Математическая биология и биоинформатика
2017. Т. 12. № 2. С. 521–535. doi: 10.17537/2017.12.521

========== **МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ** ==========

УДК: 577.3

Использование связных масок в задаче восстановления изображения изолированной частицы по данным рентгеновского рассеяния. III. Стратегии отбора решений по результатам максимизации правдоподобия

Лунина Н.Л.[1], Петрова Т.Е.[1], Уржумцев А.Г.[*2,3], Лунин В.Ю.[1]**

[1]*Институт математических проблем биологии РАН – филиал Института прикладной математики им. М.В.Келдыша РАН, Пущино, Московская область, 142290, Россия*
[2]*Институт генетики и молекулярной и клеточной биологии, 67404 Илькирш, Франция*
[3]*Лорренский университет, Вандевр-ле-Нанси, 54506, Франция*

Аннотация. Основное экспериментальное ограничение биологической кристаллографии связано с необходимостью приготовления исследуемого объекта в виде монокристалла. Ввод в эксплуатацию новых мощных источников рентгеновского излучения – рентгеновских лазеров на свободных электронах – позволяет ставить вопрос о практическом определении структуры изолированных биологических макромолекул и их комплексов. Дополнительным преимуществом работы с изолированными частицами является возможность получения информации о рассеянии во всех направлениях, а не только направлениях, ограниченных условиями дифракции Лауэ–Брэгга. Это существенно облегчает решение фазовой проблемы рентгеноструктурного анализа – определения значений фаз структурных факторов, недоступных для измерения в эксперименте. Данная работа посвящена двум направлениям развития предложенного ранее авторами метода решения фазовой проблемы, основанного на случайном сканировании конфигурационного пространства потенциальных решений фазовой проблемы. В работе предложен новый тип критерия отбора в процессе сканирования кандидатов на решение фазовой проблемы, включающий максимизацию статистического правдоподобия, и показана (в тестовых расчетах) его эффективность. Второе направление связано с выбором оптимальной стратегии сканирования. Показано, что в данном подходе постепенное расширение используемого в работе набора экспериментальных данных позволяет получать решения более высокого качества, нежели при одновременном включении в работу всех имеющихся данных. Такое расширение может осуществляться в неявном виде использованием в работе синтезов Фурье электронной плотности, взвешенных показателями достоверности имеющихся значений фаз.

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

---

[*]sacha@igbmc.fr
[**]lunin@impb.ru

1. РЕШЕНИЕ ФАЗОВОЙ ПРОБЛЕМЫ ДЛЯ ИЗОЛИРОВАННЫХ ЧАСТИЦ СКАНИРОВАНИЕМ КОНФИГУРАЦИОННОГО ПРОСТРАНСТВА ФАЗОВЫХ НАБОРОВ

Ввод в эксплуатацию источников рентгеновского излучения нового поколения – рентгеновских лазеров на свободных электронах (XFEL) – позволяет ставить новые задачи в исследовании структуры биологических объектов [1–7]. Одной из таких задач является исследование методами рентгеновской дифракции изолированных биологических частиц (Single Particles) [8–16]. Возможность исследования изолированных частиц позволяет снять основное экспериментальное ограничение рентгеновской кристаллографии – необходимость приготовления исследуемого образца в виде монокристалла. Второй важной особенностью работы с изолированными частицами является потенциальная возможность измерения интенсивности рассеянных волн для всевозможных направлений рассеяния. Напомним, что при работе с кристаллом возможность измерения ограничена дискретным набором направлений рассеяния, определяемых кристаллической решеткой.

