14.09.2001

# Описание программы GENMEM

(Это описание справедливо для версии от 21.08.2001 и позже)

**Содержание**.

**1. Назначение.**
Программа GENMEM предназначена для генерации большого числа наборов фаз при заданных модулях структурных факторов и отбора из сгенерированных наборов фаз таких, которые приводят к синтезам, обладающим нужными топологическими свойствами, а именно свойствами связности.

**2. Используемые понятия**.
*Генерируемый набор фаз*
*Эталонные фазы*
*Компонента связности.*
*Свойства связности*
*Распределение фон Мизеса*
*Кумулятивная функция синтеза Фурье.*
*Преобразование, восстанавливающее кумулятивную функцию (гистограмму)*
*Уровень срезки*
(Знание этих понятий необходимо для понимания дальнейшего текста. Краткое определение используемых понятий приведено в конце описания. Для более подробного знакомства с этими понятиями рекомендуется обратиться к используемой литературе.)

**3. Способы генерации фаз**.

*Общее замечание*. В последующих примерах управляющих данных строки, начинающиеся с восклицательного знака (!), означают строки комментариев.

Обязательной входной информацией для программы GENMEM является файл структурных факторов, где приведены экспериментальные модули. Сгенерировать набор фаз - это значит приписать каждому рефлексу некоторое допустимое для него значение фазы. Фазы могут генерироваться случайным образом как равномерно, так и вблизи некоторого пробного решения.
**Мода генерации 1**. Фазы генерируются равномерно, причем для нецентросимметричных рефлексов фаза может принимать любое значение от 0 до $2\pi$, а для центросимметричных - одно из двух допустимых значений.
**Мода генерации 2**. Фазы генерируются равномерно, но оставляются для работы только те из них, которые находятся достаточно близко от некоторого решения (а именно, корреляция которых с некоторым заданным решением не меньше указанного значения).
**Мода генерации 3**. Фазы генерируются в соответствии с распределением фон Мизеса вблизи некоторого указанного решения. Степень отклонения от этого решения определяется показателями достоверности (каждой из фаз), которые также должны быть указаны. Если информация о показателях достоверности не содержится в файле, то ее можно ввести явным образом, задав разные значения в разных зонах разрешения.
**Мода генерации 4**. Эта мода объединяет случайную генерацию фаз и их уточнение одним из методов “модификации электронной плотности”. Фазы генерируются в соответствии с модой 3. После этого строится синтез с экспериментальными модулями и сгенерированными фазами, и к нему применяется преобразование (восстанавливающее эталонную гистограмму электронной плотности). Для модифицированного синтеза вновь рассчитываются фазы, которые и используются в дальнейшей работе. Для применения преобразования, восстанавливающего гистограмму электронной плотности, задается файл с эталонной кумулятивной функцией синтеза Фурье желаемого разрешения. Это может быть кумулятивная функция, отвечающая известной кристаллической структуре, для которой значение константы Метьюза $F_{000}/V$ такое же, как и для исследуемой структуры.

**4. Способы отбора синтезов по вычисляемым характеристикам связности**.

Для каждого из сгенерированных наборов фаз строится синтез Фурье, и для области его самых высоких (или самых низких значений) вычисляются некоторые топологические характеристики. А именно, рассчитывается общее число компонент связности; определяется объем каждой из компонент, выраженный в количестве узлов сетки; компоненты одинакового объема (например, связанные операцией симметрии) объединяются в группы. Следует иметь в виду, что объединение в группы происходит формально, только на основании числового значения объема. (То есть теоретически возможно, что в одну группу будут объединены компоненты разной формы, но одинакового объема, но практически такая вероятность слишком мала).

Анализируемая область определяется:

- разрешением, при котором рассчитывается синтез Фурье;
- сеткой в элементарной ячейке для расчета синтеза Фурье;
- уровнями срезки.

Синтез рассчитывается для заданного разрешения на заданной сетке. Затем для данного синтеза выделяется область его максимальных (или минимальных) значений путем указания относительного объема этой области (см. **Определение используемых понятий**, *Уровень срезки*). Как правило, такие области образуют несколько связных областей. Область называется связной, если из каждой ее точки в любую другую точку можно пройти, делая шаг в соседний узел сетки в одном из трех направлений, при этом оставаясь внутри области. При этом можно подсчитать объем каждой такой области, выраженный в числе точек сетки. Обычно из-за наличия симметрии в кристалле можно выделить несколько одинаковых областей. И тогда синтез может быть описан, например, таким образом:

16 4*1250 4*330 8*2

Это означает, что синтез распадается на 16 областей, 4 из которых одинаковые и очень большие (по 1250 точек сетки); 4 сравнительно небольшие (330 точек сетки) и 8 совсем маленькие, на уровне “брызг” (2 точки сетки). Числа 4, 4, 8 называются мультипликаторами. Здесь первый и второй мультипликаторы равны 4, третий равен 8, а все последующие равны 0. Общее число компонент здесь равно 16.

Мы можем сформулировать требования отбора синтезов в рамках вычисляемых топологических характеристик. И тогда сгенерированные фазы, которые приводят к синтезу с желательными свойствами, будут использоваться в работе, а остальные фазы будут отброшены.

В программе GENMEM реализовано 6 способов отбора фаз в зависимости от топологических свойств соответствующих синтезов.

**Мода отбора 1**. Этой модой уместно пользоваться, когда у нас имеется некоторое грубое представление о том, как может выглядеть желательный синтез, и это представление может быть описано двумя первыми мультипликаторами.

*Возможность 1*. В независимой части ячейки содержится одна молекула, а пространственная группа имеет несколько элементов симметрии (далее для определенности в этом качестве будем рассматривать число 4). Нам хотелось бы, чтобы синтез был чистым и содержал ровно 4 компоненты связности. Тогда в качестве значений мультипликаторов следует задать числа 4 0.

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

*Возможность 2*. Все то же, но нас устраивают также и не очень чистые синтезы. Однако хотелось бы, чтобы “брызги” были не слишком большие. В этом случае нужно указать значение второго мультипликатора и границы для отношения объемов второй по величине области к первой.

```
! 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 zero for almost clear synthesis (0.0  0.2, for example)
 0.0       0.2
 1.0
```

*Возможность 3*. Если независимая часть элементарной ячейки содержит 2 молекулы, связанные некристаллографической симметрией, то мы можем ожидать, что объемы их изображений на сетке будут почти одинаковыми. Эту ситуацию в рамках используемых здесь терминов можно описать так:

```
! 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
```

Заметим, что требование отбирать “почти чистые” синтезы при наличии некристаллографической симметрии не может быть описано модой отбора 1. (См. Мода отбора 6 ниже).

**Мода отбора 2**. Этой модой уместно пользоваться, когда у нас есть лишь понимание, что общее число областей связности должно быть не слишком велико, на заранее неизвестно. В этом случае отбираются фазы, приводящие к синтезам, дающим **не более чем** указанное общее число компонент связности. При этом может быть проверен первый мультипликатор.

**Мода отбора 3**. Иногда дополнительная информация о структуре позволяет нам предполагать, что число компонент связности **в точности равно** некоторому числу. Фазы, приводящие к синтезам, дающим **указанное** число компонент связности, могут быть отобраны, используя моду 3.. При этом может быть проверен первый мультипликатор.

