14.09.2001

GENMEM : program description

(this description corresponds to the versions from 21.08.2001 and later)

Contents

1. Purpose
2. Basic concepts
3. Phase generation modes
4. Syntheses selection on the base of calculated connectivity characteristics
5. Input file of control data
5.1. Name of the calculation step
5.2. Number of records - step for the screen
5.3. Number of records - step for the screen copy
5.4. Input file of structure factors
5.5. Space group number
5.6. Unit cell parameters
5.7. Column number for the experimental magnitudes
5.8. Column number for the test flags
5.9. Column number for the reference phases
5.10. Resolution zone for the correlation calculation
5.11. Synthesis flipping
5.12. Start value for the random generator
5.13. Maximum number of generated phase sets
5.14. Maximum number of selected phase sets
5.15. Phase generation mode
5.16. Number of resolution zones for connectivity analysis
5.17. Information for a resolution zone
6. Input file FAM.ITC (International Crystallographic Tables).
7. Output message file
8. Output file of selected phase sets
9. Output screen copy
10. Recommended protocol for the first phasing steps
11. Program limitations
12. Definition of basic concepts
13. References

1. Purpose

Program GENMEM developed in order to generate a large number of random phase values for a given set of structure factor magnitudes and to select those that provide with the syntheses possessing necessary topological properties namely connectivity properties.

2. Basic concepts

Generate phase set
Reference phases
Connectivity component
Connectivity properties
Von Mises distribution
Cumulative function of a Fourier synthesis
Transformation reconstructing the cumulative function (or histogram)
Cut-off level

A knowledge of these concepts whose short definition is done at the Section 12 is necessary for understanding the following text. Indicated articles can be read in order to get more details of these definitions and their applications for connectivity-based phasing.

3. Phase generation modes

General note. In the following examples of control data, the lines starting with the exclamation (!) are the lines of comments.

Input file of structure factor magnitudes is a necessary information for the program. Phase generation is a procedure which, for every structure factor, prescribes a phase value allowed in a given space group. In the current procedure, the phases are generated randomly, either uniformly or following a known probability distribution (near an available reference value).

Generation mode 1. Phases are generated uniformly; phase of a non-centrosymmetric reflection can have any value from 0 to 2π and the phase of a centrosymmetric reflection - any of two values admitted for it.

Generation mode 2. Phases are generated uniformly; only phase sets close to the given start phase set are left for further analysis; the closeness is estimated through the phase correlation higher than a prescribed value.

Generation mode 3. Phases are generated following the von Mises distribution around the given start phase set. The phase discrepancy for every individual phase is defined by the figures of merit that also should be provided in the file of structure factors. If the latter information is not available in the file, it can be defined explicitly, as a function of resolution, by a set of values for different resolution zones.

Generation mode 4. This mode joins a random phase generation and their refinement by one of the density modification methods. Phases are generated accordingly to the mode 3. For every phase set, a synthesis with generated phases and experimental magnitudes is calculated and is subjected a non-linear transformation in order to reproduce a typical electron density histogram at a given resolution. For the modified distribution, phases of its Fourier coefficients are calculated and used for further analysis. In order to define the transformation restoring the histogram, corresponding cumulative function of a typical Fourier synthesis at a given resolution is provided in an input file. It can be a cumulative function calculated for a known crystal structure with the Matthews constant F_{000}/V close to that for a structure under analysis.

4. Syntheses selection on the base of calculated connectivity characteristics

For every generated and selected phase set a Fourier synthesis is calculated. Then, for the regions of its highest and/or lowest values some topological characteristics are calculated namely the number of connected regions and groups of regions of an equal volume (for example, those related by symmetry operations). The volume is expressed through the number of grid points which belong to the given connected region. Eventually, two regions of different form can have exactly the same volume and as a consequence will be assigned to the same group; however, in practice it is hardly possible. Connected regions depend on

- synthesis resolution;
- grid on which the synthesis is calculated;
- density cut-off level.

A synthesis is calculated at a given resolution in a given grid of the unit cell. Then, for this synthesis a region of maximal or minimal values is constructed being defined by its relative volume with respect to the unit cell volume (see **Definition of basic concepts,** *Cut-off level*). As a rule, these points form several connected regions. A region is considered as a connected if any its point can be reached from any other its point by a suite of steps parallel to coordinate axes when all these intermediate points also belong to this region. For every such region its volume expressed in the number of grid points can be calculated. Due to a crystallographic symmetry, several copies of the same region can be found. Synthesis description as

```
16   4*1250   4*330   8*2
```

means that the synthesis taken at a given cut-off level shows 16 connected regions, 4 of them are of the same large size (1250 grid points each), 4 are relatively small (330 grid points) and the rest 8 are very small drops of only 2 grid points each. These numbers 4, 4, 8 are called multipliers. In the example above, the first two multipliers are equal to 4, the third one is equal to 8 and the rest are equal to zero. The total number of component is equal to 16.

These topological characteristics can be used to express the selection rules (or selection modes in what follows). If the Fourier synthesis calculated with a generated phase set verifies the given selection rule then the phase set is selected for further analysis otherwise it will be excluded from it. Six selection modes are currently available in the program:

Selection mode 1. This mode is useful when we have some very basic ideas how the correct synthesis looks like, and these ideas can be expressed through two first multipliers.

Possibility 1. Asymmetric part of the unit cell contains a single molecule and the space group is not trivial (for example, contains 4 symmetry operations). We can believe that a good synthesis should show 4 identical molecular envelopes with no noise between them. This condition is defined by two first multipliers equal to 4 and to 0, respectively.

```
! selection mode
1
! two first multipliers
4   0
```

Possibility 2. In the same situation we are ready to accept also the syntheses with some noise but the corresponding ‘drops’ should be small. In this case we need to define the second multiplier different from 0 (for example, equal to 4, which means a single ‘noise’ region per asymmetric unit is possible) and the limits for the ratio of volumes of corresponding connected regions (‘noise’ region to the ‘signal’ region) .

```
! selection mode
1
```

```
! two first multipliers
4   4
! min & max value for ratio of components (Vol_2/Vol_1)
! this ratio has to be near zero for almost clear synthesis (0.0  0.2, for example)
0.0   0.2
```

Possibility 3. If an asymmetric unit contains 2 molecules related by a noncrystallographic symmetry, volumes of their images calculated in a grid will be close each to other. This situation can be described as following :

```
! selection mode
1
! two first multipliers
4   4
! min & max value for ratio of components (Vol_2/Vol_1)
! this ratio has to be close to 1 for a non-crystallographic symmetry (0.8 1.0, for example)
0.8   1.0
```

It should be noted that the possibilities 2 and 3 (selection of practically 'clean' syntheses when a non crystallographic symmetry exists) cannot be mixed in this selection mode; selection mode 6 can be used to do so (see below).

