COMPUTER PROGRAMS

J. Appl. Cryst. (1996). **29**, 741–744

A procedure compatible with *X-PLOR* for the calculation of electron-density maps weighted using an *R*-free-likelihood-based approach

ALEXANDR G. URZHUMTSEV,[a,b] TATIANA P. SKOVORODA[b] AND VLADIMIR Y. LUNIN[b] *at* [a]*UPR de Biologie Structurale, IGBMC, BP 163, 67404 Illkirch, CU de Strasbourg, France, and* [b]*IMPB RAS, Pushchino, Moscow Region, 142292, Russia. E-mail: sacha@igbmc.u-strasbg.fr*

(*Received 12 March 1996; accepted 30 May 1996*)

Abstract

A program has been developed that uses an *R*-free-likelihood-based technique to estimate the errors of phases calculated from atomic models. This technique allows one to obtain realistic estimates when it is applied to refined models. The program reads a file of structure factors in the *X-PLOR* format and creates a new one that contains information necessary to calculate weighted maps. The output file can be used directly by *X-PLOR*.

1. Introduction

Electron-density syntheses calculated using experimental magnitudes and phases obtained from an available atomic model are known to be strongly biased towards the model, resulting in spurious density features (Raman, 1959; Ramachandran & Srinivasan, 1970; Oppenheim, 1981; Hodel, Kim & Brünger, 1992). Significant efforts were made to find a combination of experimental and calculated magnitudes and of calculated phases that reduces this bias (Luzzati, 1953; Main, 1979; Vijayan, 1980; Read, 1986).

Different forms of synthesis coefficients were suggested based mostly on two different ideas. The first one is the atomicity of the electron-density map, which describes map distortions in terms of true and false peaks, peak heights *etc.* (Raman, 1959; Main, 1979). The second one is the statistical estimation of errors in phase values, which allows one to introduce into these coefficients the expected (mean) values of the cosine of the phase error

$$m(\mathbf{s}) = \langle \cos[\varphi^{\text{true}}(\mathbf{s}) - \varphi^{\text{calc}}(\mathbf{s})] \rangle \quad (1)$$

and the values

$$\alpha(\mathbf{s}) = \langle \cos[2\pi(\mathbf{s}, \Delta\mathbf{r})] \rangle \quad (2)$$

[$D(\mathbf{s})$ in the notation of Read (1986)], where the coordinate errors $\Delta\mathbf{r}$ are supposed to be random variables (Luzzati, 1952; Srinivasan & Parthasarathy, 1976; Lunin & Urzhumtsev, 1984; Read, 1986). The simplest form of the weighted Fourier coefficients is

$$m(\mathbf{s})F^{\text{obs}}(\mathbf{s}) \exp[i\varphi^{\text{calc}}(\mathbf{s})], \quad (3)$$

while Read (1986) has shown that better coefficients have the form

$$[2m(\mathbf{s})F^{\text{obs}}(\mathbf{s}) - \alpha(\mathbf{s})F^{\text{calc}}(\mathbf{s})] \exp[i\varphi^{\text{calc}}(\mathbf{s})] \quad (4)$$

as implemented in his program *SIGMAA* (Collaborative Computational Project, Number 4, 1994).

The use of $m(\mathbf{s})$ and $\alpha(\mathbf{s})$ implicitly requires estimates of the quality of the given atomic model (*e.g.* mean coordinate errors) or, more formally, of some parameters in the probability distributions of phase errors. Existing methods like the Luzzati plot (Luzzati, 1952) and likelihood-based estimates (Lunin & Urzhumtsev, 1984; Read, 1986) tend to overestimate the model quality when based on the comparison of experimental magnitudes with ones calculated from a refined model (Fields *et al.*, 1994; Ohlendorf, 1994; Lunin & Skovoroda, 1995). Nevertheless, the use of a control set (Brünger, 1992*a*) in the refinement protocol allows one to obtain much more realistic estimates of model quality (Lunin & Skovoroda, 1995) and therefore calculate the values of $m(\mathbf{s})$ and $\alpha(\mathbf{s})$ more correctly.

A computer program has been developed that, when used together with the *X-PLOR* system (Brünger, 1992*b*), allows one to calculate weighted electron-density maps using this improved method of model quality estimation. The kernel of the program, which concerns the likelihood maximization (Lunin, 1982), has many common features with the *SIGMAA* approach (Read, 1986). Nevertheless, the possibility to use cross-validated data and the different design of the program make it a new crystallographic tool.

2. Program description

The program *RFLEXPL* (*R*-free likelihood estimates for phase error for *X-PLOR* users) is a tool for crystallographers using the *X-PLOR* system (Brünger, 1992*b*) to calculate weighted electron-density maps.