**Мода отбора 4**. Иногда мы не можем оценить общее число областей связности, но можем предположить, каким образом сгруппированы наиболее крупные области. Тогда задаются условия на значения первых N мультипликаторов. При этом само число N тоже задается в качестве первого параметра. Отбираются такие фазы, которые приводят к синтезам, имеющим первые N мультипликаторов с указанными значениями.

**Мода отбора 5**. Используется тогда, когда мы хотим расширить требования, заданные модой 4, а именно вместе с указанными отбирать и более чистые синтезы, у которых указанные мультипликаторы равны 0. Задаются условия на значения первых N мультипликаторов, которые должны быть равны либо указанным значениям, либо нулю.

**Мода отбора 6**. Если у нас, кроме оценки величин первых мультипликаторов, имеется информация и для оценки соотношений объемов наибольших областей, то следует

использовать моду 6. Здесь задаются N первых мультипликаторов, а также границы для N-1 отношений объемов (каждой следующей области к предыдущей). Все числа для оценки отношений должны лежать в пределах от 0 до 1.
*Общее замечание*. Понятно, что могут возникнуть потребности отбора, плохо формулируемые в рамках перечисленных возможностей. В таком случае рекомендуется отбратиться к разработчикам. Вам либо помогут сформулировать свои потребности оптимальным образом, либо расширят возможности программы.

**5. Входной файл управляющих данных**.

*Общие замечания*. Строки в файле управляющих данных, которые начинаются знаком "!", воспринимаются как комментарии. Это позволяет для начинающего пользователя взять некоторый ГОТОВЫЙ ПОЛНЫЙ файл управляющих данных и адаптировать его к своим нуждам, попросту закомментарив ненужные строки. Все комментарии при этом очень полезно оставить в файле, поскольку они существенно облегчают изменение параметров и уменьшают число ошибок при их изменении.
Числовые данные вводятся по свободному формату.
Буква "Y" или "N" в ответах типа "ДА/НЕТ" может быть как прописной, так и строчной.

