18.06.2001

AVERAGE : program description

(this description corresponds to the versions from 18.06.2001 and later)

Contents

1. Purpose

The program AVERAGE allows to average several phase sets. In particular, these phase sets («variants») may be obtain previously by programs GENMEM, FAMREF, RING or in some other way and must be in .OUT-format file.

2. Basic concepts

Phase set («variant»)
Reference phases
Correlation and distances between two phase sets
Maps alignment

A knowledge of these concepts, a short definition of which is done at the Section 9, is necessary for understanding the following text. Indicated articles can be read in order to get more details of these definitions and, in particular, their applications for connectivity-based phasing.

3. Algorithme

Brief description:

Phase sets before be averaged need to be reduced to their common origin (if the space group allows several of them). First of all, the center of the "cloud" of the phase sets is determined as the phase set such that the sum of the square of distances to all other sets is minimal. Then the program shifts all variants to the central one and calculates the average phases Ph_ave and their figures of merit FOM. Two different averaged data sets are calculated corresponding to two different weighting schemes. Both results are written in the output file of structure factors. Some statistical information about averaging is written to the message file.

More formal description :

For each reflection, the input file of structure factors contains the experimental magnitude Fobs, and the phases Ph and corresponding real values T, a pair per every phase set,. Depending on the phase generation programs, such value T may be either the corresponding structure factor magnitude (FAM) or may be also considered as a corresponding weight; in many cases all these values are equal to zero and present in the file for the format compatibility with other phasing programs. In what follows, the values are called *individual magnitudes.*

Distance between phases is defined through the distance between corresponding Fourier synthesis. For this calculation, both experimental and individual magnitudes may be used (see Section **9**).

For every available variant (excluding maybe the first variant, as it is defined in the control data) the sums of the square of distances to all other variants are calculated over reflections of given resolution. The variant with the minimal sum is determinate and considered as the central one. Then every variant is reduced to the central one by all origin shifts available for the given space group.

For the phase sets reduced to the same origin the average of their phases is calculated using the experimental magnitudes; both the phase value Ph_ave ($\varphi^{ave}(\mathbf{s})$) and its figure of merit FOM ($m(\mathbf{s})$) are obtained by this operation :

$$m(\mathbf{s})\exp[i\varphi^{ave}(\mathbf{s})]=\frac{1}{M}\sum_{k=1}^{M}\exp[i\varphi_k(\mathbf{s})]$$

Then the second average is done using its individual modules which gives the values of Ph_calc and T_calc.

$$T^{calc}(\mathbf{s})\exp[i\varphi^{calc}(\mathbf{s})]=\frac{1}{M}\sum_{k=1}^{M}F_k^{calc}(\mathbf{s})\exp[i\varphi_k(\mathbf{s})]$$

The calculated values are written to the output file of structure factors. Every record in this file contains an information about one reflection :
H K L Fobs Test_flag FOM Ph_ave T_calc Ph_calc Ph_ex
if the reference phases are available in the input file or
H K L Fobs Test_flag FOM Ph_ave T_calc Ph_calc
otherwise.
Some statistical information about phases in the file and correlations between average phases with reference ones (which are calculated by different ways) are written in the messages file.
Remark. In the particular case when all individual magnitudes are equal to one, the values of T_calc coincides with FOM, and Ph_calc concide with Ph_ave.

4. Input file of control data

General remarks. The file of control data allows to use comment lines ; they are those started from exclamation "!". Therefore, a beginner can take some available complete file of control data and adapt it to his purposes by converting unnecessary lines to comments. It is strictly recommended to leave all comments in the file because this facilitates further modification of parameters and decrease errors. Free format is used for control data. Letter "Y" or "N" in switches "Yes/No" can be both majuscule and minuscule. Letter "R" or "D" in information about angles units can be both majuscule and minuscule.

