Математическая биология и биоинформатика. 2014. Т. 9. № 2. С. 543–562.
URL: http://www.matbio.org/2014/Lunin_9_543.pdf

========== МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ==========

УДК: 577.3

Использование связных масок в задаче восстановления изображения изолированной частицы по данным рентгеновского рассеяния

Лунин В.Ю., Лунина Н.Л., Петрова Т.Е.

Институт математических проблем биологии, Российская академия наук, Пущино, Московская область, 142290, Россия

Аннотация. Задача восстановления изображения изолированного макромолекулярного объекта по данным рентгеновского рассеяния может быть сформулирована как задача восстановления трехмерного распределения электронной плотности по модулю его преобразования Фурье. Эта задача может быть редуцирована к серии стандартных задач рентгеновской кристаллографии – восстановлению значений периодической функции по значениям модулей ее коэффициентов Фурье (структурных факторов), экспериментально определяемых в рентгеновском эксперименте. В данной работе предлагается новый подход к решению таких задач, основанный на использовании связных бинарных масок в качестве аппроксимации искомого распределения электронной плотности. Подход включает в себя: случайную генерацию большого числа связных масок; отбор масок, согласующихся с экспериментальной и априорной информацией об объекте; выравнивание и усреднение наборов фаз структурных факторов, соответствующих отобранным маскам. Усредненные значения фаз используются вместе с экспериментально определенными значениями модулей структурных факторов для расчета синтеза Фурье, используемого для визуализации изучаемого объекта. Предложенный подход может применяться как при исследовании изолированных частиц, так и кристаллических образцов, однако демонстрирует наибольшие перспективы при работе с изолированными объектами. Приведены результаты тестирования предложенного подхода.

Ключевые слова: *рентгеновская кристаллография, фазовая проблема, XFEL, рассеяние изолированной частицей.*

ВВЕДЕНИЕ

Метод рентгеновской дифракции (рентгеноструктурный анализ) является основным методом получения информации об атомной структуре биологических макромолекул и их комплексов. Однако на сегодняшний день существенным ограничением метода является то, что для проведения рентгеновского эксперимента образец должен быть приготовлен в виде монокристалла исследуемой молекулы. Это требование вызвано тем, что интенсивности рассеянных отдельной молекулой электромагнитных волн (вторичных) на много порядков меньше интенсивности первичного рентгеновского пучка, что чрезвычайно затрудняет их регистрацию. Интерференция волн, рассеянных множеством молекул, лежащих в узлах регулярной периодической решетки, приводит для некоторых выделенных направлений рассеяния к многократному усилению волн, что делает возможным их регистрацию. Появление новых мощных источников рентгеновского излучения – рентгеновских лазеров на свободных электронах – позволяет, в обозримой перспективе, говорить о возможности проведения

рентгеновского эксперимента с отдельными большими молекулами или комплексами молекул [1, 2]. Такая возможность делает актуальной разработку подходов к определению трехмерной структуры биологических макромолекул на основе данных, полученных в рентгеновском дифракционном эксперименте с одиночной молекулой. Данная работа посвящена обсуждению одного из таких подходов.

Главным препятствием на пути определения структуры объекта по дифракционным данным является фазовая проблема. Эксперимент позволяет определить непосредственно лишь модуль комплексной функции, являющейся преобразованием Фурье функции, описывающей пространственное распределение электронов в исследуемом объекте (функции распределения электронной плотности). Решение фазовой проблемы (расчет значений фаз преобразования Фурье) позволяет восстановить распределение электронной плотности посредством вычисления обратного преобразования Фурье – синтеза Фурье электронной плотности. Следует отметить, что решение фазовой проблемы в случае одиночной молекулы является, теоретически, более простой задачей, чем решение этой проблемы в случае кристалла. При дифракции рентгеновских лучей на кристалле, интенсивность рассеянных волн удается измерить лишь для дискретного набора выделенных направлений (Брэгговских рефлексов), в то время как в случае успешного эксперимента с отдельной молекулой можно получить информацию об интенсивности рассеяния во всех направлениях.

Детальность информации, получаемой посредством расчета синтеза электронной плотности, зависит от количества рефлексов (в случае кристалла) или размера области углов рассеяния (в случае одиночной молекулы), включенных в расчет. Эта детальность характеризуется величиной разрешения и определяется предельными значениями углов рассеяния, для которых удалось зарегистрировать интенсивности рассеянных волн. Принятой в кристаллографии практикой является поэтапное продвижение в определении распределения электронной плотности от низкого разрешения к высокому. В данной работе мы обсуждаем подходы к решению фазовой проблемы на начальной стадии исследования – в зоне низкого и среднего разрешения (малые углы рассеяния). Полученная информация позволяет определить общие очертания молекулы и является стартовой точкой для последующего использования методов повышения разрешения. Авторами данной работы был предложен ряд подходов к решению фазовой проблемы в зоне низкого разрешения при исследовании кристаллических структур [3, 4]. В этой работе мы делаем попытку распространить один из этих подходов [5] на случай рассеяния одиночной молекулой. Как будет показано ниже, задачу восстановления распределения электронной плотности в случае рассеяния одиночной молекулой можно рассматривать как серию обычных кристаллографических задач, относящихся к воображаемым кристаллам, содержащим в элементарной ячейке исследуемую молекулу и значительный объем растворителя. Увеличение доли растворителя в воображаемой ячейке потенциально облегчает решение фазовой проблемы, но приводит к существенному росту объема вычислений.

Основой предлагаемого метода является использование свойств компактности и связности области высокой электронной плотности в биологических макромолекулах. Ранее была предложена процедура Монте-Карловского типа [5], которая состоит из случайной генерации фазовых наборов и построении соответствующих синтезов Фурье; отборе тех наборов фаз, которые обеспечивают связность области высоких значений электронной плотности; выравнивания и усреднения отобранных наборов фаз. В данной работе мы исследуем альтернативный подход к использованию свойств связности. В новой процедуре случайным образом генерируются непосредственно гипотетические связные области высокой электронной плотности. Для дальнейшей работы отбираются те из них, модуль преобразования Фурье которых достаточно хорошо воспроизводит значения, полученные в эксперименте. Наборы фаз преобразования Фурье, соответствующие отобранным областям, далее выравниваются

и усредняются, что приводит к искомому решению фазовой проблемы. Такой подход требует существенно меньшего объема вычислений и обеспечивает более высокое качество получаемых наборов фаз, по сравнению с ранее использовавшимся подходом [5].

1. ОСНОВЫ МЕТОДА

1.1. Теоретические основы рентгеноструктурного анализа

Принципиальная схема рентгеновского дифракционного эксперимента изображена на рис. 1. Исследуемый объект помещается в пучок рентгеновских лучей, и регистрируются интенсивности возникающих при этом вторичных волн, расходящихся от объекта во всех направлениях. В процессе эксперимента объект вращается, что позволяет получить на выходе набор двумерных рентгенограмм, отвечающих разной ориентации исследуемого объекта относительно первичного пучка. В кинематической теории рассеяния первичный пучок рассматривается как плоская монохроматическая электромагнитная волна, и учитывается взаимодействие электронов исследуемого объекта только с этой первичной волной. Под воздействием падающей волны электроны объекта начинают осциллировать и становятся источниками новых сферических волн. Эти волны суммируются, образуя рассеянные волны.

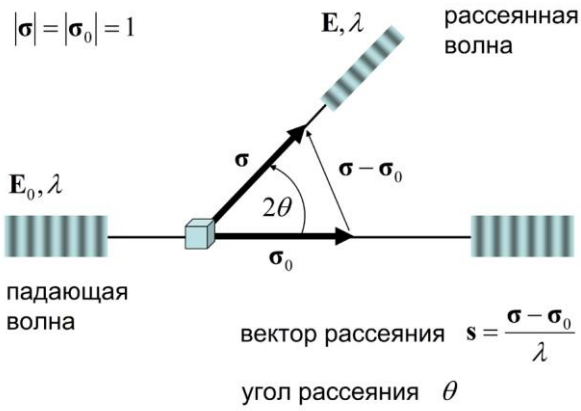

Рис. 1. Схема рентгеновского эксперимента.

Комплексная амплитуда рассеянной волны $\mathbf{E}$ отличается от амплитуды первичной волны $\mathbf{E}_0$ двумя множителями:

$$\mathbf{E} = \varepsilon \mathbf{F}(\mathbf{s}) \mathbf{E}_0 . \qquad (1)$$

Множитель ε равен доле потока энергии исходной волны, приходящей с рассеянной волной на детектор, при рассеянии одним электроном. Эта доля очень мала (она может быть оценена величиной $\varepsilon \sim 10^{-12}$), что и создает основную сложность при регистрации рассеянного излучения. Комплексный множитель $\mathbf{F}(\mathbf{s})$ называется структурным фактором, он определяется распределением электронной плотности в рассеивающем объекте.