*Пример управляющих данных*. Подробное описание каждого параметра содержится ниже.

```
! data for program GENMEM dated 27.05.2001 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. Имя шага задания**.

( ! step code (from 1 to 3 symbols) )

Это не более чем 3-символьная метка, которая будет сопровождать все выходные файлы программы. Во время работы шага s1 будут созданы выходные файлы

s1_gmem.out - отобранные наборы фаз;

s1_gmem.mes - протокол работы программы;

s1_gmem.con - копия экрана.

Рекомендуется часть символов имени сделать цифрами и увеличивать их в процессе работы.

**5.2. Количество записей - шаг экрана**.

( ! number of records - screen step)

Это некоторое целое число, регулирующее частоту **вывода на экран** информации о том, как идет процесс отбора. Если задать этот параметр равным 100, то в процессе работы на экране будут появляться записи типа:

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

иллюстрирующие, что из 100 сгенерированных наборов фаз 43 удовлетворяют первому условию отбора и 7 - обоим заданным условиям; из 200 сгенерированных наборов 89 удовлетворяют первому условию отбора и 12 - обоим условиям и т.д.

Важно понимать, что, если задать это число очень большим (больше, чем реально нужно генераций), то записи на консоли не будут появляться вообще, кроме самой последней записи, которая появляется в любом случае.

**5.3. Количество записей - шаг копии экрана**.

(! number of records - step for file <code of step> _gmem.con)

Это некоторое целое число, регулирующее частоту **вывода в файл, являющийся копией экрана,** информации о том, как идет процесс отбора. В частном случае эти 2 числа - шаг экрана и шаг копии экрана - могут совпадать, но это требование не является обязательным. Если задать этот параметр равным 500, то в файле будут содержаться записи типа:

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

иллюстрирующие, что из 500 сгенерированных наборов фаз 239 удовлетворяют первому критерию отбора и 35 - обоим заданным критериям; из 1000 сгенерированных наборов 471 удовлетворяют первому критерию отбора и 65 - обоим критериям и т.д.

Важно понимать, что, если задать это число очень большим (больше, чем реально нужно генераций), то записей в файле не будет вообще, кроме самой последней записи, которая появится там в любом случае.

**5.4. Входной файл структурных факторов**.
(! input file of structure factors (UF-format))
Далее во входном файле должна содержаться символьная информация, первые 72 символа которой интерпретируются как имя файла структурных факторов  Этот форматизованный файл организован следующим образом (так называемый формат UF):
- первая запись (72 символа) - первый заголовок файла;
- вторая запись (72 символа) - второй заголовок файла;
- третья запись (целое число, которое читается по формату I4) - LREC, размер каждой из последующих записей (количество чисел в одной записи; не должно превышать 100; как синоним понятия “количество чисел” мы будет также использовать выражение “количество колонок”);
- далее до конца файла следуют одинаковые записи, состоящие из LREC чисел (которые называются также колонками), отвечающих одному рефлексу; из этих чисел первые три - целые (это индексы H, K, L), а остальные – вещественные, содержащие разнообразную информацию о данном структурном факторе. Такие записи вычитываются по формату
    (3I4, 5g12.6)
если  LREC не больше 8, и по формату
   (3I4, 5g12.6/(6g12.6))
в противном случае.
Предполагается, что записи каждого файла в одинаковых позициях содержат информацию одинакового типа (например, модули структурных факторов в позиции 5, сответствующие sigma в позиции 6, разрешение рефлекса - в позиции 4 и т.п.)

**5.5. Номер пространственной группы**.
(! space group number )
Целое число, соответствующее номеру пространственной группы в Международных кристаллографических таблицах. Информация о такой группе должна содержаться в файле FAM.ITC, который поставляется вместе с программой. Этот файл может быть пополнен для пространственных групп, информация по которым еще не внесена авторами..

**5.6. Параметры элементарной ячейки**.
(! unit cell parameters (in A and degrees))
6 вещественных чисел - длины ребер ячейки и углы. Длины задаются в ангстремах, углы - в градусах.

**5.7. Номер колонки с экспериментальными модулями**.
(! column number for observed modules)
Это целое число, указывающее, где во входном файле находятся экспериментальные модули. Это должно быть положительное число, строго большее 3 (так как первые 3 числа в записи входного файла структурных факторов - это индексы H, K, L) и не превышающее LREC.

**5.8. Номер колонки с тест-флагами**.
(! column number for test-flag)
Это целое число, указывающее, где во входном файле находятся признаки того, рассматривается ли данный рефлекс как рабочий или контрольный. Для рабочих рефлексов этот признак установлен в 1, для контрольных - в 0. Если такая информация отсутствует, номер колонки должен быть задан 0. Если такая информация во входном файле есть, то это

должно быть положительное число, строго большее 3 (так как первые 3 числа в записи входного файла структурных факторов - это индексы H, K, L) и не превышающее LREC.
Сама программа GENMEM одинаково работает как с рабочими, так и с контрольными рефлексами. Поэтому в ней этот признак не используется. Однако предполагается, что программа GENMEM может быть использована для работы над конкретной структурой в комплексе с другими программами, где эта информация является важной. Так что информация о том, какие из рефлексов объявлены контрольными, переносится в выходной файл и становится доступной для следующих этапов работы над структурой.
Если во входном файле не содержится информации об этом признаке (номер колонки задан 0), то все рефлексы объявляются рабочими и для них этот признак устанавливается в 1.

**5.9. Номер колонки с эталонными фазами**.
(! column number for reference phases (0 if not ) and radian/degree information (R/D))
Это целое число, указывающее, где во входном файле находятся фазы, которые в рамках нашей задачи мы будем использовать как эталонные, и символьный признак - в радианах или градусах выражены эти фазы, а также фазы в остальных файлах, используемых программой. Если эталонные фазы отсутствуют, то номер колонки должен быть равнен 0. Если такие фазы есть, то это должно быть положительное число, строго большее 3 (так как первые 3 числа в записи входного файла структурных факторов - это индексы H, K, L) и не превышающее LREC. Результаты работы программы (отобранные варианты) не зависят от наличия или отсутствия во входном файле эталонных значений фаз.
Эталонные фазы используются для сбора статистики распределения сгенерированных и отобранных наборов фаз. Если имеются точные фазы (например, структура известна и мы используем программу для тестирования метода), то разумно именно их использовать в качестве эталонных. Если точных фаз нет, но имеется какое-либо решение, полученное независимым путем, то такие фазы также разумно использовать в качестве эталонных, чтобы анализировать, как получаемые с помощью программы GENMEM фазы распределены по отношению к нему. При отсутствии таких фаз (например, при работе с неизвестной структурой) для целей статистики используется первый сгенерированный набор фаз. Корреляция всех остальных сгенерированных наборов считается по отношению к нему. Заметим, что первый сгенерированный набор фаз не обязательно является первым отобранным; он попадает в выходной файл только в том случае, если удовлетворяет всем заданным условиям связности.
После целого числа по крайнем мере через один пробел (пробелов может быть и больше) может следовать символьный признак - в радианах или градусах выражены фазы. Буква может быть как прописной, так и строчной, то есть допустимыми являются буквы R, D, r, d. Предполагается, что все фазы во входных и выходных файлах выражены в одних и тех же единицах: в радианах, если указан признак R или r, и в градусах, если этот признак указан как D или d.
В любом случае, независимо от того, задана ли информация о единицах измерения углов, программа делает попытку проинтерпретировать единицу измерения фаз самостоятельно. В случае, если входные фазы вообще отсутствуют (как реперные, так и стартовые), самостоятельная интерпретация невозможна. И тогда либо будет использована входная информация о единицах измерения углов (если она есть), либо будет принято решение, что выходные углы будут выражены в градусах (если входная информация отсутствует).
Когда интерпретация возможна, она происходит так: если в программе есть какие-либо входные фазы (реперные или пробные для мод генерации 2, 3 или 4), то для реперных (а при их отсутствии - для пробных) фаз подсчитывается среднее значение абсолютных величин этих фаз. Если это среднее по абсолютной величине не превосходит 5, то фазы интерпретируются в радианах; если среднее оказывается не меньше 50, то фазы будут

интерпретируются в углах. В противном случае программа считает, что углы не могут быть проинтерпретированы.
Возможны 4 случая сочетания наличия входной информации и успешности интерпретации.
1. Входной информации нет, самостоятельная интерпретация безуспешна
   Программа заканчивает работу с аварийным сообщением.
2. Входной информации нет, самостоятельная интерпретация успешна.
   Используются единицы, полученные в результате самостоятельной интерпретации.
3. Входная информация есть, самостоятельная интерпретация безуспешна.
   Используется входная информация.
4. Входная информация есть, самостоятельная интерпретация успешна.
   И тогда возможны 2 варианта:
4а. Единицы измерения, полученные независимым образом, совпадают.
    Они и используются.
4б. Единицы измерения находятся в конфликте.
    Используются единицы, заданные во входной информации.
    Выдается предупреждающее сообщение.
Напоминаем, что все входные фазы должны быть выражены в одних и тех же единицах.
*Замечание*. Если в программе выполняется преобразование углов из градусов в радианы или обратно, то все значения 1.e+10, являющиеся признаком того, что фаза не определена, остаются без изменения.

**5.10. Диапазон разрешения для подсчета корреляции**.
(! resolution limits for correlation with reference phases )
3 вещественных числа (DMIN, DMAX, DSTEP), из которых обязательными являются первые 2. Эти 2 числа задают минимальное и максимальное разрешение, определяющие зону в обратном пространстве, в которой будет подсчитываться корреляция сгенерированных фаз с эталонным набором (если он есть) либо с первым сгенерированным набором (если эталонные фазы отсутствуют). Они могут быть заданы в любом порядке. При расчете корреляции сравниваемые наборы фаз предварительно выравниваются допустимыми для данной пространственной группы изменениями начала координат и/или энантиоморфа (информация о допустимых сдвигах и изменении энантиоморфа задается в файле FAM.ITC). Если для одной из осей возможен любой выбор начала координат (например, сдвиг вдоль оси Y для пространственной группы P21), то третье число в строке управляющих данных задает шаг, с которым будут проверяться сдвиги при выравнивании вдоль этой оси. Это число может отсутствовать в данных, и тогда программа назначает переменной DSTEP значение DMIN/4.

