Acta Crystallographica Section B
Structural Science
ISSN 0108-7681

Searching the Cambridge Structural Database for polymorphs

Jacco van de Streek* and Sam Motherwell

Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, England

Correspondence e-mail: streek@ccdc.cam.ac.uk

Received 18 March 2005
Accepted 24 June 2005

In order to identify all pairs of polymorphs in the Cambridge Structural Database (CSD), a method was devised to automatically compare two crystal structures. The comparison is based on simulated powder diffraction patterns, but with special provisions to deal with differences in unit-cell volumes caused by temperature or pressure. Among the 325 000 crystal structures in the Cambridge Structural Database, 35 000 pairs of crystal structures of the same chemical compound were identified and compared. A total of 7300 pairs of polymorphs were identified, of which 154 previously were unknown.

1. Introduction

The word 'polymorph' can be loosely defined as 'two different crystal packings of the same chemical compound'. In this work the chemical compounds are organic or organometallic, including molecular salts and solvated systems, and the polymorphism in other types of compound, such as inorganic compounds and metals, is not dealt with. As the physical properties of two polymorphs can be so different, polymorphism is an important phenomenon for *e.g.* the pharmaceutical industry, where many drugs are manufactured in crystalline form. As a result, polymorphism has been extensively studied; for references see the book by Bernstein (2002). The Cambridge Structural Database (CSD; Allen (2002), maintained by the Cambridge Crystallographic Data Centre (CCDC), contains virtually all the published crystal structures of organic and organometallic molecules, which makes the CSD an ideal tool for studying the polymorphism of organic and organometallic compounds. In order to assist such polymorphism studies, entries in the CSD that are known to be polymorphic are flagged with a 'POLYMORPH' keyword.

Currently, however, whether or not a CSD entry is flagged as a polymorph depends mainly on this being mentioned in the original publication. This then raises the question if it is possible to data-mine the CSD to find additional pairs of polymorphs, *i.e.* find compounds with two or more different crystal structures that are currently not flagged as polymorphs. Owing to the large number of entries in the CSD, more than 325 000, manual inspection is not feasible and an automated search method is required.

There are two more problems with the existing content of the CSD. First, the polymorph flag only indicates that the compound is polymorphic, it does not guarantee that the crystal structure of any of the other polymorphs is known or recorded in the CSD. Second, once a compound is known to be polymorphic, all of its entries in the CSD are flagged as polymorphic; this includes redeterminations, and given two crystal structures flagged as polymorphic it is not possible to establish if they are polymorphs or re-determinations. Para-

cetamol, for example, has two known polymorphs but 21 CSD entries – all flagged as polymorphic. The automated search procedure for missed polymorphs can also be used to examine and check the polymorphs that are already flagged to weed out the redeterminations.

It is the aim of this paper to address all three issues:

(i) To find all currently unflagged polymorphs in the CSD.

(ii) To eliminate compounds that are flagged as polymorphic, but for which only one crystal structure has been determined.

(iii) To distinguish redeterminations from true polymorphs.

The results will be presented in the form of two lists: one list containing all pairs of previously unknown polymorphs and a second list containing *all* currently known pairs of small molecule polymorphs – previously unknown or not.

We would like to stress that there is an important difference between the two lists. The first list, containing newly discovered polymorphs, should lead to the addition of polymorph flags to the CSD. As such, this list has to be accurate, *i.e.* must not include false positives (pairs of crystal structures that are in the list, but that are not polymorphs), and this list therefore requires a manual inspection after the automated screening. The second list contains too many entries to inspect manually and is necessarily an automatically generated list that may suffer from both false positives and false negatives (pairs of crystal structures that are polymorphs, but that are missing from the list).

2. Methods

In the CSD, the crystal structures are grouped per chemical compound in so-called 'refcode' families. A refcode consists of six letters that identify the chemical compound followed by two digits that identify the entry. Paracetamol, for example, has been assigned the refcodes HXACAN, HXACAN01, HXACAN02, HXACAN03 *etc*. In this paper, we will therefore equate 'chemical compound' with refcode family, ignoring the small number of errors that have been made in the assignment of refcodes. As a check, we compared the empirical sum formulae (*e.g.* $C_8H_9NO_2$ for paracetamol) and required that the count for each element agree to within 0.225 for each pair of entries. We will equate 'crystal structures' with crystal structures that are well defined, *i.e.* do not have disorder, have an R factor of less then 10% and have all the three-dimensional atomic coordinates of all non-H atoms determined. In other words, it is our aim to compile a list of polymorphs that can be readily used for further analysis by computer programs, *e.g.* force-field calculations and conformational or hydrogen-bond analysis, without requiring any further manipulation other than possibly adding H atoms in calculated positions. By 'different' we mean that the crystal structures correspond to clearly different molecular packings, *i.e.* we do not count minor order–disorder phase transitions, breaking of symmetry or pseudo-symmetry.

