descriptions of the reference structure must be considered. Reduced-symmetry descriptions (down to *P1*) can also be considered for comparing structures where full symmetry has not been recognized.

Crystal structure similarity can also be identified by comparing radial distribution functions, as implemented in the *Polymorph Predictor* program (Verwer & Leusen, 1998; van Eijck & Kroon,

1997), and by comparing computed powder patterns (Karfunkel *et al.*, 1993; de Gelder, 2001).

In this work, we describe an alternative method that in addition to identifying structures that are the same, can also be used to compare sections of crystals, such as a specific arrangement of a pair of molecules or a larger cluster such as a characteristic layer. For example, structures I and II may differ only in their stacking sequence (say stacking *ABCABC* compared with *ABABAB*) and it is desirable to identify the fact that the layers *AB* are common to both structures.

2. Basic approach

2.1. Generation of a search query

COMPACK has been developed using C++. Fig. 1 shows the basic steps adopted. To begin, a coordination shell is constructed from the reference structure composed of a central molecule together with the nearest *N* molecules. The distance between molecular centroids is used to decide which molecules are nearest. This produces a finite

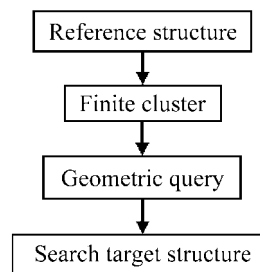


Figure 1
An overview of the steps implemented by *COMPACK* to identify crystal structure similarity. A reference structure is represented by a finite cluster containing no crystal symmetry information. This cluster is used to construct a search query containing geometric constraints which can be searched for in a target structure.

molecular cluster or 'sphere' of molecules, which is taken to represent the entire crystal. As such, the value for N needs to be sufficiently large to represent the touching nearest-neighbour molecules. As a default value, N is set to 14, but is made adjustable.

The precise arrangement of molecules in the cluster, the relative positions and orientations, is captured using a set of interatomic distances drawn between nearest-neighbour molecules. For our purposes, neighbouring molecules are determined to be those where an interatomic distance can be found that is less than the sum of van der Waals radii plus 2 Å. Fig. 2 shows the distances drawn between just two molecules in the cluster. A sufficient number of distances are required in order to represent each molecule–molecule relationship accurately. We use as many distances as there are atoms in the largest of each pair of molecules and so use every atom at least once. There is then the question of which particular distances to consider. Distances could be chosen at random or simply be based on the atom ordering. In our method we bias the selection of distances towards the use of short distances as these provide a better description of the molecule–molecule relationship. First the shortest contact distance between the two molecules is found; then the next shortest contact distance is found from the remaining atoms. This process is repeated until the specified number of distances has been reached.

The molecular cluster can now be viewed as a chemical search query that is subject to several geometric constraints. Once a crystal structure, or a section of a crystal structure, has been represented in this way, the task of searching for structural similarity becomes a matter of searching the three-dimensional coordinates of a target structure for an arrangement of molecules that match the search query to within specified tolerances. As a default value, a tolerance of $\pm 15\%$ is placed on distance constraints, but this is made adjustable.

2.2. Searching target structures

The searching of target structures is performed by *3DSEARCH*, which is an efficient general search algorithm designed to identify chemical queries within three-dimensional crystal structure information (Chisholm & Motherwell, 2004). The task is to identify the presence of a molecular cluster, the structure of which is described in terms of interatomic distances, within an infinite target crystal, the structure of which is described by the unit cell, asymmetric unit and atomic coordinates. Only a brief description of the methodology for the specific task of comparing structures is presented here.

The algorithm begins by searching the target structure for a match for the first molecule in the molecular cluster. Molecules are matched by comparing atom types and atom connectivity using a modified graph-matching algorithm (Ullmann, 1976). If a molecule match can be found, the search proceeds to search the target structure for a

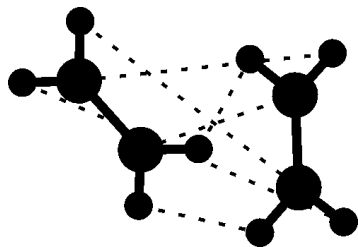


Figure 2

The arrangement of molecules in a molecular cluster is captured using scalar interatomic distances drawn between neighbouring molecules (dashed lines). For each pair of neighbouring molecules in the cluster, we draw as many distances as there are atoms in the larger of the two molecules and bias the choice of distances towards the use of short rather than long distances.

match for a molecule in the cluster. The only criterion used to choose this next molecule is that it be connected, *via* a interatomic distance constraint, to a molecule that has already been found. This process of finding the next molecule is composed of three main steps.

Step 1. Search for one connection (distance constraint).

Step 2. Find molecule.

Step 3. Check remaining connections.

The steps 1 and 2 can retrieve several matches and all matches must be visited to ensure the search is exhaustive. Step 3 is a simple check procedure and the term 'remaining' refers to all connections that connect with the current search molecule, have not yet been matched, and that connect to a molecule that has been matched. The steps 1 to 3 are given the name 'find next'. The search algorithm repeats 'find next' working outwards until a complete query match can be found. Connection searches and molecule searches can be viewed as nodes in a search tree and any combination of searches by a branch. If at any point a match cannot be found, the algorithm backtracks and searches the next branch in the search tree. The search stops once a complete match has been found or once all branches in the search tree have been traversed.