Величина структурного фактора зависит от направления рассеяния, более точно, от вектора

$$\mathbf{s} = \frac{\boldsymbol{\sigma} - \boldsymbol{\sigma}_0}{\lambda} , \qquad (2)$$

именуемого вектором рассеяния, и связана преобразованием Фурье с распределением электронной плотности $\rho(\mathbf{r})$ в исследуемом объекте:

$$\mathbf{F}(\mathbf{s}) = \int \rho(\mathbf{r}) \exp\left[2\pi i(\mathbf{s},\mathbf{r})\right] dV_{\mathbf{r}} . \qquad (3)$$

Регистрируемая в эксперименте интенсивность рассеянной волны пропорциональна квадрату модуля структурного фактора, т.е. зависит от распределения электронов в исследуемом образце. Это позволяет ставить задачу определения распределения $\rho(\mathbf{r})$, исходя из набора экспериментально измеренных интенсивностей $\{I(\mathbf{s})\}$.

Теоретически, распределение электронной плотности могло бы быть восстановлено из значений структурных факторов посредством обратного преобразования Фурье

$$\rho(\mathbf{r}) = \int \mathbf{F}(\mathbf{s}) \exp\left[-2\pi i(\mathbf{s},\mathbf{r})\right] dV_{\mathbf{s}} , \qquad (4)$$

однако этому препятствуют два обстоятельства. Во-первых, рентгеновский эксперимент позволяет получить только значения модулей структурных факторов (3) и не позволяет измерить значения фаз. Во-вторых, из (2) следует, что эксперимент позволяет получить значения модулей только для векторов $\mathbf{s}$, удовлетворяющих ограничению $|\mathbf{s}| \le 2/\lambda$. Поэтому, даже при известных фазах структурных факторов, обратное преобразование Фурье позволяет получить вместо точного распределения плотности $\rho(\mathbf{r})$ лишь синтез Фурье конечного разрешения d

$$\rho(\mathbf{r}) = \int\limits_{|\mathbf{s}| \le 1/d} \mathbf{F}(\mathbf{s}) \exp\left[-2\pi i(\mathbf{s},\mathbf{r})\right] dV . \qquad (5)$$

При этом возможное разрешение d не может быть меньше половины длины волны $\lambda/2$.

Пусть теперь исследуемый объект представляет собой кристалл, т. е. имеется множество идентичных молекул, расположенных в узлах целочисленной решетки $\Re$, построенной на базисе $\{\mathbf{a},\mathbf{b},\mathbf{c}\}$. Сдвиг положения молекулы в пространстве на вектор $\mathbf{u}$ приводит к появлению в ее структурном факторе дополнительного фазового множителя $\exp\left[2\pi i(\mathbf{s},\mathbf{u})\right]$. Предположим, что вектор $\mathbf{u}$ пробегает все узлы решетки $\Re$ и выполнены дополнительные условия на вектор $\mathbf{s}$:

$$(\mathbf{s},\mathbf{a}) = h, \quad (\mathbf{s},\mathbf{b}) = k, \quad (\mathbf{s},\mathbf{c}) = l, \qquad h,k,l \text{ – целые числа.} \qquad (6)$$

В этом случае все дополнительные множители в структурных факторах, отвечающих отдельным молекулам, будут равны единице, и произойдет многократное усиление рассеянной волны за счет суммирования идентичных вкладов от множества молекул в кристалле. Интенсивность такой волны становится достаточно большой для практической регистрации в эксперименте. Если же условия (6) не выполняются, то величины дополнительных множителей почти случайны и структурные факторы, отвечающие отдельным молекулам, аннигилируют при сложении. В таком случае интенсивность рассеянной волны становится слишком малой для экспериментального определения. Условия (6) называются условиями дифракции (Лауэ–Брэгга–Вульфа), а соответствующие рассеянные волны – рефлексами. Мы будем далее говорить о модулях и фазах структурных факторов, отвечающих тем или иным рефлексам, имея в виду структурные факторы, входящие в уравнение (1) для этих волн.

Формулы для расчета структурных факторов и электронной плотности в случае кристалла приводятся к виду

$$\mathbf{F}(\mathbf{s}) = \int_V \rho(\mathbf{r}) \exp\left[2\pi i(\mathbf{s},\mathbf{r})\right] dV_{\mathbf{r}}, \tag{7}$$

$$\rho(\mathbf{r}) = \sum_{\mathbf{s}\in\Re'} \mathbf{F}(\mathbf{s}) \exp\left[-2\pi i(\mathbf{s},\mathbf{r})\right], \tag{8}$$

где V – параллелепипед, построенный на векторах $\{\mathbf{a},\mathbf{b},\mathbf{c}\}$ (элементарная ячейка кристалла), $|V|$ – объем элементарной ячейки, суммирование в (8) идет по всем узлам целочисленной решетки, построенной на векторах базиса $\{\mathbf{a}^*,\mathbf{b}^*,\mathbf{c}^*\}$, сопряженного к базису $\{\mathbf{a},\mathbf{b},\mathbf{c}\}$:

$$\Re' = \left\{h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*;\ \ h,k,l - \text{ целые числа}\right\}. \tag{9}$$

Отметим, что это в точности те вектора рассеяния, для которых выполнены условия дифракции (6), т.е. модули структурных факторов которых могут быть измерены в эксперименте.

1.2. Рассеяние изолированной частицей. Редукция к кристаллографическим задачам

Вернемся к случаю рассеяния отдельной молекулой. Пусть $\rho_0(\mathbf{r})$ описывает распределение электронной плотности в отдельной молекуле, и $\mathbf{F}^{sp}(\mathbf{s})$ – соответствующие структурные факторы, вычисленные согласно (3). Выберем в пространстве произвольный базис $\{\mathbf{a},\mathbf{b},\mathbf{c}\}$. Пусть вектора базиса выбраны достаточно большими для того, чтобы молекула целиком помещалась в параллелепипед, построенный на этих векторах. Среди всех структурных факторов $\{\mathbf{F}(\mathbf{s})\}$ выберем подмножество, отвечающее узлам целочисленной сетки

$$\Re'_{abc} = \left\{h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*:\quad h,k,l - \text{ целые числа}\right\}, \tag{10}$$

где $\{\mathbf{a}^*,\mathbf{b}^*,\mathbf{c}^*\}$ – базис, сопряженный к базису $\{\mathbf{a},\mathbf{b},\mathbf{c}\}$ Рассмотрим, наконец, воображаемый кристалл, с параметрами элементарной ячейки $\{\mathbf{a},\mathbf{b},\mathbf{c}\}$, внутри которой находится одна молекула с распределением плотности $\rho_0(\mathbf{r})$ внутри молекулы, а плотность вне границ молекулы принята равной нулю. Структурные факторы для такого кристалла

$$\mathbf{F}^{cryst}(\mathbf{s}) = \int_{V_{abc}} \rho_0(\mathbf{r}) \exp\left[2\pi i(\mathbf{s},\mathbf{r})\right] dV_{\mathbf{r}} = \int \rho_0(\mathbf{r}) \exp\left[2\pi i(\mathbf{s},\mathbf{r})\right] dV_{\mathbf{r}} = \mathbf{F}^{sp}(\mathbf{s}). \tag{11}$$

Таким образом, задание преобразования Фурье для отдельной частицы на целочисленной сетке узлов $\Re'_{abc}$ эквивалентно заданию набора структурных факторов для воображаемой кристаллической структуры, в которой частица находится в кристаллической ячейке с параметрами ячейки $\{\mathbf{a},\mathbf{b},\mathbf{c}\}$. Соответственно задача восстановления электронной плотности по значениям модуля преобразования Фурье $\left|\boldsymbol{F}^{sp}(\boldsymbol{s})\right|$ может быть переформулирована как задача восстановления распределения плотности в воображаемой ячейке по значениям модулей структурных факторов $\left\{\mathbf{F}^{cryst}(\mathbf{s}):\ \ \mathbf{s}\in\Re'_{abc}\right\}$. В силу этого к задаче определения структуры отдельной частицы применимы все существующие методы, разработанные для решения фазовой проблемы в кристаллографии белка. Следует отметить, что при таком сведении задачи к задаче

макромолекулярной кристаллографии мы используем лишь часть потенциально доступной информации – только модули преобразования Фурье, взятые в узлах сетки (10). Как выбор базиса $\{\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3\}$, так и выбор длин сторон ячейки a,b,c может быть осуществлен множеством разных способов. Данные по рассеянию отдельной частицей содержат гораздо больше информации, нежели обычная кристаллографическая задача, и это дает больше возможностей для решения фазовой проблемы.

Хотя формально формулировка задачи, предложенная в предыдущем параграфе, похожа на обычную кристаллографическую проблему, она имеет существенное отличие. При больших значениях параметров a,b,c мы получаем элементарную ячейку, в которой лишь небольшая часть ячейки занята неизвестной нам электронной плотностью, а значения плотности в большей части элементарной ячейки нам известны (равны нулю). Это делает задачу определения структуры переопределенной [6] и создает возможность ее решения. Существенным препятствием на пути решения является то, что мы не знаем, в каких именно точках плотность равна нулю. Если эта информация (маска области молекулы) каким-то образом получена, то для решения фазовой проблемы могут применяться различные итерационные процедуры [6–8]. В данной статье мы обсуждаем способ получения этой информации. Следует отметить, что, увеличивая параметры ячейки, мы вовлекаем в работу большее количество экспериментальной информации (большее число рефлексов) и работаем (при сохранении одного и того же разрешения) с большим количеством структурных факторов воображаемого кристалла. С другой стороны, такое увеличение приводит к росту доли области растворителя в ячейке, что повышает избыточность информации и может облегчать решение фазовой проблемы.