Example of the control data. A detailed description of every parameter is given below.

```
!
!       Script for the program AVERAGE (version 1.0)
!
! Code of step  (up to 4 symbols)
a9
!
! Input file containing variants for averaging.
! In this example, an output file of the program GENMEM is used to be processed.
a8_gmem.out
!
! Space group number (in accordance with the International Crystallographic Tables)
19
!
! Unit cell parameters (in A and degrees)
34.900   40.300   42.200  90.00  90.00  90.00
!
! The key (Y/N) for the presence of reference phases in the input
!   file and the [optional] key (D/R) for phase values unit
!   (degree/radian) in the input and output files
!   e.g.   Y  D   or  N  R  or Y  etc.
Y
!
! The following data define the mode to calculate the map correlation
!   coefficient:
```

```
!
!   Whether Fobs will be used for all variants (Y) or individual
!     magnitudes will be taken from the input file for every
!     variant (N)
Y
!   Resolution limits (Dmin, Dmax) for the maps calculation and
!     [optional] the step (in A) for the search of the optimal
!     alignment (may be important in the case of monoclinic or
!     triclinic groups, e.g.)
12. 9999. 2.
!   Whether the map flip is considered as a permitted map
!     transformation(Y/N) in the alignment
N
! Number of resolution zones to calculate FOM-statistics
5
! Resolution limits (Dmin, Dmax) for every  zone
16. 9999.
13. 9999.
12. 9999.
8. 9999.
4. 9999.
!
!               the end of the script
```

4.1. Name of the calculation step

(! step code (from 1 to 4 symbols))

This is a label (4 character or less) which precedes the name of any file created at a given program run. For example, if the step code is defined as s1 , the names of crated files will be

s1_ave.out - selected phase sets;

s1_avemes - output message file;

s1_ave.con - output screen copy

It is convenient to reserve a part of this label for the number of consecutive computational step and increase this number accordingly.

4.2. Input file of phase sets (format .OUT).

(! The path for the input file containing variants for averaging)

The name of file (not more than 72 symbols) which contains the phase sets for averaging.

Its structure is the following :

First record (format A72) - text, a copy of the first title of the input file of structure factors.

Second record (format A72) - text, a copy of the second title of the input file of structure factors.

Third record (format I4) - integer NREF - the number of reflection in the input file of structure factors.

Then NREF records follow (format 3I4, 2G12.4), one per reflection, each containing Miller indices H, K, L, experimental value for the structure factor magnitude and the test flag equal to 1 for work reflections and 0 for the reflections from the test set. This information is copied from the input file of structure factors. If the column number for the test flag was defined as 0, all test flags are assigned to be equal to 1 and all reflections are considered as the reflections from the work set. In any case, program AVERAGE does not distinguish the reflections by this flag for its calculations.

Then the file contains the records (format (6G12.5) or (6G12.6)), one per selected phase set. Every record contains 2*NREF+1 real numbers. It starts from a real number (the value of the criterion by which this phases set was selected) phase correlation with the reference values), then NREF real values for corresponding phases Ph and then NREF real numbers for the corresponding values T. These latter values may be either some weights or some structure factor magnitudes, depending on the program created this file.
The phase values can be presented either in degrees or in radians (the same units for all phases in the file)); the formats ((6G12.5)) or ((6G12.6)) are used respectively.
The first record of such type may contain some special values called reference phases which are used for comparison with them. This can be some phase estimates found by alternatively; it can be noted that these values may have nothing to the true solution. The parameter "Y/N" placed in file of control data after unit cell information indicates whether the first record in file is the record with reference phases or not; "Y" means that the phases from the first record will be interpreted as the reference values. As a consequence, they will not be included in averaging.

4.3. Space group number
(! space group number)
This integer should be equal to the number of the space group in the International Crystallographic Tables. Necessary information on this space group must be presented in the file FAM.ITC a copy of which is provided with the program. It can be completed if necessary for space groups not yet included by the authors.

