КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2010 Т. 2 № 1 С. 91–101

Ки&М

АНАЛИЗ И МОДЕЛИРОВАНИЕ СЛОЖНЫХ ЖИВЫХ СИСТЕМ

УДК: 543.442.3:548.73

Применение методов кластерного анализа к исследованию множества допустимых решений фазовой проблемы биологической кристаллографии

О. В. Соболев[a], Н. Л. Лунина, В. Ю. Лунин

Учреждение Российской академии наук Институт математических проблем биологии РАН,
142290, Московская обл., г. Пущино, ул. Институтская, д. 4

E-mail: [a]oleg@impb.psn.ru

Получено 5 марта 2010 г.

Рентгеновский дифракционный эксперимент позволяет определить значения модулей комплексных коэффициентов в разложении в ряд Фурье функции, описывающей распределение электронов в исследуемом объекте. Определение недостающих значений фаз коэффициентов Фурье представляет центральную проблему метода. Результатом применения некоторых подходов к решению фазовой проблемы является множество допустимых решений. Методы кластерного анализа позволяют исследовать структуру этого множества и выделить одно или несколько характерных решений. Существенной особенностью описываемого подхода является то, что близость решений оценивается не по их формальным параметрам, а на основе корреляции предварительно выровненных синтезов Фурье электронной плотности, рассчитанных с использованием сравниваемых наборов фаз. Предлагаемый метод исследования реализован в виде интерактивной программы ClanGR.

Ключевые слова: биологическая кристаллография, фазовая проблема, кластерный анализ

The use of cluster analysis methods for the study of a set of feasible solutions of the phase problem in biological crystallography

O. V. Sobolev, N. L. Lunina, V. Yu. Lunin

Institute of Mathematical Problems of Biology RAS, 4 Institutskaya str., Pushchino, Moscow region, 142290, Russia

Abstract. – X-ray diffraction experiment allows determining of magnitudes of complex coefficients in the decomposition of the studied electron density distribution into Fourier series. The determination of the lost in the experiment phase values poses the central problem of the method, namely the phase problem. Some methods for solving of the phase problem result in a set of feasible solutions. Cluster analysis method may be used to investigate the composition of this set and to extract one or several typical solutions. An essential feature of the approach is the estimation of the closeness of two solutions by the map correlation between two aligned Fourier syntheses calculated with the use of phase sets under comparison. An interactive computer program ClanGR was designed to perform this analysis.

Keywords: biological crystallography, phase problem, cluster analysis

Citation: *Computer Research and Modeling*, 2010, vol. 2, no. 1, pp. 91–101 (Russian).

Работа поддержана грантом РФФИ 10-04-00254а и программой межакадемического сотрудничества РАН и НЦНИ (Франция).

1. Введение

1.1. Фазовая проблема в дифракционных методах

Основой рентгеновского дифракционного эксперимента является рассеяние рентгеновских лучей на электронах исследуемого объекта, что позволяет ставить задачу нахождения функции трех пространственных переменных $\rho(\mathbf{r})$, описывающей распределение электронной плотности в объекте. Интерпретация максимумов этой функции как центров положения атомов дает возможность представлять результат в виде атомной модели исследуемого объекта. Приготовление исследуемого образца в виде кристалла позволяет радикально увеличить интенсивность рассеяния, доведя ее до регистрируемого современной аппаратурой уровня.

Распределение электронной плотности в кристалле может быть представлено в виде трехмерного ряда Фурье

$$\rho(\mathbf{r}) = \frac{1}{|V|} \sum_{\mathbf{s} \in \Re'} F(\mathbf{s}) \exp\left[i\varphi(\mathbf{s}) \right] \exp\left[-2\pi i (\mathbf{s}, \mathbf{r}) \right]. \tag{1.1}$$

Здесь

- $\mathbf{r}$ – точка пространства, определяемая обычно координатами в базисе из минимальных периодов кристалла $\{\mathbf{a}, \mathbf{b}, \mathbf{c}\} : \mathbf{r} = x\mathbf{a} + y\mathbf{b} + z\mathbf{c}$;
- $\mathbf{s}$ – «вектор рассеяния», задаваемый, обычно, в базисе $\{\mathbf{a}^*, \mathbf{b}^*, \mathbf{c}^*\}$, сопряженном базису $\{\mathbf{a}, \mathbf{b}, \mathbf{c}\} : \mathbf{s} = h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*$;
- суммирование в (1.1) идет по всем векторам $\mathbf{s}$ с целочисленными координатами (h, k, l);
- V – элементарная ячейка кристалла: параллелепипед, построенный на векторах $\{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$; $|V|$ – объем элементарной ячейки.

