Journal of
Applied
Crystallography
ISSN 0021-8898

# *COMPANG*: automated comparison of orientations

Ludmila Urzhumtseva and Alexandre Urzhumtsev

Journal of
Applied
Crystallography

ISSN 0021-8898

Received 13 May 2002
Accepted 5 July 2002

# *COMPANG*: automated comparison of orientations

Ludmila Urzhumtseva and Alexandre Urzhumtsev*

LCM3B, UMR 7036 CNRS, Faculté des Sciences, Université Henry Poincaré, Nancy I, 54506 Vandoeuvre-lés-Nancy, France. Correspondence e-mail: sacha@lcm3b.uhp-nancy.fr

A search for similar orientations of a three-dimensional object is a usual task in structure analysis. An example is a comparison of peaks of rotation functions in molecular replacement. An automated comparison of orientations defined as a list or several lists of corresponding Eulerian angles can be performed using the interactive program *COMPANG*. When calculating the closeness of orientations, this program allows one to take into account the symmetry operations of the space group as well as non-crystallographic symmetry. The similarity of orientations can be considered at a given accuracy, thus allowing a user to identify the groups of close orientations, *i.e.* their clusters. The size of such clusters can be used as a criterion to choose the correct orientation in difficult cases of molecular replacement.

## 1. Introduction

Different orientations of a three-dimensional object can be presented by a list of rotation parameters, for example, by triplets of Eulerian angles. Recognition of similar orientations may be required in various applications, in particular, in macromolecular crystallography. For example, in molecular replacement (Rossmann, 1972, 1990) this can be useful in order to recognize the peaks of the rotation function linked by non-crystallographic symmetry. More importantly, in difficult cases of molecular replacement, many rotation functions are available simultaneously and a comparative analysis of their peaks allows a user to find an orientation, suggested by one function, in the list of peaks for other functions. However, such a comparison of rotation parameters is difficult to perform manually, especially for crystals with non-trivial symmetries. The interactive program *COMPANG* has been developed to automate such comparisons.

## 2. Comparison of multiple orientations

### 2.1. Basic points in automatic comparison of orientations

A manual comparison of orientations defined by their parameters, for example by triplets of Eulerian angles, is difficult for several reasons, including: (i) close orientations can correspond to two triplets of Eulerian angles which are quite different at first sight; (ii) close orientations can be presented by symmetry-related sets of parameters, a transformation of which by symmetry is not simple.

A representation of orientations in other coordinates, such as polar angles or direction cosines [different systems of rotation description in crystallography have been discussed by Urzhumtseva & Urzhumtsev (1997)], does not really simplify such an analysis.

To cope with this difficulty, an algorithm for automatic comparison of various orientations has been developed. All orientations (for example, peaks coming from several rotation functions) are taken together and the orientations which are closer to one another than a chosen threshold $D_{\min}$ are considered 'to be the same within the given limit'.

### 2.2. Main steps of the automatic comparison

The input information for such a comparison of orientations is presented as lists of triplets of Eulerian angles read from one or from several files. These lists can correspond to different initial orientations of the object; in the latter case the relative transition from one initial orientation to another should also be defined. The principal steps of the automated comparative procedure are the following.

The joint list of orientations is prepared from individual lists; the orientations in this joint list correspond now to the same common starting orientation of the objects, even if their initial orientations were different in different lists.

For each pair of orientations in this joint list, a 'distance' between them is calculated, taking symmetry operations into account.

In order to determine the mutual arrangement of orientations, a clustering procedure is applied that uses the matrix of 'distances' calculated in the previous step.

For a chosen distance limit, the orientations which are closer to one another than this value (the cluster threshold) are considered to be identical, and the groups (clusters) of such close orientations are defined; naturally, the number and the composition of clusters vary with this cluster threshold.

Two optional steps can follow if requested.

Firstly, the number of orientations inside each cluster is calculated; the diagram of the size of clusters is displayed.