**5.11. Возможность переворачивать синтез при расчете корреляции**.
(! whether the calculated synthesis may be flipped (Y/N) )
Этот параметр определяет, будет ли при выравнивании фаз рассматриваться не только “прямой”, но и “перевернутый” синтез. Если число допустимых сдвигов равно, например, 4, указана возможность смены энантиоморфа, а возможность “переворачивать” синтез не указана, то в качестве корреляции пробного синтеза с точным будет выбрано максимальное из 8 чисел, каждое из которых соответствует корреляции при некотором допустимом сдвиге и любом выборе энантиоморфа. А если возможность “переворачивать” синтез указана, то будет выбрано максимальное из 16 чисел, образующих 8 пары; в каждой паре одно число является корреляцией “прямых” синтезов, а другое - корреляцией “перевернутого” пробного и “прямого” точного синтеза.

**5.12. Стартовое значение датчика случайных чисел**.
(! start value for the random numbers generator)

Это достаточно большое (7 - 8-значное) целое число. Задает последовательность псевдослучайных чисел для генерации фаз. Задав в точности то же стартовое значение датчика случайных чисел, мы получим в точности ту же последовательность фаз, что бывает важно в ряде ситуации.
Иногда может возникнуть необходимость продолжить генерацию с текущего псевдослучайного числа. Для этой цели в конце протокола печатается последнее значение датчика случайных чисел; оно может быть использовано в качестве стартового при следующем запуске.

**5.13. Максимальное число сгенерированных фазовых наборов**.
(! max number of generated phase sets )
Целое число, которое указывает, когда следует прекратить генерацию фаз, если она не закончится при достижении другого ограничения (а именно при достижении максимального числа отобранных фазовых наборов).
Это число рекомендуется устанавливать небольшим (несколько сотен или тысяч), пока идет подбор параметров.

**5.14. Максимальное число отобранных фазовых наборов**.
(! max number of selected phase sets)
Целое число, которое указывает, при достижении какого количества отобранных фазовых вариантов работа будет прекращена (если она не будет прекращена раньше при достижении другого ограничения - максимального числа генераций).
Следует иметь в виду, что при указании ненулевого номера колонки для реперных фаз они будут перенесены в качестве первого набора фаз в выходной файл. В таком случае, если мы укажем число 100 в качестве максимального числа отобранных фазовых наборов, то в выходном файле реально будет содержаться 101 фазовый набор.

**5.15. Мода генерации фаз**.
```
! 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 figes of merit FOM)
! 3 - generation by von Mise 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 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
```

Понятие “Способы генерации фаз” вводится в разделе 3. Как следует из него, этот параметр должен быть равен целому числу, принимающему в данной версии программы значение от 1 до 4. При этом

*Мода генерации 1* (равномерная генерация) не требует дополнительных данных.

*Мода генерации 2* (генерация вблизи некоторое стартового решения) требует дополнительной информации, задающей это стартовое решение и процесс сравнения синтезов:

1) номеров колонок для фаз пробного решения;
2) номер колонки для показателей достоверности пробного решения и количество зон по разрешению для принудительного задания показателей достоверности;
3) если это количество зон не равно нулю, то для каждой зоны – границы разрешения и значение показателя достоверности для всех рефлексов этой зоны;
4) диапазон разрешения, при котором рассчитываются пробный и проверяемый синтезы для последующего расчета их корреляций;
5) возможность переворачивать синтез при подсчете корреляции;
6) минимально возможное значение корреляции для отбора фаз. Все сгенерированные фазы, корреляция которых с пробным решением ниже указанного значения, будут отброшены;
7) максимально допустимое количество попыток для выбрасывания одного варианта.

Параметры 1 - 2 задают некоторое пробное решение. Вместе с номером колонки для показателей достоверности вводится информация о принудительном задании показателей достоверности: количество зон и  для каждой зоны - ее границы и единое значение показателя достоверности для всех рефлексов этой зоны. При этом показателям достоверности СНАЧАЛА присваивается значение в соответствии с номером колонки (при ненулевом номере колонки они берутся из нее, а при нулевом все показатели  достоверности полагаются равными ЕДИНИЦЕ), а ПОТОМ они корректируются в соответствии с информацией для принудительного задания, если количество таких зон не равно нулю.

Параметры 3 - 5 определяют способ сравнения сгенерированных фаз с пробным решением.

Необходимость задания  параметра 7 вызвана следующим соображением. Мы можем задать параметры столь неудачно, что случайные фазы вообще не будут располагаться вблизи указанного пробного решения. В этом случае программа прекратит работу, исчерпав указанный ресурс (например, сгенерировав 1000 наборов и не найдя из них не одного, корреляция которого с указанным стартовым  решением была бы не менее 0.7).

*Замечание*. Важно понимать, что эти показатели достоверности используются только при расчете корреляции со взвешенным синтезом. В выходной файл эти значения не попадают.

*Мода генерации 3* (распределение фон Мизеса вблизи некоторое пробного решения) требует дополнительной информации, задающей это пробное решение. Это
1) номер колонки для фаз пробного решения;
2) номер колонки для показателей достоверности пробного решения и количество зон по разрешению для принудительного задания показателей достоверности;
Параметры 1 - 2 задают некоторое пробное решение. Вместе с номером колонки для показателей достоверности вводится информация о принудительном задании показателей достоверности: количество зон и для каждой зоны - ее границы и единое значение показателя достоверности для всех рефлексов этой зоны. При этом показателям достоверности СНАЧАЛА присваивается значение в соответствии с номером колонки (при ненулевом номере колонки они берутся из нее, а при нулевом все показатели достоверности полагаются равными НУЛЮ), а ПОТОМ они корректируются в соответствии с информацией для принудительного задания, если количество таких зон не равно нулю.
*Замечание*. Важно понимать, что эти показатели достоверности используются только при генерации фаз с использованием распределения фон Мизеса. В выходной файл эти значения не попадают.

*Мода генерации 4* (распределение фон Мизеса вблизи некоторое пробного решения и модификация этих фаз) требует дополнительной информации:
1) номеров колонок для фаз пробного решения;
2) номер колонки для показателей достоверности пробного решения и количество зон по разрешению для принудительного задания показателей достоверности;
3) диапазон разрешения D_start_min, D_start_max для расчета стартового синтеза;
4) три целых числа - сетка для расчета стартового синтеза;
5) эталонная кумулятивная функция электронной плотности (считывается из файла), использующаяся для модификации. Это может быть, например, кумулятивная функция родственного белка с известной структурой. Рекомендуется использовать кумулятивную функцию, рассчитанную при более высоком разрешении, чем D_start_min.

Параметры 1 - 2 задают некоторое пробное решение. Вместе с номером колонки для показателей достоверности вводится информация о принудительном задании показателей достоверности: количество зон и для каждой зоны - ее границы и единое значение показателя достоверности для всех рефлексов этой зоны. При этом показателям достоверности СНАЧАЛА присваивается значение в соответствии с номером колонки (при ненулевом номере колонки они берутся из нее, а при нулевом все показатели достоверности полагаются равными НУЛЮ), а ПОТОМ они корректируются в соответствии с информацией для принудительного задания, если количество таких зон не равно нулю.
Формат файла с кумулятивной функцией предполагается следующим: в первой строке должно содержаться целое число NPOINT, задающее количество бинов, для которых определена кумулятивная функция (в текущей версии это число не должно превышать 201; размер бинов предполагается одинаковым), а также 4 вещественных числа, первые два из которых задают границы интервала, для которого была рассчитана кумулятивная функция, (третье и четвертое - среднее и среднеквадратичное отклонение, в программе не используются). Начиная со второй строки в файле должны содержаться NPOINT вещественных чисел - значений кумулятивной функции (т.е. числа, расположенные в возрастающем порядке от 0 до 1). Читаются они по формату ((5G15.6)).

**5.16. Число зон по разрешению для анализа связности**.
Это должно быть целое число N, не превышающее 20. Далее должно следовать N групп строк. Каждая группа содержит информацию для одной зоны разрешения.

**5.17. Информация для одной зоны разрешения**.

1) Диапазон разрешения.

Это 2 вещественных числа, указывающих, с какими именно рефлексами будет рассчитан синтез, топологические свойства которого будут анализироваться.

2) Сетка.

Это 3 целых числа, задающие трехмерную сетку, на которой будет вычисляться синтез Фурье.

3) Уровни срезки.

Это 2 вещественных числа, лежащих в пределах между 0. и 1. Они определяют, какая часть объема синтеза будет использоваться для анализа связности (см. 4. Способы отбора синтезов по вычисляемым характеристикам связности.).

4) Признак бесконечности области.

Символьное значение “Y” или “N”, определяющее, считать ли допустимыми “бесконечные” области самой высокой плотности.

5) Мода отбора.

Понятие о возможных способах отбора дается в разделе “Способы отбора”. В текущем разделе указывается тип и возможные значения параметров для каждой моды отбора.

*Мода отбора 1*.

а) 2 целых числа, соответствующие 2 первым мультипликаторам в желаемом синтезе;

б) если второе число не равно 0 и не равно -1, то далее должны следовать два вещественных числа между 0 и 1. Они указывают границы, в которых должно находится отношение объема второй по величине области связности к самой большой. Например, требованию

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

будет удовлетворять синтез, имеющий по 4 связные области размера 1230 и 123:

4*1230 4*123

так как в этом случае оба мультипликатора равны 4 и отношение второй по размеру области к первой строго равно 0.1,

и не будет удовлетворять синтез, имеющий по 4 связные области размера 1230 и 124:

4*1230 4*124

так как в этом случае отношение объема второй по размеру области к первой превышает 0.1.

Допустимыми синтезами считаются такие, у которых третий мультипликатор равен 0.

*Мода отбора 2*. В этом случае задаются ограничение на общее число компонент и возможное значение первого мультипликатора. Отбираются фазы, приводящие к синтезам, дающим **не более чем** указанное число компонент связности. Если значение второго параметра положительно, то отбираются синтезы с первым мультипликатором, равным указанному значению. Если значение второго параметра равно -1, то такая проверка не производится.

