DD704731 **CISTI ICIST** CI-05870539-8

Document Delivery Service
in partnership with the **Canadian Agriculture Library**

Service de fourniture de Documents
en collaboration avec **la Bibliothèque canadienne de l'agriculture**

THIS IS NOT AN INVOICE / CECI N'EST PAS UNE FACTURE

ANTHONY ARTALE
MED LIB NATHAN CUMMINGS CTR (S-46)
MEMORIAL SLOAN KETTERING CANCER CTR
1275 YORK AVENUE
NEW YORK, NY 10021
UNITED STATES

ORDER NUMBER:	**CI-05870539-8**
Account Number:	DD704731
Delivery Mode:	ARI
Delivery Address:	arielsf.infotrieve.com/140.163.217.217
Submitted:	2005/11/21 16:26:42
Received:	2005/11/21 16:26:42
Printed:	2005/11/22 09:36:23

Direct **Periodical** **OPENURLOPAC** **UNITED STATES**

Client Number:	DDS36492/GANGI-DINO, RITA
Title:	**JOURNAL OF APPLIED CRYSTALLOGRAPHY.**
DB Ref. No.:	IRN10006370
ISSN:	ISSN00218898
Vol./Issue:	27
Date:	1994
Pages:	1006-1009
Article Title:	RECONSTRUCTION OF SYMMETRY RELATED MOLECULES FROM PROTEIN DATA
Article Author:	HOOFT, R, VRIEND G
Report Number:	IRN10006370
Publisher:	MUNKSGAARD INTERNATIONAL BOOKSELLERS AND PUBLISHERS.

Estimated cost for this 4 page document: $10.2 document supply fee + $26 copyright = $36.2

J. Appl. Cryst. (1994). **27**, 1006–1009

Reconstruction of Symmetry-Related Molecules from Protein Data Bank (PDB) Files

By Rob W. W. Hooft, Chris Sander and Gerrit Vriend

European Molecular Biology Laboratory, D-69012 Heidelberg, Germany

(*Received* 8 *April* 1994; *accepted* 4 *July* 1994)

Abstract

Many natural proteins are active as multimers. Crystallographic protein databases, however, generally store only part of the native multimer, the asymmetric unit, along with symmetry information. As a result of inaccuracies in the data, it is not always possible to reconstruct the native multimer. Here, a set of methods is presented that are designed to cope with inconsistencies in symmetry information. Applications include the validation of Protein Data Bank entries and the automatic generation of symmetry contacts for inspection and analysis.

Introduction

Study of protein multimers

Protein–protein interactions have a key role in many biological processes, such as intercellular signalling and antibody–antigen interactions. Many enzymes are functional as multimeric complexes only and their activity is dependent on interactions between the constituent protein monomers. Crystal structures as present in the Protein Data Bank (PDB), the primary source of macromolecular three-dimensional data, provide a good database for the study of protein–protein interactions. Many data sets contain multiple chains, so their interactions are explicit. In other cases, interactions in native multimers are only implicitly given in the form of symmetry relations, *e.g.* for hemoglobin or virus-coat proteins. Even when a molecule is not part of a multimer, *i.e.* active as a single protein chain, geometrical constraints on surface side-chain conformations are implicit in the symmetry relations from which the complete set of crystal contacts can be generated.

To be able to study these interactions, it is necessary to reconstruct the coordinates of symmetry-related neighboring molecules in the crystal structure. The PDB holds all information required to calculate the positions of symmetry-related molecules. However, administrative or scientific errors in the data sometimes prevent a study of the intermolecular interactions. Because of the redundancy in the data provided in a standard PDB file, it is often possible to detect such errors. The current paper makes a classification of the observed problems and describes tools that detect these problems and, if possible, allow corrections.

The work described was done in the context of the 'PDB verification project' funded by the European Commission. This project has as one of its goals to automate the process of data entry and validation in protein-structure databases.

Symmetry information in PDB files

The information required to generate a complete assembly and the direct neighbors in the crystallographic unit cell are stored in the PDB file in the so-called CRYST1, SCALE and MTRIX 'cards'.

The CRYST1 card in the PDB file contains the space-group name, the number of equivalent molecules in the unit cell (Z), the unit-cell axis lengths (a, b, c) and the interaxial angles (α, β, γ).

The SCALE cards contain a transformation matrix and a translation vector that together provide sufficient information to transform orthogonal coordinates into fractional coordinates (as needed for space-group symmetry transformations) (Rossman & Blow, 1962).

According to the PDB file-format specification, there is a one-to-one correspondence of the SCALE matrix and the unit-cell dimensions given on the CRYST1 card.