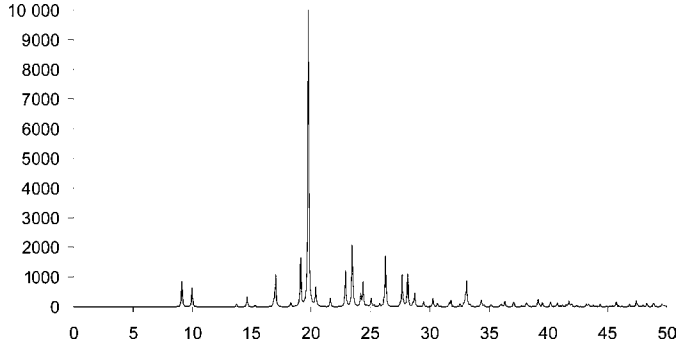

Figure 1
Step one: transformation of the crystal structure to a simulated powder diffraction pattern. The CSD entry AANHOX as simulated in the *Mercury* program is shown. Axes: x axis: diffraction angle in $^\circ$ 2θ; y axis: intensity: arbitrary units.

The problem now reduces to: given all crystal structures within a certain refcode family, which of them are different? This requires a method to compare crystal structures that, because of the vast number of crystal structures in the CSD, must be suitable for automation.

When comparing crystal structures, problems arise due to:

(i) space-group settings, *e.g.* $P2_1/n$ and $P2_1/c$;

(ii) choice of origin, *e.g.* $P2_12_12_1$ or *Fddd*;

(iii) unit-cell differences due to temperature/pressure;

(iv) missed symmetry/pseudo-symmetry;

(v) assignment of bonds and bond types.

(We will show examples of the less obvious problems in §3.) All of these problems except unit-cell differences can be solved by transforming the three-dimensional atomic coordinates to a one-dimensional function depending on interatomic distances only, namely a simulated powder diffraction pattern. Note that simulated powder diffraction patterns do not suffer from problems such as zero-point shift, the presence of a background, preferred orientation or counting statistics. Therefore, the first step in our comparisons is to transform each crystal structure to its simulated powder diffraction pattern (Fig. 1). We would like to stress that for simulating the powder diffraction patterns the space-group symmetry operators as stored in the CSD entry were used, avoiding problems with ambiguous origins for *e.g.* $P2_12_12_1$.

The X-ray powder diffraction patterns were simulated with Cu $K\alpha_1$ radiation, wavelength 1.54056 Å, from 0.0 to 50.0° 2θ, corresponding to a resolution of 1.8 Å, with a step size of 0.02° 2θ (earlier attempts using a 2θ range of 5.0–30.0° turned out to give too many problems with peaks near either of the two boundaries, and problems with very small unit cells where the range would sometimes contain only one single peak). The 000 reflection was omitted. The peak shape was pseudo-Voigt with a fixed full width at half-maximum (FWHM) of 0.1° 2θ. These are the settings used in the *Mercury* program (Bruno *et al.*, 2002). The finite peak width plays a role in reducing the sensitivity of the similarity measure to minor peak shifts, see also below. In order to reduce the effect of poorly determined or even partially absent H atoms, all H and D atoms were omitted from the calculation. This is important as there are many cases of re-determinations where the earlier crystal structure has no three-dimensional coordinates for H atoms

and the later crystal has all the H atoms determined; an example would be the entries ANTCEN02 (Mathieson *et al.*, 1950) and ANTCEN14 (Brock & Dunitz, 1990) published in 1950 and 1990, respectively.

The comparison of powder diffraction patterns is very sensitive to shifts in peak positions. Since the peak positions are directly related to the unit-cell parameters, which in turn are directly related to the unit-cell volume, the comparison of simulated powder diffraction patterns is sensitive to the temperature and pressure at which the crystal structures were determined. In order to reduce the influence of unit-cell size differences due to temperature or pressure, all unit-cell volumes were normalized to ambient temperature and pressure by scaling the unit-cell lengths *a*, *b* and *c* so as to reproduce the unit-cell volumes obtained from summing the average atomic volumes published by Hofmann (2002). The effect on the similarity of AANHOX (Ciajolo *et al.*, 1981) and AANHOX01 (Jerslev, 1987), which were determined at room temperature and at 105 K, respectively, is shown in Fig. 2. If there are atoms missing from the unit cell, *e.g.* due to the presence of a disordered or unknown solvent, the calculated expected volume is too low and the normalization to room temperature may cause distortions of the crystal structure. Therefore, the unit-cell volume as expected at the temperature at which the structure was determined (V_{cal}) was calculated and compared with the actual unit-cell volume (V_{obs}). If the ratio V_{calc}/V_{obs} was greater than 1.4 or smaller than 0.7, the crystal structure was assumed to contain an error and was skipped in the analysis.