```
! selection mode
2
! limit for the number of connected regions and the 1st multiplier
! (-1 for the 1st multiplier if any value is allowed)
20    4
```

*Мода отбора 3*. Очень похоже на моду 2. Единственное отличие - требование **“НЕ БОЛЬШЕ”** для общего числа областей заменяется требованием **“СТРОГО РАВНО”**. То есть отбираются фазы, приводящие к синтезам, дающим **указанное** число компонент

связности. При этом может быть проверен первый мультипликатор. Отбираются синтезы с первым мультипликатором, равным указанному значению, если только это значение не равно -1. В случае, если задано значение -1, первый мультипликатор не проверяется.

```
! selection mode
3
! value for the number of connected regions and the 1st multiplier
! (-1 for the 1st multiplier if any value is allowed)
20   -1
```

*Мода отбора 4*. В этом случае первое целое число определяет количество проверяемых мультипликаторов N (в версии программы от 27.05.2001 это значение не должно превышать 6). Далее следуют N целых чисел. Отбираются такие фазы, которые приводят к синтезам, имеющим первые N мультипликаторов с указанными значениями.

Так, при входных данных

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

будут отбираться фазы, приводящие к синтезам, у которых 16 и больше компонент, причем первые мультипликаторы равны 4, 4, 8.

*Мода отбора 5*. Является расширением моды 4. В этом случае первое целое число определяет количество проверяемых мультипликаторов N (в версии программы от 10.12.2000 это значение не должно превышать 6). Далее следуют N целых чисел. Отбираются такие фазы, которые приводят к синтезам, имеющим первые N мультипликаторов либо с указанными значениями, либо нулевые.

Так, при входных данных

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

будут отбираться фазы, приводящие к синтезам, у которых число компонент

- либо равно 4;
- либо равно 8, а первые мультипликаторы 4 и 4:
- либо больше или равно 16, а первые мультипликаторы 4, 4, 8.

В случае, если число компонент больше 16 является нежелательным, уместно такой запрос переформулировать как

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

*Мода отбора 6*. Совмещает проверку как значений мультипликаторов, так и отношений объемов различных областей. Первое целое число определяет количество проверяемых мультипликаторов N (в версии программы от 27.05.2001 это значение не должно превышать

6). Далее следуют N целых чисел. Отбираются такие фазы, которые приводят к синтезам, имеющим первые N мультипликаторов с указанными значениями. Далее следуют N-1 пара вещественных чисел, каждое из которых находится в пределах от 0. до 1. Происходит N-1 проверка для отношений объемов компонент связности. Первая проверка определяет, лежит ли отношение объема второй по величине области к первой в пределах, указанных первой парой вещественных чисел. Последняя проверка определяет, лежит ли отношение объема N-ой области к N-1-ой области в пределах, указанных N-1-ой парой. Так, если кристалл имеет 4 кристаллографические симметрии и вдобавок некристаллографическую симметрию, и нам хотелось бы отобрать такие синтезы, в которых имеются 2 группы по 4 области примерно одинакового объема, и если при этом нас устраивают не вполне "чистые" синтезы, а содержащие некоторые "брызги", причем количество таких маленьких областей для нас не существенно, то уместно задать данные следующим образом:

```
! 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. Входной файл FAM.ITC.**

Входной файл FAM.ITC содержит информацию о некоторых пространственных группах в виде, доступном для программы. Ниже приведена информация для группы P212121. Если файл не содержит информацию о группе, необходимой для вашей работы, вы можете сами пополнить его по приведенному здесь образцу либо обратиться к разработчикам.

```
NEWGROUP    P212121
```
заголовок, отмечающий начало нового блока информации

```
19  (the group number)
```
номер пространственной группы (по этому номеру этот блок находится программами, использующими этот файл)

```
4   (number of symmetries)
```
количество элементов симметрии в пространственной группе

```
 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
```
коэффициенты преобразований симметрии в следующем порядке $r_{11}$, $r_{21}$, $r_{31}$, $r_{12}$, $r_{22}$, $r_{32}$, $r_{31}$, $r_{32}$, $r_{33}$, $t_1$, $t_2$, $t_3$

```
3   (number of centrosymmetric zones)
```
число зон, содержащих центр симметрии

```
0 0 1 .5  0  0
0 1 0  0  0 .5
1 0 0  0 .5  0
```
каждая из центросимметричных зон должна быть определена шестью параметрами: m1, m2, m3, a1, a2, a3:

рефлекс hkl принадлежит этой зоне тогда и только тогда, когда m1*h + m2*k + m3*l = 0;

при этом допустимые значения фазы для этого рефлекса:
$\alpha = (a1*h + a2*k + a3*l)*\pi$ либо $\alpha + \pi$ .

0 (number of axes with an arbitrary origin shift along it)

определяет число осей (0, 1 или 3), таких, что любой сдвиг вдоль них приводит к эквивалентному выбору начала координат; такие оси не существуют в орторомбических группах, но существуют, например, в моноклинных (ось вращения)

8 (number of the possible discrete origin positions)

количество вариантов дискретного выбора эквивалентных начал координат; если допустимы произвольные сдвиги начала вдоль какой-то из осей, то эти сдвиги добавляются к каждому из дискретных вариантов выбора начала;

0 0 0
.5 0 0
0 .5 0
.5 .5 0
0 0 .5
.5 0 .5
0 .5 .5
.5 .5 .5

соответствующие начала координат

1 (possibility of enantiomorph switch during the phase search; 1 - if possible)

указывает, совпадает ли энантиомер данной группы с ней самой

**7. Выходной файл - протокол работы**.

Протокол работы имеет имя <имя_шага>_gmem.mes.

В протокол заносится, прежде всего, вся информация из входного файла управляющих данных.

Затем сообщается, сколько рефлексов из входного файла структурных факторов участвует в работе.

Если не была задана информация о том, в радианах или градусах приведены реперные фазы, программа сообщает о своих результатах интерпретации фаз (в приведенном ниже протоколе они проинтерпретированы в радианах; это говорит о том, что и выходные фазы будут выражены в радианах).