The inverse of the scale matrix (M_{sc}^{-1}) can be written in terms of the unit-cell dimensions and angles as

$$M_{sc}^{-1} = M_{\rm inv} \equiv \begin{pmatrix} a & b\cos\gamma & c\cos\beta \\ 0 & b\sin\gamma & -c\sin\beta\cos\alpha^* \\ 0 & 0 & c\sin\beta\sin\alpha^* \end{pmatrix} \quad (1)$$

$$\cos\alpha^* = (\cos\beta\cos\gamma - \cos\alpha)/(\sin\beta\sin\gamma) \quad (2)$$

$$\sin\alpha^* = (1-\cos^2\alpha^*)^{1/2}, \quad (3)$$

which provides the transformation between any set of fractional coordinates (p, q, r) and the corresponding set of orthogonal coordinates (x, y, z):

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = M_{\rm inv}\begin{pmatrix} p \\ q \\ r \end{pmatrix}. \quad (4)$$

Journal of Applied Crystallography
ISSN 0021-8898 © 1994

It is possible to describe the orientation of the cell with respect to the world coordinate system in the SCALE matrix. From any possible scale matrix, the cell parameters a, b, c, α, β and γ can be calculated but, if additional orientational information is expressed in the SCALE matrix (*i.e.* for any nonstandard scale matrix), the fractional coordinates can no longer be generated from the cell dimensions alone.

MTRIX cards in PDB files are used to describe noncrystallographic symmetry. They contain a matrix and a vector operating on orthogonal coordinates that can generate one molecular assembly from another.

Once the information regarding the symmetry is consistent, it is possible to generate coordinates for symmetry-related molecules and the structure can be checked for intermolecular atomic clashes ('bumps').

Computational procedures

Checking for administrative problems

To detect administrative errors in a PDB file, our procedure imposes a number of 'common sense' constraints on the values encountered. Some of the checks catch only exceptional cases, others are broadly applicable. These checks can all be performed without knowledge of atomic coordinates.

For the data present on the CRYST1 card, the following checks are performed:

(i) Is more than one CRYST1 card present? Give a warning.

(ii) Does the unit cell have the dimensions $a = 1$, $b = 1$, $c = 1$ Å, $\alpha = 90$, $\beta = 90$, $\gamma = 90°$? This is the adopted standard to indicate that a structure was determined by techniques other than X-ray crystallography. If this is the case, no symmetry operation is applicable.

(iii) Is any axis shorter than 2.0 Å? If so, the information on the CRYST1 card is rejected.

(iv) Is any cell angle smaller than 25° or larger than 155°? If so, the information on the CRYST1 card is rejected.

(v) Is the space-group name given? If not, assume that the space group is *P* 1 if the given cell is triclinic; otherwise, give an error message.

(vi) Is the space-group name a valid space group for a protein, *i.e.* contains no mirror planes, glide planes, inversion centers or $\bar{4}$ inversion axes? Is the character spacing and placement in the data file correct? If not, try to find the correct space group by applying heuristics (*e.g.* replace *R* 32 by *R* 3 2). If the correct space group cannot be determined unambiguously, use *P* 1 and give an error message.

(vii) Does the space-group name represent the standard setting as given in *International Tables for Crystallography* (Hahn, 1992)? If not, give a warning.

(viii) Is the lattice monoclinic, with the c axis unique, but the space group given as *P* 21? If yes, replace the space group by the nonstandard *P* 1 1 21 and give an error message.

(ix) Is the cell setting compatible with the symmetry transformations implied by the space group? If not (*e.g.* *R* 3 2 with a trigonal setting), the space-group symbol is rejected.

For the SCALE cards, the following checks are performed:

(i) Is more than one SCALE matrix present? If yes, give a warning.

(ii) Is the SCALE matrix the unit matrix? If the CRYST1 card gives the corresponding cube with vertices of 1 Å, this is the adopted standard for structures that were determined using techniques other than X-ray crystallography, so no symmetry is applicable. If a noncorresponding CRYST1 card is found, the SCALE matrix is discarded.

(iii) Is any value in the SCALE matrix larger than 0.5 (implying a cell axis shorter than 2 Å)? Is the determinant of the matrix either larger than 0.5 or equal to zero (implying a cell volume not within normal ranges)? In any of these cases, there is most probably a typing mistake in one or more of the values so the SCALE matrix is discarded.

(iv) Are any of the unit-cell angles as calculated for the SCALE matrix smaller than 25° or larger than 155°? In that case, there is most probably a typing error in a value so the SCALE matrix is discarded.

In case both the SCALE and CRYST1 information are present and individually valid according to all previous tests, the following tests are performed to check the consistency of the two:

(i) Do any of the cell parameters as calculated from the SCALE matrix differ by more than 0.05 Å or 0.05° from the values given on the CRYST1 card? If so, perform further checks to find the cause of the discrepancy. Otherwise, only check whether the SCALE matrix is given in accordance with the PDB specification and give a warning if this is not the case.

