Acta Cryst. (1996). D**52**, 1119–1124

Direct-Space Methods in Phase Extension and Phase Refinement. VI. *PERP* (Phase Extension and Refinement Program)

L. S. Refaat, C. Tate and M. M. Woolfson

Physics Department, University of York, York YO1 5DD, England. E-mail: mmwl@york.ac.uk

(Received 29 March 1996; accepted 25 June 1996)

Abstract

Several techniques for extending and refining phases for macromolecular structures have been incorporated into a program package *PERP*. In addition to previously employed techniques such as solvent flattening and histogram matching, *PERP* includes a new way of applying the Sayre equation [Refaat, Tate & Woolfson (1995). *Acta Cryst.* D**51**, 1036–1040], low-density elimination [Shiono & Woolfson (1992). *Acta Cryst.* A**48**, 451–456] and two double-histogram methods [Refaat, Tate & Woolfson (1996). *Acta Cryst.* D**52**, 252–256]. *PERP* is an easy-to-use package controlled by keywords and provided with default parameters that usually give near-optimum results. Examples are given of refinement, and also extension and refinement, for six known protein structures with a variety of characteristics. In each case *PERP* gives a very satisfactory outcome as measured by improvements in the mean-phase-error and conventional map-correlation coefficient. The main conclusion is that the several methods used in sequence give more effective extension and refinement than using any single method alone.

1. Introduction

An important stage of macromolecular crystallography is that of phase extension and refinement when initial phase estimates are available for all or some reflections from either anomalous scattering or isomorphous replacement data. The widely used package *SQUASH* (Zhang & Main, 1990*a*,*b*; Cowtan & Main, 1993) has been very successful in such applications where the main procedures used have been solvent flattening (Wang, 1985), histogram matching (Xu, 1984; Lunin, 1988; Mariani, Luzzati & Delacroix, 1988; Zhang & Main, 1990*a*,*b*; Main, 1990*a*,*b*; Lunin & Skovoroda, 1991; Lunin & Vernosova, 1991) and the application of the Sayre equation (Sayre, 1974). Other packages, such as *DEMON*, *SOLOMON* (Abrahams & Leslie, 1996) and *DM* (Cowtan, 1994) are also being successfully used for protein phase extension and refinement at the present time.

In a series of papers (Shiono & Woolfson, 1992; Refaat & Woolfson, 1993; Refaat, Tate & Woolfson, 1995, 1996) new density-modification methods have been described for phase extension and refinement in proteins. Each of these methods has been tested on a stand-alone basis and they have all been shown to be very effective. Further testing has shown that they are even more effective when they are used together. Here we describe a program package *PERP* (phase extension and refinement program) in which these new approaches and other previously known procedures, are used in conjunction and we give examples of the application of *PERP* to a number of known structures.

2. The procedures deployed

There are seven procedures used in *PERP viz.*

1. Solvent flattening (SF).
2. Histogram matching (HM).
3. Sayre equation (SE) in the way proposed by Refaat *et al.* (1995).
4. Low-density elimination (LE) [Shiono & Woolfson (1992); Refaat & Woolfson (1993)].
5. The double-histogram method using the local maximum density (DM) (Refaat *et al.*, 1996).
6. The double histogram using the local density variance (DV) (Refaat *et al.*, 1996).
7. Squaring of density (SQ) This process consists of taking the phases of the squared current density and is the equivalent of applying the tangent formula in parallel fashion *i.e.* to all reflections simultaneously.

For the three histogram-matching methods solvent flattening is included automatically.

PERP is driven by keywords both to input the necessary basic data to the program, to set various parameters and to control the order of use of the seven procedures listed above. An example of a keyword and parameters is,

REF 3(HM 1 DM 1 LE 1 DV 2)DM 1 SQ 1,

which indicates a refinement sequence of three macrocycles of (HM followed by DM followed by LE followed by two cycles of DV) followed by one application of DM followed by one application of SQ. The default for this keyword is,

REF 3(HM 1 DM 1 LE 1 DV 1).