The input file for the program *RFLEXPL* is a standard file of structure factors in the *X-PLOR* format that contains the values for FOBS (the magnitude F^{obs}), FCALC (both the magnitude F^{calc} and phase φ^{calc}) and TEST (the flag for the control data set). Any other standard *X-PLOR* values are allowed. No preliminary scaling of F^{obs} and F^{calc} is necessary.

The control data file contains

(i) the resolution range (high and low resolution limits) and the number of resolution zones;

(ii) the unit-cell parameters;

(iii) a list of symmetry operations entered in symbolic form.

In each of the resolution zones, the parameters of the phase probability distribution $P(\varphi)$ will be considered as a constant (see §3 below). These zones are equally spaced in $1/d^2$ in order to have an approximately equal number of reflections in each of them.

The output file contains a list of reflections of chosen resolution in the standard *X-PLOR* format. For each reflection, the record contains the values of F^{obs}, F^{calc}, φ^{calc}, σ_F (if it

exists), m(s) as FOM and α(s) as WEIGHT values of the *X-PLOR* file. F^{obs} and F^{calc} values are the same as they are in the input file and the necessary scaling is already included in WEIGHT coefficients.

To calculate the map with the coefficients (3) or (4), this file should be used as the input file of structure factors in the run of *X-PLOR*. The following commands should be included in the *X-PLOR* input file as

```
DO AMPLITUDE (FCALC = FOM * FOBS)
```

for the coefficients (3) and

```
DO AMPLITUDE (FOBS = 2 * FOM * FOBS)
```

```
remarks map calculation
{ ------------------------------------------------------ }
xrefine
   a=84.66 b=137.35 c=78.88  alpha=90. beta=90. gamma=90.
   symmetry=(x,y,z)
   symmetry=(-x,-y,z)
   symmetry=(1/2-x,y+1/2,-z)
   symmetry=(1/2+x,1/2-y,-z)
{ ------------------------------------------------------ }
nreflections=35000
   reflection
     @protref.hkl
   end
   resolution 200.0 2.3
   method=FFT
   fft
     memory=1000000
   end
 end
{ ------------------------------------------------------ }
 xrefin
{<< do amplitude ( FCALC = 2 * FOM * FOBS - WEIG * FCALC )>> }
   do amplitude ( FOBS = 2 * FOBS * FOM )
   do amplitude ( FCALC = FCALC * WEIGHT )
   do amplitude ( FCALC = FOBS - FCALC )
   map
    extend=box
    xmin -21. xmax 70. ymin -11. ymax 82. zmin -21. zmax 60.
    output=f21m.map
   end
 end
{ ------------------------------------------------------ }
 stop
```

Fig. 1. An example of an *X-PLOR* input file for the *SIGMAA*-type map calculation.

```
DO AMPLITUDE (FCALC = FCALC * WEIGHT)
```

```
DO AMPLITUDE (FCALC = FOBS - FCALC)
```

for the coefficients (4). Note that in the latter case *no recalculation* of structure factors (command 'UPDATE-FCALC') should be done during this run as well as *no scaling* between F^{obs} and F^{calc} (command 'SCALE FOBS FCALC)'. An example of the full *X-PLOR* command file is shown in Fig. 1.

The program *RFLEXPL* prints some useful information both for each resolution zone defined by control parameters (Fig. 2) and a summary one for the whole run of the program (Fig. 3).

3. Phase-quality estimates

The details of the method for phase estimation used in the program are described by Lunin & Skovoroda (1995) (LS in what follows). Only the basic features necessary to understand the program messages are explained in this paragraph.

The weights calculation is performed independently in shells of reciprocal space. The number of reflections in the test set and the total number of reflections for every shell are printed. To have reliable weight estimates, it is necessary to have a reasonable number of test-set reflections in the shell (~50 as low estimate; a recommended number is ~100 or more); the number of resolution zones should be chosen correspondingly to provide such a number of test-set reflections. If there are no test-set reflections in a shell, all structure factors from this resolution zone will get zero weights and actually will be excluded from the map calculation.