Модули комплексных коэффициентов Фурье (обычно именуемых в кристаллографии «структурными факторами») $F(\mathbf{s})$ могут быть определены непосредственно из эксперимента, а определение значений фаз $\varphi(\mathbf{s})$ представляет собой отдельную проблему (т. н. «фазовую проблему» рентгеноструктурного анализа). В настоящее время разработан ряд походов к решению этой проблемы, опирающихся либо на дополнительные эксперименты со специальным образом модифицированными объектами, либо на некоторые априорно известные свойства распределения электронной плотности [Бландел, Джонсон, 1979; International Tables, 2001].

Следует заметить, что проблема нахождения распределения $\rho(\mathbf{r})$ осложняется тем, что дифракционный эксперимент позволяет определить значения модулей структурных факторов лишь для конечного набора членов ряда (1.1). Этот набор зависит от качества используемого кристалла и условий эксперимента (длины волны используемого рентгеновского излучения, времени жизни кристалла в мощном рентгеновском пучке и т. д.). Размер набора структурных факторов с измеренными в эксперименте модулями обычно характеризуется разрешением $d_{\min}$, которое в биологической кристаллографии определяется как минимальное значение величины $d = 1/|\mathbf{s}|$ для структурных факторов, включенных в набор. Частичная сумма ряда Фурье (1.1), включающая только члены с экспериментально определенными значениями $F^{obs}(\mathbf{s})$, называется синтезом Фурье электронной плотности. В зависимости от разрешения такой синтез может содержать большие или меньшие искажения и представлять более или менее детальную информацию об исследуемой структуре. Разрешение синтеза $d_{\min}$ определяет минимальный размер деталей структуры, которые могут быть «визуально» различимы при каком-либо графическом представлении синтеза Фурье.

Достигнутый уровень развития биологической кристаллографии позволяет определять структуру многих биологических макромолекул (белков, вирусов, нуклеиновых кислот и их комплексов) почти автоматически при условии, что получены кристаллы исследуемого объекта достаточно высокого качества. Тем не менее существует значительное число случаев, когда стандартные подходы не приводят к успеху, и требуется «ручная» работа квалифицированного исследователя и специально разработанные инструменты для проведения такой работы. Особенно часто такая ситуация возникает при исследованиях больших макромолекулярных комплексов либо новых классов биологических объектов (например мембранных белков). Данная работа посвящена описанию методики сравнительного анализа множества потенциально возможных решений фазовой проблемы и реализующего ее программного средства ClanGr.

1.2. Многовариантные подходы к решению фазовой проблемы

В ряде подходов к решению фазовой проблемы на выходе процедуры расчета значений фаз получается не единственное решение, а набор «возможных решений». Мы остановимся на двух примерах, хотя область применения методики существенно шире. Первым примером является процедура прямого решения фазовой проблемы при низком разрешении [Lunin et al., 2002; Лунин, 2002]. При этом подходе просматривается большое количество случайно сгенерированных наборов фаз (вариантов) и отбираются варианты, отвечающие заранее сформулированным критериям отбора. К сожалению, все известные на настоящее время критерии отбора имеют общие недостатки:

- наилучшие значения критерия могут соответствовать вариантам, далеким от правильного решения;
- варианты, очень близкие к правильному решению, могут обладать плохими или «невыразительными» значениями критерия отбора.

В то же время было показано, что ряд критериев обладает «мягкой» селективностью, а именно, что концентрация вариантов, близких к правильному решению, возрастает при фильтрации исходной случайной «популяции» с использованием этих критериев отбора. Примерами таких критериев являются распределение частот встречаемости тех или иных значений электронной плотности («гистограмма электронной плотности» или функция Лебега) [Lunin et al., 1990] или свойства связности области высоких значений электронной плотности [Lunin et al., 2000]. Результатом такой процедуры является набор вариантов, среди которых могут встречаться как варианты, близкие к правильному решению, так и далекие от него.

В качестве второго примера мы упомянем метод молекулярного замещения, часто используемый для решения фазовой проблемы [Бландел, Джонсон, 1979; International Tables, 2001]. В этом методе приближенные значения фаз структурных факторов рассчитываются, например, по модели белка, близкого по структуре (гомологичного) исследуемому белку. Выбор гомологичной структуры обычно осуществляется на основе сравнения первичных последовательностей исследуемого белка и белков с уже известной пространственной структурой. При наличии известной гомологичной структуры проблема сводится к определению оптимального положения и ориентации молекулы гомолога в элементарной ячейке кристалла исследуемого белка. Эта задача решается просмотром шестимерного пространства параметров модели – трех углов, определяющих вращение молекулы гомолога, и трех позиционных параметров – и сравнения модулей структурных факторов, рассчитанных по текущему варианту размещения модели, с экспериментальными значениями. Указанный метод широко применяется на практике, но встречается со сложностями, когда отсутствует гомологичная структура достаточно хорошего качества. В этом случае иногда осуществляется серия поисков с разными моделями (например с серией моделей, полученных методом ЯМР), и результатом являются несколько наборов параметров, отвечающих разным вариантам гомологов. Расчет значений фаз структурных факторов по каждой такой модели с использованием соответствующих ей параметров размещения приводит к множеству вариантов решения фазовой проблемы, аналогично случаю, рассмотренному в предыдущем параграфе [Buehler et al., 2009].