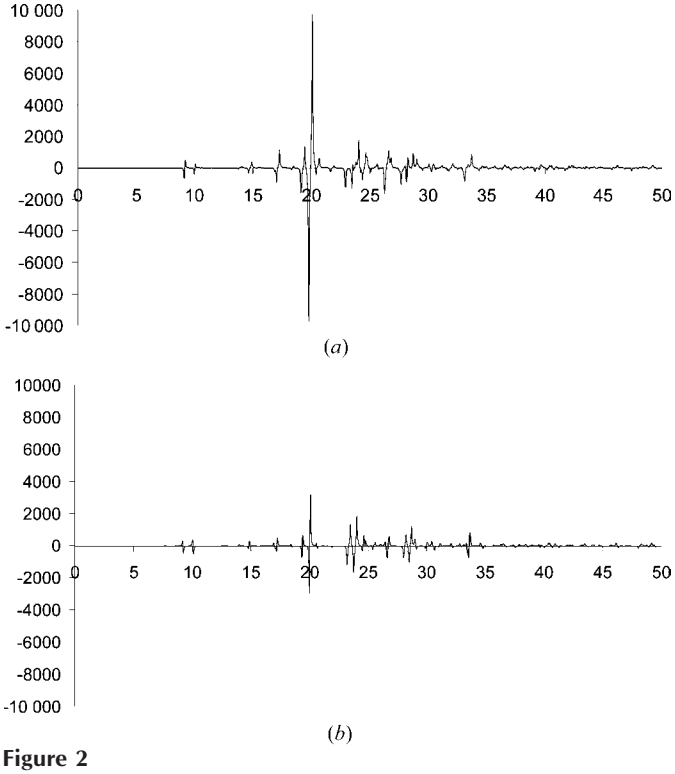

Figure 2
Step two: normalization of the unit-cell volumes. Above: difference profile of the simulated powder diffraction patterns of AANHOX (room temperature) and AANHOX01 (105 K) before volume normalization; their similarity is 0.930. Below: difference profile after volume normalization, their similarity is 0.995. Axes: *x* axis: diffraction angle in $^\circ$ 2θ; *y* axis: intensity, arbitrary units.

As it is not known if a given unit-cell angle increases or decreases as a function of temperature, the unit-cell angles were not changed. Moreover, examination of several random pairs of crystal structures determined at different temperatures tells us that the expansion of crystals of organic molecules is in general very anisotropic. For example, for the temperature difference of 190 K between AANHOX and AANHOX01 shown in Fig. 2, the unit-cell parameters *a*, *b* and *c* contract by 2.5, 1.0 and 0.6%, respectively. Therefore, even after volume-normalization there will in general still be some discrepancy between the unit-cell parameters of the two crystal structures of the same polymorph determined at different temperatures.

The similarity of two powder diffraction patterns is commonly expressed as an *R* value or an R_{wp} value. However, these point-by-point measures are very sensitive to peak shifts and for the reasons mentioned in the previous section, these point-by-point measures yield very low similarities for the two crystals determined at different temperatures, even if the crystal structures are clearly the same. Therefore, measures that are less sensitive to peak shifts have been published (see De Gelder *et al.*, 2001). Most of these measures, however, are not normalized, which makes it very difficult to use a computer program to decide if a certain similarity value indicates that the two powder diffraction patterns are similar or not. The similarity measure based on weighted cross-correlation functions published by De Gelder *et al.* (2001) is both normalized and less sensitive to peak shifts and this measure has been used throughout this work. The overlap function used was a simple triangle function of width $l = 2.0^\circ$ 2θ; this value should be adapted depending on the FWHM of the simulated powder diffraction pattern.

In order to be able to focus the search on the most interesting cases, namely those where polymorphs have been missed and are not yet flagged as such in the CSD, the presence or absence of the polymorph flag was also recorded in the list.