Depending on the quality of the model structure factors, the program decides for every shell whether the calculated phases contain information about the real structure and may be used to estimate the true phases or not. An intensity correlation coefficient OMEGA printed for every shell reflects this quality. A negative OMEGA value means that the model used to calculate structure factors is too bad to produce any reasonable

```
** LBEST ************************************************************
** LBEST ********************** SUBROUTINE LBEST **********************
** LBEST ************************************************************
** LBEST **********      zone: 0.0025< ss=< 0.0118
** LBEST ** abcd **      585 reflections are in zone,
** LBEST ** abcd **      50 of them are used for estimation.
** LBEST ** abcd **      A = 2337687. B = 37529.80
** LBEST ** abcd **      C = 203875.8 D = 0.8477243E+11
** LBEST ** abcd **      (A*B-C*C)/(A*B) = 0.5262287
** LBEST ** abcd **      OMEGA =-0.9495934E-02
** LBEST **********   ************************************************
** LBEST **********   * alpha = 0.0000000E+00 beta = 37529.80         *
** LBEST **********   * topt = 0.0000000E+00 ALLG = 0.0000000E+00     *
** LBEST **********   ************************************************



** LBEST ************************************************************
** LBEST ************************************************************
** LBEST ********************** SUBROUTINE LBEST **********************
** LBEST ************************************************************
** LBEST **********      zone: 0.0305< ss=< 0.0398
** LBEST ** abcd **     1336 reflections are in zone,
** LBEST ** abcd **      132 of them are used for estimation.
** LBEST ** abcd **      A = 652697.9          B = 34143.30
** LBEST ** abcd **      C = 141540.3          D = 0.5530552E+11
** LBEST ** abcd **      (A*B-C*C)/(A*B) = 0.1010355
** LBEST ** abcd **                OMEGA = 0.8198478
** LBEST ** solv **      n1= 2   n2= 4   fg(topt) = 0.1430511E-05
** LBEST **********   ************************************************
** LBEST **********   * alpha = 0.2037189 beta = 7055.444             *
** LBEST **********   * topt = 0.2887399E-04 ALLG = 64.97922          *
** LBEST **********   ************************************************
```

Fig. 2. An example of a *RFLEXPL* message file. OMEGA is a intensity correlation that characterizes the reliability of calculated phases at a given resolution shell. Coefficients A, B, C, D, alpha, topt and ALLG are as given in the main text. Only the results corresponding to two different resolution zones are shown. For the first shell, OMEGA is negative owing to a poor agreement between observed and calculated magnitudes, and all figures of merit in that resolution range as well as the corresponding alpha parameter have been set to zero.

```
statistics on different resolution zones : Nrefl <Fom> <estimated Dphi>

 N      resol.(A)          Total          Refl. used for refin.      Test set

 1    9.20   20.00    585  0.000  90.0    535  0.000  90.0     50  0.000  90.0
 2    6.88    9.20    942  0.683  35.9    854  0.684  35.8     88  0.675  36.1
 3    5.73    6.88   1150  0.688  36.0   1027  0.691  35.8    123  0.665  37.4
 4    5.01    5.73   1336  0.782  28.0   1204  0.784  27.9    132  0.765  29.2

    ..........................................................

17    2.49    2.57   1201  0.778  30.0   1099  0.778  30.0    102  0.780  30.0
18    2.42    2.49   1078  0.620  43.1    973  0.621  43.1    105  0.612  43.5
19    2.36    2.42   1017  0.707  36.2    902  0.708  36.2    115  0.698  36.5
20    2.30    2.36    923  0.761  31.6    829  0.761  31.7     94  0.766  31.2
```

Fig. 3. An example of a *RFLEXPL* message file (total statistics). For every resolution shell, the line contains 3 × 3 numbers. These are the number of reflections, estimated figure of merit and estimated phase error for all reflections, for ones from the work set and for ones from the test set. Note that the first line indicates zero figure of merit (see the caption to Fig. 2).

estimates of phases in this zone. All weights are set as zero in this case and reflections of this shell are not used in the weighted map calculation.

The intensity is defined as $I(\mathbf{s}) = F(\mathbf{s})^2/\varepsilon$, where ε is the number of the symmetry operations of the space group that leave the vector $\mathbf{s}$ of reciprocal space unchanged. Additional weights, equal to 2 for acentric reflections and to 1 for centric ones, are used when calculating averaged values. The averaged values of I^{calc} (A value), I^{obs} (B value), $(I^{\text{obs}}I^{\text{calc}})^{1/2}$ (C value) and $(I^{\text{obs}}I^{\text{calc}})$ (D value) are printed for every shell [see LS, equation (18), for more details].

The phase-quality estimates are based on statistical modelling of the source of phase errors. Different statistical models result in probability distributions of the same shape [LS, equations (2) and (3)], making irrelevant the problem of choice of a statistical hypothesis for a given particular case. This distribution (two-dimensional Gaussian distribution rewritten into magnitude-phase variables) depends on two parameters α and β, which are considered as constant for a given resolution shell. The meaning of α and β for some simple statistical models are presented in Table 1 (for a more common case see LS). In particular, the value of α corresponds to the mean value of (2) for a given resolution shell.

The magnitudes of test-set reflections are used to find these parameters α and β in every shell of reciprocal space; these are then applied to the calculation of figures of merit for all reflections of this shell. The α and β values and their ratio topt are printed by the program, as well as the ALLG value (the logarithm of the ratio of probabilities corresponding to the cases of the optimal choice of α and β values and of $\alpha = 0$).