The keyword and parameters,

RES 1.9 1.8 1.7 1.6 1.5 1.5,

indicate that refinement is first carried out at a resolution of 1.9 Å, followed by extension and refinement to 1.8 Å and then followed by further incremental steps of extension and refinement with 0.1 Å steps to 1.5 Å. For each stage of extension and refinement the sequence of processes used is defined by the keyword EXT with an associated list of parameters *e.g.*,

EXT DV 1 HM 1 LE 1 SQ 1 DV 1,

with default,

EXT DV 1 HM 1 LE 1 DV 1 LE 1 DV 1.

At the highest resolution the phases are automatically further refined according to the sequence indicated in REF.

The final 1.5 in the RES keyword sequence indicates either the upper limit of resolution of the data or an upper limit imposed by the user beyond which the data is ignored throughout the *PERP* application. This can be useful if the observed data is regarded as unreliable beyond some resolution limit. The list for RES must have at least two entries, R_1 and R_2 which is all that is required if it is used with only the keyword REF. The default is R_1 and R_2 found from the input data where refinement only is carried out at resolution R_1.

As is usual, and well known to be advantageous in this type of operation, phase estimates obtained by the use of any of the refinement procedures are combined with the original phase estimates, usually from MIR or anomalous scattering data.

We have described just three of the 19 keywords which are available. The user must always input necessary information such as the space group, cell dimensions and contents and the observed reflection data. Also required are the coordinates for a real or simulated known structure from which structure factors and ideal maps at the resolutions indicated by RES may be calculated. From these maps target histograms can be found for the three histogram-matching methods. This information is provided by a user-driven standard program, *STDHIS*, which is part of the *PERP* package but is separate from the program *PERP*. Later, examples will be given of the use of *STDHIS* to provide target histograms for particular structures.

For most of the operations of extension and refinement the program is provided with default parameters and it is our experience that, normally, not much improvement can be made by departing from these. The default parameters were decided upon after extensive trials; by trying various random, but sensible, alternative parameters a reduction of 1° or so in the final mean phase error may be obtained – but this would not be known in advance for an unknown structure.

Most of the development work on *PERP* was carried out on a Hewlett-Packard-730 workstation and timings for the various processes, which depend on the size of the structure, were as follows,

Process	Time (min)
DM	2–2.5
DV	2.5–3
HM	1–2
LE	1–1.5
SE	2–2.5
SF	1–1.5
SQ	1–1.5

These timings would, of course, be much less on an advanced mainframe machine.

3. Examples of application

For the purpose of illustrating the application of *PERP* we show the results of phase extension and refinement, or just refinement alone, for six protein structures with a variety of characteristics.

3.1. *The protein structures*

3.1.1. *RNApl (Bezborodova, Ermekbaeva, Shlyapnikov, Polyakov & Bezborodov, 1988).* Space group $P2_1$, $a = 32.01$, $b = 43.76$, $c = 30.67$ Å, $\beta = 115.83°$, $Z = 2$. The asymmetric unit contains 808 non-H atoms in the protein plus 83 water molecules. Solvent volume 40%.

3.1.2. *2-Zn insulin (Baker et al., 1988).* Space group $R3$, $a = 82.5$, $c = 34.0$ Å, $Z = 9$. The asymmetric unit contains 831 non-H atoms, excluding solvent but including two Zn atoms. Solvent volume 32%.

3.1.3. *RNASE (Sevcik, Dodson & Dodson, 1991).* Space group $P2_12_12_1$, $a = 64.90$, $b = 78.32$, $c = 38.79$ Å, $Z = 4$. The asymmetric unit contains 1 735 non-H atoms, including 28 S atoms. Solvent volume 41%.

3.1.4. *UTPASE (Cedergren-Zeppezauer, Larsson, Nyman, Dauter & Wilson, 1992).* Space group $R3$, $a = 86.64$, $c = 62.23$ Å, $Z = 9$. The asymmetric unit contains 1028 non-H atoms, including seven S atoms, plus 183 water molecules. Solvent volume 63%.

3.1.5. *Selenobiotinyl streptavidin (Hendrickson et al., 1989).* Space group $I222$, $a = 95.27$, $b = 105.40$, $c = 47.56$ Å, $Z = 8$. The asymmetric unit contains 1984 non-H atoms, including four Se atoms, plus 149 water molecules. Solvent volume 45%.