4.4. Unit cell parameters
(! unit cell parameters (in A and degrees))
6 real parameters – periods of the unit cell (in Angstroms) and angles (in degrees).

4.5. Information about reference phases
(! The key (Y/N) for the presence of reference phases in the input
! file and the [optional] key (D/R) for phase values unit
! (degree/radian) in the input and output files
! e.g. Y D or N R or Y etc.)
Fist phase set in the file .OUT is treated as some special set of phases which are used as a reference and are not included into averaging if the key «Y» is defined.
For the first set of phases (even if the reference phases are not defined) the program calculates the mean absolute value. If this value is smaller than 5, the phases are supposed to be in radians, if it is larger than 50, it is supposed to be in degrees. If this estimation contradicts to the input parameter, the program stops and complains. For the intermediate case, when the program cannot identify the units unambiguously, the units are used as they are defined in control data.

Remark. The value of 1.e+10 is used to indicate that the given phase is not defined. These values do not change during angle transformations from radians to degrees et vice versa.

4.6. Type of structure factor magnitudes
(! Whether Fobs will be used for all variants (Y) or individual
! magnitudes will be taken from the input file for every variant (N))
This parameter answers whether experimental magnitudes will be used for the calculation of synthesis ("Y") or every phase set will be used with its individual magnitude ("N").

4.7. Resolution zone for the calculation of distances between the phase sets

(! Resolution limits (Dmin, Dmax) for the maps calculation and
! [optional] the step (in A) for the search of the optimal
! alignment (may be important in the case of monoclinic or
! triclinic groups, e.g.)

Two first real numbers define the resolution limits (in any order) in Angstroms for a spherical shell zone of a reciprocal space where the distance between phase sets is calculated (see **9, Definition of basic concepts,** *Correlation and distances between two phase sets***).**
Before calculate the distance, phase sets are preliminary aligned by all origin shifts allowed for a given space group and the enantiomer transformation if possible (defined in the file FAM.ITC). If for one of axes any origin choice is allowed (for example, axe Y for the space group P21) then the third parameter of this line defines the step with which origin shifts along this axis will be checked. If this value is absent in the control data, the program defines it as Dmin/4.

4.8. Synthesis flipping

(! Whether the map flip is considered as a permitted map
! transformation(Y/N) in the alignment)

This parameter defines whether an extra operation of synthesis flipping is used during the phase alignment. For example, if 4 origins are possible, an enantiomer substitution is allowed and the synthesis flipping is forbidden then the correlation of the calculated synthesis with the control one is defined as the maximum of 8 numbers, one per a combinations of the origin and enantiomer choice. At the same time, if we allow also to flip the calculated synthesis, the number of combinations increases to 16 because for any synthesis its flipped image is also considered as possible.

4.9. Information for FOM-statistics

```
! Number of resolution zones to calculate FOM-statistics
5
! Resolution limits (Dmin, Dmax) for every  zone
16. 9999.
13. 9999.
12. 9999.
8. 9999.
4. 9999.
!
```

This line contains integer number NZON. If it is not equal to zero, following NZON lines everyone containing 2 real values, the resolution limits of zones for FOM-statistics (in any order).
Mean FOM values in different zones on resolution can be used as some characteristic of the quality of average phases. A high value of mean figure of merit shows that the phases are close each to other and therefore there is a hope that they all are around the true solution and as a consequence the result of averaging of these phases will be not far from the true solution too.

5. Input file FAM.ITC (International Crystallographic Tables).

Input file FAM.ITC contains the basic information on crystallographic space groups in the form convenient for the program. Below there is an information on the space group P212121.

If a necessary space group is not yet included in this file this can be done in a similar way either by the user or by the authors.