It is CCDC policy to try to report objectively what is presented in the literature as closely as possible and this means that unavoidably, especially for older entries when coordinates had to be re-keyboarded from hard copy and the error-checking software was not as good as it is today, errors are present in the database. The errors could be missing minus signs, errors in the space-group setting or misprints in atomic coordinates. These errors are always corrected when discovered and the CCDC welcomes input from CSD users about entries that might require modification. Even small misprints, such as a missing minus sign, can have a big effect on the simulated powder diffraction pattern. An incorrect crystal structure will differ from a correct one and might therefore be inadvertently identified as a 'polymorph'. This is not too important for the generation of the first list, the list of missed polymorphs, because of the manual inspection step involved. These errors are, however, a serious problem for the second

list, the automatically generated list of all pairs of polymorphs in the CSD, where they would cause false positives. In practice, almost all of these errors are in the unit-cell contents, *i.e.* the atomic coordinates of the asymmetric unit or the space group, and hardly ever in the unit-cell parameters. Therefore, identifying polymorphs based only on unit-cell parameters is more reliable than identification based on unit-cell contents, because there are fewer errors. The slight disadvantage of this method is that there will be a handful of false negatives if the unit cells of two polymorphs happen to be very similar, which is rare but not impossible (see *e.g.* Kálmán *et al.*, 2004), but this small number of additional false negatives far outweighs the number of eliminated false positives.

Therefore, in order to render the method less sensitive to minor misprints and missed minus signs in printed atomic coordinates, a second list of similarities was prepared, based only on the unit cells and not on their contents (the atoms and the space group). In a powder diffraction pattern, the unit-cell parameters determine the peak positions, whereas the combination of the asymmetric unit and the space group determines the peak intensities. Therefore, all that is needed to calculate a normalized, smoothly varying similarity measure based only on the unit-cell parameters is to eliminate the contribution of the unit-cell contents by setting the structure factors of all the reflections in the simulated powder diffraction pattern to an arbitrary constant value, including those reflections that were systematically absent, but excluding those that are systematically absent due to the centring of the space group. The numbers thus obtained are essentially similarity measures for the stable reduced unit cells of the crystal structures. Fig. 3 shows such a 'reduced unit-cell diffraction pattern' for AANHOX, which is to be compared with Fig. 1 for the simulated powder diffraction pattern of the full crystal structure. These reduced unit-cell similarities were used for the generation of the second list, the automated list of all polymorph pairs in the CSD.

Now that we have a method to decide if two crystal structures are polymorphs based on the unit cell only, the restrictions that we imposed earlier regarding all three-dimensional atomic coordinates having been determined are no longer necessary. This allows us to generate a third list that includes all entries in the CSD, even those without three-dimensional atomic coordinates. We can still normalize the unit-cell volume by multiplying the known chemical formula by Z, which still allows us to eliminate entries for which the volume of the unit-cell contents based on Hofmann's average atomic volumes disagrees with the unit-cell volume. Obviously, entries that do not have unit-cell parameters and pairs of entries whose empirical sum formulae do not match are still eliminated from the list.

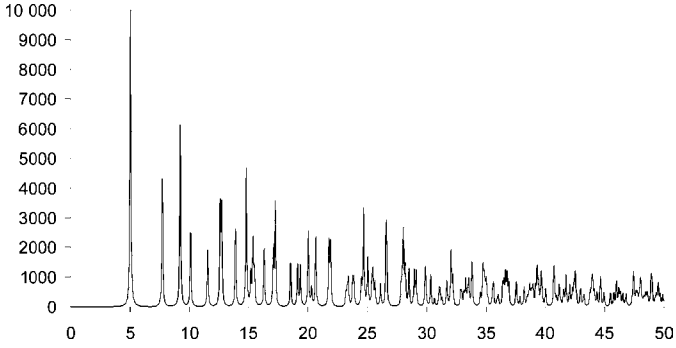

Figure 3
Simulated diffraction pattern of the reduced unit cell of AANHOX, used to calculate a smoothly varying similarity measure for reduced unit cells. The Lorentz–polarization correction provides a natural down-weighting of the high-resolution part of the pattern. Axes: x axis: diffraction angle in $^\circ$ 2θ; y axis: intensity, arbitrary units.

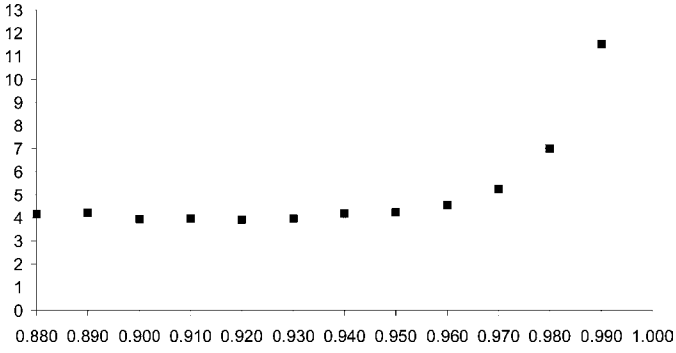

Figure 4
The number of unidentified polymorphs per 100 CSD entries (y axis) as a function of similarity cut-off (x axis). The background level is around 4% for similarities below 0.960 that unambiguously point to polymorphism, with more and more false positives being added as the similarity cut-off approaches 1.000.