Selection mode 2. This mode is useful when we know only that the total number of connected regions is relatively small but is not known exactly. In this case, a phase set is selected if it produces a Fourier synthesis where a total number of connected components is **less than or equal** to a given number. First multiplier can be verified if necessary.

Selection mode 3. Sometimes, an extra structural information allows us to know the **exact number** of connected components. The current mode 3 allows to select the corresponding phase sets. The first multiplier can be verified.

Selection mode 4. Sometimes, we cannot estimate the total number of connected regions however the composition of largest regions is known. The mode 4 allows to define first N multipliers; the number N is a parameter itself. A phase set is selected if the corresponding synthesis has first N multipliers as demanded.

Selection mode 5. This mode is used when the conditions of the previous mode 4 are extended and 'more clean' syntheses are also allowed. In this mode, the first multipliers of the connected regions of the calculated synthesis should be either equal to the given values (which, in particular, may be equal to zero).

Selection mode 6. Current mode 6 can be used for phase set selection if, on the top of knowledge of the first multipliers, an information on the volumes of largest regions is available. In this mode, N first multipliers must be defined as well as the N-1 volume ratios : volume of every next region to the volume of the preceding region. Naturally, all these ratios should vary between 0 and 1.

General remark. It is natural to expect that one can propose a selection criterion which is hardly formulated using these 6 modes. Please, contact the authors of the program who can help you either to formulate your needs or to extent the possibilities of the program.

5. Input file of control data

General remarks. The file of control data allows to use comment lines - they are those started from the exclamation '!'. Therefore, a beginner can take some available complete file of control data and adapt it by converting unnecessary lines to comments. It is strongly recommended to leave all comments in the file because this facilitates further modification of parameters and decreases eventual errors. Free format is used for the control data. Letter 'Y' or 'N' in switches 'Yes/No' can be both majuscule and minuscule.

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

```
! data for program GENMEM dated 17.05.2001 and later
!  step code (from 1 to 3 symbols)
s1
! number of records - screen step
100
! number of records - step for file <code of step>_gmem.con
500
! input file of structure factors (UF-format)
protg4a.uf
! space group number
19
! unit cell parameters (in A and degrees)
34.900   40.300   42.200  90.00  90.00  90.00
! column number for observed modules
5
! column number for test-flag
0
! column number for reference phases (0 if not available) and radian/degree information (R/D)
7 R
! resolution limits for correlation with reference phases
16. 9999. 5.
! whether the calculated synthesis may be flipped (Y/N)
N
! start value for the random numbers generator
70191321
! max number of generated phase sets
300000
! max number of selected phase sets
20
! phase generation mode (1-4)
! 1 - uniform distribution
! 2 – generation near the start phase set
!     (it is necessary to define the column number for phases Ph and their figures of merit FOM)
! 3 - generation by von Mises law
! 4 - generation by von Mises law with density modification
1
!2
!3
!4
! --- parameters for gen_mode=2, 3 or 4 ---
! column number for the start phases
!7
```

```
! column number for FOM (0 if not available)
! and NZFOM - number of zones for given FOM
!6  3
! if NZFOM is not equal to 0, then
! resolution limits and FOM_value for every zone
! 16.       9999.      .80
! 12.          16.      .50
!   4.          12.      .01
! --- parameters for gen_mode=2 ---
! resolution limits for phase comparison with start values
!16.   9999.   2.
! minimum correlation value for phase comparison
!0.7
! whether the calculated synthesis may be flipped (Y/N)
!N
! max number of attempts to generate a phase variant close in correlation
!1000
! --- parameters for gen_mode=4 ---
! resolution limits for start synthesis
!4. 9999.
! grid numbers for synthesis calculation: nx, ny, nz
!36 40 40
! file with a cumulative function of electron density
!rna3.cum
! --- parameters common for all generation modes ---
! number of resolution zones for connectivity analysis
2
! ========== parameters for the first resolution zone ==========
! resolution limits for connectivity analysis
16. 9999.
! grid numbers for synthesis calculation: nx, ny, nz
36 40 40
! cut-off levels (see Section 4 for explanations)
  0.0  0.1
! whether endless connected regions are accepted (Y/N)
N
!Y
! selection mode:
! 1 - two first multipliers  and limits for their ratio
!      (if the second multiplier is different from zero);
! 2 - common number of areas (<=) and the first multiplier;
! 3 - common number of areas (=)   and the first multiplier;
! 4 - the number of multipliers  and their values (=value);
! 5 - the number of multipliers  and their values (= value or =0);
! 6 - the number N of multipliers (one integer),
!       their values (N integer parameters) and
!       limits for the ratio of corresponding volumes (N-1 pairs of real values between 0 and 1)
1
! ---  data for selection mode 1: ---
! The accepted number of connected regions (2 integer parameters)
!   (the 2-nd parameter equal to -1 means no limitations for the second multiplier)
```

```
4 0
!4  4
! min and max values for the ratio of volumes of the 2-nd and the 1-st connected regions
! 0.9  1.0
! --- data for selection mode 2 or 3: ---
! limit for the number of connected regions  and the 1st multiplier
! (-1 if any value is allowed)
! 4 4
! --- data for selection mode 4 or 5 : ---
! the number of multipliers
! and their values:
!2
!4 0
! ========== parameters for the second resolution zone ==========
! resolution limits for connectivity analysis
12. 9999.
! grid numbers for synthesis calculation: nx, ny, nz
36 40 40
! cut-off levels (see Section 4 for explanations)
  0.0 0.1
! whether endless connected regions are accepted (Y/N)
!N
Y
! Phase selection mode:
! 1 - two first multipliers  and limits for their ratio
!     (if the second multiplier is different from zero);
! 2 - common number of areas (<=) and the first multiplier;
! 3 - common number of areas (=)   and the first multiplier;
! 4 - the number of multipliers  and their values (=value);
! 5 - the number of multipliers  and their values (= value or =0);
! 6 - the number N of multipliers (one integer),
!      their values (N integer parameters) and
!      limits for the ratio of corresponding volumes (N-1 pairs of real values between 0 and 1)
2
! --- data for selection mode 1: ---
! The accepted number of connected regions (2 integer parameters)
!   (the 2-nd parameter equal to -1 means no limitations for the second multiplier)
!4 0
!4  4
! min and max values for the ratio of volumes of the 2-nd and the 1-st connected regions
! (volume is calculated as the number of grid points which belong to the region)
! 0.9  1.0
! --- data for selection mode 2 or 3: ---
! limit for the number of connected regions and the 1st multiplier
! (-1 if any value is allowed)
 8  -1
! --- data for selection mode 4 or 5: ---
! the number of multipliers
! and their values:
!2
!4 0
```

5.1. Name of the calculation step
(! step code (from 1 to 3 symbols))
This is a label (3 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 created files will be
s1_gmem.out - selected phase sets;
s1_gmem.mes - output message file;
s1_gmem.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.