1.3. Кластерный анализ множества возможных решений

Дальнейшая обработка результатов базируется на предположении, что правильное решение должно обладать некоей «устойчивостью», т. е. в его окрестности должно концентрироваться достаточно много отобранных вариантов. В то же время варианты, отобранные критерием, но далекие от правильного решения, являются скорее случайными выбросами, нежели правилом. Для анализа картины распределения множества отобранных вариантов в многомерном «конфигурационном» пространстве могут быть использованы методы кластерного анализа. Под этим термином обычно понимается широкий набор различных алгоритмов классификации. Мы ограничимся рассмотрением одного из них – построения иерархического дерева (кластерного дерева). Кластерное дерево отражает процесс последовательного «слипания» точек многомерного пространства (вариантов решения фазовой проблемы в нашем случае) в кластеры при постепенном ослаблении требования на степень близости вариантов внутри кластера. Анализ общего вида дерева может позволить отсечь случайные выбросы и выделить одно основное решение или свести неопределенность к небольшому числу таких решений.

2. Построение кластерного дерева

2.1. Определение расстояния между наборами фаз

Основой для построения кластерного дерева является матрица попарных расстояний между вариантами. Обычно это расстояние определяется на базе евклидова расстояния либо максимального различия в координатах точек. Однако в нашем случае более удачным оказывается введение расстояния, опирающегося на содержательный смысл рассматриваемых вариантов и дополнительную экспериментальную информацию.

Пусть $\mathbf{v}_1 = \{\varphi_1(\mathbf{s})\}_{\mathbf{s}\in S}$ и $\mathbf{v}_2 = \{\varphi_2(\mathbf{s})\}_{\mathbf{s}\in S}$ – два сравниваемых набора фаз. Определим «формальную» корреляцию этих вариантов как корреляцию распределений электронной плотности $\rho_1(\mathbf{r})$, $\rho_2(\mathbf{r})$, получаемых суммированием рядов Фурье (1.1) с использованием экспериментальных значений модулей структурных факторов и сравниваемых наборов фаз [Lunin et al., 1993]:

$$\tilde{C}(\mathbf{v}_1,\mathbf{v}_2) = \frac{\int\limits_V \rho_1(\mathbf{r})\rho_2(\mathbf{r})dV_\mathbf{r}}{\sqrt{\int\limits_V \rho_1^2(\mathbf{r})dV_\mathbf{r}\int\limits_V \rho_2^2(\mathbf{r})dV_\mathbf{r}}} = \frac{\sum\limits_{\mathbf{s}\in S}\left(F^{obs}(\mathbf{s})\right)^2\cos\left(\varphi_1(\mathbf{s})-\varphi_2(\mathbf{s})\right)}{\sum\limits_{\mathbf{s}\in S}\left(F^{obs}(\mathbf{s})\right)^2} \qquad (2.1)$$

и «формальное» расстояние между вариантами

$$\tilde{d}(\mathbf{v}_1,\mathbf{v}_2) = \sqrt{2\left(1-\tilde{C}(\mathbf{v}_1,\mathbf{v}_2)\right)}. \qquad (2.2)$$

При расчете (2.1) полагается, что член $F(\mathbf{0})$, равный общему числу электронов, удален из ряда (1.1). Преимуществом такого подхода к определению расстояния является то, что учитывается информация, полученная об исследуемом объекте в эксперименте, и отражается близость наборов фаз с точки зрения их «потребительской ценности»: в обычной работе мы используем найденные значения фаз именно для того, чтобы рассчитать синтез Фурье и строить по нему модель структуры.

2.2. Выравнивание

Определение расстояния посредством (2.1)–(2.2) (впрочем, как и выбор евклидова расстояния) имеет существенный недостаток. Это связано с тем, что функция $\rho_1(\mathbf{r})$ и «сдвинутая» функция $\rho_2(\mathbf{r}) = \rho_1(\mathbf{r}-\mathbf{t})$ имеют один и тот же набор модулей структурных факторов, но раз-

личающиеся наборы фаз

$$\varphi_2(\mathbf{s}) = \varphi_1(\mathbf{s}) + 2\pi(\mathbf{s},\mathbf{t}). \tag{2.3}$$