3.1.6. *OPPAL (Glover et al., 1995).* Space group $P2_12_12_1$, $a = 110.50$, $b = 76.58$, $c = 70.67$ Å, $Z = 4$. The asymmetric unit contains 4662 non-H atoms, including five S and eight U atoms. Solvent volume 48%.

3.2. *Results of applying PERP*

We now describe the result of applying the *PERP* procedure to each of these. Where values of the map correlation coefficient (MCC) are given these are with respect to a refined structure at the highest resolution of the data, even if the data for which the MCC is given is a subset at some lower resolution. Defining the MCC this way gives a better impression of the increase in information provided by the refinement, or extension plus refinement, process. All indications of phase from *PERP* are accompanied by weights, in the range zero to unity, which are used to modify the Fourier coefficients of the maps calculated as part of the *PERP* process. These weights are also used in the final map which gives the MCC, although no actual map is calculated and all MCC's are evaluated using the analytical formulae given by Lunin & Woolfson (1993).

We also give initial and final values of the mean phase error (MPE) and the $|E|$-weighted mean phase error (WMPE). If the process has worked well, and the stronger reflections have smaller phase errors, then WMPE should be less than MPE. Similarly we quote the mean phase error for a subset of reflections with the largest values of E and this should be smaller still.

For two of the structures, RNASE and UTPASE, we have used the default sequence for the extension and refinement stages but otherwise we have used non-default, but reasonable, sequences of operations. These variations are just to illustrate the range of possible ways in which *PERP* can be used. By using default or other non-default parameters slightly different results could be obtained but we are confident that they would not be much better, or much worse, than those we give here. Again, in choosing the structures to generate target histograms we have tried to illustrate the inherent flexibility and non-critical nature of this process – as long as a reasonable choice is made.

3.2.1. *RNApl*. There are 23 853 independent reflections to 1.17 Å resolution. In this case there were no AS or multiple isomorphous replacement (MIR) phases available but we started with initial phases degraded to a MPE of 64.8 as might be found from a first rough fitting of the molecule to a trial density. The WMPE was also 64.8 so there was no tendency for the stronger reflections to have smaller phase errors which is usually the case when phases are derived from some physically based process. In addition the MPE was uncorrelated with resolution in distinction from what would normally be the case. The MCC for the initial phases was 0.339.

The structures RNASE was used to generate the target histograms. The refinement procedure used is described by the keyword and parameters,

REF 3(SE 1 HM 1 LE 1 DM 1 DV 1 LE 1).

The final MPE for all reflections was 31.5° with WMPE 27.5 . For the 2049 reflections with the largest values of $|E|$ the MPE was 12.6°. For the refined phases the MCC was 0.843.

3.2.2. *2-Zn insulin*. For this structure MIR phases were available for 6450 reflections to 1.9 Å resolution with an MPE of 61.5° and WMPE of 59.2°. The MCC for the initial MIR phases was 0.329. There were 13 289 reflection data available to a resolution of 1.5 Å.

The structure used for calculating the target histograms was OPPAL. The keyword information governing extension and refinement was,

REF HM 2 DM 2 LE 2 DV 2 LE 1

RES 1.9 1.85 1.8 1.75 1.7 1.65 1.6 1.55 1.5 1.5

EXT HM 2 DM 2 LE 1 DV 1.

The final result is that the original MIR phases had an MPE of 37.9° with WMPE of 33.3°. For the complete set to 1.5 Å resolution the MPE was 40.2° with WMPE 35.5 . The 1271 reflections with the largest values of $|E|$ had an unweighted mean phase error of 21.2°. For the refined phases the MCC was 0.773.

3.2.3. *RNASE*. There are 7216 reflections to 2.5 Å resolution with MIR phase estimates having an MPE of 56.4° with WMPE 51.6°. The MCC for these initial phases is 0.367. There are 17 211 observed reflections to 1.8 Å resolution.

The insulin structure, with zinc removed, was used to provide data for the target histograms. The extension and refinement parameters were,

REF HM 1 DM 1 SQ 1 LE 1 DV 1 LE 1 DM 1 HM 1 LE 1

RES 2.5 2.4 2.3 2.2 2.1 2.0 1.9 1.8 1.8

EXT DV 1 HM 1 LE 1 DV 1 LE 1 DM 1.