5.2. Number of records - screen step
(! number of records - screen step)
This number defines the frequency with which the current status of the phase selection will be **displayed at the screen**. For example, if this parameters is equal to 100, then during the run the number of selected phase sets after every 100 set generation will be reported as, for example :

```
  100/    43/     7
  200/    89/   12
  300/   139/  20
```

This means that 43 phase sets from the first 100 generated verify the first selection condition and that 7 of them verify both conditions; for the first 200 generated phase sets these numbers are 89 and 12, respectively, etc.
It is important to note that a too large value for this parameter leads to an absence of any message at the screen except the last one which appears in any case.

5.3. Number of records - step for the screen copy
(! number of records - step for file <code of step> _gmem.con)
This number defines the frequency with which the current status of the phase selection will be **printed in the file (screen copy)**. Eventually, this parameter can be equal to the screen step but not necessarily. For example, if the step for the screen copy is equal to 500, then during the run the number of selected phase sets after every 500 set generation will be reported as :

```
  500/  239/    35
 1000/  471/    65
 1500/  704/    91
```

This means that 239 phase sets from the first 500 generated verify the first selection condition and that 35 of them verify both conditions; for the first 1000 generated phase sets these numbers are 471 and 65, respectively, etc.
It is important to note that a too large value for this parameter leads to an absence of any message at the screen except the last one which appears in any case.

5.4. Input file of structure factors.
(! input file of structure factors (UF-format))
This line contains a suite of characters first 12 of which are considered as the file name. This formatted file is organised in the following way (so called UF format) :
- first record (72 characters) – first file title
- second record (72 characters) – second file title
- third record (format I4) - integer number LREC (equal or less than 100) which defines the length of following records (the number of data per record; we will also use ‘the number of columns’ as a synonym) ;

- all following records are similar, composed everyone of LREC numbers per reflection (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)

for LREC less or equal to 8, and

(3Ii4, 5G12.6/(6G12.6))

otherwise.

It is supposed that for a given file all records contain the same type of information in the same position (for example, experimental magnitudes in the position 5, corresponding sigmas in the position 6, resolution in the position 4, etc.)

5.5. Space group number

(! space group number)

This integer number 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 which are not yet included by the authors.

5.6. Unit cell parameters

(! unit cell parameters (in A and degrees))

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

5.7. Column number for the experimental magnitudes

(! column number for observed modules)

This integer number indicates the position (the number of the corresponding column) of experimental structure factor magnitudes in the input file. This number can be between 4 (first 3 positions of every record are reserved for Miller indices H, K, L) and LREC.

5.8. Column number for the test flags

(! column number for test-flag)

This integer number indicates the position (the number of the corresponding column) of the test flag in the input file. This flag is equal to 1 for work reflections and is equal to 0 for test reflections. This number can be between 4 (first 3 positions of every record are reserved for Miller indices H, K, L) and LREC.

Program GENMEM does not distinguish work and test reflections and treat them similarly. Nevertheless, this program can be a used as a step in a chain of structure determination where other programs do use test reflections. Therefore, this information is not lost and still be available for further steps of structure determination and analysis.

If the input file does not contain this information (defined by column number equal to 0) then all reflections are declared as work reflections and the corresponding flag is assigned to be equal to 1.

5.9. Column number for the reference phases

(! column number for reference phases (0 if not) and radian/degree information (R/D))

The integer number indicates the position in the input file (the number of the corresponding column) of the phases which can be considered as the reference ones, and the character R or D which indicate the units in which these phases are expressed (radians or degrees). If the input file does not contain this information, corresponding column number must be defined as 0 otherwise this number can be between 4 (first 3 positions of every record are reserved for Miller indices H, K, L) and LREC. The results of the program do not depend on these phases in any case.

Reference phases are used for analysis of the distribution of generated and selected phase sets. When they are not available (usual situation for an unknown structure) the first generated phase

set is used for such statistical analysis and the correlation of all other phase sets is calculated with respect to this first phase set. It should be noted that the first generated phase set is not necessary a selected one.

When some phase estimates are available before the connectivity search, using of these estimates as reference phases will facilitate to compare GENMEM results with these values.

The character for the radian/degree information follows the column numbers being separated by one space at least. Valid characters are R, D, r, d. The phases in all input and output files must be defined in the same units: in radians when R or r are indicated and in degrees when it is D or d.

In order to avoid a mistake an explicit definition of the phase unit is highly recommended. It is reminded that all input phases should be expressed in the same units.

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

Remark 2.

In any case, the program tries first to identify the units itself. If there is no input phases, neither reference nor trial – see below 5.15), the output phases will be written in degrees if otherwise is not declared by this control parameter.

If input phases exist, the program calculates mean absolute value for reference phases (or for trial values for the modes 2,3, or 4 when the reference phases are absent). If this mean 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. For the intermediate case, the program stops with the message that a complementary information is necessary.

In the case of any contradiction between the units, defined in the control data and estimated by the program, a warning will be printed.

5.10. Resolution zone for the correlation calculation

(! resolution limits for correlation with reference phases)

Two first real numbers (DMIN, DMAX) define the resolution limits in Angstroms for a spherical shell zone of a reciprocal space where the phase correlation is calculated (with reference phases or with the phases from the first generated phase set, see the previous Section). The order of these numbers is arbitrary. For correlation calculation, phase sets are preliminary aligned by all origin shifts allowed for the 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. By default, DSTEP is assigned as DMIN/4 .

5.11. Synthesis flipping

(! whether the calculated synthesis may be flipped (Y/N))

This parameter defines whether an operation of synthesis flipping, which also does not influence the structure factor magnitudes, is allowed during the phase alignment. If the number of origin choices is equal to 4, 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 everyone corresponding to syntheses correlation with one of origin and enantiomer choices. At the same time, if we allow to flip the calculated synthesis, the number of combinations increases to 16 because for any synthesis its flipped image is also considered as possible.

5.12. Start value for the random generator

(! start value for the random numbers generator)

This is a large (7 - 8-digits) integer number which defines the sequence of pseudorandom numbers for phase generation. The same start value defines exactly the same phase values which can be important in some situations.
Sometime, one would like to continue phase generation starting from the current values of the generator. In order to permit this, the message file contains the latest value of this parameter so that it can be used as a start value for the next run.

5.13. Maximum number of generated phase sets

(! max number of generated phase sets)
This integer indicates the iteration when the phase generation should be interrupted if other limitation are not yet reached.
We recommend to use a small value (few hundred or thousand) for initial tuning of parameters.

