CRYSTALCMP: EXAMPLES OF USE

J. Rohlíček

Institute of Physics of the CAS Na Slovance 2, Prague 8, Czechia
rohlicek@fzu.cz

Keywords: molecular packing, comparison, similarity

Abstract

The program CrystalCMP can be used for comparison of crystal structures containing separated molecular objects as it is, for instance, in molecular crystals or some metal-organic compounds. The program can be useful in many different situations. This article shows the use of the program CrystalCMP on three basic examples for which the program was primarily intended.

Introduction

Probably every crystallographer experienced a situation where he needed to compare two or more crystal structures of organic or organometallic compounds. All the time that has elapsed in front of the monitor, when crystallographers have tried to find similarities between two crystal structures, is a silent witness to the fact that it is appropriate to master some tool for this purpose.

Several methods for comparing crystal structures have been published in the past [1-6]. Among the user-friendly tools, I would include the Crystal Packing Similarity tool in Mercury [7], and COMPSTRU [8], xPac [9] and CrystalCMP [10]. The Crystal Packing Similarity tool is only available to the user in the paid version of Mercury. The COMPSTRU program is created as an online tool on the Bilbao Crystallographic Server website [11] and the xPac and CrystalCMP programs can be freely downloaded.

The Crystal Packing Similarity in Mercury, xPac, and CrystalCMP tools use a similar approach to compare the packing of molecules in crystal structures. The programs select a representative molecular cluster for each crystal structure being compared. They then perform comparisons based on the differences in the positions of the molecules in both clusters. The individual implementations differ in the way how the two clusters are compared and in the speed of comparison. The comparison method used by COMPSTRU differs significantly from the other three methods. The program finds the best transformation of the unit cells of the compared structures and then compares the atomic positions.

As already mentioned, the CrystalCMP method is based on the comparison of a representative molecular cluster in which one type of molecule, usually the largest one, can be included. During the comparison, molecular clusters of the individual crystal structures are overlapped, and the similarity is calculated according to this formula:

$$Ps_{a,b} = D_c + wA_d \qquad (1)$$

where D_c is the average distance (in Å) of the centroids of the overlapping molecules and A_d represents the deviation of the rotation (in °) of the overlapped molecules in space. The value of w is chosen by the user and represents the weight between D_c and A_d.

By default, the difference in the rotation of molecules is more weighted ($w > 1$). This is because the same packing is not that much conditioned by the same position of the molecules in space, but rather by their similar rotation. The volume difference between compared crystal structures has only little effect on the change of the $Ps_{a,b}$ function. It is, therefore, possible to compare the packing similarity in crystal structures whose expansion is caused, for example, by temperature or by the presence of solvent molecules of different sizes, as it is in the so-called solvatomorphic series.

The comparison method in CrystalCMP has recently been improved [12]. First of all, the automatic procedure for selecting atoms, that are needed for overlapping the compared crystal structures was introduced. Secondly, the A_d term in the formula of the $Ps_{a,b}$ value representing the angular difference between molecules in related pairs was changed by calculation of the RMSD function, see more details in the recent publication about CrystalCMP [12]. Both changes led to the more user-friendly *black-box* method, which in few seconds can sort the list of crystal structures to the similarity groups just by clicking on one button.