1.3. Фазовая проблема. *Ab initio* методы решения

Одной из центральных проблем при определении структуры кристаллического вещества является потеря в эксперименте половины необходимой информации. Для восстановления распределения электронной плотности посредством (8) необходимы и модули, и фазы структурных факторов, а эксперимент позволяет получить лишь значения модулей. Этот недостаток информации должен быть компенсирован какой-либо другой информацией об объекте. Как правило, в качестве такой дополнительной информации выступают либо экспериментальные данные, полученные в дополнительных рентгеновских экспериментах (с изоморфными производными, либо при других длинах волн рентгеновского излучения), либо известная структура гомологичного белка. Особую группу методов составляют так называемые прямые или *ab initio* методы, в которых используется дополнительная информация общего типа, не привязанная к конкретному исследуемому объекту. Примером такой информации является связность областей высокой электронной плотности в исследуемом объекте [5, 9]. Для использования этого свойства была разработана процедура Монте-Карловского типа и создан комплекс программ GENMEM, реализующий эту процедуру. Процедура состоит из нескольких шагов. На первом этапе генерируется случайно гипотетический набор возможных значений фаз структурных факторов. Сгенерированные фазы совместно с экспериментально полученными значениями модулей структурных факторов используются для расчета синтеза Фурье электронной плотности. По синтезу Фурье выделяется область высоких значений плотности (уровень "срезки" является параметром метода) и определяется количество компонент связности этой области. Если число таких компонент не превосходит указанное число, то сгенерированный набор фаз рассматривается как допустимый и запоминается для дальнейшего анализа. Генерация случайных наборов фаз повторяется до тех пор, пока количество отобранных допустимых наборов фаз не достигнет заданной величины. На

втором этапе отобранные наборы фаз выравниваются [10, 11]. Необходимость выравнивания связана с тем, что модули структурных факторов не фиксируют выбор начала координат. Распределения электронной плотности для исходного объекта и этого же объекта, сдвинутого на некоторый вектор, имеют идентичные наборы модулей структурных факторов (но разные наборы фаз). Поэтому при случайной генерации наборов фаз возможна ситуация, когда два сгенерированных набора приводят к одинаковым (или близким) изображениям молекулы, но имеют формально разные значения фаз. Эта разница может быть устранена сдвигом второго изображения (и соответствующей трансформацией фаз). После того, как все фазовые наборы выровнены, для каждого структурного фактора осуществляется усреднение значений фаз и вычисляется показатель разброса значений фазы в разных отобранных наборах. Средние значения и показатели разброса (в виде корректирующих множителей) используются для расчета синтеза Фурье (8), являющегося текущим приближением к решению задачи нахождения распределения электронной плотности в объекте. Найденное приближение может использоваться для модификации первого шага процедуры. Случайные значения фаз могут теперь генерироваться с учетом полученной информации об их наиболее вероятных значениях. Процедура может быть повторена несколько раз, с меняющимися индивидуальными распределениями вероятностей при генерации значений фаз структурных факторов, до достижения сходимости.

Дополнительным критерием отбора в рамках этой процедуры могут являться ограничения на размер молекулы, например, в виде требования, чтобы область высоких значений помещалась в параллелепипед заданных размеров.

1.4. Использование масок области молекулы для решения фазовой проблемы

При работе на низком и среднем разрешении основной информацией, извлекаемой из полученного распределения электронной плотности $\rho(\mathbf{r})$, является маска области молекулы (характеристическая функция). Эта маска определяется как область наиболее высоких значений на синтезе Фурье электронной плотности (далее: область высоких значений, ОВЗ). При заданном уровне срезки ρ_{crit} мы определяем маску области как бинарную функцию

$$\rho^{mask}(\mathbf{r}) = \begin{cases} 1, & \text{если } \rho(\mathbf{r}) \geq \rho_{crit}, \\ 0, & \text{если } \rho(\mathbf{r}) < \rho_{crit}. \end{cases} \qquad (12)$$

Альтернативный путь – задать желаемый размер области высоких значений и подобрать уровень срезки ρ_{crit} так, чтобы функция (12) выделяла область желаемого размера. Желаемый объем области может указываться разными путями. Это может быть, к примеру, *абсолютный объем* области трехмерного пространства (в $Å^3$) или число узлов сетки, если область строится на некоторой сетке в элементарной ячейке. Это также может быть *относительный объем* – отношение объема области к объему элементарной ячейки. В ряде случаев удобно характеризовать размер области тем, какой объем приходится на некую структурную единицу. Мы будем использовать для этих целей *удельный объем*, вычисляемый как объем, приходящийся, в среднем, на один аминокислотный остаток структуры.

На начальных этапах исследования структуры объектом поиска может являться не само распределение плотности, а бинарная маска, наилучшим образом описывающая молекулу. Более того, само свойство бинарности может быть использовано как дополнительная информация об искомом объекте для восстановления значений фаз [12].

В этой работе мы предлагаем новый подход к определению *ab initio* структуры биологических макромолекул при низком и среднем разрешении. Этот подход может

применяться как в случае кристаллических образцов, так и в случае отдельной частицы, но его вычислительные преимущества становятся особенно значимыми при исследовании отдельных частиц.

Объектом поиска является маска области молекулы, заданная на некоторой сетке в элементарной ячейке.

$$\{b_{mnp}\}, \quad 0 \le m < N_x,\ 0 \le n < N_y,\ 0 \le p < N_z. \tag{13}$$

Здесь N_x, N_y, N_z обозначают количество делений сторон ячейки при построении сетки. На первом этапе случайным путем генерируются связные множества точек, состоящие из заданного количества точек. Каждому сгенерированному множеству точек Ω ставится в соответствие бинарная функция

$$\rho^{mask}(\mathbf{r}) = \begin{cases} 1 \text{ для } \mathbf{r} \in \Omega, \\ 0 \text{ для } \mathbf{r} \notin \Omega. \end{cases} \tag{14}$$

По этой функции рассчитываются модули и фазы структурных факторов. Если рассчитанные модули оказались достаточно близки к экспериментальным (например, имеют корреляцию с ними выше заданной), то соответствующий набор фаз рассматривается как допустимый и запоминается для дальнейшей работы. Процесс генерации повторяется до тех пор, пока не отбирается нужное число допустимых наборов фаз. Последующие два этапа работы – выравнивание и усреднение – происходят так же, как это было описано ранее в разделе 1.3. При отборе сгенерированных масок могут налагаться дополнительные условия отбора, сформулированные в терминах характеристик масок, например, ограничиваться протяженности масок в различных направлениях или задаваться требуемый радиус инерции масок. При начале работы над структурой добавление узлов в генерируемую маску может происходить с равной вероятностью для всех точек сетки. По мере продвижения в работе синтез Фурье, получаемый после усреднения отобранных наборов фаз, может быть конвертирован в распределение вероятностей для присутствия в маске тех или иных узлов сетки. После этого генерация масок на последующем цикле может вестись уже с учетом этого распределения вероятностей.

2. РЕЗУЛЬТАТЫ ТЕСТОВ

2.1 Тестовый объект

В качестве тестового объекта был выбран тример мембранного белка AcrB [13], состоящий из 3129 аминокислотных остатков (23811 неводородных атомов). Общий вид тримера показан на рис. 2.

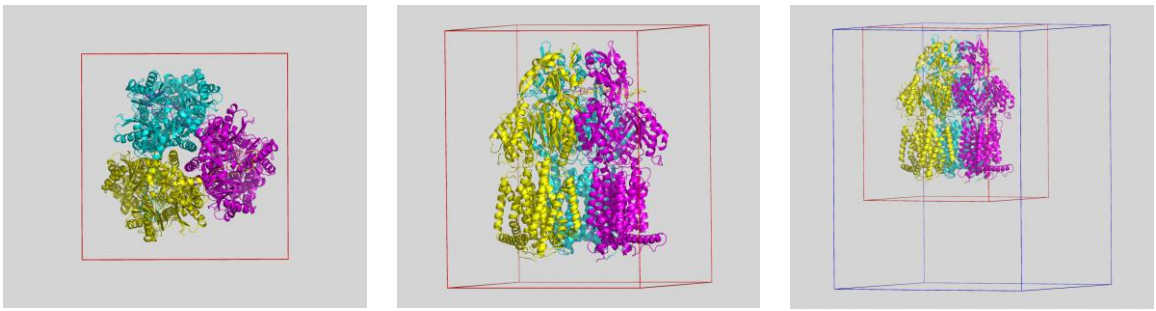

Рис. 2. Структура белка AcrB. Три мономера, образующих молекулу, показаны разными цветами. Показаны границы малой (красный цвет) и расширенной (синий цвет) элементарной ячейки.