5.14. Maximum number of selected phase sets

(! max number of output records)
This integer shows the maximal possible number of selected phase sets if the run is not interrupted by other reasons, for example, by reaching the maximum number of generated phase sets.
It should be noted that if reference phases are available (corresponding column number is defined to be different from 0) they will be kept as the first phase set in the output file; therefore, if for example, 100 is defined as the maximum number of selected phase sets, the file will contain 101 phase set instead of 100

5.15. Phase generation mode

```
! phase generation mode (1-4)
! 1 - uniform distribution
! 2 - generation near the start phase set
!     (it is necessary to define the column number for phases Ph and their figures of merit FOM)
! 3 - generation by von Mises law
! 4 - generation by von Mises law with density modification
1
!2
!3
!4
! --- parameters for gen_mode= 2, 3 or 4 ---
! column number for the start phases
!7
! column number for FOM (0 if not available)
! and NZFOM - number of zonez for given FOM
!6  3
! if NZFOM is not equal to 0, then
! resolution limits and FOM_value for every zone
! 16.       9999.      .80
! 12.          16.     .50
!   4.         12.     .01
! --- parameters for gen_mode=2 ---
! resolution limits for phase comparison with start values
!16. 9999. 2.
! minimum correlation value for phase comparison
!0.7
! whether the calculated synthesis may be flipped (Y/N)
!N
```

```
! max number of attempts to generate a phase variant close in correlation
!1000
! --- parameters for gen_mode=4 ---
! resolution limits of start synthesis
!4. 9999.
! grid numbers for syntheses calculation: nx, ny, nz
!36 40 40
! file with a  cumulative function of electron density
!rna3.cum
```

This concept of 'Phase generation modes' is defined in Section 3. As it follows, this parameter is an integer which varies from 1 to 4 for the current version of the program.

Generation mode 1 (uniform distribution) does not need any extra information.

Generation mode 2 (generation near the start phase set) needs to define the phase set called start phases and the procedure of synthesis comparison :

1) column number for the start phases;
2) column number for corresponding figures of merit and the number of zones where these values will be assigned in the control data below;
3) if the number of zones for manual assignment of figures of merit is different from zero, then the corresponding lines follow; every line contains the resolution limits in Angstroms and the universal value of the figure of merit which will be assigned to all reflections from this resolution zone;
4) resolution limits for syntheses calculation (with generated and with the trial phases) in order to calculate their correlation;
5) possibility to flip the synthesis for correlation calculation;
6) minimal correlation value for which a generated phase set is accepted; a generated phase sets whose correlation (calculated as defined before) with the trial phase set is below this threshold will be skipped;
7) maximal possible number of essays to get a phase variant;

Parameters 1 - 2 defines the start phase set. At the beginning, values for all figures of merit are assigned accordingly to the first parameter, the column number for the figures of merit ; if this parameter is positive, the values are taken from corresponding column, otherwise they are assigned to be equal to 1 (and not to 0 as it is for the Generation Modes 3 and 4). Then, for the resolution zones indicated, an unique given figure of merit is assigned to all reflections of this zone. Parameters 3 - 5 define the way to compare the generated phase set with the start values. The reason to introduce the parameter 7 is the following. Occasionally, the generation parameters can be chosen such that the random phases are always far from the start phase set. In this case the program will be interrupted after this limit it exhausted (for example after it failed to select at least one phase set from 1000 with the correlation higher than 0.7 with the start phase set).

Remark : It is important to note that the modified figures of merit are used only for the map correlation; the output file always contains initially calculated non corrected values.

Generation mode 3 (generation by von Mises law near some phase set) needs to define the start phase set :

1) column number for the start phases;
2) column number for corresponding figures of merit and the number of zones where these values will be assigned in the control data below;
3) if the number of zones for manual assignment of figures of merit is different from zero, then the corresponding lines follow; every line contains the resolution limits in Angstroms and the

universal value of the figure of merit which will be assigned to all reflections from this resolution zone;

Parameters 1 - 2 defines the start phase set. At the beginning, values for all figures of merit are assigned accordingly to the first parameter, the column number for the figures of merit; if this parameter is positive, the values are taken from corresponding column, otherwise they are assigned to be equal to 0 (and not to 1 as it is for the Generation Mode 2). Then, for the resolution zones indicated, an unique given figure of merit is assigned to all reflections of this zone.

Remark : It is important to note that the modified figures of merit are used only for phase generation; the output file always contains initially calculated non corrected values.

Generation mode 4 (generation by von Mises law with following density modification) needs an extra information :

1) column number for the start phases;
2) column number for corresponding figures of merit and the number of zones where these values will be assigned in the control data below;
3) if the number of zones for manual assignment of figures of merit is different from zero, then the corresponding lines follow; every line contains the resolution limits in Angstroms and the universal value of the figure of merit which will be assigned to all reflections from this resolution zone;
4) resolution interval D_start_min, D_start_max to calculate initial Fourier synthesis;
5) grid numbers (three integers); the Fourier synthesis will be calculated at this grid and modified in order to recover the correct histogram ;
6) known cumulative function of the Fourier synthesis used for the density modification (defined in an input file); for example, this can be a cumulative function for a homologous protein with a known structure; it is recommended to use cumulative function calculated at a resolution higher than D_start_min.

Parameters 1 - 2 defines the start phase set. At the beginning, values for all figures of merit are assigned accordingly to the first parameter, the column number for the figures of merit; if this parameter is positive, the values are taken from corresponding column, otherwise they are assigned to be equal to 0 (and not to 1 as it is for the Generation Mode 2). Then, for the resolution zones indicated, an unique given figure of merit is assigned to all reflections of this zone.

File of a cumulative function has the following format. First record contains an integer NPOINT, which defines the number of bins for cumulative function (for the current version this number should not exceed 201; bins are supposed to be of an equal size), and also 4 real numbers ; first two define interval limits for which the cumulative function has been calculated, last two define the mean and rms deviations (not used in the program). Then NPOINT real numbers follow (values of the cumulative function in the increasing order, from 0 to 1) written by the format ((5G15.6)).

Remark : It is important to note that the modified figures of merit are used only for phase generation; the output file always contains initially calculated non corrected values.

5.16. Number of resolution zones for connectivity analysis

This is a positive integer N equal of less than 20. Then N groups of control data should follow every one containing an information for a single resolution zone.

5.17. Information for a resolution zone

1) resolution limits for the connectivity analysis

These 2 values define the reflections to be used in order to calculate the Fourier synthesis topological properties of which will be analysed.

2) grid numbers for syntheses calculation

These 3 integers define the grid at which the syntheses with generated phases will be calculated

3) cut-off levels (see Section 4 for more explanations)

There are two real numbers between 0.0 и 1.0 . They define the share of the unit cell volume which is used for the connectivity analysis.

4) Condition of endless regions

Character 'Y' or 'N' which defines whether endless regions of a high density values are allowed.

5) Phase selection mode.

Various selection modes are described in the Section 4. Syntheses selection is done on the base of calculated connectivity characteristics. Here the type and possible parameter values are given for different selection modes.