Затем печатается таблица, иллюстрирующая, как устроены (с точки зрения вычисляемых характеристик связности) невзвешенные синтезы, рассчитанные с точными фазами (если они есть) , а также со всеми отобранными наборами фаз. Эта таблица имеет заголовок:

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

В ней содержится информация следующего вида:

- разрешение, на котором устраивается последняя проверка по связности;
- корреляция сгенерированных фаз с эталонными (вычисленная, вообще говоря, при другом разрешении, указанном в начале файла управляющих данных); напомним, что в качестве эталонных фаз могут быть использованы как входные фазы (если номер соответствующей колонки был отличен от 0), так и просто первый сгенерированный набор фаз (в противном случае);
- уровень срезки;
- общее число областей;
- мультипликаторы и объемы шести наиболее крупных компонент связности.

В конце этой таблицы сообщается, сколько вариантов отобрано после скольких генераций. Если эталонные фазы содержались во входной файле, то именно они находятся в выходном файле в качестве первого набора фаз, даже если они и не удовлетворяют критериям отбора.

Затем приводится 4 гистограммы распределения вариантов по их корреляциям с эталонными фазами. Первые 2 гистограммы показывают распределение количества вариантов для всех сгенерированных (1) и для отобранных (2) фазовых наборов, последние 2 - распределение соответствующих частот. Для каждой гистограммы указаны некоторые статистические данные - минимальное, максимальное, среднее значение корреляции и среднеквадратичное отклонение от среднего.
Протокол заканчивается строчкой с последним значением датчика случайных чисел (на тот случай, если понадобится продолжить генерацию с такими же условиями).

Ниже приведен пример протокола. После него имеются некоторые комментарии, позволяющие понять, на какие именно величины следует обратить внимание в первую очередь.

```
------------------------------------------------------------------------------
                      ***  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
```

*Комментарии*

В процессе генерации нам удавалось получить фазы, которые на разрешении 16A дают корреляцию 84,20% с точными. Однако эти фазы не были отобраны, поскольку не прошли одну из двух проверок на связность. Наилучшие отобранные фазы дают корреляцию 80.29% с точными. Тем не менее, следует признать, что отбор был успешным, так как он позволил поднять среднюю корреляцию с 47.82% до 59.29%.

По гистограммам для относительных частот видно, что имеется “обогащение набора”: некоторое уменьшение самого правого значения (500 вместо 526), но заметное увеличение предпоследнего (1000 вместо 526).

**8. Выходной файл отобранных наборов фаз**.
Выходной файл отобранных наборов фаз имеет имя <имя_шага>_gmem.out. Структура его следующая:
первая запись (формат A72) - символьная информация, она формируется следующим образом: первые 22 символа – это текст ”Output data of GENMEM”, далее следуют параметры элементарной ячейки и номер пространственной группы;
вторая запись (формат A72) - символьная информация; это одна из двух строк:
This is file in .OUT-format. Phases are in degrees
либо
This is file in .OUT-format. Phases are in radians
третья запись (формат I4) - целое число NREF - количество рефлексов, находившихся во входном файле структурных факторов; все они будут включены и в выводной файл;
далее следуют NREF записей (формат 3I4, 2G12.4), одна на рефлекс, в каждой из которых содержатся индексы рефлексы H, K, L, экспериментальное значение модуля и тест-флаг, принимающий значение 1 для рабочих рефлексов и 0 для свободных рефлексов. Если был указан ненулевой номер колонки для тест-флага, последняя информация информация переписывается из входного файла структурных факторов, в противном случае всем рефлексам устанавливается значение флага, равное 1, в знак того, что в дальнейшем они будут расцениваться как рабочие. Программа GENMEM одинаково работает как с рабочими, так и с контрольными рефлексами;
далее следуют записи формата ((6G12.5)) или ((6G12.6)), по одной на каждый отобранный набор фаз. В одной такой записи содержится в начале одно вещественное число - значение корреляции набора с эталонным набором, затем NREF вещественных чисел - значения фаз для каждого рефлекса, затем NREF вещественных чисел - веса, приписываемые фазам каждого рефлекса. Программа GENMEM с весами не работает, поэтому на этих местах всегда стоят 1. Если входные фазы были заданы в градусах или их не было, то выходные фазы выражаются в градусах, и тогда используется формат ((6G12.5)). Если входные фазы были заданы в радианах, то выходные фазы выражаются в радианах, и тогда используется формат ((6G12.6)).
Данный формат согласуется с форматами некоторых других фазовых программ (FAMREF, RING) и позволяет обрабатывать полученные программой GENMEM результаты с помощью имеющихся программ усреднения и кластерного анализа.

**9. Выходной файл - копия экрана**.
Выходной файл копии экрана имеет имя <имя_шага>_gmem.con.
Файл “копия экрана” используется для того, чтобы наблюдать за ходом отбора в ситуации, когда постоянный контроль экрана затруднен (например, ночной счет или счет, продолжающийся несколько дней). Двумя первыми параметрами во входном файле управляющих параметров можно регулировать частоту вывода таких строк как на экран, так и в файл <имя_шага>_gmem.con.
В процессе выполнения задания на экран выводится некоторая информация. Она позволяет определить, как идет процесс отбора. В одну строку через разделители выводятся несколько чисел - сколько всего вариантов сгенерировано, сколько из них удовлетворяют первой проверке на связность, сколько из отобранных - второй и так далее. Например, при двух проверках эти строки могут выглядеть так:

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

Например, вторая строка означает, что из 1000 сгенерированных фазовых наборов первой проверке (в нашем случае на 16 А) удовлетворяют 239 вариантов, а второй проверке (в

нашем случае на 12 А) только 35 вариантов из ранее отобранных 239. В большинстве случаев, если параметры отбора заданы удачно, то процент отбираемых вариантов примерно одинаков для каждой проверки. И, наоборот, резкое отклонение от средней доли отбираемых вариантов может свидетельствовать о неудачно заданных параметрах. Например, строка вида

10000/ 8000/ 20/ 16

свидетельствует о том, что из трех проверок две - первая и третья - скорее .всего, сформулированы слишком мягко, а вторая задает слишком жесткие условия отбора.
Эта информация выводится на экран во время выполнения задания.

**10. Рекомендуемая стратегия начальных этапов работы**.
*Общее замечание*. Естественно, каждый белок обладает своей спецификой. Поэтому все сказанное ниже носит лишь рекомендательный характер. В каждом конкретном случае могут найтись другие, более эффективные решения проблемы