Два таких распределения электронной плотности дают одно и то же изображение структуры и различаются лишь выбором начала координат. Поэтому с практической точки зрения они представляют собой одно и то же решение структуры. Однако с точки зрения формального критерия (2.2) они должны быть рассмотрены как различные решения. В связи с этим мы разбиваем все возможные наборы фаз на классы эквивалентности, полагая два варианта эквивалентными, если они связаны соотношением вида (2.3). С учетом этого мы определяем расстояние между двумя вариантами как расстояние между соответствующими классами эквивалентности равенством

$$\begin{aligned} C(\mathbf{v}_1,\mathbf{v}_2) &= \max_{\mathbf{t}} \frac{\int_V \rho_1(\mathbf{r})\rho_2(\mathbf{r}-\mathbf{t})dV_{\mathbf{r}}}{\sqrt{\int_V \rho_1^{\,2}(\mathbf{r})dV_{\mathbf{r}} \int_V \rho_2^{\,2}(\mathbf{r}-\mathbf{t})dV_{\mathbf{r}}}} = \\ &= \max_{\mathbf{t}} \frac{\sum_{\mathbf{s}\in S} (F^{obs}(\mathbf{s}))^2 \cos(\varphi_1(\mathbf{s}) - \varphi_2(\mathbf{s}) + 2\pi(\mathbf{s},\mathbf{t}))}{\sum_{\mathbf{s}\in S} (F^{obs}(\mathbf{s}))^2}, \end{aligned} \tag{2.4}$$

$$d(\mathbf{v}_1,\mathbf{v}_2) = \sqrt{2\left(1 - C(\mathbf{v}_1,\mathbf{v}_2)\right)} = \min_{\mathbf{t}} \tilde{d}\left(\mathbf{v}_1,\mathbf{v}_2^{\mathbf{t}}\right), \mathbf{v}_2^{\mathbf{t}} = \left\{\varphi_2(\mathbf{s}) - 2\pi(\mathbf{s},\mathbf{t})\right\}. \tag{2.5}$$

Введенное таким образом расстояние можно интерпретировать как формальное расстояние между предварительно выровненными распределениями $\rho_1(\mathbf{r})$, $\rho_2(\mathbf{r})$, где выравнивание осуществляется посредством минимизации формального расстояния (2.1)–(2.2).

На начальных стадиях работы по расшифровке структуры дополнительная неопределенность может быть связана с проблемой выбора энантиомера. Суть проблемы в том, что функции $\rho_1(\mathbf{r})$ и $\rho_2(\mathbf{r}) = \rho_1(-\mathbf{r})$ имеют один и тот же набор модулей структурных факторов и не различаются на основе стандартного рентгеновского эксперимента. В то же время функция $\rho_2(\mathbf{r})$ отвечает инвертированному в начале координат изображению $\rho_1(\mathbf{r})$ и имеет отличный от $\{\varphi_1(\mathbf{s})\}$ набор фаз $\varphi_2(\mathbf{s}) = -\varphi_1(\mathbf{s})$. Переход от распределения $\rho_1(\mathbf{r})$ к энантиомеру соответствует замене L-аминокислот на D-аминокислоты и наоборот. Все найденные в живой природе белки состоят исключительно из L-аминокислот. Поэтому выбор правильного энантиомера может быть сделан по этому признаку, если структура определена с точностью, позволяющей определять такие детали, как конформация отдельных аминокислот. (Существуют и другие пути, связанные с проведением более тонких рентгеновских экспериментов). Однако на начальных стадиях исследования такая точность не всегда доступна, поэтому в этом случае мы вынуждены включать энантиоморфные решения $\rho_1(\mathbf{r})$ и $\rho_2(\mathbf{r}) = \rho_1(-\mathbf{r})$ в один класс эквивалентности, считать переход к энантиомеру допустимой операцией при выравнивании и определять расстояние между вариантами как

$$d(\mathbf{v}_1,\mathbf{v}_2) = \min_{\mathbf{t}\in V,\kappa=\pm1} \tilde{d}\left(\mathbf{v}_1,\mathbf{v}_2^{\mathbf{t},\kappa}\right),\ \mathbf{v}_2^{\mathbf{t},\kappa} = \left\{\kappa\varphi_2(\mathbf{s}) - 2\pi(\mathbf{s},\mathbf{t})\right\}. \tag{2.6}$$

Практическое нахождение оптимального выравнивания и вычисление расстояния (2.6) могут быть эффективно осуществлены с использованием алгоритма быстрого преобразования Фурье [Lunin, Lunina, 1996].

Произвол в выборе начала координат и энантиомера может быть ограничен, если исследуемый кристалл обладает некоторой группой симметрий. В этом случае начало координат может быть привязано к некоторой выделенной точке элементарной ячейки (например к точке

пересечения трех осей симметрии, если они имеются). Это не снимает проблему полностью, поскольку в силу периодичности кристалла в элементарной ячейке может быть несколько таких выделенных точек. Однако набор способов выбора начала координат (и энантиомера) может быть существенно ограничен требованием сохранения уравнений симметрии для сдвинутого распределения $\rho(\kappa\mathbf{r}-\mathbf{t}), \kappa=\pm 1$. Такие способы выбора называются «допустимыми». Допустимый сдвиг начала координат $\mathbf{u}$ определяется условиями

$$\mathbf{R}_\nu\mathbf{u}=\mathbf{u}\big|_{\mathrm{mod}\,\Gamma^t} \text{ при всех } \nu=1,...n,$$

и условие допустимости смены энантиомера есть

$$2\mathbf{t}_\nu=\mathbf{0}\big|_{\mathrm{mod}\,\Gamma^t} \text{ при всех } \nu=1,...n,$$

где $\Gamma=\{(\mathbf{R}_\nu,\mathbf{t}_\nu)\}_{\nu=1}^n$ – группа кристаллографических симметрий кристалла, Γ^t – ее трансляционная подгруппа [Lunin, Lunina, 1996]. Выравнивание в таком случае осуществляется с использованием только допустимых способов выбора начала координат или энантиомера.