Secondly, the rotation parameters corresponding to the chosen cluster are provided in the same rotation convention as the original molecular-replacement package uses and can be converted to the rotation matrix either by the same package or, for example, by *CONVROT* (Urzhumtseva & Urzhumtsev, 1997).

### 2.3. Use of cluster analysis for orientation comparison

A study of the distribution of orientations and the search for their groups can be carried out by cluster analysis, which is based on the notion of 'distance' between two orientations. Any orientation is represented by rotation parameters that define the corresponding rotation matrix $M$ to be applied to the search model. For two such matrices $M_m$ and $M_n$, there are many ways to define such a 'distance' between them. For example, a root-mean-square (r.m.s.) difference between all matrix elements can be taken (Brünger, 1990). The current version of *COMPANG* uses a distance expressed through the effective rotation angle $\kappa$ between corresponding orientations.

Such an angle $\kappa$ is calculated using the matrix $M_mM_n^{-1}$ of the rotation from one of these orientations, $M_n$, to another, $M_m$:

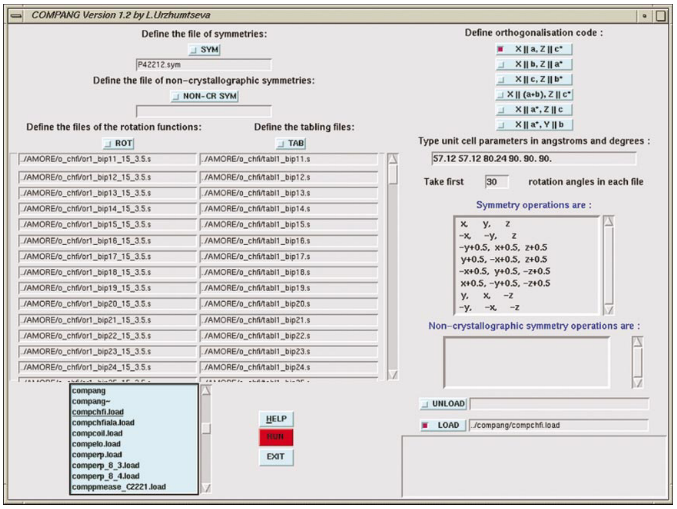
$$\kappa = \arccos\{[\operatorname{trace}(M_mM_n^{-1}) - 1]/2\}.$$

In the simplest case, the distance $D$ between two orientations is defined to be equal to this angle $\kappa$. The chosen distance 'does not know' the search model (or models) and does not take its (their) shape or atomic composition into consideration; it acts directly with the results of the rotation search carried out previously by other programs.

If the space group contains several symmetry operations, all symmetry-related orientations are equivalent and the distance between two orientations is defined as the minimal distance calculated for all symmetry-related pairs of these orientations. When a non-crystallographic rotation is present in the crystal and its order and the direction of the axis are known, this operation can also be considered at this step of the distance calculation.

After the distances are calculated, the orientations are merged one by one into clusters starting from the closest ones. In the developed procedure, the distance between two clusters is defined as the minimal distance between all pairs of orientations, one from cluster one and the second from cluster two. Such a choice is made in order to obtain a high-speed computational algorithm in comparison with other possible approaches to the definitions of the distance between clusters.

The process of formation of clusters is reflected in a cluster tree. The tree shows the closest orientations as neighbours along the abscissa axis; this needs to rearrange the orientations in their joint list. Merging of two orientations (or two clusters) is shown by the intersection of lines issued from the neighbouring points. The closer are the orientations to be merged, the lower is the point of this intersection.