Selection mode 1.

a) 2 integers corresponding to 2 first multipliers in the searched synthesis;

b) if the second value is different from 0 and from -1, then 2 real numbers between 0 and 1 should follow; they define the limits for the ratio of volumes of the second and the first (by size) connected regions. For example, the condition

```
! selection mode:
1
4   4
0.0   0.1
```

allows to accept the synthesis with 2 groups of connected regions (4 regions in each) composed of 1230 and 123 points, respectively :

4*1230 4*123

because in this case both multipliers are equal to 4 and the volume ratio is exactly 0.1, and does not accept the synthesis with 2 groups of connected regions (4 regions in each) composed of 1230 and 124 points, respectively

4*1230 4*124

because in this case the volume ratio 124/1230 is higher than 0.1 .

It should be stressed that the third multiplier in the selected synthesis must be equal to 0.

Selection mode 2. In this case, a phase set is selected if it produces a Fourier synthesis where the total number of connected components is **less or equal** to a given number. First multiplier can be verified if necessary otherwise its value should be defined as –1 (no verification).

```
! selection mode
2
! total number of components
! and  first multipliers (-1 if any value is admissible)
20     4
```

Selection mode 3. Very similar to the previous selection mode 2. The only difference is that the condition '**LESS OR EQUAL**' is substituted by the condition '**EXACTLY EQUAL**'. A phase set is selected if it produces a Fourier synthesis where a total number of connected components is **equal** to a given number. First multiplier can be verified if necessary otherwise its value should be defined as –1 (no verification).

```
! selection mode
3
! total number of components
! and first multipliers (-1 if any value is admissible)
20    -1
```

Selection mode 4. In this case the first integer N defines the number of first verified multipliers (in the version from 27.05.2001 this number should not exceed 6). Then N integers follow corresponding to these multipliers. For example, a selection with parameters

```
! selection mode
4
! number of multipliers
3
! values of multipliers
4    4     8
```

will choose phase sets which provide with the syntheses with 16 or more connected regions three largest of which have multipliers 4, 4 and 8, respectively.

Selection mode 5. This mode is an extension of the selection mode 4. In this case, the first integer N defines the number of first verified multipliers (in the version from 27.05.2001 this number should not exceed 6). Then N integers follow corresponding to these multipliers. Syntheses which have corresponding first multipliers equal either to these values or to zero are selected. For example, this criterion with control data

```
! selection mode
5
! number of multipliers
3
! values of multipliers
4    4     8
```

will select phase sets which lead to the syntheses such that verify one of the following conditions :
- the number of the connected regions is equal to 4;
- the number of the connected regions is equal to 8, and the first multipliers are 4 and 4:
- the number of the connected regions is equal to or higher than 16, and the first multipliers are 4, 4 and 8.

If the total number of connected regions higher than 16 is not allowed, this can be defined as

```
! selection mode
5
! number of multipliers
4
! values of multipliers
4    4     8    0
```

Selection mode 6. This mode joins a verification of the multipliers and the volume ratio for the connected regions. In this case the first integer N defines the number of first verified multipliers (in the version from 27.05.2001 this number should not exceed 6). Then N integers follow. Phase sets are selected if the connected regions of the corresponding syntheses have corresponding number of multipliers. Then N-1 pairs of real values, everyone between 0 and 1, follow in control data. N-1 volume ratios are checked. The first pair defines the limits for the ratio of the size of the second to the first (by size) connected regions, the last N-1-th pair defines the limits for the ratio of the size of the smallest region (from the N defined connected regions) to the previous one. For example, for a crystal with 4 crystallographic symmetries and a non-crystallographic symmetry one could accept the syntheses with 2 groups of connected regions of roughly equal size and with some small ‘drops’ the number of which is not important. This condition can be defined as :

```
! selection mode
6
! number of multipliers N
3
! values of multipliers
4    4    -1
! N-1 pair of real values between 0 and 1
```

```
! limits for the ratio VOL(j)/VOL(jj-1)
 0.9     1.0
 0.0     0.2
```

6. 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, as an example, 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 the 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 transformations are written as following :
                              r11, r21, r31, r12, r22, r32, r31, r32, r33, t1, t2, t3
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 the given zone if the following
                                    equality m1*h + m2*k + m3*l = 0 is verified;
                                    in this case, the allowed phase values are :
                                    α = (a1*h +a2*k + a3*l)*π   or α + π .
0   (number of axes with an arbitrary origin shift along it)
                              defines the number of axes (0, 1 or 3) such that any shift along them
                              gives an allowed position for the unit cell origin; such axes do not
                              exist in orthorhombic space groups but, for example, exist in
                              monoclinic space groups (rotation axe);
8   (number of the possible discrete origin positions)
                              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 choice for the origin (origin shifts)
1 (possibility of enantiomorph switch during the phase search; 1 - if possible)
this parameter defines whether this group and its enantiomer coincide or not

7. Output message file

Output message file has the name <step_code>_gmem.mes.
This file contains, first of all, a replica of the control data followed by an information about the number of reflections taken for phasing.
If the control data do not contain an information on the units in which the reference phases are expressed, the program informs the user on automatic choice (in the example below, the phases were recognised to be in radians ; as a consequence, the phases in the output file will be also expressed in radians).
Then the table is printed providing with the connectivity information of unweighted syntheses calculated with the reference phases if available and with all selected phase sets. The title of this table is :

```
Res !cor!cut-off!#area! multipliers and volumes for 6 largest areas
```

The table contain the following information:
- resolution for the last connectivity check;
- correlation of the generated phases with the reference values calculated at the resolution, generally speaking, different from the defined above; the first generated phase set is used for comparison if the reference phases are not available;
- cut-off level
- total number of connected regions
- multipliers and the size of six largest connected regions .

At the end of the table, the numbers of generated and the selected phase sets are printed. If the reference phases are defined they are considered as the first selected set even when they do not verify the selection criteria.
Four following histograms show the distribution of generated phase sets with their correlation with the reference phases. Two former histograms show the distribution for generated (histogram 1) and selected (histogram 2) phase sets, and two latter show corresponding frequencies for the same sets. Some basic characteristics are printed also for every histogram such as minimal, maximal and mean value and rms deviation of the variants.
The message file is terminated by the current value for the random numbers generator that allows to continue the work with the same conditions.