Для тестов были рассчитаны два набора комплексных структурных факторов. Модули структурных факторов рассматривались в тестах как известные, экспериментально определенные значения $F^{obs}(\mathbf{s})$. Рассчитанные фазы структурных факторов считались точными значениями фаз $\varphi^{exact}(\mathbf{s})$ и использовались только для контроля результатов. В процессе тестового определения фаз они считались неизвестными. Два тестовых набора структурных факторов соответствовали двум вариантам введения воображаемой кристаллической структуры. В первом случае молекула белка AcrB предполагалась размещенной в прямоугольной элементарной ячейке с длинами сторон 120, 120, 150 Å (малая ячейка), а во втором 180, 180, 225 Å (расширенная ячейка). В обоих случаях предполагалась пространственная группа *P*1 (отсутствие кристаллографической симметрии). Первый случай близок к обычной задаче определения кристаллической структуры при наличии большого объема растворителя (около 75%). Второй случай соответствует воображаемому кристаллу (с нереально большим объемом растворителя), рассматриваемому при рентгеноструктурном исследовании изолированных частиц.

Следует заметить, что молекула AcrB обладает осью симметрии третьего порядка. Эта симметрия никак не учитывалась ни при выборе элементарной ячейки, ни в процессе тестового восстановления значений фаз структурных факторов. Поэтому проявление этой симметрии на получаемых изображениях молекулы являлось одним из подтверждений правильности решения.

Работа по *ab initio* определению значений фаз структурных факторов велась в зоне разрешения 25 Å. При этом для контроля рассчитывались также значения различных показателей в зонах более низкого разрешения 40 и 30 Å. Распределение числа используемых структурных факторов по зонам разрешения приведено в табл. 1.

Таблица 1. Распределение числа структурных факторов по зонам разрешения

	длины сторон ячейки [Å]	зона разрешения (d_{min}) [Å]		
		40	30	25
		число структурных факторов		
малая ячейка	120, 120, 150	69	170	288
расширенная ячейка	180, 180, 225	247	555	976

2.2 Контрольные критерии

Выбор наиболее адекватных критериев для сравнения наборов модулей или фаз структурных факторов или синтезов Фурье электронной плотности является предметом отдельных дискуссий в белковой кристаллографии [14]. Мы ограничимся в данной статье упоминанием двух использованных критериев. Для сравнения рассчитанных по бинарной маске модулей структурных факторов $\{F^{calc}(\mathbf{s})\}$ с их экспериментальными значениями $\{F^{obs}(\mathbf{s})\}$ использовался нецентрированный коэффициент корреляции

$$CM = \frac{\sum_{\mathbf{s}} F^{obs}(\mathbf{s}) F^{calc}(\mathbf{s})}{\sqrt{\sum_{\mathbf{s}} \left(F^{obs}(\mathbf{s})\right)^2 \sum_{\mathbf{s}} \left(F^{calc}(\mathbf{s})\right)^2}}. \quad (15)$$

Критерий для сравнения двух наборов фаз (например, найденного *ab initio* решения и точных фаз) вводится в два этапа [11, 15]. На первом этапе мы определяем коэффициент корреляции двух фазовых наборов в виде

$$FCP\left(\left\{\varphi_1(\mathbf{s})\right\},\left\{\varphi_2(\mathbf{s})\right\}\right)=FCP\left(\rho_1(\mathbf{r}),\rho_2(\mathbf{r})\right)=$$

$$=\frac{\int\rho_1(\mathbf{r})\ \ \rho_2(\mathbf{r})dV_{\mathbf{r}}}{\sqrt{\int\left(\rho_1(\mathbf{r})\right)^2 dV_{\mathbf{r}}\int\left(\rho_2(\mathbf{r})\right)^2 dV_{\mathbf{r}}}}=\frac{\sum_{\mathbf{s}}\left(F^{obs}(\mathbf{s})\right)^2\cos\left(\varphi_1(\mathbf{s})-\varphi_2(\mathbf{s})\right)}{\sum_{\mathbf{s}}\left(F^{obs}(\mathbf{s})\right)^2}. \qquad (16)$$

Здесь $\rho_1(\mathbf{r})$ и $\rho_2(\mathbf{r})$ – синтезы Фурье, вычисляемые с экспериментальными модулями $F^{obs}(\mathbf{s})$ и сравниваемыми наборами фаз. Следует заметить, что введенная метрика привязана к конкретному рентгеновскому эксперименту посредством введения в расчет экспериментальных значений модулей структурных факторов.

Введенное выше формальное сравнение наборов фаз не всегда приемлемо в задачах восстановления значений фаз по значениям модулей структурных факторов. Это связано с тем, что значения модулей структурных факторов сами по себе не фиксируют выбор начала координат. Более точно, две функции $\rho_1=\rho(\mathbf{r})$ и $\rho_2=\rho(\mathbf{r}-\mathbf{u})$, отличающиеся сдвигом на вектор $\mathbf{u}$, имеют одни и те же значения модулей структурных факторов, но разные значения фаз

$$\varphi_2(\mathbf{s})=\varphi_1(\mathbf{s})+2\pi(\mathbf{s},\mathbf{u}). \qquad (17)$$

Эти две функции приводят к одному и тому же изображению структуры, и, в рамках задачи определения структуры, должны рассматриваться как эквивалентные. В процессе случайной генерации масок или случайной генерации фаз возможна ситуация, когда сгенерированные изображения (или синтезы Фурье, построенные со сгенерированными фазами) отличаются лишь сдвигом, или становятся очень похожими после надлежащим образом подобранного сдвига. Такие наборы фаз в рамках нашей задачи должны рассматриваться как близкие. Поэтому в качестве второго этапа в процедуру определения степени близости двух наборов фаз мы вводим выравнивание: подбор вектора сдвига $\mathbf{u}$, обеспечивающего максимальное значение формального критерия (16):

$$CP\left(\left\{\varphi_1(\mathbf{s})\right\},\left\{\varphi_2(\mathbf{s})\right\}\right)=\max_{\mathbf{u}}FCP\left(\rho_1(\mathbf{r}-\mathbf{u}),\rho_2(\mathbf{r})\right)=$$

$$=\max_{\mathbf{u}}\frac{\sum_{\mathbf{s}}\left(F^{obs}(\mathbf{s})\right)^2\cos\left(\varphi_1(\mathbf{s})+2\pi(\mathbf{s},\mathbf{u})-\varphi_2(\mathbf{s})\right)}{\sum_{\mathbf{s}}\left(F^{obs}(\mathbf{s})\right)^2}. \qquad (18)$$

При работе в пространственной группе $P1$ максимизация идет по всем векторам сдвига $\mathbf{u}$. Набор допустимых векторов сдвига может быть ограничен при наличии кристаллографической симметрии [11].

Еще одним преобразованием функции распределения электронной плотности, не меняющим значения модулей структурных факторов, является переход к энантиомеру: функции $\rho(\mathbf{r})$ и $\rho(-\mathbf{r})$ обладают одинаковыми наборами модулей структурных факторов. При работе при низком и среднем разрешении выбор правильного энантиомера решения, как правило, невозможен. Поэтому в процедуру выравнивания вводится дополнительно возможность смены энантиомера в тех случаях, когда такая смена допустима симметрией (например, при работе в группе $P1$).

$$CP\left(\left\{\varphi_1(\mathbf{s})\right\},\left\{\varphi_2(\mathbf{s})\right\}\right)=\max_{\kappa=\pm1}\max_{\mathbf{u}}FCP\left(\rho_1(\kappa\mathbf{r}-\mathbf{u}),\rho_2(\mathbf{r})\right). \qquad (19)$$

2.3. Усреднение наборов фаз. Взвешенные синтезы Фурье

В обсуждаемых в данной работе процедурах *ab initio* определения фаз результатом первого этапа работы является множество (популяция) отобранных наборов фаз. Простейшей процедурой обработки отобранных фазовых наборов является их усреднение, которое применяется к предварительно выровненным относительно друг друга наборам фаз $\{\varphi_1(\mathbf{s})\},\{\varphi_2(\mathbf{s})\},...,\{\varphi_M(\mathbf{s})\}$ [4]. Более строго, усреднение применяется к синтезам Фурье, вычисленным с экспериментальными модулями структурных факторов $F^{obs}(\mathbf{s})$ и различными наборами фаз. Это приводит к взвешенному синтезу Фурье, который может быть вычислен как

$$\rho(\mathbf{r})=\frac{1}{|V|}\sum_{\mathbf{s}} m(\mathbf{s})F^{obs}(\mathbf{s})\exp\left[i\varphi^{best}(\mathbf{s})\right]\exp\left[i2\pi(\mathbf{s},\mathbf{r})\right]. \quad (20)$$

При этом для каждого структурного фактора определяются наилучшая фаза $\varphi^{best}(\mathbf{s})$ и показатель достоверности (ее определения) $m(\mathbf{s})$:

$$m(\mathbf{s})\exp\left[i\varphi^{best}(\mathbf{s})\right]=\frac{1}{M}\sum_{j=1}^{M}\exp\left[i\varphi_j(\mathbf{s})\right]. \quad (21)$$