COMPANG Version 1.2 by L.Urzhumtseva

Define the file of symmetries:
SYM
P42212.sym

Define the file of non-crystallographic symmetries:
NON-CR SYM

| Define the files of the rotation functions: ROT | Define the tabling files: TAB |
|---|---|
| ./AMORE/o_chfi/or1_bip11_15_3.5.s | ./AMORE/o_chfi/tabl1_bip11.s |
| ./AMORE/o_chfi/or1_bip12_15_3.5.s | ./AMORE/o_chfi/tabl1_bip12.s |
| ./AMORE/o_chfi/or1_bip13_15_3.5.s | ./AMORE/o_chfi/tabl1_bip13.s |
| ./AMORE/o_chfi/or1_bip14_15_3.5.s | ./AMORE/o_chfi/tabl1_bip14.s |
| ./AMORE/o_chfi/or1_bip15_15_3.5.s | ./AMORE/o_chfi/tabl1_bip15.s |
| ./AMORE/o_chfi/or1_bip16_15_3.5.s | ./AMORE/o_chfi/tabl1_bip16.s |
| ./AMORE/o_chfi/or1_bip17_15_3.5.s | ./AMORE/o_chfi/tabl1_bip17.s |
| ./AMORE/o_chfi/or1_bip18_15_3.5.s | ./AMORE/o_chfi/tabl1_bip18.s |
| ./AMORE/o_chfi/or1_bip19_15_3.5.s | ./AMORE/o_chfi/tabl1_bip19.s |
| ./AMORE/o_chfi/or1_bip20_15_3.5.s | ./AMORE/o_chfi/tabl1_bip20.s |
| ./AMORE/o_chfi/or1_bip21_15_3.5.s | ./AMORE/o_chfi/tabl1_bip21.s |
| ./AMORE/o_chfi/or1_bip22_15_3.5.s | ./AMORE/o_chfi/tabl1_bip22.s |
| ./AMORE/o_chfi/or1_bip23_15_3.5.s | ./AMORE/o_chfi/tabl1_bip23.s |
| ./AMORE/o_chfi/or1_bip24_15_3.5.s | ./AMORE/o_chfi/tabl1_bip24.s |

compang
compang~
compchfi.load
compchfiala.load
compcoil.load
compelo.load
comperp.load
comperp_8_3.load
comperp_8_4.load
comppmease_C2221.load

Define orthogonalisation code :
■ X || a, Z || c*
□ X || b, Z || a*
□ X || c, Z || b*
□ X || (a+b), Z || c*
□ X || a*, Z || c
□ X || a*, Y || b

Type unit cell parameters in angstroms and degrees :
57.12 57.12 80.24 90. 90. 90.

Take first 30 rotation angles in each file

Symmetry operations are :

```
x,       y,       z
-x,      -y,      z
-y+0.5,  x+0.5,   z+0.5
y+0.5,   -x+0.5,  z+0.5
-x+0.5,  y+0.5,   -z+0.5
x+0.5,   -y+0.5,  -z+0.5
y,       x,       -z
-y,      -x,      -z
```

Non-crystallographic symmetry operations are :

□ UNLOAD
■ LOAD ./compang/compchfi.load

**Figure 1**
First screen: input data. The left side of the screen contains the file names for the lists of orientations (rotation functions) and eventually for the corresponding 'tabling' files. File names can be typed directly or chosen by browsing the directory contents. The files with crystallographic and non-crystallographic symmetry operations can be defined in the same way. At the right, the orthogonalization options are displayed, followed by the unit-cell parameters. The maximal number of selected orientations per list can also be defined here. Other windows display the symmetry operations and the program messages. This and all other figures display CHFI data (Behnke *et al.*, 1998).

In order to analyse the size of the clusters when requested, some cluster threshold $D_{\min}$ is chosen (which can be varied by user); all orientations which are closer than this distance are considered to belong to the same cluster (they indicate the same model orientation within the chosen accuracy). The sizes of the clusters are represented in a diagram in the same order as the clusters are found in the tree. The multiple-rotation-functions method (Urzhumtsev & Urzhumtseva, 2002) searches for the most populated cluster, which often corresponds to the correct answer.

An illustration of the different steps of *COMPANG* is given in the following section.