NEWGROUP P212121
Title for a block of information (space group)

19 (the group number)
number of the space group (programs which use this file use this number to identify the block of information)

4 (number of symmetries)
number of the symmetry operation for a given space group

```
 1 0 0 0  1 0 0 0  1  0  0  0
-1 0 0 0 -1 0 0 0  1 .5  0 .5
 1 0 0 0 -1 0 0 0 -1 .5 .5  0
-1 0 0 0  1 0 0 0 -1  0 .5 .5
```

elements of the symmetry equations written as following :
r_{11}, r_{21}, r_{31}, r_{12}, r_{22}, r_{32}, r_{31}, r_{32}, r_{33}, t_1, t_2, t_3

3 (number of centrosymmetric zones)
number of centrosymmetric zones in reciprocal space

```
0 0 1 .5  0  0
0 1 0  0  0 .5
1 0 0  0 .5  0
```

every of centrosymmetric zone is defined by 6 parameters :
m1, m2, m3, a1, a2, a3:
reflection hkl belongs to a given zone if the following equality is verified : m1*h + m2*k + m3*l = 0;
then allowed phase values are :
$\alpha = (a1*h + a2*k + a3*l)*\pi$ or $\alpha + \pi$.

0 (number of axes with any shift of origin)
defines the number of axes (0, 1 or 3) such that any shift along them gives an allowed position for the unit cell origin

8 (number of the possible origin choices)
the number of variants for the discrete choice of the origin; if a continuous shift along a coordinate axis is allowed, it is applied to any of the discrete choices of the origin;

```
 0  0  0
.5  0  0
 0 .5  0
.5 .5  0
 0  0 .5
.5  0 .5
 0 .5 .5
.5 .5 .5
```

corresponding choices of the origin (origin shifts)

1 (possibility of enantiomorph choice, 1 - if possible)
this parameter defines whether this group and its enantiomer coincide

==========

6. Output file of structure factors

The output file of the program is the file of structure factors named as
<code_of_step>_ave.uf

This formatted file is organised by the following way (so called UF format) :
- first record (72 characters) – first file title; its first 22 symbols are «Output data of AVERAGE», followed by the unit cell parameters and the space group number;
- second record (72 characters) – second file title; it is

H K L Fobs Test_flag FOM Ph_ave T_calc Ph_calc

when reference phasesare absent and

H K L Fobs Test_flag FOM Ph_ave T_calc Ph_calc Ph_ex

otherwise.
- third record (format I4) - integer number LREC (equal or less than 100) which defines the length of following records (number of data per record; we will also use "the number of columns" as a synonym) ;
- all following records are similar; every record corresponds to one reflection and consists of LREC numbers (like columns of a matrix) ; first 3 of them are integer (Miller indices H,K,L of the reflection) while others are real and contain different information on the reflection. Corresponding record format is

(3I4, 5G12.6) or (3I4, 5G12.6)

for LREC less or equal to 8, and

(3I4, 5G12.5/(6G12.5)) or (3I4, 5G12.5/(6G12.5))

otherwise.
In this file all records contain the same type of information in the same position, according to the second title (experimental magnitudes in the position 4, averaged phases in the column 7, their figures of merit in the column 6, etc.).
Remark. Format G12.6 is used when the angles in the input file (therefore in the output file) are given in radians. If they are given in Angstroms, format G12.5 is used.