3. Results and discussion

The first result, a hand-checked list of 308 previously unknown polymorphs (154 pairs), is given in the supplementary material.[1] All of these newly discovered polymorphs are now flagged as such in the November 2004 release of the CSD.

For the automated lists we need to choose a cut-off value for the similarity, below which the two crystal structures are most likely to be polymorphs rather than a re-determination. We can find this value by making the assumption that the number of unflagged (missed) polymorphs in the CSD is independent of the similarity and then plotting the percentage of unflagged polymorphs as a function of cut-off. If our assumption is correct, we would expect such a plot to show a base level for very low similarities, where our method is almost certainly discriminating enough, followed by a steady rise above a certain similarity value as more and more false positives are being mixed into the list. This graph is shown in Fig. 4. There is a plateau of around 4% and this percentage increases rapidly for similarity values greater than 0.970. Two other possible cut-offs are 0.960 or 0.980, each with a different compromise of false negatives *versus* false positives. For a similarity cut-off of 0.970, the combined number of false

[1] Supplementary data for this paper are available from the IUCr electronic archives (Reference: NS5003). Services for accessing these data are described at the back of the journal.

positives and unflagged polymorphs is 5.2%, suggesting that each contributes approximately 3%. In other words, we estimate that this cut-off introduces approximately 3% of false positives (and an unknown number of false negatives) and that approximately 3% of the polymorphs in the CSD are currently unflagged, *i.e.* 97% of the polymorphs in the CSD are currently flagged as such.

The automatically generated lists of all pairs of polymorphs in the CSD based on the cut-off value of 0.970 are the second and third list in the supplementary material. We would like to stress that these lists are automatically generated and will contain both false negatives and false positives. An example of part of the lists is shown in Fig. 5.

List 3 contains 7300 pairs of polymorphs, or 14 600 polymorphic entries, with an unknown number of duplicates (redeterminations of the same pair of polymorphs, see the caption to Fig. 5), an unknown number of multiple polymorphs of the same compound and an unknown number of false negatives. Of these, 3142 pairs or 6284 entries are up to the high quality standards that we set. It is difficult to compare our numbers to the number of polymorphic compounds found in the CSD by Bernstein in 2002, as described in chapter 7.2.1 of his book (Bernstein, 2002), because the aims of the two searches were so different. Bernstein's aim was to get a feel for the frequency of polymorphism and he set out to find all compounds known to be polymorphic, regardless of the quality of the crystal structure determinations. Our aim was to compile a list of reliable pairs of polymorphs and we have not taken any precautions to reduce the number of false negatives. With those restrictions in mind, we find that list 3 contains 2862 unique refcode families ('chemical compounds'), to be compared with 303 733 chemical compounds in the whole CSD, giving an estimate of 1%. List 2 contains 1625 unique refcode families, from which we estimate that for only 0.5% of all chemical compounds in the CSD the crystal structures of at least two polymorphs have been determined.

```
ABIKEB     ABIKEB01
ABUCUP     ABUCUP01
ACACCE01   ACACCE03
ACACMN     ACACMN21
ACACMN02   ACACMN21
ACACRU     ACACRU02
ACACRU02   ACACRU03
ACBNZA     ACBNZA01
ACEDAN     ACEDAN01
ACEMID01   ACEMID03
ACEMID01   ACEMID05
ACEMID01   ACEMID06
ACEMID02   ACEMID06
ACEMID03   ACEMID06
ACEMID05   ACEMID06
etc.
```

Figure 5
Example of polymorph pair entries in list 2 of the supplementary material. The pairs ACACMN/ACACMN21 and ACACMN02/ACACMN21 are duplicates: ACACMN and ACACMN02 are the same structure. ACEMID01/ACEMID03 and ACEMID01/ACEMID05 are false positives due to a very large temperature difference of 270 K. There are several false negatives, but these are all caused by entries being rejected by our strict criteria, not by problems with the similarity measure.

Comparison with the similarities of the full crystal structures (not just the reduced unit cells) and the hand-checked list indicates that similarities > 0.990 generally point to the two crystal structures being the same. This criterion can be used to detect and remove pairs of duplicates to reduce the number of pairs of polymorphs for paracetamol, for example, from 56 to one. This would also enable us to determine which compounds have more than two polymorphs.