Мы упомянем еще одну степень свободы при осуществлении выравнивания, включение которой иногда оказывается полезным при работе при низком разрешении. Обычно около половины объема кристаллов белков заняты растворителем. При этом геометрические очертания областей, занятых растворителем и белком, могут быть довольно похожи. Это приводит к тому, что ряд методов решения фазовой проблемы при низком разрешении может давать решения, отвечающие перевернутому изображению $\rho_2(\mathbf{r})=-\rho_1(\mathbf{r})$ или, что то же, $\varphi_2(\mathbf{s})=\varphi_1(\mathbf{s})+\pi$. В таких случаях может являться оправданным включение преобразования $\rho_2(\mathbf{r})=-\rho_1(\mathbf{r})$ в список допустимых операций при выравнивании.

Отметим еще, что одним из параметров, определяющих вычисление расстояния между вариантами, является разрешение синтезов Фурье, используемых при вычислении расстояний.

2.3. Расстояние между кластерами

На старте процедуры построения кластерного дерева каждый вариант рассматривается как отдельный кластер. Каждый последующий шаг состоит в объединении двух наиболее близких кластеров, что требует задания правила вычисления расстояния между кластерами. В наших расчетах мы используем три наиболее распространенных способа определения такого расстояния

$$D_{\min}(A,B)=\min_{a\in A,b\in B} d(a,b),$$

$$D_{\max}(A,B)=\max_{a\in A,b\in B} d(a,b),$$

$$D_{ave}(A,B)=\frac{\sum_{a\in A,b\in B} d(a,b)}{n_A n_B},$$

где n_A, n_B – число элементов соответствующих кластеров A и B.

Существенным достоинством этих трех способов является то, что пересчет на каждом шаге матрицы попарных расстояний между кластерами не требует возврата к исходной матрице расстояний между вариантами, а ограничен модификацией строки и столбца в текущей матрице межкластерных расстояний [Дюран, Оделл, 1977]. Например, если расстояние между кластерами A, B определяется как

$$D(A,B)=\min_{a\in A,b\in B} d(a,b),$$

то

$$D(A\cup B,C)=\min_{u\in A\cup B,c\in C}d(u,c)=\min\left\{\min_{a\in A,c\in C}d(u,c)\ ,\ \min_{b\in B,c\in C}d(u,c)\right\}=\min\left\{D(A,C),D(B,C)\right\}.$$

Т. е. расстояние от нового объединенного кластера до прежних вычисляется через межкластерные расстояния предыдущего шага без обращения к исходной матрице межэлементных расстояний. Аналогично осуществляется модификация матрицы расстояний для двух других типов расстояний.

3. Обработка кластеров

3.1. Центральный вариант и множественное выравнивание

После того как визуальный анализ кластерного дерева привел к выделению одного или нескольких кластеров, возникает задача преобразования отобранного подмножества вариантов в вид, пригодный для дальнейшей работы. Первая проблема, с которой мы здесь сталкиваемся, задача множественного выравнивания отобранных вариантов. Мы используем подход, основанный на выравнивании всех вариантов относительного одного «центрального» варианта $\mathbf{v}_C$. При этом в качестве центрального варианта берется вариант, обеспечивающий минимальный радиус кластера A, вычисляемый как

$$r_{\mathbf{v}^*}=\left\{\sum_{v\in A}d^2\left(\mathbf{v},\mathbf{v}^*\right)\right\}^{1/2},\ r_{\mathbf{v}_C}=\min_{\mathbf{v}^*\in A}r_{\mathbf{v}^*}.$$

Сам центральный вариант может при этом рассматриваться как «типичный» представитель кластера, и соответствующие ему значения фаз структурных факторов могут быть использованы для расчета синтеза Фурье.