(ii) Are all cell parameters equal within 0.3 Å and 0.3° or the angles within 3° of 90 or 120°? In these cases, the difference is most probably caused by rounding of the parameters on the CRYST1 card, so these are replaced with the SCALE cell parameters.

(iii) Are exactly two of the three reciprocal-axis lengths and at least their enclosed angle equal in the two sets? If so, the difference is most likely caused by a mistyped value in the SCALE matrix because an error in one value in the SCALE matrix often corresponds to an error in three values in the cell dimensions.

(iv) Is the crystal class of the SCALE cell different from that of the CRYST1 cell? If so, accept the cell that agrees with the space group.

(v) If the only difference between the given SCALE matrix and the one calculated from CRYST1 as in (1) is that values are equal to 0.0 in one of the two and not in the other, that values are inverted in sign or that any values are off by a factor of 10.0, then most likely the SCALE matrix is wrong so the CRYST1 information is used.

For any MTRIX record, we verify whether the determinant of the matrix is equal to 1.0 and whether all pairs of columns are orthogonal. If for either of these this is not the case, a warning is given.

In any case where the automated procedure decides to use the cell specified on the CRYST1 card, the problem of the rotational information incorporated in the SCALE matrix has to be solved. This is done by extracting the rotational information (in the form of a proper rotation matrix M_r) from the given SCALE matrix using

$$M_{sc} = M_r * U \qquad (5)$$

with U an upper triangular matrix representing the SCALE cell dimensions in standard orientation. From M_r and the cell parameters from the CRYST1 card (upper triangular matrix M_{cr}), a corrected scale matrix (M_{new}) is calculated:

$$M_{new} = M_r * M_{cr}. \qquad (6)$$

Checking for scientific problems

After verification of symmetry information, checks on nonadministrative errors in coordinate data are performed. An example is errors in interpretation of electron density revealed by clashes between symmetry-related molecules.

The conventional cell

For all X-ray structures, a primitive cell with minimal axis lengths (Buerger reduced cell) (Buerger, 1957, 1960; Le Page, 1992) is calculated from the cell dimensions and the Bravais-lattice type of the cell as given on the CRYST1 card. Subsequently, the conventional cell is constructed by deriving the lattice symmetry from the distribution of twofold axes determined from coincident real-space and reciprocal axes (Le Page, 1982), allowing for a measurement error of 1.0°.

The given space group is now validated using the derived lattice symmetry. In cases where the lattice symmetry is higher than expected from the space group and more than one protein chain is present, a check for higher symmetry is performed. For this, the sequences of the protein chains are aligned in pairs and for each pair that has sufficient identity a three-dimensional superposition (Vriend & Sander, 1991) is performed, taking into account that terminal residues could be absent in one of the chains. If the superposition results in a r.m.s. deviation of less than 0.5 Å for the Cα atoms, the superposition matrix is checked for validity as a space-group transformation. If all of the protein chains can be superimposed on at least one other chain using a valid transformation, this is taken as an indication of missed symmetry.

If the lattice symmetry of the conventional cell and of the given cell are identical, the unit cell given on the CRYST1 card is checked for compliance with the conventions put forward in chapter 9 of *International Tables for Crystallography* (Hahn, 1992). If this is not the case, a transformation matrix is calculated that converts the given cell to the conventional cell.

Intermolecular clashes

An analysis of short intermolecular contacts can be performed using the reconstructed SCALE and CRYST1 symmetry (any MTRIX cards as well as any transformations that are present in a computer-unparsable way in the REMARK records of the PDB file are ignored at this stage; incorporating these could increase – in no case decrease – the number of problems located). A list is printed of all residues that are involved in excessively close intermolecular contacts (atomic clashes or 'bumps'). A bump is signalled whenever any intermolecular atom pair is closer than the sum of their respective van der Waals radii minus 1 Å. This criterion was chosen such that no bumps would be found in a number of 'good' reference protein structures. A 'severe bump' is signalled whenever two atoms are closer than the sum of their respective van der Waals radii minus 2 Å.

A number of cases remain where valid SCALE as well as CRYST1 cards are found but inconsistencies between the two are detected and cannot be unambiguously resolved by the rules described in the section on *Administrative problems*. In those cases, an analysis for close contacts is used to decide between the two. The analysis is performed twice: once using the information contained on the CRYST1 card and once using the SCALE transformation. If only one of the two reveals a severe collision between neighboring protein molecules, or either one reveals more than three times as many collisions than the other, this is taken to be a significant difference between the results obtained by the two analyses. If there are fewer than three 'severe bumps' in both cases, the same criterion is used to decide whether the total number of bumps is significantly different. If a significant difference is found, it is used to make a final decision on which information is to be chosen: CRYST1 or SCALE.