The final result is that the original MIR reflections have an MPE of 45.0° with WMPE 36.4°. For the complete set of reflections to 1.8 Å resolution the MPE and WMPE were 51.6° and 43.9°, respectively. For the 2203 reflections with the largest values of $|E|$ the MPE was 30.9°. For the refined phases the MCC was 0.693.

3.2.4. *UTPASE*. There are 8290 reflections to 2.24 Å resolution for which MIR phase estimates are available with MPE and WMPE 56.6° and 53.0°, respectively. These give an MCC of 0.435. There are 13 640 observed reflections to 1.89 Å resolution.

The RNASE data was used to provide the target histograms. The keyword information provided to control the extension and refinement was,

REF HM 1 DM 1 LE 1 DV 1 LE 1 HM 1 DM 1 LE 1
RES 2.24 2.20 2.15 2.10 2.05 2.00 1.95 1.89 1.89
EXT DV 1 HM 1 LE 1 DV 1 LE 1 DM 1.

The final result was that the MIR phases had an MPE and WMPE of 35.8° and 27.5°, respectively. The corresponding mean phase errors for the complete set of data to 1.89 Å were 36.9° and 28.9°, respectively. For the 1593 reflections with the largest values of $|E|$ the MPE was 15.4°. The MCC for the refined phases was 0.856.

In Fig. 1 there is shown the projected total density between $z = 0/96$ and $z = 3/96$ for the initial MIR phases and also for the final refined phases for this structure together with indicated atomic positions. The improvement in the continuity and definition of the protein density and in the reduction of noise in the solvent region is self evident.

3.2.5. *Selenobiotinyl streptavidin*. There are 4571 reflections to 3.0 Å resolution for which multiple-wavelength anomalous-scattering phase estimates are available. The initial MPE was 57.9° with WMPE 51.5°. The 323 reflections with the largest values of $|E|$ had an unweighted error of 44.0°. With the initial phases the MCC was 0.560.

For the target histograms the structure OPPAL was used but with the U atoms replaced by Se atoms. This