3.2. Усреднение вариантов в кластере

Синтезы Фурье, построенные с экспериментально определенными модулями и каким-то образом определенными фазами структурных факторов, могут содержать существенные артефакты, вызванные ошибками в значениях фаз структурных факторов. В биологической кристаллографии для снижения величины ложных сигналов принято вводить в расчет синтеза Фурье индивидуальные весовые множители (показатели достоверности) $m(\mathbf{s})$, отражающие надежность определения значения соответствующей фазы

$$\rho_S(\mathbf{r})=\frac{1}{|V|}\sum_{\mathbf{s}\in S}m(\mathbf{s})F(\mathbf{s})\exp\left[i\varphi(\mathbf{s})\right]\exp\left[-2\pi i(\mathbf{s},\mathbf{r})\right]. \qquad (3.1)$$

В идеальном случае показатели достоверности совпадают с ожидаемым значением косинуса фазовой ошибки [Blow, Crick, 1959]. В нашем случае, мы приходим к этой весовой схеме, вычисляя среднее значение для синтезов Фурье, рассчитанных со всеми вариантами фаз в кластере, предварительно выровненными относительно центрального варианта:

$$\frac{1}{M}\sum_{j=1}^{M}\left\{\frac{1}{|V|}\sum_{\mathbf{s}}F^{obs}(\mathbf{s})\exp\left[i\varphi_j(\mathbf{s})\right]\exp\left[2\pi i(\mathbf{s},\mathbf{r})\right]\right\}=$$

$$=\frac{1}{|V|}\sum_{\mathbf{s}}F^{obs}(\mathbf{s})m(\mathbf{s})\exp\left[i\varphi^{ave}(\mathbf{s})\right]\exp\left[2\pi i(\mathbf{s},\mathbf{r})\right],$$

где

$$m(\mathbf{s})\exp\left[i\varphi^{ave}(\mathbf{s})\right]=\frac{1}{M}\sum_{j=1}^{M}\exp\left[i\varphi_j(\mathbf{s})\right],$$

$$m(\mathbf{s})=\frac{1}{M}\sum_{j=1}^{M}\cos\left(\varphi_j(\mathbf{s})-\varphi^{ave}(\mathbf{s})\right)$$

и M – количество вариантов в кластере. Значения показателей достоверности близки к 1, если во всех вариантах из рассматриваемого кластера значения фазы для данного структурного фактора близки между собой. Наоборот, показатель достоверности мал, если значения фазы рассматриваемого структурного фактора сильно расходятся от варианта к варианту. Вычисление взвешенного синтеза Фурье (3.1) является наиболее распространенным приемом обработки кластера фазовых вариантов. Следует отметить, что указанное усреднение не является простым покоординатным усреднением вариантов в кластере. Выполняя эту операцию, мы принимаем во внимание цель, для которой это делается – расчет синтеза Фурье, что и определяет специфику усреднения.

3.3. Аппроксимация эмпирических распределений вероятностей

Более подробную информацию о распределении значения фазы какого-либо структурного фактора среди вариантов кластера дает сам набор значений $\left\{\varphi_j(\mathbf{s})\right\}_{j=1}^{M}$ и соответствующее ему эмпирическое распределение вероятностей. Это распределение может быть аппроксимировано унимодальным распределением вероятностей (распределение фон Мизеса или «круговое нормальное распределение»)

$$P(\varphi)\propto\exp\left[A\cos\varphi+B\sin\varphi\right]=\exp\left[T\cos\left(\varphi-\varphi^{*}\right)\right]$$

или более общим бимодальным распределением (распределение Хендриксона–Латтмана) [Hendrickson, Lattman, 1970]

$$P(\varphi)\propto\exp\left[A\cos\varphi+B\sin\varphi+C\cos 2\varphi+D\sin 2\varphi\right].$$

Расчет параметров этих распределений позволяет хранить в сжатом виде информацию о расхождении значений фаз внутри кластера и использовать эти распределения для генерации случайных значений фаз в монте-карловских процедурах.

4. Программа ClanGR

4.1. Описание программы

Программа ClanGR предназначена для анализа множества допустимых решений фазовой проблемы в биологической кристаллографии. Программа разработана на языке Python с использованием библиотеки wxWidgets для того, чтобы обеспечить максимально дружественный пользовательский интерфейс и возможность переноса на любые операционные системы. В доступной на данный момент версии программы имеются все необходимые возможности для интерактивной работы с наборами фаз, полученных при решении фазовой проблемы. Для того чтобы начать использовать программу ClanGR, необходимо иметь текстовый файл специального формата, который содержит набор экспериментальных данных (модули структурных факторов) и наборы фаз вместе со структурными факторами, рассчитанными по разным моделям, полученным в процессе моделирования (если набор структурных факторов отвечает некоторой модели). После этого требуется ввести необходимые параметры в специальном окне (рис. 1).