Нетрудно видеть, что $m(\mathbf{s})$ характеризует разброс значений фазы в разных наборах относительно усредненного значения фазы $\varphi^{best}(\mathbf{s})$:

$$m(\mathbf{s})=\frac{1}{M}\sum_{j=1}^{M}\cos\left(\varphi_j(\mathbf{s})-\varphi^{best}(\mathbf{s})\right). \quad (22)$$

При работе с взвешенными синтезами Фурье (20) мы, помимо коэффициента корреляции фаз (19), будем рассматривать также взвешенный коэффициент корреляции CP_w. Этот коэффициент определяется аналогично коэффициенту CP как результат оптимизации по всевозможным сдвигам

$$CP_w\left(\{\varphi_1(\mathbf{s})\},\{\varphi_2(\mathbf{s})\}\right)=\max_{\kappa=\pm1}\max_{\mathbf{u}} FCP_w\left(\rho_1(\kappa\mathbf{r}-\mathbf{u}),\rho_2(\mathbf{r})\right), \quad (23)$$

но вместо формального коэффициента корреляции (16) здесь используется взвешенный формальный коэффициент

$$FCP_w\left(\{\varphi_1(\mathbf{s})\},\{\varphi_2(\mathbf{s})\}\right)=\frac{\sum_{\mathbf{s}} m(\mathbf{s})\left(F^{obs}(\mathbf{s})\right)^2\cos\left(\varphi_1(\mathbf{s})-\varphi_2(\mathbf{s})\right)}{\sqrt{\sum_{\mathbf{s}}\left(F^{obs}(\mathbf{s})\right)^2\sum_{\mathbf{s}}\left(m(\mathbf{s})F^{obs}(\mathbf{s})\right)^2}}. \quad (24)$$

2.4. Усреднение случайных наборов фаз. Эффективность *ab initio* процедур

При работе при низком разрешении коэффициент фазовой корреляции CP должен быть использован с осторожностью в качестве критерия оценки успеха. Дело в том, что наборы модулей структурных факторов для биологических макромолекул часто имеют аномально большие значения для сравнительно небольшого числа рефлексов в зоне низкого разрешения (рис. 3). Поэтому изображения, проявляющиеся на картах низкого разрешения, могут определяться, в основном, небольшим числом этих сильных рефлексов. При этом высокие значения показателя CP могут означать консенсус значений фаз лишь для небольшого числа доминирующих рефлексов. Кроме того, надо учитывать, что при сравнении случайных значений фаз с точными интуитивно

ожидаемое нулевое значение коэффициента корреляции появляется лишь при использовании формального критерия FCP. В случае же, когда сравниваемые наборы фаз предварительно выравниваются, получаемые значения коэффициентов корреляции могут быть существенно выше. На рис. 4 приведены распределения коэффициентов корреляции CP, вычисленных для случайно сгенерированных наборов фаз, а в табл. 2 характеристики этих распределений.

Усреднение (предварительно выровненных) фазовых наборов может приводить к еще большим значениям корреляции CP_w (табл. 3 и 4). Так, для тестового объекта в зоне разрешения 40 Å корреляция 0.7 с точными фазами для малой ячейки может быть получена просто усреднением случайно сгенерированных фазовых наборов. Это базовое значение должно быть принято во внимание при оценке успешности той или иной процедуры решения фазовой проблемы. Процедура может считаться успешной лишь в том случае, когда она позволяет получить более высокие значения корреляции, нежели это позволяет сделать усреднение случайных наборов.

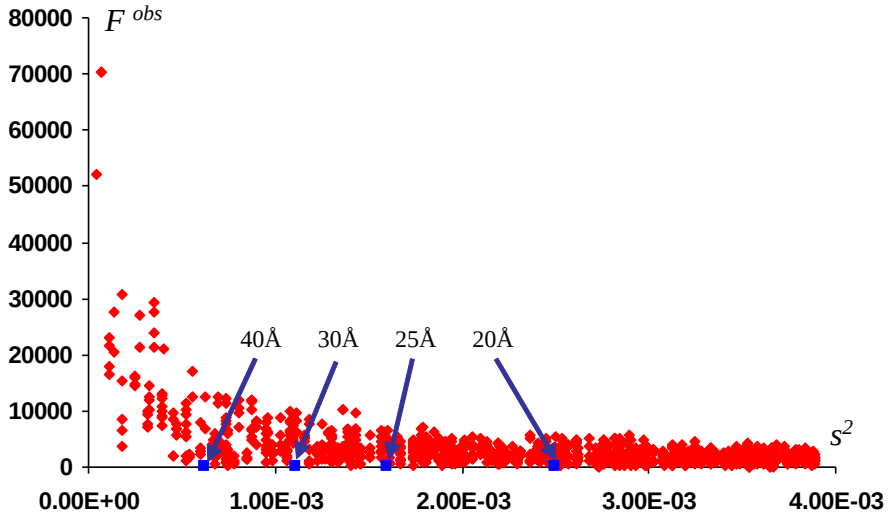

Рис. 3. Модули структурных факторов для AcrB (малая ячейка). Для каждого рефлекса модуль структурного фактора показан как функция s^2.

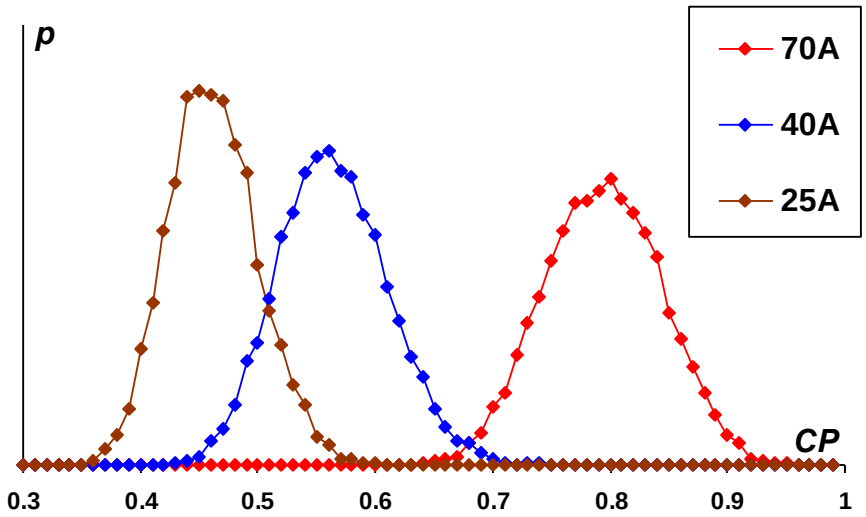

Рис. 4. Эмпирические распределения коэффициента корреляции CP точных и случайно сгенерированных фаз структурных факторов для разных зон разрешения (AcrB, малая ячейка).

Таблица 2. Корреляция *CP* случайно сгенерированных и точных значений фаз (AcrB, малая ячейка)

разрешение d_{min} [Å]	число рефлексов	*CP*				
		min	max	ave	sigma	ave+3sigma
40	69	0.4457	0.7739	0.5791	0.0464	0.7183
30	170	0.3863	0.6971	0.512	0.0412	0.6355
25	288	0.3585	0.6433	0.4767	0.0379	0.5904

Таблица 3. Корреляция CP_w точных значений фаз и фаз, полученных усреднением 100 случайно сгенерированных наборов (AcrB, малая ячейка, пять независимых расчетов)

разрешение d_{min} [Å]	вариант				
	a	*b*	*c*	*d*	*e*
	CP				
40	0.696	0.713	0.679	0.713	0.718
30	0.646	0.653	0.643	0.626	0.665
25	0.625	0.632	0.623	0.647	0.644

Таблица 4. Корреляция CP_w точных значений фаз и фаз, полученных усреднением 100 случайно сгенерированных наборов (AcrB, расширенная ячейка, пять независимых расчетов)

разрешение d_{min} [Å]	вариант				
	a	*b*	*c*	*d*	*e*
	CP				
40	0.50	0.54	0.55	0.56	0.49
30	0.48	0.52	0.52	0.54	0.46
25	0.46	0.40	0.50	0.52	0.45

Сравнение таблиц 3 и 4 показывает, что качество усредненных фаз понижается с увеличением размеров ячейки воображаемого кристалла.

2.5. Генерация случайных наборов фаз. Критерий конечности области высоких значений электронной плотности. Ограничение числа связных компонент

Численные эксперименты были организованы следующим образом. Каждый эксперимент проводился при заданном объеме V_{HDR} области высоких значений на синтезе Фурье электронной плотности. Генерировалось большое количество (до сотен миллионов) случайных наборов фаз. Каждой сгенерированный набор фаз использовался вместе с экспериментальным набором модулей для расчета синтеза Фурье. На синтезе Фурье выделялась область высоких значений заданного объема, и проверялось, помещается ли эта область целиком в элементарной ячейке без пересечения границы. Если это условие выполнялось, сгенерированный набор фаз считался допустимым и запоминался для дальнейшего анализа. Генерация продолжалась до тех пор, пока не отбиралось 100 допустимых наборов фаз. Отобранные 100 допустимых наборов фаз выравнивались и усреднялись. Точность значений фаз, полученных таким образом, отражена в табл. 5 и 6. Для каждого заданного значения объема V_{HDR} эксперимент был повторен 5 раз с разными стартовыми константами датчика случайных чисел. Следует отметить, что при больших величинах объема V_{HDR} процедура расчета начинала требовать значительного времени (суток работы процессора).