2.3. Superimposing matched structures

Once a match has been found, the two structures can be superimposed to obtain a visual impression of their similarity. To obtain the best superposition, *i.e.* the best agreement between atomic positions, we use an algorithm for overlaying points (Kabsch, 1976, 1978). By default, the best overlay is found by considering all the atom positions in the molecular cluster. An alternative is to obtain the best overlay for the central molecule only. Once structures have been superimposed, the distances between matched atoms in each molecular cluster are used to derive a root-mean-square (RMS) deviation value, which provides a numerical measure of the degree of similarity. This particular method of determining the RMS value is a function of the number of molecules compared as the discrepancy between matching structures necessarily increases with increasing distance. Note that the RMS value is different from the RMS deviations determined in the *CRYCOM* program, which considers deviations between coordinates of atoms in the asymmetric unit.

3. Applications

3.1. Crystal structure prediction

The goal of crystal structure prediction is to predict the experimentally observed crystal structure given information only on the molecular diagram. Recent blind tests have shown that current methods of prediction meet with only limited success (Lommerse *et al.*, 2000; Motherwell *et al.*, 2002). *COMPACT* has been used to analyse lists of putative structures obtained from CSP runs to answer two important questions. Firstly, do generation schemes successfully generate the observed structure somewhere in the list? Secondly, at what ranked position does the observed structure appear?

In recent work, an assessment has been made of the performance of empirical force field methods for the prediction of 50 small organic molecules in which predicted structures were ranked by minimized lattice energy (Day, Chisholm *et al.*, 2004). In this study, *COMPACT* was used to cluster predicted lists and identify the energy-ranked position of observed crystal structures. A search through the predicted crystal structures shows that the experimentally observed structure was always generated successfully, being present somewhere in the predicted list.

The comparison method can be quick. A search through the top 100 predicted formamide structures takes 3 s and identifies a match with the experimentally observed structure at rank 27. Thus, the formamide prediction was unsuccessful. However, such analysis shows that the performance of the force field predictions is generally better than this and that about half of such molecules can be expected to be found in the lowest five predicted structures, or within 1 kJ mol^{-1} of the global minimum in lattice energy.

Because the searches can be quick, the comparison method can also be used to cluster predicted lists by identifying and removing duplicate structures. This is an important stage in CSP as much computational effort can be saved by avoiding the optimization of similar structures that would end up in the same minimum. For example, the clustering of 1000 predicted structures for the 2-amino-3-nitropyridine molecule involved 156 439 comparisons and took 53 min to complete, running on a 1.8 GHz Pentium 4 processor. 410 structures were matched, which reduced the size of the list to 590 structures. For this case, 410 optimizations are avoided, which would have taken 196 min to complete.

There is the question of what size of molecular cluster is required to identify 'same structures'. The clustering process described above was repeated, this time using a coordination shell of 30 molecules instead of the default value of 15. This led to identical cluster results and suggests that a molecular cluster containing 15 molecules is of sufficient size for identifying 'same structures'. However, it is difficult to show this conclusively and the size of cluster is left as an adjustable parameter set by the user.

3.2. Partial matching

Predicted model structures that are not the same often show a degree of similarity in their packing environments. In such cases, *COMPACK* provides the number of molecules in the molecular sphere that were matched successfully. For example, if only 6–9 molecules are matched, we can say that structures display some similarities, and if say 12–14 molecules have been matched, this indicates that the structures are 'almost the same'.

Fig. 3 shows two crystal structures that display many similarities. The structure on the left is the experimentally observed structure for molecule II, the second molecule chosen for the third blind test of CSP (CSP2004; Day, Motherwell *et al.*, 2004), and the structure on the right is a predicted structure found to match the observed structure partially, with 13 molecule matches. In all, 14 groups submitted predictions for molecule II in the recent third blind test of CSP. However, only one participant successfully predicted the structure of molecule II. A search using *COMPACK* shows that the packing arrangement in the central region, marked A in Fig. 3, is frequently predicted and occurs in 40% of participants top three predictions. This indicates that the difficulty in predicting molecule II is in

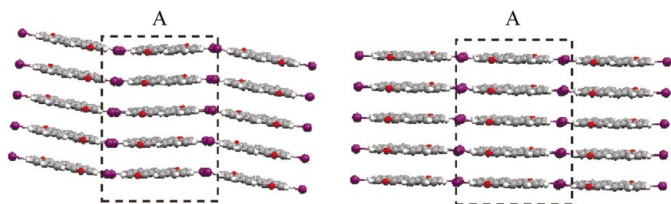


Figure 3
Blind-test molecule II: observed (left) compared with a predicted structure (right). The two structures show a degree of similarity as the packings of molecules in the central region, marked A, match.

faithfully modelling the more subtle interactions between molecules oriented 'end to end' that is in the horizontal direction in Fig. 3.

