W718–W722 *Nucleic Acids Research, 2007, Vol. 35, Web Server issue*
doi:10.1093/nar/gkm225

CrystTwiV: a webserver for automated phase extension and refinement in X-ray crystallography

Trias Thireou[1], Vassilis Atlamazoglou[1], Manolis Levakis[1], Elias Eliopoulos[2], Athanassios Hountas[1], George Tsoucaris[3] and Kostas Bethanis[1,*]

[1]Physics Lab, Department of Science, [2]Department of Biotechnology, Agricultural University of Athens, 75 Iera Odos, Votanikos, Athens 118-55, Greece and [3]Centre de recherche de restauration des musées de France, C2RMF-U.M.R. 171 du C.N.R.S., Palais du Louvre, 75001 Paris, France

Received January 30, 2007; Revised March 20, 2007; Accepted March 28, 2007

ABSTRACT

An important stage in macromolecular crystallography is that of phase extension and refinement when initial phase estimates are available from isomorphous replacement or anomalous scattering or other methods. For this purpose, an alternative method called the twin variables (TwiV) method has been proposed. The algorithm is based on alternately transferring the phase information between the twin variable sets. The phase extension and refinement is evaluated with the crystallographic symmetry test by deliberately sacrificing the space-group symmetry in the starting set, then using its re-appearance as a criterion for correctness. Here we present a software program (CrysTwiV) that runs on the web (freely available at: http://btweb.aua.gr/crystwiv/) implementing the above-mentioned method.

INTRODUCTION

An important stage in macromolecular crystallography is that of phase extension and refinement when initial phase estimates are available from isomorphous replacement or anomalous scattering or other methods. In most cases, it is necessary to extend the phases either from lower to higher resolution or within the same resolution range.

For this purpose, an alternative method called the twin variables (TwiV) method has been proposed (1). The TwiV concept consists of the use of a set of auxiliary complex variables $\Psi_{\mathbf{K}}$ which are related to the normalized structure factors $E_{\mathbf{H}}$ by means of the following Equations (1) and (2):

$$E_{\mathbf{H}} = \sum_{\mathbf{K}} \Psi_{\mathbf{K}}(\Psi_{\mathbf{K}-\mathbf{H}})^* \overset{\text{FT}}{\longleftrightarrow} \rho(\mathbf{r}) = \left|\psi(\mathbf{r})\right|^2 \quad \mathbf{1}$$

$$\Psi_{\mathbf{K}} = \sum_{\mathbf{H}} E_{\mathbf{H}}\Psi_{\mathbf{K}-\mathbf{H}}, \text{ where } \Psi_{\mathbf{H}} = \text{FT}[\psi(\mathbf{r})] \quad \mathbf{2}$$

The couple $(E_{\mathbf{H}},\Psi_{\mathbf{H}})$ is called twin variables. At this point, it is appropriate to stress the relevance of the above equations to the fundamental quantum-mechanical principles. We wish to find an approximate wave function in direct space $\psi(\mathbf{r})$ such that its squared modulus $\rho(\mathbf{r})$ 'behaves' like a crystallographic electron-density function. In a crystallographic context, dominated by the reciprocal-space data, it appears useful to introduce the Fourier Transform (FT) of the wave function $\psi(\mathbf{r})$ denoted here by $\Psi_{\mathbf{H}}$. We wish then that the FT of $\rho(\mathbf{r})$, denoted here by $\mathrm{E}_{\mathbf{H}}$, satisfies the observed moduli criterion.

This FT has a precise physical meaning as the momentum space wave function: its square modulus represents the probability distribution over the momentum in the same way that $\rho(\mathbf{r})$ represents the probability distribution over the position of a quantum-mechanical particle (2). In the present context, this physical quantum mechanical meaning is not directly involved and the set of $\Psi_{\mathbf{H}}$ plays the role of an auxiliary set of variables that determines $E_{\mathbf{H}}$ via the left part of Equation (1). In addition, the $E_{\mathbf{H}}$ and $\Psi_{\mathbf{H}}$ sets are linked together via Equation (2), the so-called regression equation whose direct-method meaning has been given in paragraph 2.3 of (1). Thus, the TwiV algorithm aims at determining the phases of the E values through a very large Ψ set, by satisfying a battery of constraints expressed by minimization functions (3).