Below there is an example of such message file followed by some comments indicating most important parts of it.

```
------------------------------------------------------------------------------
                       ***  GENMEM  ***                  21.08.2001

          Generation of phase sets and their selection on the base
                          of the connectivity analysis

 Screen  output:           every    100 generations
 s21_gmem.con-file output: every    500 generations
 Input file of structure factors:
   protg4a.uf
 Titles:
   protg   34.9 40.3 42.2 90. 90. 90.   P212121
   h k l d Fobs F(mod) Phi(mod)
```

```
Lrec:    7
Output file in FAM_OUT format: s21_gmem.out
Space group number:   19
Unit cell:   34.90   40.30   42.20    90. 90. 90.
Column number for observed modules  5
Column number for test-flag         0
Column number for reference phases  7 (no RADIAN/DEGREE information)
Resolution/step/flip for calculation
    of phase correlation:  16.00- 9999.00/ 5.00/N
Start value for random generator numbers 70191321
Max number of generated phase sets:    300000
Max number of selected phase sets:         20
Phase generation mode:  1
Number of zones for the connectivity analysis:  2
*** zone  1 ***
    resolution for the connectivity analysis:  16.00- 9999.
    grid numbers nx,ny,nz:    36   40   40
    cut-off levels:  0.000 0.100
    whether a region may be endless (Y/N): N
    selection mode (1-6):  1
    multipliers:   4  0
*** zone  2 ***
    resolution for the connectivity analysis:  12.00- 9999.
    grid numbers nx,ny,nz:    36   40   40
    cut-off levels:  0.000 0.100
    whether a region may be endless (Y/N): Y
    selection mode (1-6):  2
    limit for # of areas:    8, first mult:  -1

     580 reflections are selected
Phases are interpreted to be in radians !!!

-------------------------------------------------------------------------------
 Res !cor!cut-off!#area! multipliers and volumes for 6 largest areas
-------------------------------------------------------------------------------
 16.0 100!.00-.10!    4! 4* 1442! 0*    0! 0*    0! 0*    0! 0*    0! 0*    0
 12.0 100!.00-.10!    6! 2* 2508! 4*  185! 0*    0! 0*    0! 0*    0! 0*    0
 12.0  66!.00-.10!    2! 2* 2884! 0*    0! 0*    0! 0*    0! 0*    0! 0*    0
 12.0  65!.00-.10!    8! 4* 1161! 4*  280! 0*    0! 0*    0! 0*    0! 0*    0
 12.0  61!.00-.10!    2! 2* 2872! 0*    0! 0*    0! 0*    0! 0*    0! 0*    0
 12.0  54!.00-.10!    4! 4* 1437! 0*    0! 0*    0! 0*    0! 0*    0! 0*    0
 12.0  47!.00-.10!    4! 4* 1442! 0*    0! 0*    0! 0*    0! 0*    0! 0*    0
 12.0  75!.00-.10!    4! 4* 1440! 0*    0! 0*    0! 0*    0! 0*    0! 0*    0
 12.0  80!.00-.10!    8! 4* 1284! 4*  164! 0*    0! 0*    0! 0*    0! 0*    0
 12.0  51!.00-.10!    8! 4* 1326! 4*  112! 0*    0! 0*    0! 0*    0! 0*    0
 12.0  55!.00-.10!    8! 4* 1207! 4*  243! 0*    0! 0*    0! 0*    0! 0*    0
 12.0  65!.00-.10!    2! 2* 2886! 0*    0! 0*    0! 0*    0! 0*    0! 0*    0
 12.0  68!.00-.10!    6! 2* 2474! 4*  205! 0*    0! 0*    0! 0*    0! 0*    0
 12.0  20!.00-.10!    8! 4* 1195! 4*  239! 0*    0! 0*    0! 0*    0! 0*    0
 12.0  53!.00-.10!    6! 2* 1586! 4*  652! 0*    0! 0*    0! 0*    0! 0*    0
 12.0  55!.00-.10!    6! 2* 2712! 4*   82! 0*    0! 0*    0! 0*    0! 0*    0
 12.0  62!.00-.10!    8! 4* 1280! 4*  160! 0*    0! 0*    0! 0*    0! 0*    0
 12.0  56!.00-.10!    4! 4* 1436! 0*    0! 0*    0! 0*    0! 0*    0! 0*    0
 12.0  69!.00-.10!    8! 4*  969! 4*  469! 0*    0! 0*    0! 0*    0! 0*    0
 12.0  79!.00-.10!    8! 4*  852! 4*  586! 0*    0! 0*    0! 0*    0! 0*    0
 12.0  35!.00-.10!    8! 4* 1246! 4*  190! 0*    0! 0*    0! 0*    0! 0*    0
 12.0  59!.00-.10!    4! 4* 1439! 0*    0! 0*    0! 0*    0! 0*    0! 0*    0
     20 variants are selected after       57 generations

          ***    Distribution of variants    ***
         with their correlation with reference phases

*** Number of variants ***
for generated variants:
min,max,ave,rms:   0.1186      0.8420      0.4782      0.2022
     0      0      0      0      0      0      0      0      0      0
     0      0      0      0      0      0      0      0      0      0
     0      0      3      1      6      4      4      6      1      4
```

```
     4      5      5      5      3      3      3      0      0      0
 for selected variants:
 min,max,ave,rms:   0.2085      0.8029      0.5929      0.1375
      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      0      1      0      1
      3      4      2      5      0      2      1      0      0      0

 *** Relative frequencies ***
 for generated variants:
      0      0      0      0      0      0      0      0      0      0
      0      0      0      0      0      0      0      0      0      0
      0      0    526    175   1052    701    701   1052    175    701
    701    877    877    877    526    526    526      0      0      0
 for selected variants:
      0      0      0      0      0      0      0      0      0      0
      0      0      0      0      0      0      0      0      0      0
      0      0      0      0    500      0      0    500      0    500
   1500   2000   1000   2500      0   1000    500      0      0      0
 last value of RANDOM NUMBER GENERATOR:      1816728505
```

Remarks

During phase generation, a phase set with the 84,20% correlation with the reference phases at 16 Å resolution was obtained however these phases were rejected by one of the connectivity conditions. The best selected phases give lower correlation of 80.29% . Nevertheless, the selection was successful because it allowed to increase the mean correlation for all selected phase sets from 47.82% up to 59.29%.

Histograms of frequencies show that the selected phase sets contain more ‘good phase sets’ than the initially generated : while for the selected sets the number in the last column became slightly less, 500 instead of 526 before selection, at the same time the number became much higher in the previous column, 1000 instead of 526.

8. Output file of selected phase sets

Output file of selected phase sets has the name <step_code>_gmem.out. Its structure is the following :

First record (format A72) – a text; first 22 symbols are ‘output data of GENMEM’, then unit cell parameters and the space group number follow;

Second record (format A72) – a text; it is one of two following lines :

This is file in .OUT-format. Phases are in degrees

or

This is file in .OUT-format. Phases are in radians

Third record (format I4) – an integer NREF - the number of reflection in the input file of structure factors; all of them are taken into this output file.

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 GENMEM does not distinguish the reflections for the calculations by this flag.

Then the file contains the records in the format ((6G12.5)) or ((6G12.6)), one per selected phase set. Every record contains in the beginning a real number (phase correlation with the reference values), then NREF real values for corresponding phases and NREF numbers for corresponding weights (current version of the program GENMEM does not use weights and these values are always substituted by 1). If the input phases are defined in degrees or absent, the output phases are defined in degrees and the format ((6G12.5)) is used. If the input phases are defined in radians then the output phases are defined also in radians and the format ((6G12.6)) is used.

This format allows the results to be compatible with files created by other phasing programs (FAMREF, RING) and therefore be treated by existing programs of cluster analysis.

9. Output screen copy

Output file of the screen copy has the name <step_code>_gmem.con.