*Максимальное число генераций.* Рекомендуется подбирать параметры на небольшом числе генераций, например, несколько десятков или сотен.
*Максимальное число отобранных записей.* Когда параметры генерации и отбора подобраны, рекомендуется получить выходной файл с несколькими десятками наборов фаз для усреднения и получения предварительного решения.
*Мода генерации*. Генерировать фазы для первого шага следует в моде 1, то есть равномерно.
*Возможность переворачивать синтез*. Этот параметр рекомендуется устанавливать в "N".
*Диапазон разрешения*. Когда мы начинаем работу с новым объектом, полезно выделить 2 области разрешения - область D1, которая включает приблизительно 10-15 рефлексов, и D2, включающую приблизительно 30-40 рефлексов (см. ниже об их использовании).
*Сетка*. Рекомендуется задавать числа nx, ny, nz таким образом, чтобы шаги по осям ячейки составляли примерно 1/4 максимального разрешения, используемого при анализе связности. Кроме того, эти числа должны раскладываться на возможно большее число простых множителей, как можно меньших по величине. При этом максимальный из простых сомножителей не должен превышать 17. Так, число 48 = 2*2*2*2*3 годится, а 47 - нет; 64 = 2*2*2*2*2*2 предпочтительнее, чем 65 =5*13.
*Признак бесконечности области*. Рекомендуется на низком разрешении отказаться от бесконечных областей, то есть установить соответствующий символьный признак в "N".
*Уровень срезки*. Уровень срезки рекомендуется использовать такой, который дает 25 кубических ангстрем на один аминокислотный остаток. Т.е. рекомендуется использовать уровень_срезки = ( 25* число остатков в молекуле * число молекул в ячейке) / объем_ячейки
Часто это значение, близкое к 0.1.
*Мода отбора*. На первом шаге для отбора рекомендуется задавать 2 зоны разрешения - D1 и D2, причем в каждой зоне задавать моду отбора 1. Если структура не обладает некристаллографической симметрией, то количество выделяемых линией уровня "чистых" областей должно быть равно количеству элементов симметрии в пространственной группе (например, "4 0" для одной молекулы в независимой части ячейки и 4 элементов симметрии). Если же, например, имеется некристаллографическая ось второго порядка, то при 4 элементах симметрии и 2 молекулах в независимой части ячейки наиболее естественные условия на связность будут выглядеть как "4 4" при ограничениях на отношение объемов "0.9 1.0"

*Дальнейшая работа с отобранными фазами*. Отобранное множество фазовых наборов можно усреднить, чтобы получить набор фаз для расчета синтеза. Это делает программа

усреднения AVERAGE. В качестве входных данных для этой программы нужно задать зону разрешения для выравнивания фаз и подсчета расстояний. Наш опыт показывает, что в качестве такой зоны наиболее эффективно использовать D1.
Отобранное множество может быть проанализировано и более сложным образом на взаимное расположение в многомерном пространстве наборов фаз. Все множество может быть разбито на подмножества фазовых наборов, которые находятся недалеко друг от друга. Это делают программы кластерного анализа. Однако эти программы содержат графическую часть, которая работает только на PC. Поэтому пользователи UNIX'а могут считать такой анализ дополнительной возможностью, а работать со всем отобранным множеством фазовых наборов.
Программа усреднения AVERAGE дает на выходе файл структурных факторов в формате UF, описанном выше, где экспериментальные модули находятся в 4-ой колонке, усредненные фазы - в 7-ой и показатели достоверности - в 6-ой. Эталонные фазы, если они были указаны, переносятся в 10-ую колонку.

*Итерационный характер процедуры*. Для второго шага GENMEM генерировать фазы следует вблизи полученного на предыдущем шаге решения. Например, это можно делать с помощью моды генерации 3 (распределение фон Мизеса), указав "7" в качестве номера колонки для фаз и "6" в качестве номера колонки с показателями достоверности. Принудительным заданием показателей достоверности на первых этапах работы лучше не пользоваться – безусловно, это некоторая дополнительная возможность. Так что количество зон для принудительного задания показателей достоверности лучше задавать нулем. Что касается отбора, то полезно выбрать третий диапазон по разрешению D3, включающий 70-80 рефлексов. И тогда возможны варианты:
1) отбор в зонах D1, D2, D3; при усреднении выравнивание по зоне D1;
2) отбор в зонах D1, D2, D3; при усреднении выравнивание по зоне D2;
3) отбор в зонах D1, D3; при усреднении выравнивание по зоне D1;
4) отбор в зонах D2, D3; при усреднении выравнивание по зоне D2.
И, конечно же, всегда существует возможность повторить шаг GENMEM с предыдущими параметрами, что следует рассматривать как пятый равноправный вариант действий.
Полезнее всего реально провести все эти 5 шагов, построить 5 синтезов и рассмотреть 5 кандидатов на решение с помощью графических программ (CAN, O). На этом этапе полезно использовать любую дополнительную информацию о структуре для более адекватного выбора лучшего решения на текущем шаге.
Для дальнейшего продвижения по разрешению нужно будет ввести в рассмотрение зону D4, и т.д. Количество новых рефлексов в очередной зоне разрешения $D_n$ не должно составлять слишком большую часть от рефлексов в предыдущей зоне $D_{n-1}$.
Часто бывает так, что на разрешении, включающем 70-80 рефлексов, уже почти не генерируются фазы, приводящие к "чистым" синтезам с очень маленьким числом компонент связности. В этом случае представляется более разумным снижать требования к связности синтеза, чем значительно увеличивать число генераций. Например, можно использовать моду отбора 2, задав такое ограничение на общее число областей, которое потребует разумного (не слишком большого) числа генераций. Представляется логичным, чтобы на первым шагах больше времени уходило не на генерацию, а на анализ получаемых решений.
Иногда возникает необходимость отказаться от показателей достоверности, вычисляемых программами кластерного анализа, и ввести новые показатели достоверности. Оценить их можно из правдоподобности синтезов электронной плотности на разных разрешениях. Так, если построенный синтез дает осмысленную картину на низком разрешении и имеет некоторые правдоподобные черты на среднем, то в таком случае логично указать 3 зоны для "ручного" задания показателей достоверности и для самой нижней ввести значение порядка 0.8 - 0.9, для среднего - порядка 0.4 - 0.5, а для более высокого - ниже 0.1. Такая

модификация позволяет временами избежать застревания в локальных минимумах процедуры.

**11. Имеющиеся ограничения**

В программе приняты некоторые ограничения, изменить которые можно, перетранслировав программу. Это

- максимально возможное число рефлексов во входном файле - 5000;
- максимально возможное число симметрий в пространственной группе - 48;
- максимально возможное число областей связности - 10 000;
- размер буфера - 1 000 000. Для всех заданных решеток число (nx+2)*ny*nz не должно превышать этот размер. Кроме того, если используется мода генерации 4, то для решетки nxre, nyre, nzre, используемой в этой моде для расчета стартового синтеза, и параметра "максимальное число бинов кумулятивной функции kptmax", должно быть выполнено соотношение: (nxre+2)*nyre*nzre+3*kptmax не должно превышать указанный размер буфера lbuf;
- максимально возможное число центросимметричных зон в файле FAM.ITC - 20;
- максимально возможное число сдвигов начала координат в файле FAM.ITC - 8;
- максимально возможный размер массива, используемый при подсчете корреляции с использованием "выравнивания" карт для моноклинных или триклинных групп - 1000. Размер этого массива вычисляется как acell/dstep для случая моноклинной группы при выделенной оси **a** (или, соответственно, bcell/dstep для случая моноклинной группы при выделенной оси **b**) и acell*bcell*ccell/dstep$^3$ для случая триклинной группы. Если ваши данные таковы, что это ограничение оказалось нарушенным, рекомендуется увеличить задаваемую величину шага dstep;
- максимальное число бинов для расчета гистограмм - 40;
- максимальное число проверяемых условий связности при отборе - 20;
- максимальное число мультипликаторов, распечатываемых в протоколе - 9;
- максимальное число бинов для кумулятивной функции - 201.

**12. Определение используемых понятий**.

*Генерируемый набор фаз*.