Как и в традиционной рентгеновской кристаллографии, рентгеновский эксперимент с изолированной частицей позволяет получить лишь значения модуля $\{F(\mathbf{s})\}_{\mathbf{s}\in R^3}$ комплексного преобразования Фурье функции $\rho(\mathbf{r})$, описывающей распределение в трехмерном пространстве электронов исследуемого объекта.

$$F(\mathbf{s})\exp[i\varphi(\mathbf{s})] = \int_{R^3} \rho(\mathbf{r})\exp[2\pi i(\mathbf{s},\mathbf{r})]dV_{\mathbf{r}} . \tag{1}$$

Значения же фаз $\{\varphi(\mathbf{s})\}_{\mathbf{s}\in R^3}$ в процессе измерения теряются, что препятствует непосредственному восстановлению распределения электронной плотности посредством расчета обратного преобразования Фурье:

$$\rho(\mathbf{r}) = \int_{R^3} F(\mathbf{s})\exp[i\varphi(\mathbf{s})]\exp[-2\pi i(\mathbf{s},\mathbf{r})]dV_{\mathbf{s}} , \tag{2}$$

Это приводит к центральной проблеме дифракционных методов – фазовой проблеме, т.е. к проблеме восстановления фазовой части преобразования Фурье. Аналогичная проблема (восстановление функции по модулю преобразования Фурье) встречается также в задачах оптики [17] под именем Phase Retrieval Problem. Развитие методов решения этой проблемы до недавнего времени шло в значительной мере независимо в кристаллографии и оптике, и сходные разработанные подходы фигурируют в этих областях под разными названиями. В данной работе мы будем придерживаться кристаллографической терминологии.

Схема рентгеновского эксперимента приведена на рисунке 1. В теоретическом рассмотрении изменяемыми параметрами могут считаться направление рентгеновского пучка и направление от объекта на детектор (задаваемые векторами единичной длины $\boldsymbol{\sigma}_0, \boldsymbol{\sigma}$). На практике же направление исходного пучка $\boldsymbol{\sigma}_0$ фиксировано, а дополнительное варьирование условий осуществляется вращением исследуемого объекта относительно исходного пучка. В рамках кинематической теории рассеяния интенсивность рассеянной волны оказывается пропорциональной квадрату модуля преобразования Фурье (2), вычисленного в точке

$$\mathbf{s} = \frac{\boldsymbol{\sigma} - \boldsymbol{\sigma}_0}{\lambda} , \tag{3}$$

именуемой вектором рассеяния и определяемой как комбинация векторов $\boldsymbol{\sigma}_0, \boldsymbol{\sigma}$ и длины волны рентгеновского излучения λ. Следует заметить, что даже в идеальных условиях

Математическая биология и биоинформатика. 2017. Т. 12. № 2. doi: 10.17537/2017.12.521

эксперимент не позволяет измерить значения функции $\{F(\mathbf{s})\}_{\mathbf{s}\in R^3}$ вне ограниченной зоны значений $|\mathbf{s}| \le 2/\lambda$. На практике эта зона значений еще уже, $|\mathbf{s}| \le s_{\max} < 2/\lambda$, и ограничена особенностями конкретного объекта и экспериментальными условиями, в частности, характеристиками исходного пучка и детектора. Величина $d_{\min} = 1/s_{\max}$ называется разрешением собранного набора данных. Приближенное распределение электронной плотности, рассчитанное по формуле (2) с использованием ограниченного набора данных, называется синтезом Фурье электронной плотности разрешения $d_{\min}$. По сравнению с идеальным распределением $\rho(\mathbf{r})$, оно может содержать искажения, такие как волны обрыва ряда и расплывание пиков плотности. Разрешение синтеза Фурье определяет размер минимальных деталей, различимых визуально при анализе синтеза [18].

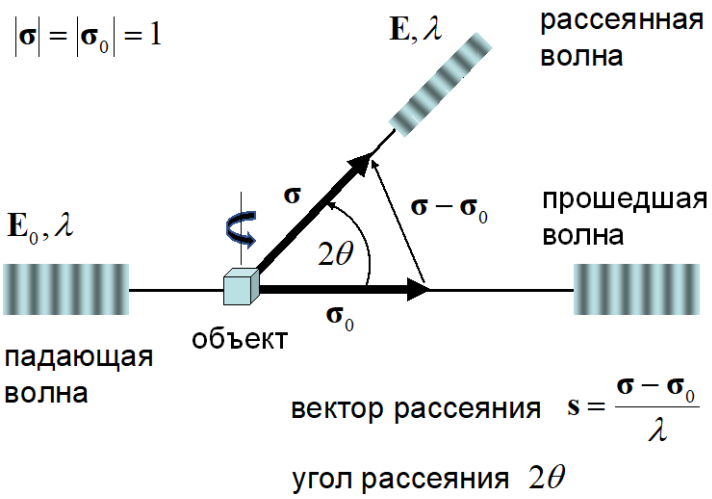

Рис. 1. Схема рентгеновского дифракционного эксперимента.