Анализ таблиц 3–4 и 5–6 позволяет сделать ряд выводов. Во-первых, изложенная в данном разделе процедура позволяет получать более высокое качество усредненных

фаз, нежели усреднение случайно сгенерированных наборов фаз без предварительного отбора. Тем самым, данная процедура может рассматриваться как работающая процедура *ab initio* определения фаз. Во-вторых, увеличение объема ячейки при сохранении величины физически осмысленного удельного объема области высоких значений приводит к заметному ухудшению качества получаемых наборов фаз. Т.е. рассмотренная процедура представляется более пригодной к работе с реальными кристаллами, нежели для использования с воображаемыми кристаллами при исследовании изолированных частиц.

Попытка наложения дополнительных ограничений на число компонент связности в области высоких значений при отборе сгенерированных наборов фаз привело лишь к незначительному повышению качества получаемых наборов усредненных фаз.

Таблица 5. Результаты усреднения случайных наборов фаз, приводящих к конечной области высоких значений (ОВЗ) на синтезе Фурье. Для пяти независимых экспериментов приведены значения фазовой корреляции CP_w, рассчитанной по разным зонам разрешения (40, 30, 25 Å). Результаты приведены для малой ячейки

	Объем ОВЗ: удельный [Å^3 на остаток]/относительный/число точек						
	50 0.072 334	75 0.109 501	100 0.145 668	125 0.181 834	150 0.218 1002	175 0.254 1168	200 0.290 1336
CP_w*100 40/30/25 Å	69/65/63	78/72/70	76/71/69	81/75/73	84/78/75	87/81/78	
	74/69/67	76/72/69	81/75/73	82/77/74	82/77/74	86/79/76	
	75/70/67	76/71/68	79/74/71	80/75/72	86/80/77	86/80/77	
	74/69/66	78/73/70	80/75/72	83/78/75	86/80/77	84/78/75	
	75/70/68	75/68/66	81/76/75	83/78/75	85/79/75	86/80/78	

Таблица 6. Результаты усреднения случайных наборов фаз, приводящих к конечной области высоких значений (ОВЗ) на синтезе Фурье. Для пяти независимых экспериментов приведены значения фазовой корреляции CP_w, рассчитанной по разным зонам разрешения (40, 30, 25 Å). Результаты приведены для расширенной ячейки

	Объем ОВЗ: удельный [Å^3 на остаток]/относительный/число точек						
	75 0.0322 436	100 0.0429 582	125 0.0636 727	150 0.0644 873	175 0.0751 1018	200 0.0859 1164	225 0.0966 1309
CP_w*100 40/30/25Å	53/51/49	55/53/51	50/48/46	62/59/57	65/62/60	73/69/67	75/71/69
	57/54/53	55/52/50	54/51/50	64/61/60	66/63/61	72/68/66	74/70/68
	53/51/49	65/62/60	66/63/61	67/64/62	67/64/62	72/69/66	73/70/67
	56/53/52	56/54/52	63/60/58	65/62/60	67/64/62	70/67/65	73/69/67
	49/47/46	51/49/48	61/58/56	61/58/56	63/59/58	69/66/64	73/69/67

2.6. Генерация случайных связных масок. Усреднение без отбора

В этой серии экспериментов изучалось, насколько полезными, с точки зрения решения фазовой проблемы, могут служить такие ограничения как связность и бинарность области высоких значений. Эксперимент ставился следующим образом. В элементарной ячейке была введена равномерная сетка с шагом, равным примерно трети номинального разрешения 25 Å (т.е. примерно 8.3 Å). Каждый эксперимент проводился при заданном объеме V_{HDR} области высоких значений на синтезе Фурье электронной плотности, которому соответствовало число N_{mask} узлов сетки, попадающих в маску (табл. 7, 8). Генерировались 100 случайных связных масок, состоящих из N_{mask} узлов. Каждая сгенерированная маска использовалась для расчета модулей и фаз структурных факторов. Полученные 100 наборов фаз выравнивались и усреднялись. Точность значений фаз, полученных таким образом, отражена в табл. 7, 8. Для каждого заданного

значения объема V_{HDR} эксперимент был повторен 5 раз с разными стартовыми константами датчика случайных чисел.

Анализ таблиц позволяет сделать ряд выводов. Во-первых, изложенная в данном разделе процедура позволяет получать более высокое качество усредненных фаз, нежели усреднение случайно сгенерированных наборов фаз без предварительного отбора. Данная процедура может рассматриваться как работающая процедура *ab initio* определения фаз. Во-вторых, сравнение таблиц 5 и 7 показывает, что при работе с малой ячейкой качество получаемых в результате усреднения наборов фаз примерно одинаково в обоих случаях – как при случайной генерации фаз и отборе вариантов по требованию конечности ОВЗ, так и в случае генерации связных масок и прямого усреднения рассчитанных по маскам фаз. В то же время вычислительные затраты существенно меньше при втором подходе. В-третьих, переход к расширенной ячейке приводит к существенному повышению качества получаемого решения фазовой проблемы. Таким образом, новый подход реализует преимущества работы с изолированной частицей, получаемые за счет возможности использования большего количества экспериментальной информации (большего количества модулей структурных факторов, включенных в расчет при работе с большой ячейкой воображаемого кристалла).

Таблица 7. Результаты усреднения наборов фаз, рассчитанных по случайно сгенерированным связным маскам. Для пяти независимых экспериментов приведены значения фазовой корреляции CP, рассчитанной по разным зонам разрешения (40, 30, 25 Å). Результаты приведены для малой ячейки

	Объем маски: удельный [Å^3 на остаток]/относительный/число точек						
	50 0.072 334	75 0.109 501	100 0.145 668	125 0.181 834	150 0.218 1002	175 0.254 1168	200 0.290 1336
CP_w*100 40/30/25 Å	61/57/55	70/65/63	77/73/70	77/72/69	85/80/77	85/80/77	85/80/77
	64/60/57	70/65/63	75/71/68	81/77/74	83/78/75	85/79/77	85/80/77
	66/61/58	72/67/65	78/74/71	80/75/72	84/80/77	87/82/79	84/79/76
	63/59/57	72/68/66	78/73/71	80/76/73	82/77/74	86/81/78	87/81/78
	62/57/55	71/65/63	75/70/68	81/76/73	87/82/79	83/78/75	86/80/77

Таблица 8. Результаты усреднения наборов фаз, рассчитанных по случайно сгенерированным связным маскам. Для пяти независимых экспериментов приведены значения фазовой корреляции CP, рассчитанной по разным зонам разрешения (40, 30, 25 Å). Результаты приведены для расширенной ячейки

	Объем маски: удельный [Å^3 на остаток]/относительный/число точек						
	75 0.0322 436	100 0.0429 582	125 0.0636 727	150 0.0644 873	175 0.0751 1018	200 0.0859 1164	225 0.0966 1309
CP_w*100 40/30/25 Å	81/77/75	86/82/79	86/82/79	86/82/80	89/86/83	87/83/80	85/81/79
	82/78/75	84/80/78	87/83/81	88/84/81	88/84/81	88/84/81	86/82/79
	81/77/75	85/81/78	86/82/79	85/82/80	88/83/81	87/83/80	85/81/79
	83/79/77	84/80/78	86/82/80	88/84/81	89/85/83	86/82/80	88/83/81
	81/77/75	84/80/78	88/84/82	86/83/80	88/84/81	87/83/81	87/83/80

2.7. Генерация случайных связных масок. Отбор по корреляции модулей структурных факторов

В этой серии экспериментов в работу было включено дополнительное требование к генерируемым случайно маскам ОВЗ: соответствие рассчитанным по маскам модулей структурных факторов информации, полученной в эксперименте. Эксперимент ставился следующим образом. В элементарной ячейке была введена равномерная сетка

с шагом, равным примерно трети номинального разрешения 25 Å (т.е. примерно 8.3 Å). Каждый эксперимент проводился при заданных двух параметрах: объеме V_{HDR} области высоких значений на синтезе Фурье электронной плотности и требуемой точности соответствия эксперименту модулей структурных факторов, рассчитанных по маске. Генерировалось большое число случайных связных масок, состоящих из N_{mask} узлов. Каждая сгенерированная маска использовалась для расчета модулей и фаз структурных факторов. Для рассчитанных значений модулей структурных факторов вычислялся коэффициент их корреляции с экспериментальными значениями (15). Если этот коэффициент оказывался выше заданного порогового значения, то сгенерированный набор фаз считался допустимым и запоминался для дальнейшего анализа. Генерация продолжалась до тех пор, пока не отбиралось 100 допустимых наборов фаз. Отобранные наборы фаз выравнивались и усреднялись.