7. Output message file
The massage file is named <step_code>_ave.mes.
Here there is an example of message file of the program AVERAGE:

```
                    AVERAGE               Version 1.0   18.06.2001
 Input file name: a8_gmem.out
 Input file titles are:
 protg   34.9 40.3 42.2 90. 90. 90.   P212121
 h k l d Fobs F(mod) Phi(mod)
 Space group number:  19
 Unit cell:   34.90   40.30   42.20    90. 90. 90.
 Whether reference phase values are present in the input files: Y
 RADIAN/DEGREE information is not defined in the control data
 Maps alignment mode:
   Whether Fobs will be used for all maps: Y
   resolution limits:   12.00 - 9999.00
   minimal trial step (in A) for the origin shift:    2.00
   whether the map flip is a permitted transformation: N
 Number of resolution zones for FOM-statistics:  5
   the corresponding resolution limits:
      16.00- 9999.00
      13.00- 9999.00
      12.00- 9999.00
       8.00- 9999.00
       4.00- 9999.00

 ********** Input file statistics ******************************
```

```
Phases are interpreted to be in radians
     20 phase variants were found in the input file
The search criterion values (qmin, qmax, ave, rms):
     0.1801       0.8705       0.5748       0.1907

Map correlation with respect to the reference phases:
  cmin,cmax,ave,rms:   0.2098  0.6915  0.3880  0.1223

   0    0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0    0
   0    0    0    0    1    3    5    6    2    0
   0    0    1    2    0    0    0    0    0    0

Mean radius of the cloud:
rmin,rmax,ave,rms:     1.00    1.13    1.05    0.04
   0    0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0    1
  11    6    2    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0    0

********** Averaging ****************************************

The program has defined the variant   18 as the middle point

Map correlation of phase variants with respect to the middle phases:
  cmin,cmax,ave,rms:   0.2487  1.0000  0.5047  0.1613

   0    0    0    0    0    0    0    0    0    0
   0    0    0    0    0    0    0    0    0    0
   0    0    0    0    1    0    2    1    5    3
   1    4    1    0    1    0    0    0    0    1

Map correlation with respect to the reference phases for maps:
(F_obs,P_ave)-map:       0.2429
(F_obs*FOM,P_ave)-map:   0.4825
(F_obs,P_calc)-map:      0.2429
(T_calc,P_calc)-map:     0.3194

FOM-statistics:

resolution   ref   <FOM>
 16.-9999.    15  0.3426
 13.-9999.    22  0.3407
 12.-9999.    28  0.3364
  8.-9999.    85  0.2459
  4.-9999.   580  0.2005
```

All information from the input file of control data are reflected in the message file.
If the information on the phase units is not defined in the control data, the program types the result of its own interpretation of phase units (the message file above says that the phases are interpreted in radians; this means that the output phases will be in radians too). Then the message file contains:
1. Some statistics of input criteria (minimal, maximal, mean and r.m.s. values) given as the first real number in the phase set records in the input file (see 4.2).

2. Statistics of correlation between input phases and reference phases (minimal, maximal, mean, r.m.s. values and the histogram of its values).
3. Characteristics of the “cloud” of input phase sets (mean radius). For every phase set, the mean distance to all other phases sets is calculated. The statistical characteristics of the ensemble of these values as well as the histogram of its distribution is given.
4. Number of variant which as chosen as the “middle point” for this “cloud”.
5. Statistics of correlation respect to the “middle” phases (minimal, maximal, mean, r.m.s. values and the histogram of its values).
6. The correlation value between the map, calculated the reference phases if they are available, and four maps calculated with the coefficients : (F_obs,P_ave), (F_obs*FOM,P_ave), (F_obs,P_calc), (T_calc,P_calc). Our experience shows that the second value seems to be the most significant among these values.
If in the input file all individual modules are equal to 1, the first and the third values must be the same.
7. The mean value of the figure of merit, calculated in given resolution zones.

8. Program limitations
- maximal number of structure factors in the input file - 5000;
- maximal number of symmetry operation for a given space group - 48;
- maximal number of the centrosymmetric zones in the file FAM.ITC - 20;
- maximal number of the discrete origin shifts in the file FAM.ITC - 8;
- maximal size of the array used to calculate the correlation for the map alignment for continuous origin shift along an axis – 10000; acell/dstep for a monoclinic group (with the rotation axis **a**), bcell/dstep for a monoclinic group (with the rotation axis **b**) or acell*bcell*ccell/dstep^3 for a triclinic group should not exceed this value; otherwise either the parameter should be increased and the program recompiled or simply the parameter ‘step’ can be increased;
- maximal number of variants – 1000 ;
- maximal number of zones for FOM-statistics – 20 ;
- maximal number of bins for the histogram calculation – 40.