Дискретизация экспериментальных данных на равномерную сетку с шагом a^{-1} (где величина a превосходит линейные размеры исследуемой частицы) позволяет рассматривать их как набор модулей коэффициентов Фурье (именуемых в кристаллографии структурными факторами) для воображаемого кристалла с кубической элементарной ячейкой с ребром a. Это, в свою очередь, позволяет применять для решения фазовой проблемы весь инструментарий биологической кристаллографии [19–20]. Более подробно, пусть V обозначает элементарную ячейку (куб со стороной a), выбранную так, что исследуемая частица находится внутри нее, а $|V|$ обозначает объем этой ячейки. Пусть $\rho(\mathbf{r})$ – функция, описывающая распределение электронной плотности внутри изучаемой частицы и равная нулю вне частицы. Обозначим через $\rho^{\text{cryst}}(\mathbf{r})$ периодическую функцию, обладающую периодом a по осям элементарной ячейки, и совпадающую с $\rho(\mathbf{r})$ внутри V. В таком случае функция $\rho^{\text{cryst}}(\mathbf{r})$ может быть рассчитана как сумма ряда Фурье

$$\rho^{\text{cryst}}(\mathbf{r}) = \frac{1}{|V|}\sum_{\mathbf{s}\in\Re'_a} F(\mathbf{s})\exp\left[i\varphi(\mathbf{s})\right]\exp\left[-2\pi i(\mathbf{s},\mathbf{r})\right], \tag{4}$$

где

$$F(\mathbf{s})\exp\left[i\varphi(\mathbf{s})\right] = \int_V \rho(\mathbf{r})\exp\left[2\pi i(\mathbf{s},\mathbf{r})\right]dV_{\mathbf{r}} \tag{5}$$

– структурные факторы, а $\Re'_a$ – решетка с шагом a^{-1} в пространстве векторов рассеяния, называемом также обратным пространством. Узлы этой решетки называются далее рефлексами. Отличие от случая с обычным кристаллом состоит в том, что величину a мы можем выбрать произвольно и, выбирая ее достаточно большой, мы получаем функцию $\rho^{\text{cryst}}(\mathbf{r})$, равную нулю в большей части элементарной ячейки. Это создает значительную избыточность экспериментальных данных и является фундаментом для решения фазовой проблемы [19, 21–26].

Основным используемым подходом к практическому использованию избыточности данных, создаваемой наличием в элементарной ячейке большой области нулевых значений электронной плотности, является широкий класс итерационных методов [27–33]. При удачном выборе параметров метода и стартовой точки эти методы могут приводить к достаточно точным решениям. Однако эти методы работают локально, и их устойчивость и сходимость к правильному решению могут требовать не всегда очевидного выбора параметров. Дополнением к локальным процедурам уточнения достаточно хороших начальных приближений могут служить глобальные процедуры сканирования конфигурационного пространства возможных решений, помогающие в выборе начального приближения. В данной работе мы исследуем одну из таких процедур – процедуру случайного сканирования конфигурационного пространства, предложенную первоначально в работе [34] и развитую далее в серии работ [35–37,20]. Схема процедуры показана на рисунке 2.