As another example, consider the blind-test molecule IV, which was not successfully predicted by any group. Fig. 4 shows a section of the experimentally observed structure which contains a hydrogen-bonded chain motif. This chain can only be described by considering structures with Z' greater than 1. The arrangement of molecules in the chain is selected and used to form a search query containing six molecules that represent the chain. A search for this cluster in predicted structures, using a tolerance of 25% on distances, shows that only 10 out of 969 predicted structures contain the observed hydrogen-bond chain motif. A more commonly predicted motif is shown on the right in Fig. 4. This chain was found to appear in 20% of all structures submitted. This indicates that molecule IV was difficult to predict due to difficulties in generating sufficient candidate structures with $Z' = 2$.

4. Conclusions

A method of comparing crystal structures has been presented which uses distance constraints to represent molecular packing. This method has been implemented in a computer program *COMPACK*. The method can be used to determine whether any given two crystal structures are the same, or whether they show a degree of similarity. *COMPACK* has applications for crystal structure prediction work, for identifying experimentally observed structures from lists of predicted structures and for clustering predicted crystal structure lists. In addition, structures can be compared to identify characteristic packing arrangements, such as layers or chains of molecules. Such searches have been performed to analyse the results of the third blind test for crystal structure prediction. A characteristic packing subunit for molecule II was shown to occur frequently in predicted structures. The *COMPACK* program is currently being integrated with software in the distributed CSD system.

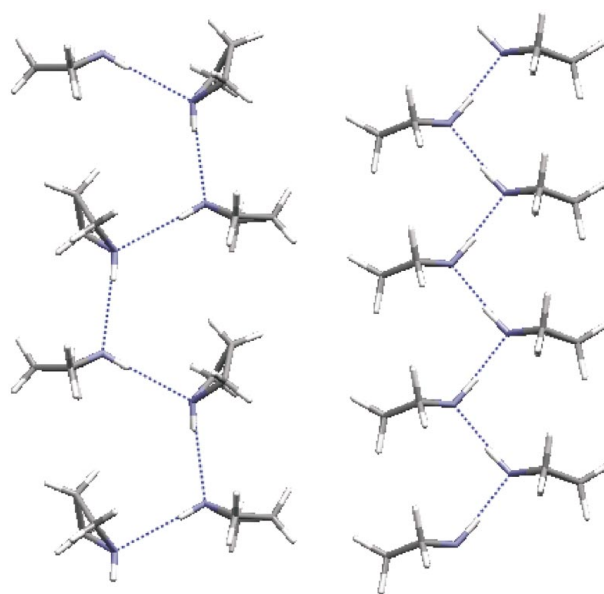


Figure 4
The hydrogen-bond chain motif present in the experimentally observed structure of molecule II chosen for the third blind test of crystal structure prediction (left). This chain motif requires $Z' > 1$. The chain motif on the right shows a commonly predicted motif in structures with $Z' = 1$.

JAC acknowledges Pfizer Inc. for financial support.

References

- Chisholm, J. A. & Motherwell, S. (2004). *J. Appl. Cryst.* **37**, 331–334.
- Day, G. M., Chisholm, J., Shan, N., Motherwell, W. D. S. & Jones, W. J. (2004). *Cryst. Growth Des.* In the press.
- Day, G. M., Motherwell, W. D. S., Ammon, H., Boerrigter, S. X. M., Della Valle, R. G., Venuti, E., Dunitz, J., Dzyabchenko, A., van Eijck, B. P., Erk, P., Facelli, J. C., Bazterra, V. E., Ferraro, M. B., Hofmann, D. W. M., Leusen, F. J. J., Liang, C., Pantelides, C., Karamertzanis, P. G., Price, S. L., Lewis, T. C., Torrissi, A., Nowell, H., Scheraga, H., Arnautova, Y., Schmidt, M. U., Schweizer, B. & Verwer, P. (2004). In preparation.
- Dzyabchenko, A. V. (1994). *Acta Cryst.* **B50** 414–425.
- Eijck, B. P. van & Kroon, J. (1997). *J. Comput. Chem.* **18**, 1036–1042.
- Gelder, R. de (2001). *J. Comput. Chem.* **22**, 273–289.
- Kabsch, W. (1976). *Acta Cryst.* **A32** 922–923.
- Kabsch, W. (1978). *Acta Cryst.* **A34** 827–828.
- Karfunkel, H. R., Noordik, J. H., Leusen, F. J. J., Gdanitz, R. J. & Rihs, G. (1993). *J. Comput. Chem.* **14**, 1125–1135.
- 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.* **B56**, 697–714.
- 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., Scheraga, H., Schweizer, B., Schmidt, M. U., van Eijck, B. P., Verwer, P. & Williams, D. E. (2002). *Acta Cryst.* **B58**, 647–661.
- Ullmann, J. R. (1976). *J. Assoc. Comput. Machinery*, **23**, 31–42.
- Verwer, P. & Leusen, F. J. J. (1998). *Rev. Comput. Chem.* **12**, 327–365.