Results and discussion

From a study of 1996 structures and prerelease structures in the Protein Data Bank in October 1993

(Bernstein *et al.*, 1977), 76 structures (3.8%) can be identified as containing a total of 80 severe errors in either CRYST1 or SCALE cards. 65 of these errors can be corrected by our automatic procedure. Some examples are:

(i) In 54 cases, a serious discrepancy exists between the CRYST1 and SCALE information. 44 of these can be resolved.

(ii) In nine cases, the space-group name cannot be recognized. Seven of these can be corrected.

(iii) In eight cases, the setting of the unit cell is incompatible with one or more of the space-group transformations. This can be corrected for seven cases.

(iv) In nine cases, SCALE matrices or CRYST1 cards are missing (all of these concern nuclear-magnetic-resonance structures).

Other, less severe, problems located by our procedure are:

(i) 137 (6.9%) of PDB entries did not conform to the PDB specification because the SCALE matrix had an orientation other than that prescribed in the PDB file standard.

(ii) In 39 structures (2.0%), there is a significant difference between the SCALE and CRYST1 information, where it is clear that the CRYST1 card contains rounded values.

(iii) In 372 structures, the unit cell given could be identified as having a unit cell not conforming to the rules in *International Tables for Crystallography*: in these cases, either an equivalent unit cell with angles closer to 90° could be found or the conventional cell could be obtained by interchanging the cell axes.

(iv) 111 entries were identified as possibly having a higher lattice symmetry; in six of these, the protein chains were found to possibly obey this higher symmetry as well. We have informed the depositors of these structures of our findings.

(v) Four structures contain MTRIX cards (describing noncrystallographic symmetry) containing matrices that have a determinant that is significantly different from 1.000, four others contain MTRIX matrices with nonorthogonal vectors. Both of these problems will lead to deformations if a reconstruction of the symmetry-related partner(s) is attempted.

In 228 files, the CRYST1 cell is given as a cube with vertices of 1 Å; in 233 cases, the SCALE matrix is the unit matrix. In six structures, the SCALE cards were missing, in two of those, no CRYST1 card was present either. In two cases, the SCALE information was present more than once. In all, 232 structures contained neither a valid CRYST1 nor a valid SCALE, so no symmetry could be applied (most of these are nuclear-magnetic-resonance structures).

For 44 structures where the information on CRYST1 and the SCALE matrix was inconsistent and other methods could not decide which of the two was correct, an analysis for close contacts was performed. In seven cases, this analysis decided between the two.

In 297 out of 1533 entries where the symmetry could be deduced unequivocally from the CRYST1/SCALE cards, unrealistically close contacts were found between symmetry-related molecules. In many cases, this amounts to a single amino-acid side chain (*e.g.* Lys or Phe) colliding with a neighboring molecule. In other cases, complete loops of neighboring molecules actually occupy the same space.

The automated final analysis of these 1996 PDB entries took about 1.5 h of workstation CPU time.

Concluding remarks

Most commonly used programs use only the SCALE matrix, or only the CRYST1 information, to reconstruct the molecular aggregate by symmetry operations. The lack of consistency checks between the two can lead to errors. By combining a number of validation steps into a consistent set, we circumvent most of these errors. By correctly reconstructing molecular aggregates in most cases, our method can be a valuable tool for anyone studying interactions between protein molecules. It can also prove useful for producing updated versions of older database entries.

Availability

The procedure described has been incorporated as a module into the *WHAT IF* program and is available from the authors under the normal conditions for this program package (a nominal fee for the academic community). Information about possible problems in one's own entries can be obtained from us following requests by e-mail (hooft@embl-heidelberg.de) or regular mail. We also offer to run this check and our complete set of other checks on files that are to be submitted to the PDB.

References

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

Buerger, M. J. (1957). *Z. Kristallogr.* **109**, 42–60.

Buerger, M. J. (1960). *Z. Kristallogr.* **113**, 52–56.

Hahn, T. (1992). Editor. *International Tables for Crystallography*, Vol. A: *Space-Group Symmetry*. Dordrecht: Kluwer Academic Publishers.

Le Page, Y. (1982). *J. Appl. Cryst.* **15**, 255–259.

Le Page, Y. (1992). *J. Appl. Cryst.* **25**, 661–662.

Rossman, M. G. & Blow, D. M. (1962). *Acta Cryst.* **15**, 24–31.

Vriend, G. & Sander, C. (1991). *Proteins*, **11**, 52–58.