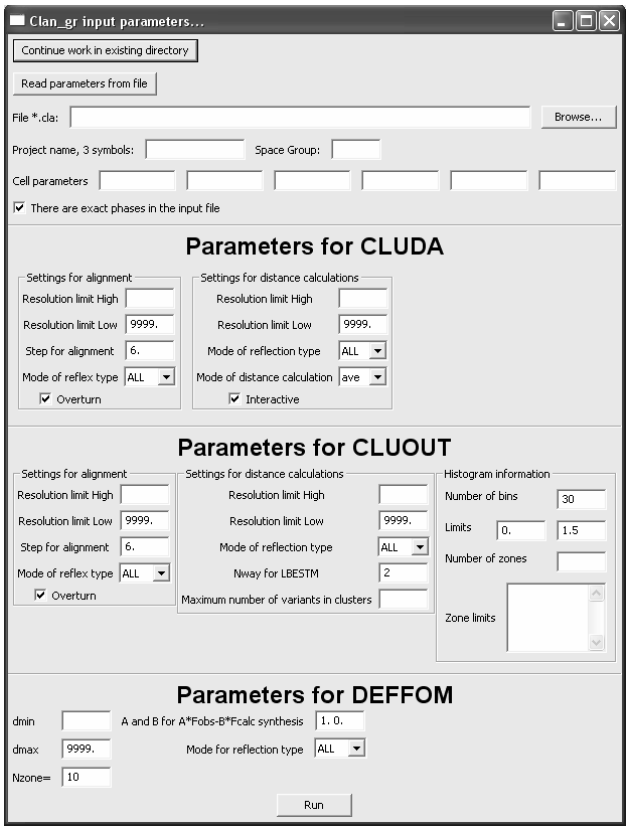

Рис. 1. Окно для ввода параметров ClanGR

Большинство параметров в этом окне уже имеют приемлемые значения по умолчанию. Основными параметрами, которые необходимо ввести, являются путь к файлу с вариантами наборов фаз и параметры элементарной ячейки кристалла. Кроме того, можно включить или выключить возможность «переворачивания» синтеза $\rho(\mathbf{r}) \to -\rho(\mathbf{r})$ при выравнивании синтезов Фурье. Информация о допустимых сдвигах начала координат при выравнивании и имеющихся симметриях извлекается из специального файла-словаря, содержащего описание существующих пространственных групп. Множество других параметров влияют на сбор статистической информации, которая будет отображена в выходных текстовых файлах. Перед запуском расчетного модуля все параметры проверяются на правильность, и если в каком-нибудь окне допущена ошибка, выполнение программы не происходит, а это окно отмечается красным цветом, и выдается сообщение об ошибке. Существует возможность считать все параметры из уже готового файла, который был создан ранее для обработки другого файла с наборами фаз, или продолжить работу с ранее рассчитанным деревом. Для этого можно воспользоваться стандартным диалогом выбора файла или папки.

После ввода всех параметров и нажатия кнопки «Run» будет запущена программа, которая вычислит матрицу попарных расстояний между всеми наборами фаз из входного файла и подготовит информацию о кластерном дереве. Когда она закончит свою работу, на экране появится кластерное дерево (рис. 2). Это основное рабочее окно программы.

Слева показана шкала, по которой можно определить, на каком расстоянии друг от друга находятся отдельные решения или кластеры. Работа с кластерным деревом осуществляется с помощью мыши. Выбранный кластер выделяется красным цветом, и меняются маркеры вершин, входящих в кластер. Для выбранного кластера можно подготовить коэффициенты для расчета взвешенного усредненного синтеза и разностных синтезов. Ответ будет записан в отдельный файл формата UF [Urzhumtsev et al., 1989; Vernoslova, Lunin, 1993], который можно перевести в другие форматы, используемые в биологической кристаллографии, например, с по-

мощью программы CONFOR [Urzhumtseva, Urzhumtsev, 1996]. Имеется возможность сохранить изображение кластерного дерева в формате «portable network graphics» (.png) без потери качества изображения и с минимальным размером файла. Во входном файле с набором фаз также находятся значения критерия отбора для каждого варианта. Максимальное, минимальное и среднее значения этого критерия отображаются для всего дерева и для выбранного кластера.