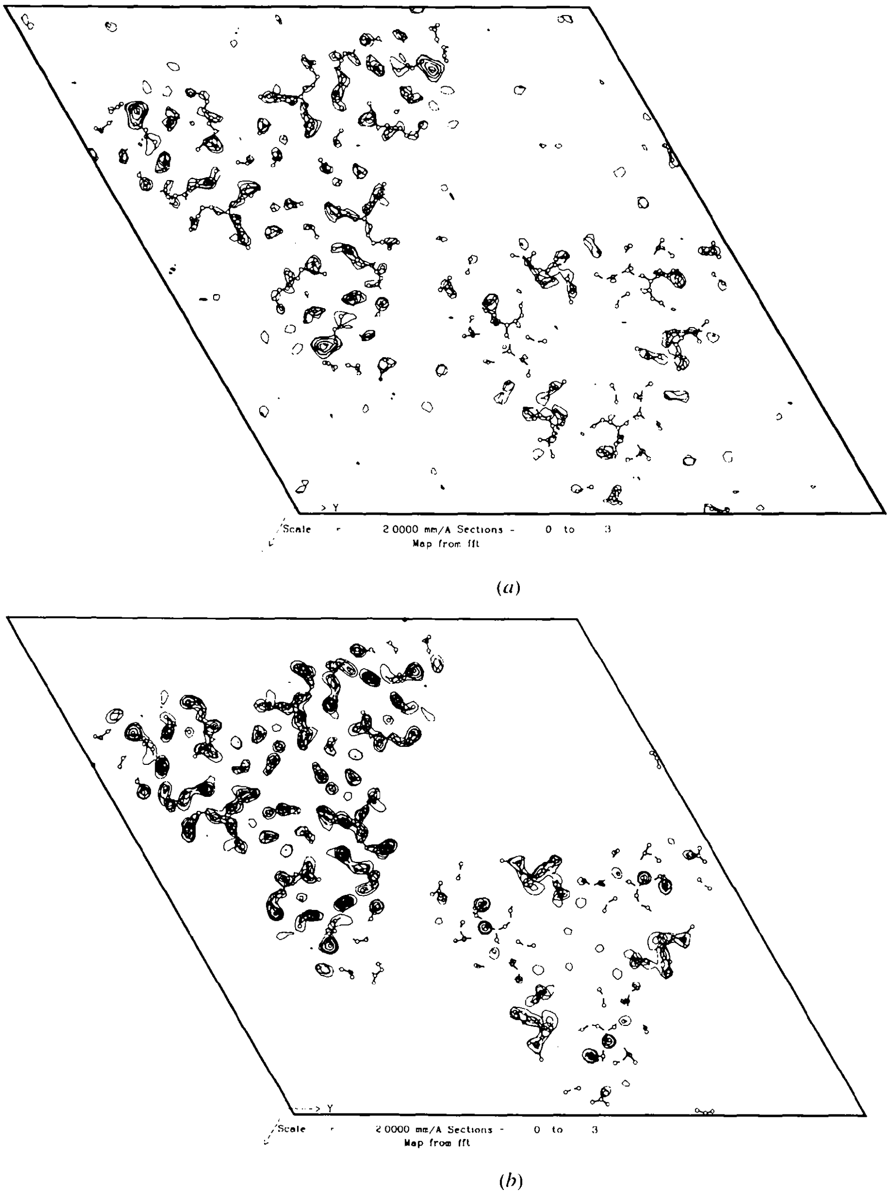

Fig. 1. (*a*) Overlapped sections $z = 0/96$ to $z = 3/96$ for the map with the original MIR phases of deoxyuridine triphosphotase (UTPASE), mean phase error 56.6°, resolution 2.24 Å. (*b*) Overlapped sections $z = 0/96$ to $z = 3/96$ for the map with the final refined phases, mean phase error 36.9°, resolution 1.89 Å.

case required only refinement and the refinement sequence was defined by,

REF 2(DV 1 HM 1 DM 1 HM 1 DV 1 DM 1 DV 1
HM 1 DM 1 DV 1 HM 1).

The final MPE and WMPE were 50.9 and 42.0 , respectively. For the 323 reflections with the largest values of $|E|$ the unweighted mean phase error was 29.9 . The value of MCC with the refined phases was 0.710.

3.2.6. *OPPAL*. Estimates of phase from anomalous scattering were available for the 30 538 reflections to 2.15 Å resolution. These gave an MPE and WMPE of 58.4 and 54.1 , respectively. The 3695 reflections with the largest values of $|E|$ had an MPE 49.1 . The MCC for the initial phase estimates was 0.541.

The data from 2-Zn insulin was used to provide the target histograms. This case required only refinement and the refinement sequence was defined by,

REF 2(DV 1 HM 1 DM 1 LE 1 DV 1
HM 1 LE 1 DM 1 DV 1).

The final MPE and WMPE for the complete set of reflections were 49.8 and 41.5 , respectively. The subset of 3695 reflections with the largest values of $|E|$ had an MPE of 26.4 . For the refined phases the MCC was 0.739.

4. Conclusions and comments

PERP is an easy-to-use and very flexible program system for phase extension and refinement which is capable of tackling most day-to-day problems that arise in protein crystallography. The main decision to be made by the user is that of choosing, and perhaps modifying, the known structure which provides the information for calculating the standard histograms but the program is very tolerant to this choice. In original proving trials of the procedure we used the histograms from the actual structure under investigation, which would not be possible in practice, but when we substituted the histograms from other structures, as described above, the results were little different. The changes in MPE were of the order of 0.5 , surprisingly more often better than worse, and the MCC's changed in the range −0.003 to +0.011. As an illustration of the tolerance to the target histograms it should be noted that for OPPAL we used the unmodified 2-Zn insulin structure without replacing Zn atoms with U atoms and, similarly, for 2-Zn insulin we used the OPPAL structure without change of the heavy atoms. Apart from providing the target histograms and other essential standard information about the structure, the user can leave remaining decisions to the defaults provided, with confidence that the results will usually be quite close to the best possible. For users wishing to make individual decisions on parameters there is a comprehensive write-up describing the program and giving advice about actions to be taken if it does not perform as expected.

PERP has a 'module' construction which enables new or improved procedures to be added readily, so giving the possibility of steady improvement. Other procedures currently under development will be added in due course, including a histogram-moments method (Gu, Woolfson & Yao, 1996). It is also intended to add an option for the use of non-crystallographic symmetry.