Similarity values between 0.970 and 0.990 are a grey area, caused by problematic pairs, *e.g.* due to large temperature differences.

Some interesting cases of the similarities of pairs of crystal structures will now be discussed.

3.1. Same crystal structure, different molecules: MEACCU01 (Wu *et al.*, 1996) and MEACCU02 (Koval *et al.*, 2003)

This is an example of two crystal structures that are not polymorphs, even though various algorithms might have come to that conclusion. MEACCU01 is a room-temperature study, whereas MEACCU02 was determined at 110 K. The unit cell contracts anisotropically, with a and b almost the same, but c changing by 0.3 Å (1.7%). The authors of the second determination chose to describe the structure as a polymer and there is no longer a one-to-one correspondence between the molecules in the two unit cells (even though the crystal structures are the same). Even the chemical formulae of these two structures are different: $C_{12}H_{24}Cu_4O_{12}$ and $(C_6H_{12}Cu_2O_6)_n$. Without volume correction, the similarity of the powder patterns is 0.964 and the similarity of the reduced unit cells is 0.992. After applying the volume correction, the similarity of the powder patterns is 0.990 and the similarity of the reduced unit cells is 0.997.

3.2. Pseudo-symmetry and misprint: DOTJEB (Gerard *et al.*, 1986) and DOTJEB01 (Korte *et al.*, 1988)

DOTJEB was reported in the space group $P2_1$ and DOTJEB01 in $C222_1$. Their unit-cell parameters are also substantially different and the similarity of their powder diffraction patterns is only 0.944. However, closer inspection shows that the x coordinate of one of the Cl atoms in DOTJEB01 is missing a minus sign in the original publication, which was faithfully copied into the CSD. After correction, the similarity of the powder diffraction patterns is 1.000. This

turns out to be a case of pseudo-symmetry, the two unit cells being related by the transformation matrix

$$\begin{bmatrix} 1 & 0 & 0 \\ 2 & 0 & -2 \\ 0 & 1 & 0 \end{bmatrix}.$$

The similarity of their reduced unit cells is 1.000. Obviously, excluding the reflections generated by the unit-cell centring is crucial to detect this similarity.

3.3. Extremely anisotropic unit-cell contraction: BIJWAS01 (Rogers & Green, 1986) and BIJWAS02 (Rogers & Richards, 1987)

This is an example of an extremely anisotropic unit-cell contraction, making it very difficult to automatically detect that these two crystal structures are the same (Fig. 6). For a temperature difference of 170 K, the contraction of the unit-cell parameters is −2.5, 5.1 and 2.8% for *a*, *b* and *c*, respectively; note that the **a** axis *contracts* with increasing temperature. Without volume normalization, the similarities of the full powder patterns and the reduced unit cells are 0.905 and 0.972, respectively. After volume normalization, these values become 0.938 and 0.979, respectively: clearly, this pair of structures falls in the grey area, where our algorithm is not able to tell if the two crystal structures are the same or not. However, it should be noted that such extremely anisotropic contractions are rare and that in spite of the extreme anisotropic contraction, the similarities are still higher when the volume normalization is included than when it is not.

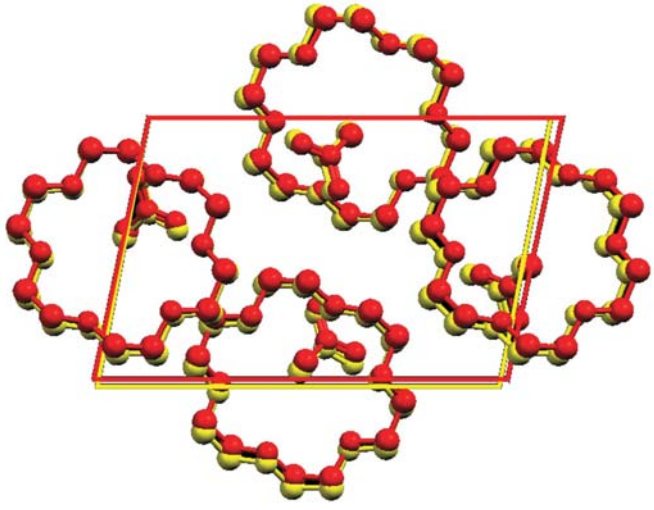

Figure 6
The crystal structure of 18-crown-6 bis(nitromethane) clathrate at room temperature (red) and at 123 K (yellow). The two crystal structures are clearly the same, but note how at lower temperature the crystal contracts in one direction, whereas it expands in the other.