On the other hand, efficient testing during the process of phase extension is a crucial part of direct methods. The TwiV method offers the possibility to introduce a new overall evaluation test for successful development of the phase-determining algorithm, based upon symmetry considerations. This possibility stems from the decoupling between the E values, bearing the observed $E_{\mathbf{H}}$-moduli information, and the auxiliary variables Ψ which alone control the phasing procedure (1). The new criterion consists of testing the phase-extension and refinement

*To whom correspondence should be addressed. Tel: +30 210 5294211; Fax: +30 210 5294233; Email: kbeth@aua.gr

algorithm by deliberately ignoring the space-group symmetry in the starting set, then using its progressive re-establishment as a criterion for correctness.

The TwiV algorithm has been upgraded and used for protein phase extension from a small set of 200 reflections at low resolution to a large one of 10 000 reflections at high resolution (3,4). It has also been adapted to handle similar problems often encountered in supramolecular structures where an inherent disorder impairs the determination of the positions of the sites of all atoms (5).

IMPLEMENTATION

The CrysTwiV web server is built on the basis of the TwiV method. In macromolecular crystallography, a number of reflections are often approximately phased *a priori* by (insufficiently) isomorphous replacement or other methods. The problem of 'phase extension/refinement' is treated by CrysTwiV in an automated manner.

The values of the initial coordinates borrowed from a known very roughly isomorphous structure are, in general, corrupted by a considerable error. The program reads these initial coordinates given by the user in a pdb file and calculates as usual the corresponding normalized structure factors which contain the necessary information to start the procedure. It has to be noted that the program retains only a small set of the phases of the strongest E-values at lower than the observed resolution, and attempts phase extension to the rest observed reflections. Thus, it is able to handle the poor information of the initial coordinate set. In this way, the large initial error associated with atomic positions is now reflected in the very limited number at low resolution of accepted E-values to be introduced as initial information.

The information of the observed moduli $|F|$ is also given by the user in a file which contains the experimental X-ray diffraction data in a proper format. At the first stage of the program, the observed moduli $|F|$ are converted to normalized structure factors moduli using the subroutine NORMAL of program MULTAN88 (6).

The phase-extension algorithm is based on alternately transferring the phase information between the twin variable sets of E and Ψ values. From the very beginning it is used a very large auxiliary Ψ set. The Miller indices of the Ψ set are taken to be identical to those of the observed E. However, the Ψ set can be extended beyond the resolution of the observed E to the so-called 'super-resolution' shell.

In addition, the Ψ set is not restricted by theory to obey the symmetry constraints and, therefore, the re-appearance of the crystallographic symmetry in the E set calculated by Equation (1) can be used as a criterion for correctness. We denote by: S_MPE = Overall symmetry mean phase error—discrepancy index for the phases of a set of reflections and S_R_{mod} = Overall symmetry mean modulus error—discrepancy index for the moduli of a set of reflections. These indices for the calculated structure factors can be used throughout the iterations as overall indices that are likely to reflect the correctness of the phasing procedure. The examination of the symmetry of the Ψ set (Ψ_S_MPE and Ψ_ S_R_{mod} indices) enriches the evaluation of the correctness of the phase extension.

The Ψ variables control 'alone' the whole procedure; they are allowed to change both in modulus and phase (or real and imaginary part) throughout the procedure.

A web browser provides the user interface for the CrysTwiV application by sending via Hypertext Transfer Protocol (HTTP), a request to the application's web server. Application logic and data reside on the server side (resembling the traditional client-server paradigm). Apache 2 HTTP with Apache Tomcat 5.0.28 (Apache Software Foundation information from: http://www.apache.org/) allows the web server to work with servlets and JavaServer Pages. Registration and authentication information is stored in a relational database (using MySQL 5.0.24 database management system) and accessed via a JDBC driver to identify valid users of the web application. Job submission, ZIP archive creation of the results-output files, status monitoring and email notification is implemented with Java (version 1.5). A Perl script is used to serve job requests on a first-in first-out (FIFO) basis, calling the main program (that implements the twin variables method), which is written in Fortran 90 (compiled with g95 compliler) and runs on SunBlade2000 (Solaris 10, 64-Bit UltraSparc III+ dual processor at 1.05 GHz, 3 GB RAM).

USING THE WEB SERVER

A job submission to the CrysTwiV web server must include:

- A PDB file containing an initial very roughly isomorphous structure. The file should have a typical pdb format. The records TITLE, CRYST1 (unit cell parameters and space group), ATOM and END are mandatory to proceed.
- A RFL file which contains the observed reflections. The records of this file should be h, k, l, $|F|$, $\sigma(F)$. The default format is (3i4,2f8.2). If the uploaded file has a different format this should be stated in RFL FORMAT field.