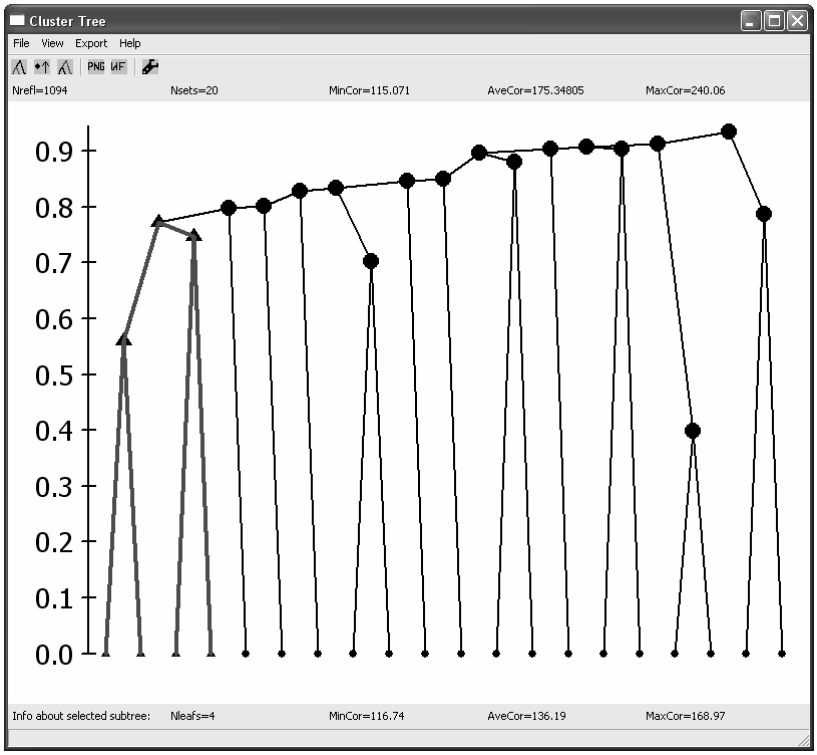

Рис. 2. Главное окно программы ClanGR с примером кластерного дерева

Время, которое требуется для расчета кластерного дерева из тысячи вариантов фаз, не превышает одной минуты на персональном компьютере, а расчет усредненного синтеза требует еще меньшего времени. Это обеспечивает комфортную работу с программой в интерактивном режиме.

4.2. Реализация программы

Программа ClanGR является отдельным приложением и не требует установки каких-либо дополнительных платных или бесплатных библиотек. Программа ClanGR состоит из графической оболочки, обеспечивающей ввод параметров и интерактивную работу с деревом, и двух отдельно компилируемых программ для проведения расчетов. Это позволяет при необходимости отказаться от графического интерфейса ввода параметров и просмотра дерева и использовать пакетный режим для обработки большого объема однородных данных. Графический интерфейс программы ClanGR разработан с использованием языка программирования Python 2.6 и графической библиотеки wxWidgets 2.8. Вычислительные модули программы написаны на языке Fortran77. Данная версия программы может быть запущена под управлением операционных систем семейства Microsoft Windows, включая Windows 2000 и Windows XP. Тем не менее примененные средства разработки позволяют создать отдельные версии ClanGR для работы на других платформах, в том числе MacOS и Linux, что будет сделано в ближайшее время.

Программа, руководство пользователя и исходные коды доступны по запросу.

Список литературы

Blow D. M., Crick F. H. C. The treatment of errors in Isomorphous replacement method // Acta Crystallographica. 1959. Vol. 12. P. 794–802.

Buehler A., Urzhumtseva L., Lunin V. Y., Urzhumtsev A. Cluster analysis for phasing with molecular replacement: a feasibility study // Acta Crystallographica. 2009. D65. P. 644–650.

Hendrickson W. A., Lattman E. E. Representation of phase probability distributions for simplified combination of independent phase information // Acta Crystallographica. 1970. B26. P. 136–143.

International Tables for Crystallography Volume F: Crystallography of Biological Macromolecules / Ed. Rossmann M.G., Arnold E. Dordrecht: Kluwer Academic Publishers, 2001.

Lunin V. Yu., Lunina N. L., Urzhumtsev A. G. Connectivity properties of high-density regions and ab initio phasing at low resolution // Acta Crystallographica. 2000. A56. P. 375–382.

Lunin V. Yu., Lunina N. The Map Correlation Coefficient for Optimally Superposed Maps // Acta Crystallographica. 1996. A52. P. 365-368.

Lunin V. Yu., Lunina N., Podjarny A., Bockmayr A., Urzhumtsev A. Ab initio phasing starting from low resolution // Zeitschrift für Kristallographie. 2002. Vol. 217. P. 668–685.

Lunin V. Yu., Urzhumtsev A. G., Skovoroda T. A. Direct low-resolution phasing from electron-density histograms in protein crystallography // Acta Crystallographica. 1990. A46. P. 540–544.

Lunin V. Yu., Woolfson M. M. Mean Phase Error and the Map Correlation Coefficient // Acta Crystallographica. 1993. D49. P. 530–533.

Urzhumtsev A. G., Lunin V. Yu., Vernoslova E. A. FROG - high-speed restraint-constraint refinement program for macromolecular structure // Journal of Applied Crystallography. 1989. Vol. 22. P. 500–506.

Urzhumtseva L. M., Urzhumtsev A. G. Tcl/Tk-based programs. I. CONFOR: user-friendly converter for crystallographic data files // CCP4 Newsletter. 1996. Number 32. P. 22–24.

Vernoslova E. A., Lunin V. Yu. The FROG PC series: programs for electron-density and model investigation for proteins // Journal of Applied Crystallography. 1993. Vol. 26. P. 291–294.

Бландел Т., Джонсон Л. Кристаллография белка. М.: МИР, 1979.

Дюран Б., Оделл П. Кластерный анализ. М.: Статистика, 1977.

Лунин В. Ю. Определение пространственной структуры биологических макромолекул // Компьютеры и суперкомпьютеры в биологии / Под ред. В. Д. Лахно, М. Н. Устинина. Москва–Ижевск: Институт компьютерных исследований, 2002.