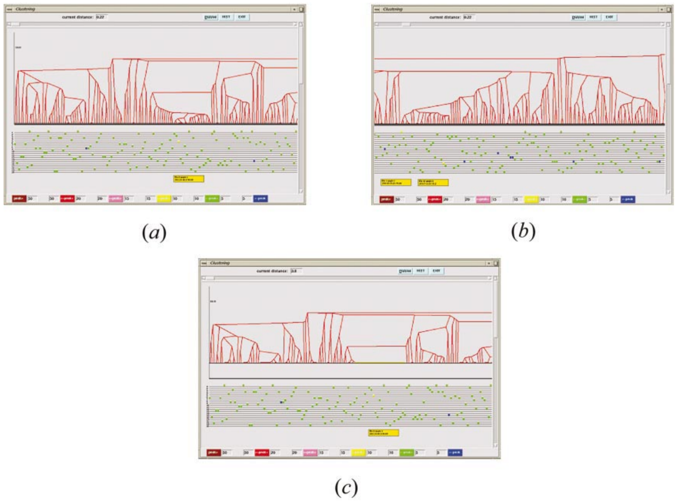


**Figure 2**
Second screen: cluster tree. A comparison of several rotation functions for CHFI data is illustrated. The left (*a*) and right (*b*) sides of the same cluster tree are shown in the case when the number of peaks is large. Parallel lines with squares below the cluster tree show the rotation peaks in different rotation functions, with their heights indicated by a colour code; the peaks are reordered to simplify the tree representation. Merging of two close peaks into a cluster is represented by two intersected lines in the tree. The height at which these lines intersect corresponds to the distance between these peaks: the closer the peaks, the lower the point of intersection. Yellow squares show the highest peaks. The yellow square in the left part of screen (*a*) indicates the correct orientation, while such squares in (*b*) indicate spurious peaks. Yellow frames below contain the values of the corresponding Eulerian angles. Screen (*c*) shows the same part of the screen as (*a*) after the cluster threshold $D_{\min}$ has been increased and the largest cluster chosen (indicated by a yellow horizontal line).

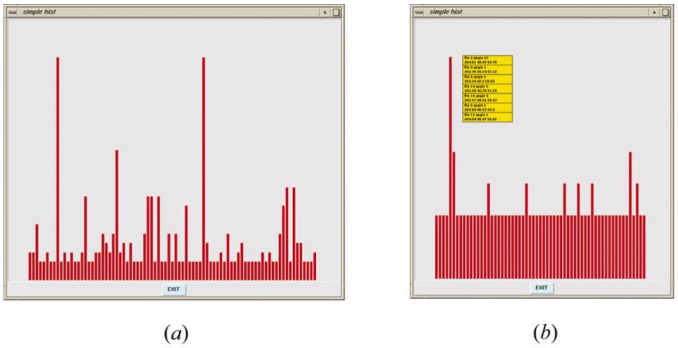


**Figure 3**
Third screen: diagram of cluster size. The diagram of cluster size is shown at two different cluster thresholds, applied for the cluster tree displayed in Fig. 2. The left-hand diagram (*a*) was calculated with $D_{\min} = 3.5°$, while diagram (*b*) was calculated with $D_{\min} = 2°$. The correct orientation corresponds to the largest cluster in (*b*) and to one of the two largest clusters in (*a*). The yellow frame in (*b*) contains the Eulerian angles for all peaks included in the largest cluster.

## 3. *COMPANG* presentation

### 3.1. First screen: input parameters

Input parameters for comparative analysis of lists of rotations are defined in the main screen (Fig. 1), which contains several windows with the following information.

(*a*) A list of files with orientations. A typical application of *COMPANG* is the study of the peaks of the rotation functions; therefore the formats used by the programs *CNS* (Brünger *et al.*, 1998), *AMoRe* (Navaza, 1994) and *MOLREP* (Vagin & Teplyakov, 1997) were chosen as appropriate for these files; both the independent version of *AMoRe* and that included in the *CCP4* suite (1994) are supported; for *CNS* and *MOLREP* formats, the initial orientation of the object should be the same. When the orientation search is performed using different objects, they should be optimally superimposed before the search.