We wish to express our appreciation of support given by SERC, BBSRC and EPSRC which enabled this work to be carried out. We are also grateful to Dr K. D. Cowtan for his help and useful advice. Some suggestions of the referees, which we are pleased to acknowledge, also led to improvements.

References

Abrahams, J. P. & Leslie, A. G. W. (1996). *Acta Cryst.* **D52**, 30–42.

Baker, E. N., Blundell, T. L., Cutfield, J. F., Cutfield, S. M., Dodson, E. J., Dodson, G. G., Hodgkin, D. M. C., Hubbard, R. E., Isaacs, N. W., Reynolds, C. D., Sakabe, K., Sakabe, N. & Vijayan, N. M. (1988). *Philos. Trans. R. Soc. London Ser. B*, **319**, 456–469.

Bezborodova, S. I., Ermekbaeva, L. A., Shlyapnikov, S. V., Polyakov, K. M. & Bezborodov, A. M. (1988). *Biokhimiya*, **53**, 965–973.

Cedergren-Zeppezauer, E. S., Larsson, G., Nyman, P. O., Dauter, Z. & Wilson, K. S. (1992). *Nature (London)*, **355**, 740–743.

Cowtan, K. D. (1994). *Jnt CCP4 ESF-EACBM Newslett. Protein Crystallogr.* **31**, 34–38.

Cowtan, K. D. & Main, P. (1993). *Acta Cryst.* **D49**, 148–157.

Glover, I. D., Denny, R. C., Nguti, N. D., McSweeney, S. M., Kinder, S. H., Thompson, A. W., Dodson, E. J., Wilkinson, A. J. & Tame, J. R. H. (1995). *Acta Cryst.* **D51**, 39–47.

Gu, Y.-X., Woolfson, M. M. & Yao, J.-X. (1996). *Acta Cryst.* **D52**, 1114–1118.

Harrison, R. W. (1988). *J. Appl. Cryst.* **21**, 949–952.

Hendrickson, W. A., Pähler, A., Smith, J. L., Satow, Y., Merrit, E. A. & Phizackerley, R. P. (1989). *Proc. Natl Acad. Sci. USA*, **86**, 2190–2194.

Lunin, V. Yu. (1988). *Acta Cryst.* **A44**, 144–150.

Lunin, V. Yu. & Skovoroda, T. P. (1991). *Acta Cryst.* **A47**, 45–52.

Lunin, V. Yu. & Vernoslova, E. A. (1991). *Acta Cryst.* **A47**, 238–243.

Lunin, V. Yu. & Woolfson, M. M. (1993). *Acta Cryst.* **D49**, 530–533.

Main, P. (1990*a*). *Acta Cryst.* **A46**, 372–377.

Main, P. (1990*b*). *Acta Cryst.* **A46**, 507–509.

Mariani, P., Luzzati, V. & Delacroix, H. (1988). *J. Mol. Biol.* **204**, 165–189.

Refaat, L. S., Tate, C. & Woolfson, M. M. (1995). *Acta Cryst.* **D51**, 1036–1040.

Refaat, L. S., Tate, C. & Woolfson, M. M. (1996). *Acta Cryst.* D**52**, 252-256.

Refaat, L. S. & Woolfson, M. M. (1993). *Acta Cryst.* D**49**, 367-371.

Sayre, D. (1974). *Acta Cryst.* A**30**, 180-184.

Sevcik, J., Dodson, E. J. & Dodson, G. G. (1991). *Acta Cryst.* B**47**, 240-253.

Shiono, M. & Woolfson, M. M. (1992). *Acta Cryst.* A**48**, 451-456.

Wang, B. C. (1985). *Methods Enzymol.* **115**, 90-112.

Xu, S. B. (1984). *Phase extension using histogram and SIR phase information with single isomorphous replacement data*. A proposal for the PhD requirement, University of Pittsburgh, USA.

Zhang, K. Y. J. & Main, P. (1990*a*). *Acta Cryst.* A**46**, 41-46.

Zhang, K. Y. J. & Main, P. (1990*b*). *Acta Cryst.* A**46**, 377-381.