A typical CrysTwiV run takes between 4 h and 5 days, depending on the protein size, the space group, number of molecules in the asymmetric unit and the server load. However, important algorithm improvements which have been the object of preliminary tests showed that the computing time can be reduced to less than the half of that needed for a complete run of the present program version. Moreover, aiming to reduce the computation time and serve many requests concurrently, the application will be enabled to run on the grid (HellasGrid).

Job status can be monitored via the CrysTwiV's web interface. The user can be informed about the current position of the submitted job in the queue, and during job execution, the percentage of completion can be displayed. Finally, a message is displayed informing about the success or not of job's termination. Submitted files as

well as data regarding program execution (log files, etc.) are kept confidential and cannot be accessed by other users.

The OUTPUT files (compressed in a Zip file) can be downloaded from the URL link sent via email to the user. In case of successful program run, the following files are available:

Inform.log

This file consists of: (1) confirmation of input as read from the datafiles. (2) Preparation of data including parameter estimation and calculation of statistics. (3) A detailed monitoring of the criteria for the correctness of the extended set of phases at each phase-extension step and phase-refinement cycle.

Calculated phases files

phs: an output file named by the name of the pdb file and having the 'phs' extension is given. The file contains $h\ k\ l\ |E|^{obs}$ and calculated phases in order to produce a map file with a simple FFT.

phd: output file named by the name of the pdb file with the 'phd' extension. The file contains h, k, l, $2|E|^{\circ bs} - |E|^{calc}$ and calculated phases in order to produce a difference electron density map at double height by using a simple FFT.

In a forthcoming version, a FFT subroutine will be included in the program and the corresponding map files will be given along with the other output files in a proper format to be plotted using the graphics programs for macromolecular crystallography.

Graph files

grp: an output file named by the name of the pdb file and having the 'grp' extension is given. The file contains the values used to generate the five given figures showing the progress of the phase calculation and the variation of several 'symmetry indicators' during the phase-extension process.

Five figures

The first figure (calculatedPhases.png) shows the number of calculated phases at each extension step of the procedure and the other four figures show the following symmetry indicators plotted at every extension step (S_MPE plotted to sMPE.png; S_R_{mod} to sRmod.png; Ψ_S_MPE to psiSMPE.png; $\Psi\_S\_R_{mod}$ to psiSRmod.png).

In case of program failure, a file named *crystwiv.stdout* is available containing the same information as inform.log file.

An overview of CrysTwiV's web interface is presented in Figure 1.

RESULTS

Several test cases using data retrieved from the Protein Data Bank (PDB) (7) or corresponding to experimental data produced in Laboratory of Structural and Supramolecular Chemistry of NCSR Demokritos and in Laboratory of Protein Structure and Function IMBB, FORTH, have been used to validate the CrysTwiV through different phasing problems of various levels of complexity. Details are available as Supplementary Data on the CrysTwiV website.

In summary, the success of the phasing process depends mainly on two factors: resolution and quality of the starting model. In general, the better the resolution and the starting model are, the more chances CrysTwiV has to succeed. In the examined cases, where the data resolution was higher than ~1.7 Å and the starting model was more than half of the final model, the phasing procedure by CrysTwiV was successful (protein structures: 1BKR, Data Resolution (DR): 1.1 Å (8), Figure 2.; Rnase Ap1, DR: 1.17 Å (9); 1SDB, DR: 1.65 Å (10)). If the data are between 1.7 and 2.5 Å, the starting model should be more complete, ~75–85% of the final model to succeed (1TMY, DR: 1.9 Å (11); 1BBC, DR: 2.2 Å (12)). For data lower than 2.5 Å, we have examined only one case (experimental data given by the IMBB, FORTH). In this case the CrysTwiV hasn't improved the starting model. However, in this case it proved that conventional methods were not applicable as well.

Moreover, the CrysTwiV program used to solve a problem often encountered in supramolecular (SM) structures where an inherent disorder impairs the determination of the positions of the sites of all atoms. The structure examined is a cyclodextrin (CD) host-guest compound: β-CD-indole-3-butyric acid complex (102 independent non-hydrogen atoms). The guest molecules, the water molecules (filling the space between and within the host molecules) and some of the host atoms of this crystal structure are highly disordered. The values of the initial coordinates borrowed from a known very roughly isomorphous structure were corrupted by a considerable error. However, an envelope of the host CD molecule could be obtained from these coordinates, containing the necessary information to start the procedure. The electron density map obtained by the calculated structure factors revealed 17 water molecule sites and all 15 atoms of the guest molecule (Figure 3), with a final symmetry mean phase error $S\_MPE = 5^{\circ}$. The final R factor (anisotropic refinement using *SHELXL97* program (13)) is 0.12, which is about normal for this type of supramolecular compound on account of the usually occurring disorder. The final result has shown that CrysTwiV has been very efficient for the determination of all atomic positions of the host, guest and water molecules.