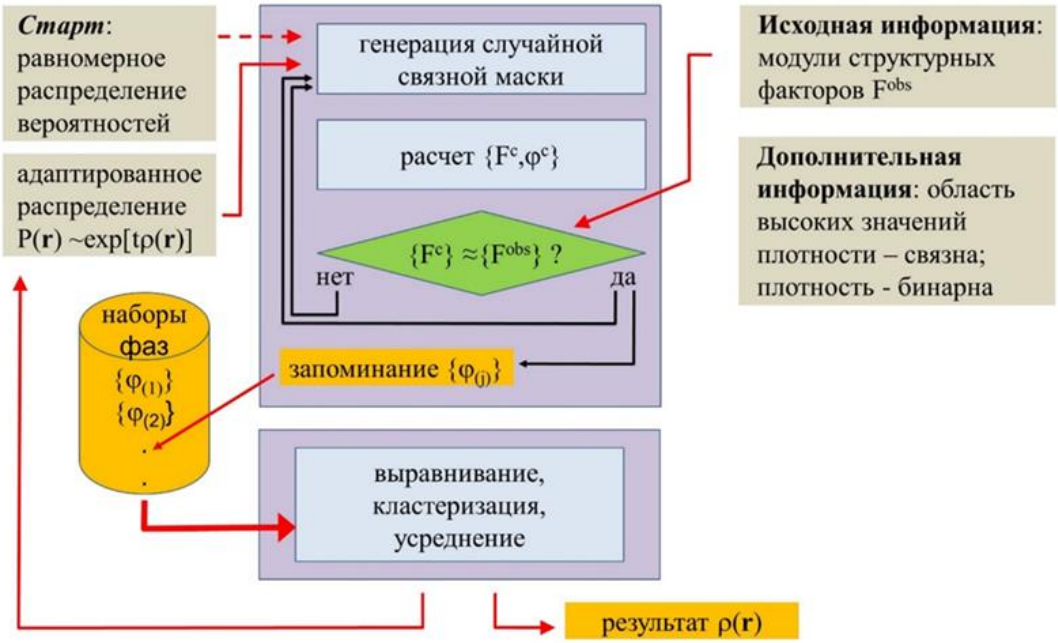

Рис.2. Схема процедуры случайного сканирования конфигурационного пространства фазовых наборов.

Основной цикл процедуры состоит в генерации большого числа случайных фазовых наборов, отбора тех из них, которые приводят к приемлемым распределениям электронной плотности, выравнивании и усреднении отобранных вариантов (см. раздел 2 ниже). При этом выбор критерия приемлемости является центральным вопросом при реализации подхода. Генерация случайных вариантов может проходить как без использования дополнительной информации о значениях фаз (на первом цикле использования процедуры), так и с использованием информации о вероятных значениях фаз, полученной на предыдущих циклах. При генерации случайных наборов

фаз значения фаз могут генерироваться как независимо, так и рассчитываться с использованием моделей, отражающих ожидаемые свойства реальных распределений электронной плотности в биологических объектах. В этом случае случайно генерируются некоторые предварительные модели, с которых далее рассчитываются фазы структурных факторов. В данной работе мы рассматриваем процедуру, в которой генерируемыми объектами являются связные маски области, занимаемой объектом [20,37]. Эти маски представляют собой бинарные функции $M(\mathbf{r})$ (со значениями 0 или 1) на сетке в элементарной ячейке, обладающие свойством связности области маски (носителя функции), т.е. множества точек, в которых значение $M(\mathbf{r})$ равно единице. Такие маски отражают два важных свойства реальных распределений электронной плотности в биологических макромолекулах. Во-первых, наличие протяженных полипептидных цепей или цепей ДНК приводит к связности области высоких значений электронной плотности или, по крайней мере, к наличию небольшого числа связных компонент [38]. Во-вторых, в основе интерпретации распределений электронной плотности лежит бинарность, когда точки пространства разделяются на принадлежащие объекту и не принадлежащие. Ранее было показано, что оба этих свойства позволяют осуществлять отбор приемлемых решений и по отдельности, и вместе. Особенно перспективным оказалось использование этих свойств при генерации моделей.