(*b*) If the *AMoRe* format is chosen, the list of corresponding 'tabling' files should be given. These files contain information about possibly different initial model orientations. For the *CCP4*-based version, this information can be found in the same file as the rotation function; it will be read automatically from therein and the duplication of the file names is not necessary.

(*c*) Unit-cell parameters.

(*d*) Orthogonalization agreement. The PDB agreement (Bernstein *et al.*, 1977) is taken by default.

(*e*) A file with space-group symmetry operations presented in symbolic format.

(*f*) A file that contains the direction cosines and the rotation angle for each of the known non-crystallographic rotation axes if this information is available.

(*g*) The number of the orientations that will be taken consecutively from each list.

Once prepared, this input information can be saved in a file and can be recovered from it for the next program session by pressing the button 'load', thus avoiding the need to retype the input parameters. The parameterization of a non-crystallographic symmetry by the direction cosines of the rotation axis with respect to the orthogonal coordinate axes has been chosen in order to minimize possible confusion in its definition by different programs.

### 3.2. Second screen: cluster analysis

After the input information is prepared, the cluster analysis is started by pressing the button 'run'. Its result is shown in the second screen in the form of a cluster tree (Fig. 2). Below the tree, all individual orientations are represented in the form of coloured squares; each square is under the corresponding point in the cluster tree. Each line of squares represents the orientations from the same input list; the orientations are reordered to avoid the intersection of lines in the cluster tree. The colour of a square corresponds to the weight of the orientation; for the rotation function, this can be the peak height. The colours are variable and can be redefined by the user.

The cluster threshold $D_{\min}$ is represented by a horizontal line and is initially equal to zero (any individual orientation is a single-point cluster). The cluster threshold can be varied by a simple move of this line up or down, using a mouse; the composition of the clusters varies correspondingly in real time. All merged variants are indicated by a continuous horizontal interval with the length proportional to the number of included points (Fig. 2*c*).

The program displays the information relevant to each individual orientation: corresponding rotation angles as contained in the file and the prescribed weights (for example, the peak height for the peaks of the rotation function). For the *AMoRe* format, the 'absolute' or 'final' angle values obtained as a combination of the preliminary rotation ('tabling') and the current orientation ('roting') can be displayed as well. The type of displayed information can be switched by the button 'param'.

### 3.3. Third screen: cluster size diagram

In order to facilitate the analysis of the size of the obtained clusters of orientations, when this is requested, the corresponding diagram is shown in the third screen (Fig. 3) which appears when the button 'diag' (or 'hist' in older versions of the program) is pressed. When calculating the size of clusters, the orientations can be weighted with the prescribed values or be taken with unit weights. The size of each cluster that is different from a single point is represented by a bar with height proportional to the size of the cluster. The composition of a cluster can be shown by clicking the mouse on the corresponding bar. This produces a list of included orientations (their sequential numbers in the corresponding input file and the file number, and the values of their Eulerian angles) and simultaneously changes the colour of this cluster in screen two (Fig. 2*c*). Work with both these screens at a time facilitates the study of the situation.

## 4. Some possible applications

The program *COMPANG* can be used to resolve various problems.

First of all, the program allows one to analyse the list (or lists) of orientations. This can assist in the recognition of a given orientation in the list produced by some program. In particular, for known non-crystallographic rotation symmetry, this can identify the pairs of rotation peaks related by this symmetry.

The latter allows the identification of the parameters of the rotation function. When several rotation functions are calculated, a function that does not give peaks linked by the known non-crystallographic symmetry is rather suspicious and should be removed from further consideration. Inversely, for a given rotation function and a non-crystallographic symmetry, the absence of clusters at low $D_{\min}$ indicates some problems in the parameters of the non-crystallographic symmetry. A scanning around the initial position of the rotation axis can be performed in order to find the optimal direction that should group the rotation peaks by pairs (by triplets, *etc.*, depending on the order of the axis) at the lowest possible value of $D_{\min}$.