In every case, the indices based on the crystallographic symmetry enable us to establish a reliable consistency criterion for the correctness of the phasing trials.

Finally, it is noted that the CrysTwiV program works solely in the reciprocal space. In a forthcoming version of the program, the method will be combined with density modification methods to produce better results. In addition, preliminary tests showed that the advantage of the extension of the flexible auxiliary Ψ-set beyond the resolution of the observed data, enhances the phase extension in a so-called 'super-resolution' sphere.

Figure 1. An overview of CrysTwiV's web interface. The application's (**a**) home page, (**b**) introduction page, (**c**) help reference and (**d**) the input-data form.

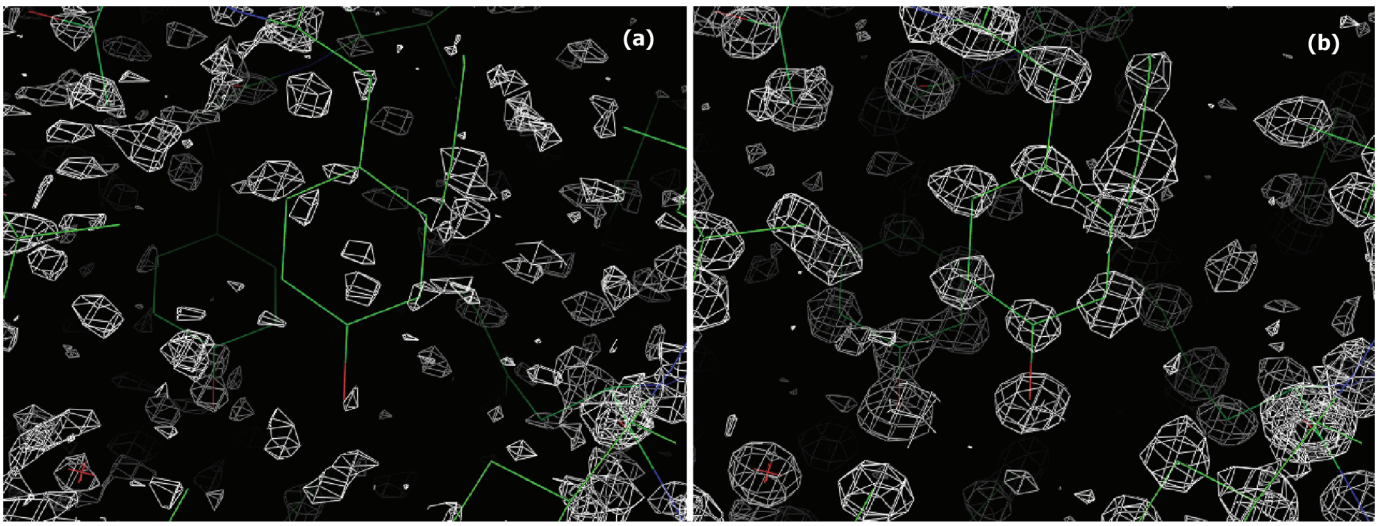

Figure 2. In both maps the drawn wireframes correspond to the truncated residues of protein 1BKR (**a**) electron density corresponding to the initial information used by CrysTwiV to start the procedure and (**b**) final electron density obtained from the calculated extended set of phases at 1.1 Å. Maps are contoured at 2.0σ.