Маска $M(\mathbf{r})$ может быть использована для расчета наборов величин модулей $\left\{F_{\mathbf{s}}^{\text{mask}}\right\}$ и фаз $\left\{\varphi_{\mathbf{s}}^{\text{mask}}\right\}$ соответствующих ей структурных факторов согласно (5). Степень соответствия рассчитанных и экспериментальных $\left\{F_{\mathbf{s}}^{\text{obs}}\right\}$ значений модулей структурных факторов может служить индикатором качества маски и, соответственно, индикатором точности соответствующего маске набора фаз в условиях реального исследования структуры, когда точные значения фаз неизвестны. В предыдущих работах [20, 37] в качестве критерия соответствия двух наборов модулей использовался коэффициент корреляции этих значений

$$CM = \frac{\sum_{\mathbf{s}} F_{\mathbf{s}}^{\text{mask}} F_{\mathbf{s}}^{\text{obs}}}{\sqrt{\sum_{\mathbf{s}} \left(F_{\mathbf{s}}^{\text{mask}}\right)^2 \sum_{\mathbf{s}} \left(F_{\mathbf{s}}^{\text{obs}}\right)^2}}. \tag{6}$$

В данной работе мы исследуем возможность использования нового типа критерия отбора – величины математического ожидания E_{CP} фазовой корреляции при использовании тестируемой маски в качестве априорного распределения координат атомной модели структуры.

2. ИНДИКАТОРЫ КАЧЕСТВА МАСКИ, ОСНОВАННЫЕ НА МАКСИМИЗАЦИИ ПРАВДОПОДОБИЯ

Использование величины статистического правдоподобия в качестве критерия отбора решений обсуждалось ранее, в работе [36], где для расчета правдоподобия использовалась трудоемкая имитационная процедура. В данной работе предлагается существенно менее трудоемкая процедура, основанная на использовании диагональной аппроксимации функции правдоподобия [39–42].

Пусть внутри выбранной маски случайным образом сгенерирована атомная модель. В рамках аппроксимации центральной предельной теоремы теории вероятностей совместное распределение вероятностей модуля и фазы комплексного структурного фактора $\mathbf{F}$, соответствующего вектору рассеяния **s,** может быть представлено в виде

$$P(F,\varphi)=\frac{F}{\pi\beta}\exp\left[-\frac{F^2+\left(\alpha F^{\text{mask}}\right)^2}{\beta}\right]\exp\left[2\frac{\alpha F F^{\text{mask}}}{\beta}\cos\left(\varphi-\varphi^{\text{mask}}\right)\right]. \quad (7)$$

Здесь α и β – параметры, аккумулирующие неопределённости в выборе модели, индивидуальные для каждого рефлекса. Для некоторых простых случаев эти параметры могут быть выписаны в явном виде [41], но в общей ситуации они неизвестны. Как правило, эти параметры мало меняются внутри групп рефлексов, принадлежащих узким слоям пространства вида $\mathbf{S}_j=\{\mathbf{s}: s_j\le s\le s_j+\Delta\}$ с малым значением Δ (оболочек), где их можно считать постоянными. При наличии экспериментальных данных $\left\{F_{\mathbf{s}}^{\text{obs}}\right\}$, значения параметров α_j и β_j оболочки $\mathbf{S}_j$ могут быть определены из принципа максимума правдоподобия посредством максимизации функции

$$L(\alpha,\beta)=\prod_{\mathbf{s}\in\mathbf{S}_j}P_F\left(F_{\mathbf{s}}^{\text{obs}};\alpha,\beta\right)\quad\Rightarrow\quad\max. \quad (8)$$

Здесь $P_F(F;\alpha,\beta)$ – распределение модуля структурного фактора, получаемое из (7) интегрированием по фазе

$$P_F(F;\alpha,\beta)=\frac{2F}{\beta}\exp\left[-\frac{F^2+\left(\alpha F^{\text{mask}}\right)^2}{\beta}\right]I_0\left(2\frac{FF^{\text{mask}}}{\beta}\right), \quad (9)$$

а I_0 – модифицированная функция Бесселя нулевого порядка. При получении оценок предполагается независимость распределений, отвечающих различным точкам $\mathbf{s}$ (т.н. диагональная аппроксимация функции правдоподобия [39–42]).