When the program runs, some statistical information is communicated to the user. This information allows to control the selection process. For a long calculation when it is not possible to follow continuously the screen messages, a file with a screen copy can be used. The frequency of messages in the screen and the corresponding screen copy file are defined by two first parameters in the control data.

A screen line contains several numbers indicating the total number of generated phase sets, number of those which satisfy the first connectivity condition, number of those which verify both the first and the second, etc. For example, when two conditions are verified, these lines look like :

```
 500/  239/   35
1000/  472/   65
```

In particular, the second line means that 239 phase sets from 1000 generated phase sets verify the first condition (in the described example, at 16 A) and only 35 phase sets from them verify also the second condition (in the described example, at 12 A). In many cases, the optimal choice of parameter gives roughly equal percentage of selection by every condition and a significant difference in these values means rather unoptimal choice of parameters. For example, a line

```
10000/  8000/   20/   16
```

shows that the first and the third selection conditions are 'too soft' while the second is, on contrary, 'too strong'.

10. Recommended protocol for first phasing steps.

General remark. Naturally, all recommendations below are general and in any particular case may be not optimal.

Maximal number of generations. A small number of trial generation, few tens or hundred, is recommended to tune the parameters.

Maximal number of selected phase sets. When the parameters of the generation and the selection are established, an output file with a few tens of phase sets can be useful to check the situation by phase averaging and the synthesis calculation.

Generation mode. Generation mode for the first step is 1 (uniform distribution).

Synthesis flipping. Option 'N' is recommended for this parameter.

Resolution zone. When starting a research, it is useful to carry out a connectivity analysis in two resolution zones, the first one, D1, with about 10-15 reflections, and the second zone, D2, with approximately 30-40 reflections (see below on their use).

Grid. It is recommended to define a grid step equal approximately to ¼ of the resolution (choose appropriate nx, ny, nz). Moreover, the grid numbers should be presented as a product of prime numbers, as many as possible, with no multiplier superior to 17. For example, 48 = 2*2*2*2*3 is a good choice, 47 is not accepted at all, 64 = 2*2*2*2*2*2 is more preferable than 65 = 5*13 .

Condition of endless regions. It is recommended do not accept endless regions , use 'N'.

Cut-off level. It was found that the cut-off level which cuts the volume of approximately 25 cubic A per residue is appropriate for such connectivity analysis. This value can be roughly estimated as

25*number-of-residues-par-molecule*number-of-molecules-per-unit-cell)/unit-cell-volume

Usually, this value is close to 0.1.

Selection mode. Two resolution zones, D1 are D2, are recommended to start with. Selection mode 1 in each of them is a good choice. If the crystal does not have a non-crystallographic symmetry then the number of connected regions should be equal to the number of symmetry operations (for example, the multipliers are 4 and 0 when the crystal has 4 crystallographic symmetries and 1 molecule per asymmetric part of the unit cell). Otherwise, for the structure with a non-crystallographic symmetry (2 molecules per asymmetric part) these multipliers will be 4 and 4 when the volume ratio limits are something like 0.9 and 1.

Further analysis of the selected phase sets. Selected phase sets can be averaged to get a single set of phases in order to calculate a Fourier synthesis. This can be done using the program AVERAGE. This program needs to define the resolution zone for phase alignment and distance calculation. Our experience shows that the zone D1 can be a good choice.

The ensemble of selected phase sets can be analysed in a more detailed way where subgroups of similar phase sets can be found by programs of cluster analysis. This analysis is more complicated and the corresponding interactive software runs only at the PC computers; therefore the details of this techniques are not discussed here.

The program AVERAGE provides a user with the file of structure factors in the UF format described above where the experimental magnitudes are in the column 4, the average phases are in the column 7 and their figures of merit are in the column 6. If the reference phases are available, they are in the column 10.

Procedure iterations. At the next steps, it is better to generate new phases near the phase set obtained previously. This can be done, for example, in the generation mode 3 (von Mises generation) with the file of structure factors obtained after the first step where the column number is equal to 7 and the column number for the figures-of-merit is equal to 6. A manual assignment of figures of merit is a complementary option which is better to avoid at the beginning of the work. Selection can be done already with 3 resolution zones whit the new zone D3 containing about 70-80 reflections. Different possibilities can be tried:

1) selection in D1, D2, D3; synthesis alignment in D1;
2) selection in D1, D2, D3; synthesis alignment in D2;
3) selection in D1, D3; synthesis alignment in D1;
4) selection in D2, D3; synthesis alignment in D2.
5) a new run with the parameters used previously.

It is worthy to try all these 5 options, to calculate 5 maps and to analyse them by graphics (CAN, O). Any available structural information is useful at this stage to choose the best synthesis and therefore the phasing strategy.

For further steps, a new resolution zone, D4, can be introduced in order to increase the resolution, and so on. Each time, the number of added reflections in D_n should not be too large in comparison with the number of reflections in the previous zone D_{n-1}.

It happens that for 70-80 reflections it is practically impossible to generate phase sets corresponding to ‘clean’ synthesis with a very small number of connected regions. In this case it is more practical to relax the connectivity conditions and not to increase too much the number of generated phases. In particular, selection mode 2 can be used with such condition that needs a reasonable, not too large number of generations. Experience shows that specially at the beginning more efforts should be devoted rather to analyse the possible solution and not to waste the time for phase generation.

Sometimes, the figures of merit calculated by clustering programs can be ambiguous and one would wish to replace them by some alternative values. They can be calculated, for example, by visual analysis of maps at different resolution. Let’s suppose that the synthesis looks quite well at some low resolution, reasonably good at some middle resolution and is very doubtful at a higher resolution. Then corresponding figures of merit can be defined as 0.8 - 0.9, 0.4 - 0.5, 0.0 - 0.1 in these zones, respectively. Such updates from time to time allows to avoid a local trap in the phasing.

11. Program limitations

There are some current restraints on program parameters; corresponding source can be modified by a text editor and the program can be recompiled . These restraints are :
- maximal number of structure factors in the input file – 10 000;
- maximal number of symmetry operation for a given space group - 48;
- maximal number of connected regions - 10 000;
- the size of the buffer array - 1 000 000 ; for all used density grids the number (nx+2)*ny*nz should not exceed this value ; moreover, if the generation mode 4 is used then for the grid nxre, nyre, nzre, used to calculate the initial synthesis and the parameter kptmax 'maximal number of bins for the cumulative' the value (nxre+2)*nyre*nzre+3*kptmax should not also exceed the indicated limit ;
- 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 for correlation calculation through map alignment for monoclinic and triclinic space groups – 1000; 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 bins for the histogram calculation - 40;
- maximal number of conditions for the connectivity analysis - 20;
- maximal number of multipliers printed by the program - 9;
- maximal number of bins for the cumulative function - 201.