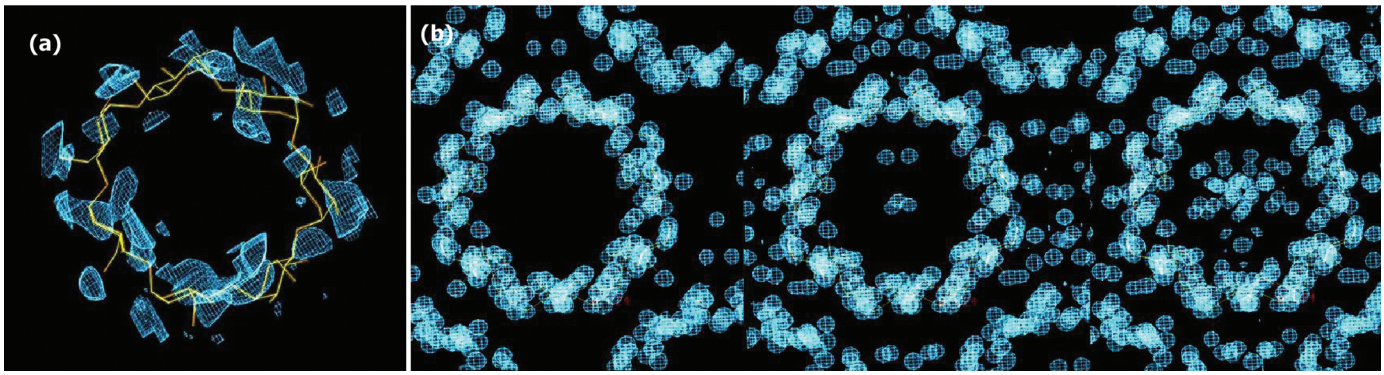

Figure 3. Solution of β-CD-indole-3-butyric acid SM complex with CrysTwiV. (**a**) Electron density corresponding to the initial information and (**b**) electron density obtained at the end of each CrysTwiV run.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Data of protein models used as test cases were retrieved from the PDB (7).
We are grateful to Dr EMavridis for collecting the X-ray data for the CD complex and to Chrysa Meramveliotaki for collecting data for mod2 protein structure. We also gratefully acknowledge the use of the subroutine NORMAL of program MULTAN88 (6). The Open Access publication charges for this paper were waived by Oxford Journals.

Conflict of interest statement. None declared.

REFERENCES

1. Hountas,A. and Tsoucaris,G. (1995) Twin variables and determinants in direct methods. *Acta Cryst*., **A51**, 754–763.
2. Bethanis,K., Tzamalis,P., Hountas,A. and Tsoucaris,G. (2002) Ab initio determination of a crystal structure by means of the SchroÈdinger equation. *Acta Cryst*., **A58**, 265–269.
3. Bethanis,K., Tzamalis,P., Hountas,A., Mishnev,A.F. and Tsoucaris,G. (2000) Upgrading the twin variables algorithm for large structures. *Acta Cryst*., **A56**, 105–111.
4. Tzamalis,P., Bethanis,K., Hountas,A. and Tsoucaris,G. (2003) The crystallographic symmetry test for the correctness of a set of phases. *Acta Cryst*., **A59**, 28–33.
5. Bethanis,K., Tzamalis,P., Hountas,A., Tsoucaris,G., Kokkinou,A. and Mentzafos,D. (2000) New developments of the *TWIN* algorithm for phase extension and refinement in disordered supramolecular structures. *Acta Cryst*., **A56**, 606–608.
6. Debaerdemaeker,T., Tate,C. and Woolfson,M.M. (1988) On the application of phase relationships to complex structures. XXVI. Developments of the Sayre-equation tangent formula. *Acta Cryst*., **A44**, 353–357.
7. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res*., **28**, 235–242.
8. Banuelos,S., Saraste,M. and Djinovic Carugo,K. (1998) Structural comparisons of calponin homology domains: implications for actin binding. *Structure*, **6**, 1419–1431, PDB ID: 1BKR.
9. Bezborodova,S.I., Ermekbaeva,L.A., Shlyapnikov,S.V., Polyakov,K.M. and Bezborodov,A.M. (1988) Ribonuclease Ap1 of Aspergillus pallidus: purification, determination of primary structure and crystallization. *Biokhimiia*, **53**, 965–973.
10. Diao,J.S., Wan,Z.L., Chang,W. R. and Liang,D.C. (1997) Structure of Monomeric Porcine DesB1-B2 Despentapeptide (B26-B30) Insulin at 1.65 Å Resolution. *Acta Cryst*., **D53**, 507–512, PDB ID: 1SDB.
11. Usher,K.C., De La Cruz,A., Dahlquist,F.N., Swanson,R.V., Simon,M.I. and Remington,S.J. (1998) Crystal structures of CheY from Thermotoga maritima do not support conventional explanations for the structural basis of enhanced thermostability. *Protein Sci*., **7**, 403–412, PDB ID: 1TMY.
12. Wery,J.P., Schevitz,R.W., Clawson,D.K., Bobbitt,J.L., Dow,E.R., Gamboa,G., Goodson,T.Jr, Hermann,R.B., Kramer,R. M. *et al.* (1991) Structure of recombinant human rheumatoid arthritic synovial fluid phospholipase A2 at 2.2 A resolution. *Nature*, **352**, 79–82, PDB ID: 1BBC.
13. Sheldrick,G. M. (1997) SHELXL97: Program for the Refinement of Crystal Structures University of Göttingen, Germany.