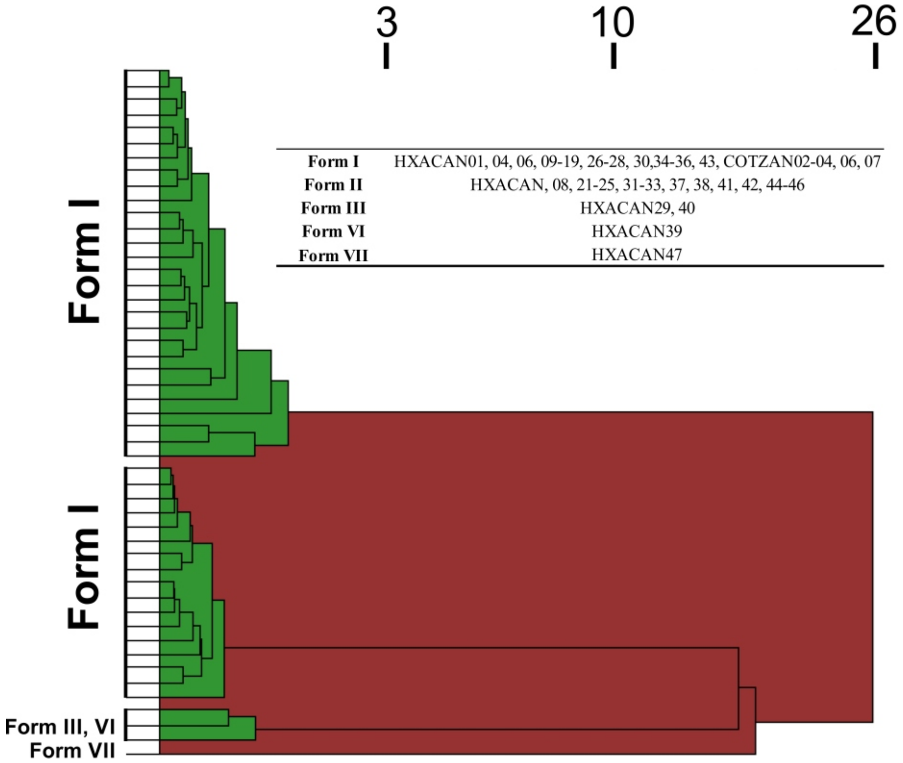
Examples of program use

The program can be used for comparison of crystal structures containing separated molecular objects as it is for example in molecular crystals or some metal-organic compounds. I believe, that the program can be useful in many different situations, however, I selected three problems for which the program was mainly designed – (i) identification of the same molecular packing independently of the temperature of the measurement, (ii) identification of the similar molecular packing in the different solvates of the same compound and (iii) identification of the same results in the large result list created during the crystal structure determination from the powder diffraction data by direct-space methods.

Every example has its specific requirement on the method. While the first and second examples require a low sensitivity to the small and relatively large expansions of the crystal structures, the third example requires the high speed and the possibility to process a high number of results.

Settings of the method

All the examples here were performed with the CrystalCMP of version 113 (2020/09/09) and with the Cambridge Structure Database (CSD) version 5.41 (November 2019) [13]. The calculation of similarity was done by automatic procedure with 14 surrounding molecules and with the weight factor $w = 3$. Only the packing similarity of the heaviest molecule in the crystal structure was studied.



Form I	HXACAN01, 04, 06, 09-19, 26-28, 30,34-36, 43, COTZAN02-04, 06, 07
Form II	HXACAN, 08, 21-25, 31-33, 37, 38, 41, 42, 44-46
Form III	HXACAN29, 40
Form VI	HXACAN39
Form VII	HXACAN47

Figure 1. Dendrogram calculated from the similarity matrix as a result of the packing comparison of paracetamol entries in CSD. The horizontal axis represents the $Ps_{a,b}$ value (in square root scaling). Individual entries are sorted and connected to the similarity groups on the vertical axis.

Example 1: How many polymorphs of paracetamol contain CSD?

According to the literature [14], there exist nine polymorphs of paracetamol: six ambient (**I**, **II**, **III**, **VII**, **VIII** and **IX**), two high pressure (**IV** and **V**) and one low temperature (**VI**). Crystal structure of only five of them has been determined either from the powder or from the single-crystal diffraction data. The (CSD) contains 56 entries with only paracetamol molecule in the crystal structure. All these entries are grouped in two ref. code groups that start either by COTZAN or by HXACAN codes. Three entries (COTZAN, COTZAN01 and COTZAN05) contains wrongly solved crystal structures because the bonding of several atoms in these crystal structures breaks basic rules of the organic chemistry. Additional four entries (HXACAN02, HXACAN03, HXACAN05 and HXACAN20) do not contain atomic coordinates. The rest of 49 entries with coordinates were used by CrystalCMP. The crystal packing similarity has been performed with default parameters and by using the automatic comparison mode and the overall computation time was around 30 s on a standard office PC.

After the comparison, the dendrogram shows that there are four distinguishable different molecular packings, see Fig. 1. The largest similarity group corresponds to the Form **I** with the monoclinic symmetry. There are significant differences in the unit cell volumes in this set of entries. The largest volume is 776.3 Å^3, and the smallest is 611.3 Å^3 corresponding to HXACAN01 and HXACAN43, respectively. No matter what is the reason for the difference, the comparison method was able to identify the same molecular packing despite a 126 % expansion of the unit cell.