A more advanced use of *COMPANG* is for multiple-rotation-function analysis (Urzhumtsev & Urzhumtseva, 2002). When no individual rotation function gives a clear answer, the available rotation functions can be studied together. Tests with several difficult cases of molecular replacement, including those involving NMR models, proved that the largest cluster found by *COMPANG* using several rotation functions corresponds to the correct model orientation. Figs. 1–3 illustrate such an application of the program to the orientation search for the CHFI molecule (Behnke *et al.*, 1998), a difficult case for a conventional molecular-replacement analysis (Chen *et al.*, 2000). Rotation functions were calculated independently for 20 NMR models (Strobl *et al.*, 1995); none of them showed a clear signal. The *COMPANG* analysis (Fig. 2) at the cluster threshold $D_{\min} = 2^\circ$ gave a single large cluster, which unambiguously indicated the correct orientation (Fig. 3*b*).

Additionally, *COMPANG* can be used for model selection: when several models are tested, the program can exclude from further analysis all the models which do not contribute to major clusters of the rotation peaks.

## 5. Technical parameters

The program *COMPANG* is the latest program in the suite of Tcl/Tk-based crystallographic programs developed by our group (Urzhumtseva & Urzhumtsev, 1998, 2000). Its computationally heavy part is written in standard Fortran 77 and depends neither on the computer nor on the operating system. The calculation of the ‘distance matrix’ between orientations and cluster analysis takes a few seconds on a standard computer, even for several thousand orientations.

The interactive part and the block of data processing are written in Tcl/Tk (Ousterhout, 1993) and are also system independent. The distributed package includes a ‘readme’ file explaining how to install the program. Detailed ‘help’ is available at any time by pressing the corresponding program button and gives necessary comments on the use of the program. No special manual is necessary.

The program is available from the authors upon request.

## 6. Conclusion

Automated analysis of multiple orientations does not only facilitate the work of crystallographers, but makes it possible to resolve some problems of structure solution. In particular, such an analysis is important for the molecular-replacement method. The program *COMPANG* allows one to perform such a comparison of orientations easily and in real time. The program is universal and can be easily installed and run on any computer. This program is compatible with the formats of widely used molecular-replacement packages.

The authors thank L. Torlay for technical help, LIPHA (Lyon) for financial support, V. Y. Lunin for a very fruitful discussion of the project, D. Teller for experimental data for CHFI, and C. Lecomte for his support of the project.

## References

Behnke, C. A., Yee, V. C., Le Trong, I., Pedersen, L. C., Stenkamp, R. E., Kim, S.-S., Reeck, G. R. & Teller, D. C. (1998). *Biochemistry*, **37**, 15277–15288.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Brünger, A. T. (1990). *Acta Cryst.* A**46**, 46–57.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.

Chen, W., Kleywegt, G. & Dodson, E. (2000). *Structure*, **8**, 213–220.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.

Ousterhout, J. K. (1993). *Tcl and the Tk Toolkit.* New York: Addison-Wesley.

Rossmann, M. G. (1972). *The Molecular Replacement Method.* New York: Gordon and Breach.

Rossmann, M. G. (1990). *Acta Cryst.* A**46**, 73–82.

Strobl, S., Muhlhahn, P., Bernstein, R., Wiltscheck, R., Maskos, K., Wunderlich, M., Huber, R., Glockshuber, R. & Holak, T. A. (1995). *Biochemistry*, **34**, 8281–8293.

Urzhumtseva, L. & Urzhumtsev, A. (1997). *J. Appl. Cryst.* **30**, 402–410.

Urzhumtseva, L. & Urzhumtsev, A. (1998). *CCP4 Newsl. Protein Crystallogr.* **35**, 22–24.

Urzhumtseva, L. & Urzhumtsev, A. (2000). *J. Appl. Cryst.* **33**, 992–992.

Urzhumtsev, A. & Urzhumtseva, L. (2002). *Acta Cryst.* D. Submitted.

Vagin, A. & Teplyakov, A. (1997). *J. Appl. Cryst.* **30**, 1022–1025.