Таблица 9. Результаты усреднения наборов фаз, рассчитанных по обладающим заданным уровнем корреляции (25) модулей структурных факторов случайно сгенерированным связным маскам. Приведены значения фазовой корреляции CP_w, рассчитанной по разным зонам разрешения (40, 30, 25 Å). Результаты приведены для малой ячейки. Выделен наилучший результат

CP_w*100 40/30/25 Å		Объем маски: удельный [Å^3 на остаток]/относительный/число точек						
		50 0.072 334	75 0.109 501	100 0.145 668	125 0.181 834	150 0.218 1002	175 0.254 1168	200 0.290 1336
корреляция модулей z_{crit} (25 Å)	1.5	60/57/54	69/64/62	72/67/64	80/75/72	87/82/79	88/83/80	87/82/79
	2.0	61/57/55	68/63/61	73/68/66	83/78/75	88/83/80	**90/84/81**	89/83/81
	2.5	59/55/53	65/60/57	73/68/66	81/76/73	88/83/80	89/83/80	89/83/79
	3.0	59/55/53	62/58/56	71/66/64	81/76/73	88/83/79	83/78/74	

Таблица 10. Результаты усреднения наборов фаз, рассчитанных по обладающим заданным уровнем корреляции (25) модулей структурных факторов случайно сгенерированным связным маскам. Приведены значения фазовой корреляции CP_w, рассчитанной по разным зонам разрешения (40, 30, 25 Å). Результаты приведены для расширенной ячейки. Выделены наилучшие результаты

CP*100% 40/30/25Å		Объем маски: удельный [Å^3 на остаток]/относительный/число точек						
		50 0.0215 291	75 0.0322 436	100 0.0429 582	125 0.0636 727	150 0.0644 873	175 0.0751 1018	200 0.0859 1164
корреляция модулей z_{crit} (25 Å)	1.5	72/69/67	78/74/71	86/82/79	89/85/83	91/86/84	91/87/85	90/86/83
	2.0	72/68/66	77/73/71	85/81/79	90/86/84	91/87/85	92/88/85	91/87/84
	2.5	71/68/66	77/73/70	88/84/82	91/87/84	92/88/85	92/88/86	90/86/83
	3.0	71/69/67	81/77/74	89/85/83	92/88/85	**94/90/87**	**94/90/87**	93/88/86

Точность значений фаз, полученных таким образом, отражена в табл. 9 и 10. В указанных таблицах жесткость отбора допустимых масок приведена в виде нормированной величины z_{CM}, которая определялась следующим образом. В качестве предварительного этапа работы выполнялась генерация большого числа (10 000 генераций в наших экспериментах) случайных связных масок заданного размера, для каждой из которых вычислялся коэффициент корреляции между рассчитанными и точными модулями структурных факторов. Для полученных 10 000 значений величины CM вычислялось среднее значение $\langle CM \rangle$ и среднеквадратичное отклонение от среднего σ_{CM}. Далее, при генерации масок с отбором, для текущего набора рассчитанных структурных факторов вычисленная корреляция CM преобразовывалась в нормализованную величину

$$z_{CM} = \frac{CM - \langle CM \rangle}{\sigma_{CM}}, \qquad (25)$$

которая и сравнивалась с заданным пороговым значением z_{crit}.

Таблицы показывают, что введение отбора по степени соответствия рассчитанных модулей структурных факторов эксперименту существенно улучшает качество полученных фаз по сравнению с усреднением без отбора. На рис. 5 показаны изображения объекта, полученные на основе синтеза Фурье, рассчитанного с экспериментальными значениями модулей и наилучшими найденными фазами.

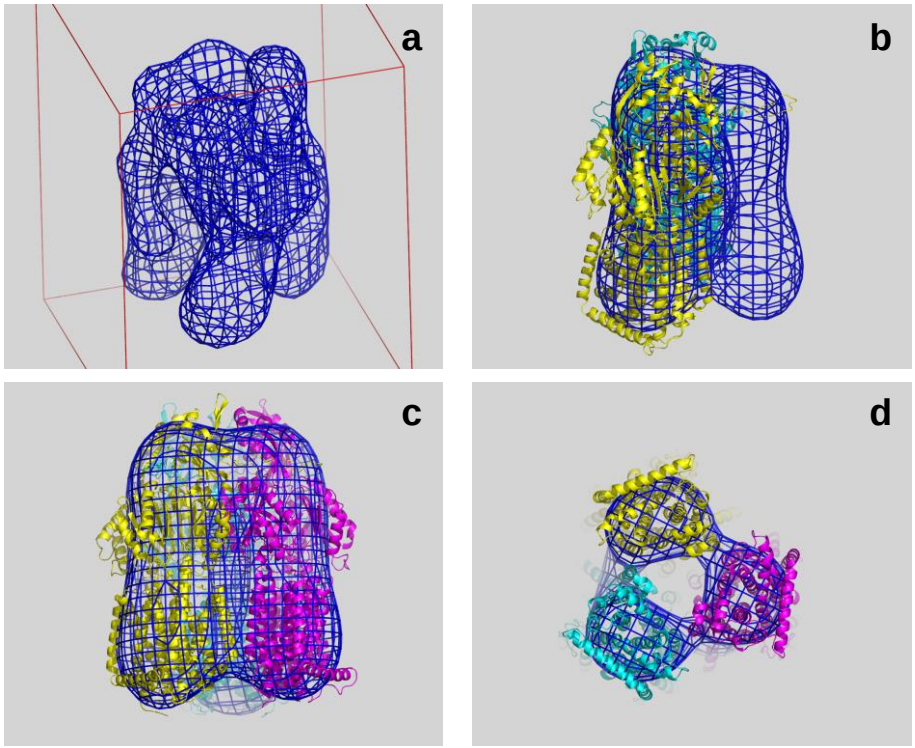

Рис. 5. Модель белка AcrB и области высоких значений электронной плотности на невзвешенных синтезах Фурье, рассчитанных с экспериментальными значениями модулей и определенными *ab initio* значениями фаз структурных факторов: a - разрешение 25 Å, срезка 3σ; b - разрешение 40 Å, срезка 3σ, один из мономеров модели не показан; c. - разрешение 40 Å, срезка 2σ; d - разрешение 40 Å, срезка 3σ.

ДОПОЛНИТЕЛЬНЫЕ ЗАМЕЧАНИЯ И ЗАКЛЮЧЕНИЕ

Проведенное тестирование нового подхода к *ab initio* решению фазовой проблемы биологической кристаллографии показало его применимость как при традиционном исследовании кристаллических образцов, так и при работе с изолированными биологическими макромолекулярными объектами.

Основой подхода является Монте-Карловского типа процедура [3, 4], состоящая из нескольких стадий. На первой стадии объектом поиска являются бинарные маски, служащие аппроксимацией области высоких значений функции распределения электронной плотности в исследуемом объекте. Для этой цели случайным образом генерируется большое число связных бинарных масок, и для каждой из них рассчитываются модули и фазы структурных факторов. Маска (и соответствующие ей фазы структурных факторов) считаются допустимыми, если степень совпадения рассчитанных по маске модулей структурных факторов и их экспериментальных значений превышает заданный предел. Генерация ведется до получения заданного количества допустимых наборов фаз. В качестве критерия близости модулей

структурных факторов в тестах использовался коэффициент их корреляции, хотя могут быть также использованы и другие критерии (например, статистическое правдоподобие [16]). Помимо критерия близости модулей структурных факторов, на этапе отбора масок могут применяться и другого типа критерии, связанные с ожидаемыми параметрами масок или основанные на результатах других экспериментов (например, малоуглового рентгеновского или нейтронного рассеяния). Генерация случайных масок может проводиться как в предположении равновероятности всех узлов сетки в элементарной ячейке, так и с использованием априорных предпочтений, установленных на предыдущих этапах работы. При работе с большими размерами элементарной ячейки и жестких ограничениях на степень соответствия маски экспериментальным данным генерация нужного количества допустимых масок может выливаться в существенные времена работы компьютера. Однако естественная параллельность процесса отбора позволяет кардинально уменьшить временные затраты при работе на компьютерах с параллельной архитектурой.

На втором этапе осуществляется выравнивание отобранных наборов фаз и, в простейшем случае, их усреднения. Более аккуратная процедура состоит в использовании методов кластерного анализа, позволяющих выделить компактные кластеры среди отобранных вариантов для последующего усреднения внутри этих кластеров. Следует отметить, что процессы выравнивания и кластерного анализа проводятся на основе специальной метрики, связанной с конкретным рентгеновским экспериментом. Полученные в результате усреднения значения фаз структурных факторов вместе с экспериментально определенными модулями используются для построения синтезов Фурье.

Тестирование метода было проведено в двух режимах. Малая элементарная ячейка моделировала кристаллический образец с большим содержанием растворителя в ячейке. Расширенная элементарная ячейка моделировала редуцированную задачу определения структуры по рассеянию изолированной частицей.