The second-largest similarity group corresponds to the orthorhombic Form **II**. In this group of entries, the difference between the largest and smallest unit cell is not that significant and corresponds to the 104% expansion.

Entries HXACAN29, 39 and 40 creates the third similarity group. Their unit cells are similar, but they differ in space groups. While HXACAN29 and 40, both corresponding to the Form **III**, have the orthorhombic space group $Pca2_1$, the HXACAN39 entry has monoclinic space group $Pc11$. HXACAN39 entry is a low-temperature form of paracetamol (Form **VI**) and its high similarity to Form **III** was already described when it was published [15]. Its similarity to the Form **III** led to the original label of Form **III-m**.

The last entry HXACAN47 is not similar to any other entry in the CSD database and corresponds to the Form **VII.**

Example 2: The similarity of trospium chloride solvates

The CSD contains 17 entries with trospium chloride with different solvates. The calculated dendrogram shows six different molecular packings of the trospium entity in the crystal structures, see Fig. 2. The largest similarity group contains either 5 or 7 entries depending on how the criterion for the similarity group would be defined. There are five highly similar solvates represented by (IPIKOJ, IPILIE, EZOLUC, IPILAW and KIXYUN) with acetonitrile, propionitrile, acetone, methanol and ethanol solvates. These entries are in the dendrogram connected at the similarity level of approx. $Ps_{a,b} = 8$ with two other entries (IPIKUP and IPILEA) containing isopropanol and nitromethane solvates. If we look at the overlapped clusters

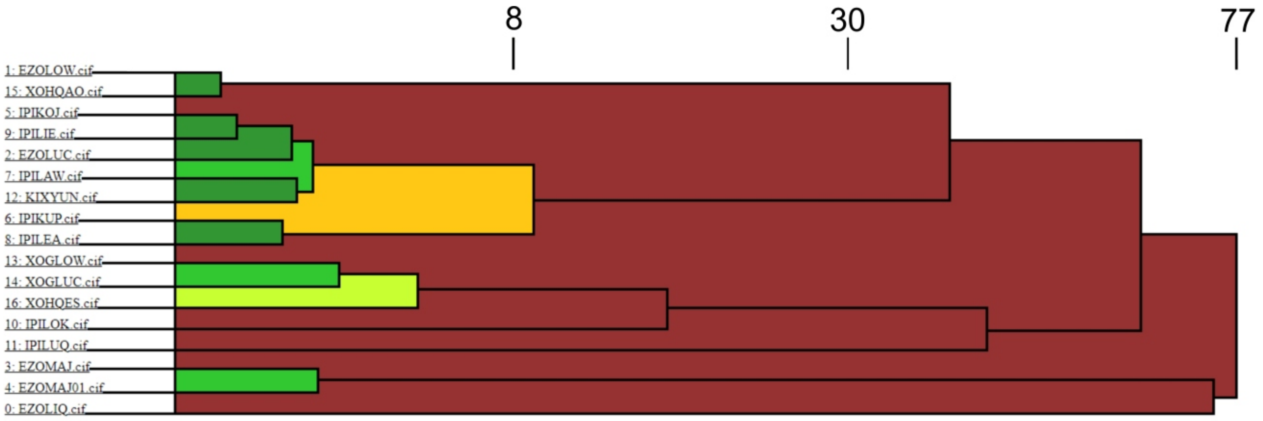


Figure 2.: Dendrogram represents a similarity of trospium chloride entries obtained from CSD. The horizontal axis represents the $Ps_{a,b}$ value (in square root scaling). Individual entries are sorted and connected to the similarity groups on the vertical axis.