3.4. Isostructural compounds: JAPRUN and JAPTAV (both from Schebler *et al.*, 1998)

These are two isostructural compounds. JAPRUN contains cobalt, JAPTAV contains nickel: only one electron difference. These are different compounds and therefore they are neither the same crystal structure nor polymorphs. The similarity of their powder patterns is 0.998, the similarity of their reduced unit cells is 0.999.

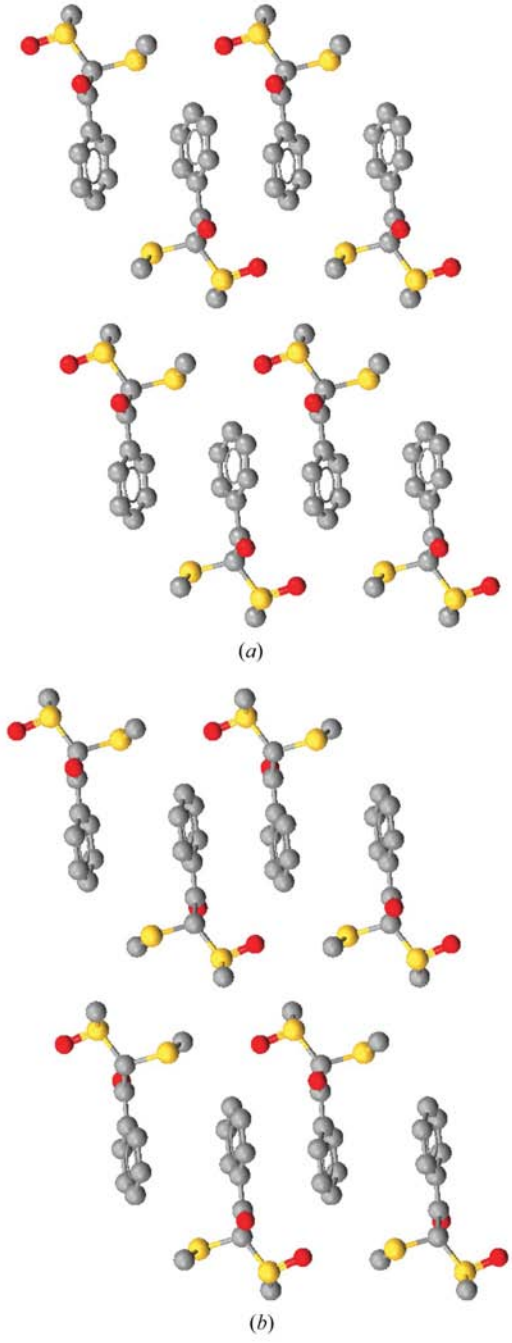

Figure 7
View along the **b** axis of two crystal structures of α-methylthio-α-methylsulfinylacetophenone; (*a*) optically pure (TITNER), (*b*) racemate (TITNEP).

3.5. Missed symmetry: ANTCEN14 (Brock & Dunitz, 1990) and ANTCEN17 (Marciniak & Pavlyuk, 2002)

The two crystal structures are clearly identical, with a similarity of 1.000, but the authors of ANTCEN17 chose to describe the crystal structure as $Z' = 1$ rather than $Z' = \frac{1}{2}$. The result is that the unit cell is twice as large (but still primitive) and the similarity of the reduced unit cells is therefore only 0.841. The authors of ANTCEN17 claim, based on a very weak symmetry-breaking reflection, that their crystal structure is a polymorph and it is therefore flagged as such in the CSD. Inspection of the anisotropic displacement parameters, valence angles and bond lengths, however, suggests numerical instabilities in the refinement. Rather than being a polymorph, it is more likely that ANTCEN17 should have been described as $Z' = \frac{1}{2}$ like ANTCEN14.

3.6. Chiral compounds: TITNEP (Distefano *et al.*, 1996) *and TITNER* (Wladislaw *et al.*, 1999)

TITNER is optically pure, TITNEP is the corresponding racemate. As such, they are two different compounds. When projected onto the **ac** plane, however (Fig. 7), the two crystal structures appear to be identical. (In principle, this can be verified quantitatively by calculating De Gelder's similarity measure for simulated powder diffraction patterns consisting of *h*0*l* reflections only, but that does not work very well in this case, and the similarity of the **ac** projection is only 0.927. Perhaps the crystal structures are not as similar as they at first appear to be, but this is hard to decide because one of the crystal structures has an *R* factor of 0.13.) On closer inspection, the carbonyl groups in TITNER are all pointing out of the **ac** plane, whereas in TITNEP their directions alternate. After applying the volume correction, the similarity of the powder patterns is 0.958 and the similarity of the reduced unit cells is 0.932.