Для входного файла структурных факторов, содержащего NREF рефлексов, это NREF чисел, каждое из которых представляет собой фазу очередного рефлекса (в радианах или градусах, в зависимости от используемого признака, описанного в **5.9**). В качестве синонима к словосочетанию "генерируемый набор фаз" может быть использованы слова "фазовый вариант" или "пробное решение".

*Эталонные фазы*.

Некоторый набор фаз, по отношению к которым рассчитывается корреляция фаз. Это могут быть как фазы, рассчитанные по известной модели, так и полученные некоторым независимым способом. Эталонные фазы используются не для определения фаз текущего шага, а лишь для статистики распределения корреляции и аналогичной информации. Фактически они могут быть очень далеки от правильного решения.

*Компонента связности*.

Когда в трехмерной ячейке (x,y,z) на некоторой сетке nx, ny, nz задана функция $\rho(x,y,z)$ и имеется некоторое значение $\rho crit$, то можно выделить такую область, где $\rho(x,y,z) >= \rho crit$. Точки этой области могут быть по-разному расположены друг по отношению к другу. Если из любой такой точки можно добраться до любой другой, двигаясь на один шаг по сетке в направлении координатных осей, оставаясь при этом внутри области, то говорят, что все эти

точки принадлежат одной компоненте связности. В программе GENMEM, изучающей периодические функции 3 переменных, для точки с координатами решетки (x,y,z) соседними являются точки с координатами

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

и соответствующие точки, связанные условием периодичности в случае точек на границе ячейки.

*Уровень срезки*.
Свойства связности мы изучаем не для всего синтеза Фурье, а для определенной части точек, где значения этой функции максимальны. Эта часть и задается в уровне срезки. Более точно: пусть в элементарной ячейке значения синтеза Фурье меняются от $\rho min$ до $\rho max$. Значение $\rho max$ обладает тем свойством, что для него не существует точек ячейки, для которых $\rho(x,y,z) > \rho max$. Говорят, что ему соответствует уровень срезки 0. Аналогично, уровню срезки 0.1 соответствует число $\rho crit\_0.1$ такое, что количество точек решетки, для которых $\rho(x,y,z) > \rho crit\_0.1$, составляет 0.1 объема всей ячейки, выраженного в узлах сетки.

*Свойства связности*.
Речь идет о количестве и размерах компактных областей синтеза Фурье при заданном уровне срезки. В программе GENMEM эти области определяются по описанной выше схеме для синтезов Фурье, рассчитанных с экспериментальными модулями и случайно сгенерированнымим фазами.

*Распределение фон Мизеса*.
Пусть у нас есть некоторое предварительное решение, то есть имеется набор фаз $\theta_h$ и набор показателей достоверности $m_h$ (в каждом наборе по NREF чисел). На очередном шаге процедуры было бы естественно генерировать фазы вблизи этого пробного решения с использованием показателей достоверности, то есть генерировать их равномерно на отрезке $[0, 2\pi]$ для абсолютно недостоверных фаз (у которых показатель достоверности равен 0), а чем выше показатель достоверности, тем генерировать их ближе к имеющемуся значению. Для этих целей можно использовать распределение фон Мизеса. В описываемой программе фазы генерируются в соответствии с распределением

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

при этом $t_h$ находится из условия

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

т.е.

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

где $I_0(t_h)$, $I_1(t_h)$ - модифицированные функции Бесселя соответствующих порядков.

*Кумулятивная функция электронной плотности*.

Когда имеется синтез Фурье, рассчитанный в элементарной ячейке на некоторой сетке, мы можем определить минимальное и максимальное значение ρmin и ρmax этой функции, а также подсчитать, как часто встречается каждое значение между ρmin и ρmax. Точнее, можно разбить этот диапазон на несколько участков (для определенности - на 10) и подсчитать, в скольких точках сетки значение функции находилось между

ρmin и ρmin+ (ρmax-ρmin)/10,

в скольких - между

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

и так далее. В нашем случае мы получим 10 чисел, и сумма этих чисел должна совпадать с количеством точек сетки. Эти 10 чисел называются гистограммой электронной плотности. Они показывают, как часто функция электронной плотности принимает те или иные значения. Например, если ρmin = -1000, ρmax = 1500, а гистограмма

10 30 50 230 320 220 110 20 10

то

функция принимает значение между -1000 и –750 в 10 точках,
функция принимает значение между -750 и –500 в 30 точках,

и т.д. Однако по ряду причин бывает удобнее работать не с гистограммой электронной плотности, а с ее кумулятивной функцией (интегралом гистограммы), которая однозначно определяется гистограммой электронной плотности. Так, приведенной здесь гистограмме соответствует кумулятивная функция

10 40 90 320 640 860 970 990 1000,

что означает, что

в 10 точках функция не превосходит -750,
в 40 точках -500,

и так далее.
Кумулятивная функция может быть выражена как в целых числах (как здесь), так и в вещественных числах, показывающих долю, которую составляют точки каждого бина по отношению к полному числу точек сетки:

0.010 0.040 0.090 0.320 0.640 0.860 0.970 0.990 1.000

Известно, что для родственных белков гистограммы выглядят похожим образом и поэтому могут быть использованы как источник дополнительной информации. Программа GENMEM использует кумулятивную функцию, выраженную в относительных частотах. Это происходит, когда значение **моды генерации** равно 4.

*Преобразование, восстанавливающее кумулятивную функцию (гистограмму*).
Пусть имеется некоторый сгенерированный набор фаз $\phi_h$ (NREF чисел) и известны экспериментальные значения модулей. С этими величинами можно рассчитать в

элементарной ячейке синтез Фурье $\rho^{calc}(\mathbf{r})$, а затем определить для этого синтеза гистограмму или (что эквивалентно) кумулятивную функцию $K^{calc}(\rho)$.

Если заранее приблизительно известна эталонная кумулятивная функция $K^{exact}(\rho)$, отвечающая синтезу рассчитанному с точными фазами, то всегда можно подобрать такую функцию $\lambda(\rho)$ что после преобразования

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

модифицированная функция $\rho^{m}(\mathbf{r})$ будет иметь кумулятивную функцию, совпадающую с $K^{exact}(\rho)$ [6]. Это преобразование $\rho \rightarrow \lambda(\rho)$ называется "преобразованием, восстанавливающим кумулятивную функцию". Подчеркнем, что при одной и той же эталонной функции $K^{exact}(\rho)$ модифицирующая функция $\lambda(\rho)$ будет разной для разных $\rho^{calc}(\mathbf{r})$, т.е. существенность модификации зависит, вообще говоря, от качества исходного синтеза (если для исходного синтеза $\rho^{calc}(\mathbf{r})$ рассчитанная кумулятивная функция совпадает с эталонной, то модификация превращается в тождественную модификацию $\rho^{m}(\mathbf{r}) = \rho^{calc}(\mathbf{r})$ ).

Получив модифицированную функцию $\rho^{m}(\mathbf{r})$ , можно рассчитать модули и фазы ее коэффициентов Фурье (структурных факторов). Эти фазы $\phi_h$_mod (а не начальные фазы $\phi_h$) используются затем для расчета синтеза, связность которого будет анализироваться. Мы вправе ожидать, что по некоторым характеристикам фазы $\phi_h$_mod будут лучше, чем случайно сгенерированные фазы $\phi_h$.

**13**. **Литература**.


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.