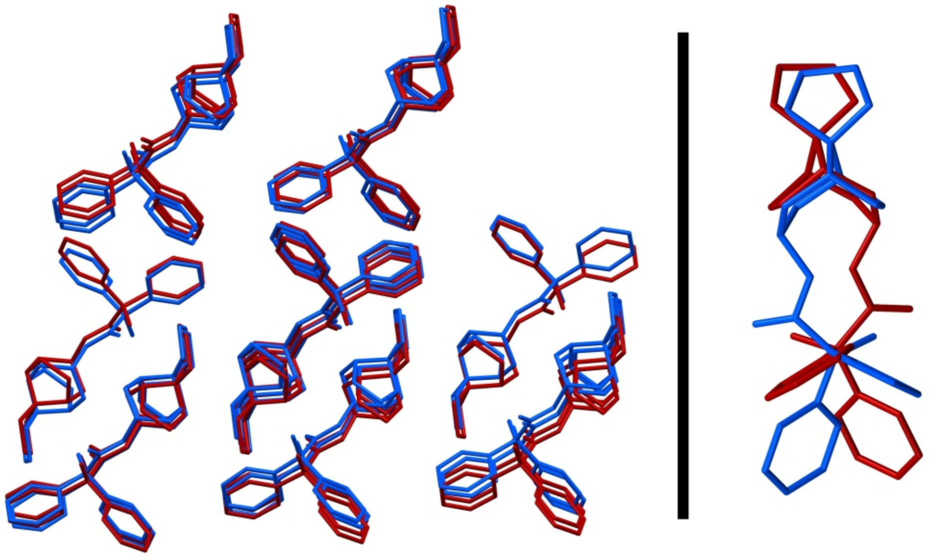


Figure 3. Comparison of molecular packing of the trospium entity in IPIKUP (red; isopropanol solvate) and IPILIE (blue; propionitrile solvate). Left – overlapped clusters show identical position s of trospium entities in the space. Right – Detail on the difference - almost perfectly mirrored every second trospium entity in the molecular clusters.

of these two groups in Fig. 3, we can conclude, that trospium entities are placed in the same positions, but every second molecule is mirrored.

The second-largest similarity group shows three entries (XOGLOW, XOGLUC and XOHQES) with glutaric acid, adipic acid and oxalic acid solvates. The fourth possible similar entry IPILOK (sesquihydrate) is connected with the similarity of approx. $Ps_{a,b}$ = 12. The overlapped clusters showed a similarity of the three entries (XOGLOW, XOGLUC and XOHQES), but the similarity with sesquihydrate can be only hardly found. In the dendrogram, the other two similarity groups contain only two entries: two solvates with benzoic and salicylic acid (EZOLOW with XOHQAO) and two saccharinate monohydrate (EZOMAJ with EZOMAJ01). The molecular packings of trospium entity in the other two entries EZOMAJ and EZOLIQ are not similar to any other entry in the list..

Example 3: How many different solutions are present in the result list?

Direct-space methods are often used for the crystal structure determination from powder diffraction data. These methods are based on global optimizations approaches. They are improving the initial model of the crystal structure by changing free torsion angles and fragment positions in the unit cell, and, in an optimal case, they find the correct solution. During this process, many possible solutions are generated, and one of them, usually the one with the best agreement factors, is taken to the final refinement process. Comparison of molecular packing in CrystalCMP gives an idea of how many different solutions is in the result list and what is the difference between them.

The advantages of knowing the composition of the result list will be shown by the example of crystal structure determination of capecitabine. During this process, two hundred results were generated and then only one hundred