4. Conclusions

For automated comparisons, comparing reduced unit cells is more reliable for detecting if two crystal structures are different or not than is comparing the simulated powder patterns of the full crystal structures. The combination of volume normalization based on Hofmann's average atomic values (Hofmann, 2002) and the normalized similarity measure published by De Gelder *et al.* (2001) is crucial for the elimination of the problems associated with unit-cell differences due to temperature and pressure. After these corrections, similarities over 0.990 indicate a redetermination, similarities below 0.970 indicate two polymorphs. The number of false negatives is unknown; the number of false positives is around 3%.

The total number of pairs of polymorphs in the CSD is probably around 7300, based on list three, among these an unknown number of duplicates. Of these, 3142 pairs (6284 entries) satisfy our criteria regarding three-dimensional atomic coordinates, reasonable unit-cell volumes and matching chemical formulae. We estimate that for only 0.5% of all the chemical compounds in the CSD have the full crystal structures of at least two polymorphs been determined. Approximately 97% of all clearly recognisable polymorphs (*i.e.* where at least two CSD entries with substantially different crystal structures are available) in the CSD are now flagged as such in the November 2004 release.

5. Possible future work

Efforts are on the way to compile a list of redeterminations in the CSD, so as to reduce statistical bias due to multiple inclusions of the same crystal structure, and to compile lists of crystal structures for which both a solvated (including hydrated) form and an unsolvated form exists.

Mrs S. Barrett is acknowledged for ensuring that all the newly discovered polymorphs from list 1 were changed in the CSD in time for the November 2004 release.

References

Allen, F. H. (2002). *Acta Cryst.* B**58**, 380–388.
Bernstein, J. (2002). *Polymorphism in Molecular Crystals.* Oxford: Clarendon Press.
Brock, C. P. & Dunitz, J. D. (1990). *Acta Cryst.* B**46**, 795–806.
Bruno, I. J., Cole, J. C., Edgington, P. R., Kessler, M., Macrae, C. F., McCabe, P., Pearson, J. & Taylor, R. (2002). *Acta Cryst.* B**58**, 389–397.
Ciajolo, M. R., Lelj, F., Tancredi, T. & Temussi, P. A. (1981). *Acta Cryst.* B**37**, 1130–1131.
De Gelder, R., Wehrens, R. & Hageman, J. A. (2001). *J. Comput. Chem.* **22**, 273–289.
Distefano, G., Dal Colle, M., de Palo, M., Jones, D., Bombieri, G., Del Pra, A., Olivato, P. R. & Mondino, M. G. (1996). *J. Chem. Soc. Perkin Trans. 2*, pp. 1661–1669.
Gerard, C., Lucken, E. A. C. & Bernardinelli, G. (1986). *J. Chem. Soc. Perkin Trans. 2*, pp. 879–884.
Hofmann, D. W. M. (2002). *Acta Cryst.* B**58**, 489–493.
Jerslev, B. (1987). *Acta Chem. Scand. B*, **41**, 184–193.
Kálmán, A., Fabian, L., Argay, G., Bernath, G. & Gyarmati, Z. C. (2004). *Acta Cryst.* B**60**, 755–762.
Korte, L., Mootz, D., Scherf, M. & Wiebcke, M. (1988). *Acta Cryst.* C**44**, 1128–1130.
Koval, I. A., Gamez, P., Roubeau, O., Driessen, W. L., Lutz, M. & Spek, A. L. (2003). *Inorg. Chem.* **42**, 868–872.
Marciniak, B. & Pavlyuk, V. (2002). *Mol. Cryst. Liq. Cryst.* **373**, 237–250.
Mathieson, A. M., Robertson, J. M. & Sinclair, V. C. (1950). *Acta Cryst.* **3**, 245–250.
Rogers, R. D. & Green, L. M. (1986). *J. Incl. Phenom. Macrocycl. Chem.* **4**, 77–84.
Rogers, R. D. & Richards, P. D. (1987). *J. Incl. Phenom. Macrocycl. Chem.* **5**, 631–638.
Schebler, P. J., Riordan, C. G., Guzei, I. A. & Rheingold, A. L. (1998). *Inorg. Chem.* **37**, 4754–4755.
Wladislaw, B., Marzorati, L., Biaggio, F. C., Vargas, R. R., Bjorklund, M. B. & Zukerman-Schpector, J. (1999). *Tetrahedron*, **55**, 12023–12030.
Wu, L.-P., Kuroda-Sowa, T., Maekawa, M., Suenaga, Y. & Munakata, M. (1996). *J. Chem. Soc. Dalton Trans.* pp. 2179–2180.