It is always possible to modify these restrictions and recompile the program ; however it is recommended first to adapt your data to these restrictions because they are derived from a large number of numerical experiments.

9. Definition of basic concepts

Phase set (variant)
This is a set of real numbers, one per reflection; every number is the phase for the corresponding structure factor; they can be expressed in radians or in degrees (see **4.5**). A "phase variant" can be used as synonyms for the "phase set".

Reference phases
This is a phase set with respect to which all phase correlations are calculated. These phases can be calculated from a known model or in some other way. Reference phases are used for the statistical analysis of the correlation distribution for the input phase sets. Eventually, reference phases can have nothing to the exact solution of the phase problem for the given crystal.

Correlation and distance between two phase sets

The closeness of two phase sets $\{\varphi_1(\mathbf{h})\}$ and $\{\varphi_2(\mathbf{h})\}$ is defined through the correlation between two corresponding maps:

$$C = \frac{\int \rho_1(\mathbf{r})\rho_2(\mathbf{r})dV_\mathbf{r}}{\sqrt{\int \rho_1(\mathbf{r})^2 dV_\mathbf{r} \int \rho_2(\mathbf{r})^2 dV_\mathbf{r}}} = \frac{\sum_\mathbf{h} F_1(\mathbf{h})F_2(\mathbf{h})\cos(\varphi_1(\mathbf{h}) - \varphi_2(\mathbf{h}))}{\sqrt{\sum_\mathbf{h} F_1(\mathbf{h})^2 \sum_\mathbf{h} F_2(\mathbf{h})^2}}$$

The structure factor magnitudes used for the map calculation can be either experimental values, the same for both maps, or individual magnitudes, different for these two phase set. Weighting of structure factor magnitudes is also possible. It is necessary to note that the value of correlation depends on the set of reflections (*i.e.*, on zone of resolution) used for its calculation.

Sometime, the distance between two phase set is more convenient measure than the phase correlation. The distance can be defined as

$$D = \sqrt{\int \left(\frac{\rho_1(\mathbf{r})}{\sqrt{\int \rho_1(\mathbf{r})^2 dV_\mathbf{r}}} - \frac{\rho_2(\mathbf{r})}{\sqrt{\int \rho_2(\mathbf{r})^2 dV_\mathbf{r}}} \right)^2 dV_\mathbf{r}} = \sqrt{2(1-C)}$$

The less is this value the closer two phase sets are.

Maps alignment

Two similar maps, calculated with the different choice of the origin, correspond to quite different, at the first glance, phase of sets, and vice versa. Therefore, all origin shifts allowed for the given space group should be checked in order to find the maximal map similarity before the formal maps correlation is calculated.

If the space group coincides with its enantiomer , this operation should be also included in the list of allowed operations for the map alignment.

Since the operation of density flipping $\rho(\mathbf{r})$ -> $-\rho(\mathbf{r})$ does not change the structure factor magnitudes, this operation can be also included in the list of allowed map transformations if necessary; this can be useful at some early stages of phasing at low resolution.

All phase or map correlations or distances calculated in the program are calculated always for the optimally aligned phase sets.

10. **References**

1. Lunin, V.Yu. & Woolfson, M.M. (1993) “The mean Phase Error and the Map Correlation Coefficient”. Acta Cryst. **D49**, 530-533.

2. Lunin, V.Yu. & Lunina, N.L. (1996) “The Map Correlation Coefficient for Optimally Superposed Maps”. Acta Cryst. **A52**, 365-368.

3. Lunin V.Y., Lunina N.L., Petrova T.E., Skovoroda T.P., Urzhumtsev A.G. & Podjarny A.D. (2000) “Low-resolution ab initio phasing: problems and advances”. Acta Cryst. D**56**, 1223-1232.