12. Definition of basic concepts

Generated phase set

For a set of NREF reflections, this is NREF real numbers, one per reflection, representing a phase for the corresponding structure factor expressed in radians or in degrees (see **5.9**). A 'phase variant' or 'a test solution' can be used as synonyms for the 'generated 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 use for the statistical analysis of the correlation distribution for the generated phase sets and not for the choice of new phase set. Eventually, reference phases can have nothing to the exact solution of the phase problem for a given crystal.

Connected region (or connected compound)

For a function ρ(x,y,z) calculated at some three-dimensional grid (nx , ny, nz) and for a given cut-off value ρcrit a set of points can be defined such that ρ(x,y,z)>= ρcrit . This set is defined as a connected one if from any its point any other its point can be reached, step by step, by translation on one grid point following the coordinate axes staying at any moment in the same set. In the program GENMEM which studies periodic functions of three variables, the neighbours of the point (x,y,z) whose coordinates are indices of the grid are

x+1,y,z
x-1,y,z
x,y+1,z
x,y-1,z
x,y,z+1
x,y,z-1

and corresponding points found by periodicity in the case of unit cell border.

Cut-off level

A definition of a cut-off level ρcrit is crucial for the concept of 'connectivity region' ; it selects some subset of the unit cell for which the connectivity analysis is applied. Let's suppose that the synthesis values vary from ρmin to ρmax. There is no points such that ρ(x,y,z)> ρmax and the volume of the region selected by this value is 0 and we call it as a 0 cut-off level. Similarly, cut-off level 0.1 defines such value ρcrit_01 that the volume of the region ρ(x,y,z)> ρcrit_01 (number of the grid point which belong to this region) composes 0.1 of the total volume of the unit cell, etc.

Connectivity properties

In the current analysis, there are the number and the size of connected regions (compounds) in the unit cell as they defined above. In the program GENMEM, these regions are defined for the Fourier syntheses calculated with the experimental moduli and the trial phase sets.

Von Mises distribution

Let's suppose that a set of phases θ_h and a set of corresponding figures of merit m_h are given (NREF values for each of them). It is natural to search for a better phase set near these given phases. The higher its figure of merit, the closer should be the phase to the known value; a uniform distribution on $[0, 2\pi]$ should be a limit case for the figure of merit equal to 0. To do so, the program GENMEM uses the von Mises distribution

$$P(\phi) \sim \exp[t_h \cos(\phi-\theta_h)],$$

with the parameter t_h taken from the condition

$$< \cos(\phi-\theta_h) > = m_h,$$

in other words, from

$$I_1(t_h) / I_0(t_h) = m_h,$$

where $I_0(t_h)$, $I_1(t_h)$ are modified Bessel functions of the order 0 and 1, respectively.

Cumulative function of the Fourier synthesis

For a given function (Fourier synthesis) calculated at a given three-dimensional grid, its minimal and maximal values ρmin и ρmax can be found as well as the frequency of every value between ρmin and ρmax. More precisely, the interval (ρmin , ρmax) can be divided in several bins (for example, ten) and the number of grid points can be computed such that the function value varies between

ρmin and ρmin+ (ρmax-ρmin)/10 .

The same can be done for another interval, between :

ρmin+(ρmax -ρmin)/10 and ρmin+2*(ρmax-ρmin)/10,

etc. In this example, 10 numbers are computed and their sum is equal to the total number of the grid points. These 10 values are called histogram of the Fourier synthesis (or of the electron density). They show how often the function has such or such values. For example, if ρmin = -1000 , ρmax = 1500, and the histogram is

10 30 50 230 320 220 110 20 10

this means that

the function values between -1000 and –750 can be found in 10 points,
the function values between -750 and –500 can be found in 30 points,

etc. Often, it is more convenient to work with the cumulative function of the histogram and not with the histogram itself. The cumulative function and histogram define unambiguously each other. For example, for the histogram given below its cumulative function is

10 40 90 320 640 860 970 990 1000,

which means that

the function values are less than –750 in 10 points,
the function values are less than –500 in 40 points,

etc.
Cumulative function can be expressed either directly by the number of points or by a relative frequency which shows the share of the points in every bin with respect to the total number of points:

0.010 0.040 0.090 0.320 0.640 0.860 0.970 0.990 1.000

It is known that the cumulative functions for the homologous proteins are similar and can be used as a source of extra information. The program GENMEM in the **generation mode 4** uses a cumulative function expressed through the frequency like in the latter example.

Transformation reconstructing the cumulative function (histogram)
Let's suppose that there are phase values ϕ_h (NREF numbers) for known structure factor modules. A Fourier synthesis $\rho^{calc}(\mathbf{r})$ can be calculated with these values, and for the synthesis its histogram or, which is the same, its cumulative function $K^{calc}(\rho)$ can be defined.

If a reference cumulative function $K^{exact}(\rho)$ is known which corresponds to the synthesis calculated with the exact phases, then a non linear function lambda(ρ) can be found [6] such that after the transformation

$\rho(\mathbf{r}) \rightarrow \rho^{m}(\mathbf{r}) = \lambda(\rho(\mathbf{r}))$

the modified function $\rho^{m}(\mathbf{r})$ will have a cumulative function exactly coinciding with the function $K^{exact}(\rho)$. This transformation $\rho \rightarrow \lambda(\rho)$ is called a 'transformation reconstructing the cumulative function'. In should be noted that for the same $K^{exact}(\rho)$ the function $\lambda(\rho)$ is different for different $\rho^{calc}(\mathbf{r})$. Therefore, the scale of modification depends on the quality of the synthesis. For example, if the cumulative function for the initial synthesis corresponds exactly to $K^{exact}(\rho)$ then the modification does not change the function: $\rho^{m}(\mathbf{r}) = \rho^{calc}(\mathbf{r})$.

For the modified function $\rho^{m}(\mathbf{r})$, the modules and phases of its Fourier coefficients (structure factors) can be calculated. These phases ϕ_h_mod and not the phases ϕ_h are used for synthesis calculation connectivity properties of which are analysed. It can be expected that by some properties these new phases ϕ_h_mod are better that the original phases ϕ_h generated randomly.

13. **References**

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

2. Lunin V.Y., Lunina N.L.. Urzhumtsev A.G. (1999) "Seminvariant density decomposition and connectivity analysis and their application to very low resolution macromolecular phasing", Acta Cryst. A**55**, 916-925.

3. Lunin V.Y., Lunina N.L. & Urzhumtsev A.G. (2000) “Connectivity properties of high-density regions and ab initio phasing at low resolution”. Acta Cryst. A**56**, 375-382.

4. 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.

5. Urzhumtsev A.G., Lunina N.L., Skovoroda T.P., Podjarny A.D. & Lunin V.Y. (2000) “Density constraints and low-resolution phasing”. Acta Cryst. D**56**, 1233-1244.

6. Lunin, V.Yu. & Vernoslova, E.A. (1991) “Frequencies-Restrained Structure Factor Refinement. II. Comparison of Methods”. Acta Cryst. **A47**, 238-243.