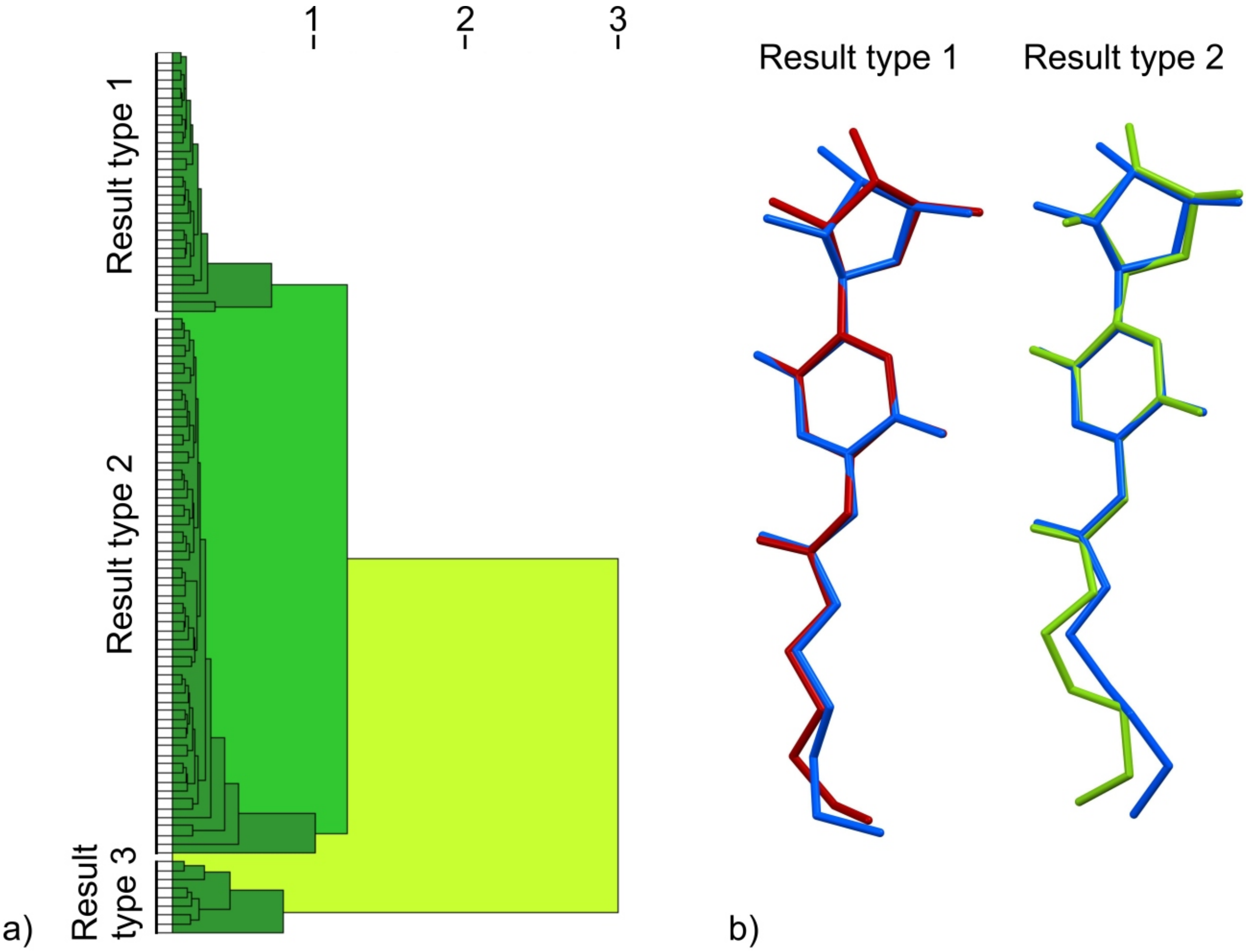


Figure 4. a) Comparison of the result list containing 100 results. The horizontal axis represents the $Ps_{a,b}$ value (linear scaling). Individual entries are sorted and connected to the similarity groups on the vertical axis. The dendrogram shows three similarity groups. b) Comparison of result type 1 (red) and result type 2 (green) with the individual disordered parts (blue) of the final refined crystal structure (BOVDUM).

with the best agreement factors were analyzed by CrystalCMP. The comparison took approximately 2 minutes, and it sorted results to the three similarity groups. Two of them contains results with similar agreement factors (Result type 1 and Result type 2). In the third group (Result type 3), results have a little bit higher agreement factors, see Fig. 4a. Individual results inside every similarity groups are almost identical. From the comparison of the Result type 1 and Result type 2 representatives, it is evident that they differ mostly by the different positions of the alkyl chain, see Fig. 4b. Since these two groups contain results with the comparable agreement factors, the representative result of both these groups have to be taken to the final refinement. In the end, the refinement revealed a disorder of the alkyl chain, and both different solutions represent both positions of the disorder, see Fig 4b. The third similarity group of the solutions was not taken to the final refinement due to its higher agreement factors.

Conclusions

The selected three examples illustrate the possible use of CrystalCMP and the benefits of using this program in various situations. The comparison process is fast and user-friendly. User can use this software as a black box with the default settings to get valuable results in a short time. As it was shown, the result of the comparison by the CrystalCMP program is a similarity matrix and a dendrogram which groups individual entries according to the similarity in a readable form.