The final table includes the mean values of figures of merit and expected phase error for both work and test sets and for the whole data set.

4. Program distribution

The program *RFLEXPL* is written in standard Fortran77 . The source code and an example of the command file are available by request from sacha@igbmc.u-strasbg.fr.

This work was supported in part by Russian Foundation for Basic Researches grant 94-04-12884 and ISF grant RMZ000. VYL was supported by a Travel grant of MENESRIP, France. AGU was supported by the CNRS through the UPR 9004, by the Institut National de la Santé et de la Recherche Médicale and the Centre Hospitalier Universitaire Régional. The authors thank Drs. F. Gomes and A. Podjarny for useful remarks and Mme N. Lunina for help in the programming.

Table 1. *Expressions for α and β parameters for different statistical models*

Individual temperature factors of atoms are included in scattering factors $f_j(s)$

Type of model	Values of (α, β) parameters
Independent coordinate errors with $p(\Delta r)$ distribution; F^{obs} and F^{calc} are scaled together; all N atoms are included in the model.	$\alpha(\mathbf{s}) = \langle\cos 2\pi(\mathbf{s}, \Delta\mathbf{r})\rangle_p$ $\beta(\mathbf{s}) = (1-\alpha^2)\sum_{j=1}^{N} f_j^2(s)$
Independent coordinate errors with Gaussian distribution with $\langle\lvert\Delta\mathbf{r}\rvert\rangle = \omega$; F^{obs} and F^{calc} are scaled together; all N atoms are included in the model.	$\alpha(\mathbf{s}) = \exp(-\pi^3\omega^2 s^2/4)$ $\beta(\mathbf{s}) = (1-\alpha^2)\sum_{j=1}^{N} f_j^2(s)$
Partial model with exact coordinates of M atoms; unknown atoms ($j = M+1, \ldots, N$) are considered as uniformly distributed in the unit cell; F^{obs} and F^{calc} are scaled together.	$\alpha(\mathbf{s}) = 1$ $\beta(\mathbf{s}) = \sum_{j=M+1}^{N} f_j^2(s)$
Partial model with independent coordinates errors with $p(\Delta r)$ distribution; unknown atoms ($j = M+1, \ldots, N$) are considered as uniformly distributed in the unit cell; scale factor $1/\kappa$ must be applied to F^{obs} to reduce them to the F^{calc} scale.	$\alpha(\mathbf{s}) = \kappa\langle\cos 2\pi(\mathbf{s}, \Delta\mathbf{r})\rangle_p$ $\beta(\mathbf{s}) = \kappa^2\left[(1-\alpha^2)\sum_{j=1}^{N} f_j^2(s) + \sum_{j=M+1}^{N} f_j^2(s)\right]$

References

Brünger, A. T. (1992*a*). *Nature (London)*, **355**, 472–474.

Brünger, A. T. (1992*b*). *X-PLOR: a System for X-ray Crystallography and Nuclear Magnetic Resonance*. Yale University Press.

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

Fields, B. A., Bartsch, H. H., Bartunik, H. D., Cordes, F., Guss, J. M. & Freeman, H. C. (1994). *Acta Cryst.* D**50**, 709–730.

Hodel, A., Kim, S.-H. & Brünger, A. T. (1992). *Acta Cryst.* A**48**, 851–858.

Lunin, V. Yu. (1982). *The Use of Maximum Likelihood Approach to Estimate Phase Errors in Protein Crystallography*. Russia: Pushchino.

Lunin, V. Yu. & Skovoroda, T. P. (1995). *Acta Cryst.* A**51**, 880–887.

Lunin, V. Yu. & Urzhumstev, A. G. (1984). *Acta Cryst.* A**40**, 269–277.

Luzzati, V. (1952). *Acta Cryst.* **5**, 802–810.

Luzzati, V. (1953). *Acta Cryst.* **6**, 142–152.

Main, P. (1979). *Acta Cryst.* A**35**, 779–785.

Ohlendorf, D. H. (1994). *Acta Cryst.* D**50**, 808–812.

Oppenheim, A. V. (1981). *Proc. IEEE*, **69**, 529–541.

Raman, S. (1959). *Acta Cryst.* **12**, 964–975.

Ramachandran, G. N. & Srinivasan, R. (1970). *Fourier Methods in Crystallography.* New York: Wiley.
Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.
Srinivasan, R. & Partharasarathy, S. (1976). *Some Statistical Applications in X-ray Crystallography.* Oxford: Pergamon Press.
Vijayan, M. (1980). *Acta Cryst.* A**36**, 295–298.