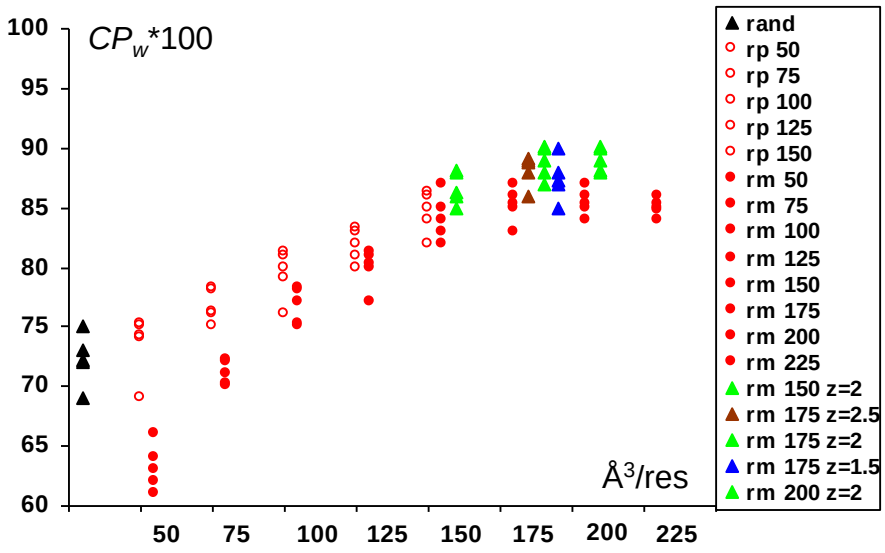

Рис. 6. Фазовая корреляция CP_w*100 по зоне разрешения 40 Å для усредненных наборов фаз при разном размере области высоких значений: rand – случайные фазы без отбора (раздел 2.4); rp – случайные фазы с отбором по свойству конечности ОВЭ (раздел 2.5); rm – случайные маски без отбора (раздел 2.6); rmz – случайные маски с отбором по корреляции модулей при разных уровнях отбора (раздел 2.7). По оси абсцисс показан удельный объем областей высоких значений. При каждых условиях показаны результаты пяти независимых экспериментов для малой ячейки.

Результаты тестирования позволили найти оптимальные параметры использования процедуры и показали, что предложенная процедура действительно позволяет

продвинуться в задаче определения фаз структурных факторов. При этом точность найденных значений фаз возрастает с ростом размера ячейки, что делает эту методику особенно перспективной при исследовании изолированных частиц.

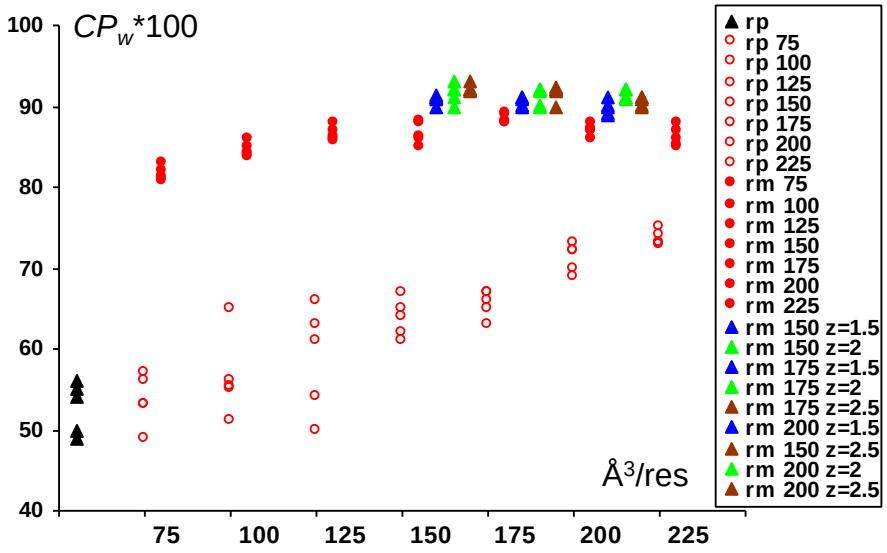

Рис. 7. Фазовая корреляция CP_w*100 по зоне разрешения 40 Å для усредненных наборов фаз при разном размере области высоких значений: rand – случайные фазы без отбора (раздел 2.4); rp – случайные фазы с отбором по свойству конечности ОВЗ (раздел 2.5); rm – случайные маски без отбора (раздел 2.6); rmz – случайные маски с отбором по корреляции модулей при разных уровнях отбора (раздел 2.7). По оси абсцисс показан удельный объем областей высоких значений. При каждых условиях показаны результаты пяти независимых экспериментов для расширенной ячейки.

Сравнение новой методики с ранее предложенной [5] в случае традиционной кристаллографической задачи показало сопоставимую эффективность решения фазовой проблемы при более высокой вычислительной эффективности нового подхода. В то же время, новая методика приводит к существенно лучшему решению фазовой проблемы при работе с ячейками, содержащими искусственно увеличенную долю растворителя.

Рис. 6 и 7 иллюстрируют точность определения фаз структурных факторов при использовании различных подходов и параметров метода.

Работа выполнена при поддержке гранта РФФИ № 13-04-00118.

СПИСОК ЛИТЕРАТУРЫ

1. Крупянский Ю.Ф., Балабаев Н.К., Петрова Т.Е., Синицын Д.О., Грызлова Е.В., Терешкина К.Б., Абдулнасыров Э.Г., Степанов А.С., Лунин В.Ю., Грум-Гржимайло А.Н. Фемтосекундные рентгеновские лазеры на свободных электронах: новый метод изучения нанокристаллов и одиночных макромолекул. *Химическая физика*. 2014. Т. 33. № 7. С. 7–20.
2. Синицын Д.О., Лунин В.Ю., Грум-Гржимайло А.Н., Грызлова Е.В., Балабаев Н.К., Лунина Н.Л., Петрова Т.Е., Терешкина К.Б., Абдулнасыров Э.Г., Степанов А.С., Крупянский Ю.Ф. Новые возможности рентгеновской нанокристаллографии биологических макромолекул с использованием рентгеновских лазеров на свободных электронах. *Химическая физика*. 2014. Т. 33. № 7. С. 21–28.
3. Lunin V.Y., Lunina N.L., Petrova T.E., Skovoroda T.P., Urzhumtsev A.G., Podjarny A.D. Low-resolution *ab initio* phasing: problems and advances. *Acta Crystallographica Section D: Biological Crystallography*. 2000. V. 56. P. 1223–1232.

4. Lunin V.Y., Urzhumtsev A.G., Podjarny A. *Ab initio* phasing of low-resolution Fourier syntheses. In: *International Tables for Crystallography.* Volume F. Crystallography of Biological Macromolecules. Second Edition. Ed. Arnold E., Himmel D.M., Rossmann M.G. John Wiley & Sons, 2011. P. 437–442.
5. Lunin V.Y., Lunina N.L., Urzhumtsev A.G. Connectivity properties of high-density regions and *ab initio* phasing at low resolution. *Acta Crystallographica Section D: Biological Crystallography*. 2000. V. 56. P. 375–382.
6. Bricogne G. Methods and programs for direct-space exploitation of geometric redundancies. *Acta Crystallographica Section A: Foundations of Crystallography*. 1976. V. 32. P. 832–847.
7. Fienup J.R. Reconstruction of an object from the modulus of its Fourier transform. *Optics Letters*. 1978. V. 3. № 1. P. 27–29.
8. Elser V. Solution of the crystallographic phase problem by iterated projections. *Acta Crystallographica Section A: Foundations of Crystallography*. 2003. V. 59. P. 201–209.
9. Baker D., Krukowski A.E., Agard D.A. Uniqueness and the *ab initio* phase problem in macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography*. 1993. V. 49. P. 186–192.
10. Lunin V.Y., Urzhumtsev A.G., Skovoroda T.P. Direct low-resolution phasing from electron-density histograms in protein crystallography. *Acta Crystallographica Section A: Foundations of Crystallography*. 1990. V. 46. P. 540–544.
11. Lunin V.Y., Lunina N.L. The map correlation coefficient for optimally superposed maps. *Acta Crystallographica Section A: Foundations of Crystallography*. 1996. V. 52. P. 365–368.
12. Lunin V.Y., Urzhumtsev A., Bockmayr A. Direct phasing by binary integer programming. *Acta Crystallographica Section A: Foundations of Crystallography*. 2002. V. 58. P. 283–291.
13. Murakami S., Nakashima R., Yamashita E., Yamaguchi A. Crystal structure of bacterial multidrug efflux transporter AcrB. *Nature*. 2002. V. 419. P. 587–593.
14. Urzhumtsev A., Afonine P.V., Lunin V.Y., Terwilliger T.C., Adams P.D. Metrics for comparison of crystallographic maps. *Acta Crystallographica Section D: Biological Crystallography*. 2014. V. 70. P. 2593–2606.
15. Lunin V.Y., Woolfson M.M. Mean phase error and the map correlation coefficient. *Acta Crystallographica Section D: Biological Crystallography*. 1993. V. 49. P. 530–533.
16. Lunin V.Y., Afonine P.V., Urzhumtsev A.G. Likelihood-based refinement. I. Irremovable model errors. *Acta Crystallographica Section A: Foundations of Crystallography*. 2002. V. 58. P. 270–282.

Материал поступил в редакцию 08.12.2014, опубликован 19.12.2014.