The program is written in C / C ++, uses OpenBabel libraries [16] to generate SMILES definitions and uses wxWidgets and OpenGL for the graphical interface. The program is free to download at http://sourceforge.net/projects/crystalcmp/, where its source code can also be found.

1. de Gelder, R.; Wehrens, R.; Hageman, J.A. A generalized expression for the similarity of spectra: application to powder diffraction pattern classification. *J. Comput. Chem.* **22** (2001) 273–289.

2. Hundt, R.; Schön, J.C.; Jansen, M. CMPZ – an algorithm for the efficient comparison of periodic structures. *J. Appl. Crystallogr.* **39** (2006) 6–16.

3. Karfunkel, H.R.; Rohde, B.; Leusen, F.J.J.; Gdanitz, R.J.; Rihs, G. Continuous similarity measure between nonoverlapping X-ray powder diagrams of different crystal modifications. *J. Comput. Chem.* **14** (1993) 1125–1135.

4. Valle, M.; Oganov, A.R. Crystal fingerprint space – a novel paradigm for studying crystal-structure sets. *Acta Crystallogr.* **A66** (2010) 507–517.

5. Van Eijck, B.P.; Kroon, J. Fast clustering of equivalent structures in crystal structure prediction. *J. Comput. Chem.* **18** (1997) 1036–1042.

6. Willighagen, E.L.; Wehrens, R.; Verwer, P.; de Gelder, R.; Buydens, L.M.C. Method for the computational comparison of crystal structures. *Acta Crystallogr.* **B61** (2005) 29–36.

7. Chisholm, J.A.; Motherwell, S. COMPACK: a program for identifying crystal structure similarity using distances. *J. Appl. Crystallogr.* **38** (2005) 228–231.

8. de la Flor, G.; Orobengoa, D.; Tasci, E.; Perez-Mato, J.M.; Aroyo, M.I. Comparison of structures applying the tools available at the Bilbao Crystallographic Server. *J. Appl. Crystallogr.* **49** (2016) 653–664.

9. Gelbrich, T.; Threlfall, T.L.; Hursthouse, M.B. XPac dissimilarity parameters as quantitative descriptors of isostructurality: the case of fourteen 4,5'-substituted benzenesulfonamido-2-pyridines obtained by substituent interchange involving CF3/I/Br/Cl/F/Me/H. *CrystEngComm* **14** (2012) 5454.

10. Rohlicek, J.; Skorepova, E.; Babor, M.; Cejka, J. CrystalCMP: an easy-to-use tool for fast comparison of molecular packing. *J. Appl. Crystallogr*. **49** (2016) 2172–2183.

11. Aroyo, M.I.; Kirov, A.; Capillas, C.; Perez-Mato, J.M.; Wondratschek, H. Bilbao Crystallographic Server. II. Representations of crystallographic point groups and space groups. *Acta Crystallogr.* **A62** (2006) 115–128.

12. Rohlíček, J.; Skořepová, E. CrystalCMP?: automatic comparison of molecular structures. *J. Appl. Crystallogr.* **53** (2020) 841–847.

13. Groom, C.R.; Bruno, I.J.; Lightfoot, M.P.; Ward, S.C. The Cambridge Structural Database. *Acta Crystallogr. B* Struct. Sci. Cryst. Eng. Mater. **72** (2016) 171–179.

14. Shtukenberg, A.G.; Tan, M.; Vogt-Maranto, L.; Chan, E.J.; Xu, W.; Yang, J.; Tuckerman, M.E.; Hu, C.T.; Kahr, B. Melt Crystallization for Paracetamol Polymorphism. *Cryst. Growth Des.* **19** (2019) 4070–4080.

15. Reiss, C.A.; Mechelen, J.B. van; Goubitz, K.; Peschar, R. Reassessment of paracetamol orthorhombic Form III and determination of a novel low-temperature monoclinic Form III-m from powder diffraction data. *Acta Crystallogr. Sect. C Struct. Chem.* **74** (2018) 392–399.

16. O'Boyle, N.M.; Banck, M.; James, C.A.; Morley, C.; Vandermeersch, T.; Hutchison, G.R. Open Babel: An open chemical toolbox. *J. Cheminformatics* **3** (2011) 33.

The crystallographic part was supported by the project LO1603 under the Ministry of Education, Youth and Sports National sustainability programme I of Czech Republic