После того, как оптимальные значения параметров $\hat{\alpha}_j$ и $\hat{\beta}_j$ оценены для всех оболочек, можно вычислить значение правдоподобия (8) с учетом всех рефлексов и использовать его (или, что более удобно на практике, его логарифм) как индикатор качества маски. Однако представляется более удобным использовать для этой цели производную характеристику – математическое ожидание коэффициента фазовой корреляции. Для каждого рефлекса выражение (7) позволяет выписать распределение значений фазы при условии, что модуль принял наблюденное в эксперименте значение, в виде

$$P_\varphi\left(\varphi\middle|F=F^{\text{obs}}\right)=\frac{1}{2\pi I_0\left[2\left(\hat{\alpha}/\hat{\beta}\right)F^{\text{obs}}F^{\text{mask}}\right]}\exp\left[2\frac{\hat{\alpha}F^{\text{obs}}F^{\text{mask}}}{\hat{\beta}}\cos\left(\varphi-\varphi^{\text{mask}}\right)\right]. \quad (10)$$

Последнее распределение позволяет вычислить условное математическое ожидание величины косинуса фазовой разницы

$$c_{\mathbf{s}}=\left\langle\cos\left(\varphi_{\mathbf{s}}-\varphi_{\mathbf{s}}^{\text{mask}}\right)\middle|F_{\mathbf{s}}=F_{\mathbf{s}}^{\text{obs}}\right\rangle=\frac{I_1(T_{\mathbf{s}})}{I_0(T_{\mathbf{s}})}, \quad (11)$$

где

$$T_{\mathbf{s}}=2\frac{\hat{\alpha}_{\mathbf{s}}F_{\mathbf{s}}^{\text{obs}}F_{\mathbf{s}}^{\text{mask}}}{\hat{\beta}_{\mathbf{s}}}. \quad (12)$$

Зная (11), можно вычислить условное математическое ожидание коэффициента корреляции фаз атомной модели, случайно размещаемой в области маски, с фазами

Математическая биология и биоинформатика. 2017. Т. 12. № 2. doi: 10.17537/2017.12.521

маски при условии совпадения рассчитанных по модели модулей с экспериментальными значениями

$$E_{CP} = \frac{\sum_{\mathbf{s}} \left(F_{\mathbf{s}}^{\text{obs}}\right)^2 c_{\mathbf{s}}}{\sum_{\mathbf{s}} \left(F_{\mathbf{s}}^{\text{obs}}\right)^2}. \tag{13}$$

В данной работе мы обсуждаем возможность решения фазовой проблемы с использованием полученного выражения (13) для отбора приемлемых масок при различных протоколах организации сканирования конфигурационного пространства.

Ниже приводятся результаты тестирования нескольких таких протоколов работы. Первая исследуемая проблема связана со стратегией подключения данных к работе. По традиции, в биологической кристаллографии подключение данных к работе идет постепенно, с постепенным расширением набора используемых данных. Альтернативой является одновременное включение в работу всех имеющихся данных.

Второй вопрос связан с построением априорного распределения вероятностей для генерации масок. Следуя работе [20], это распределение строится в виде

$$p(\mathbf{r}) = C \exp\left[\frac{\ln t}{\rho_{\max} - \rho_{\min}} \rho(\mathbf{r})\right], \tag{14}$$

где C – нормализующий множитель, $\rho(\mathbf{r})$ – приближенное распределение электронной плотности, полученное на предыдущем этапе, t – параметр, регулирующий контраст распределения $p(\mathbf{r})$:

$$t = p_{\max} / p_{\min}. \tag{15}$$

В данной работе сравнивались два простейших способа построения такого распределения по результатам отбора приемлемых масок из случайно сгенерированных. Пусть $\left\{\varphi_1^{\text{mask}}(\mathbf{s})\right\}, \left\{\varphi_2^{\text{mask}}(\mathbf{s})\right\}, \ldots, \left\{\varphi_K^{\text{mask}}(\mathbf{s})\right\}$ – наборы фаз структурных факторов, отвечающих маскам, отобранным в процессе генерации и затем оптимально совмещенным посредством сдвигов и смен энантиомеров (в дальнейшем эту операцию мы будем называть оптимальным выравниванием [5]). Определим наилучшие фазы $\varphi^{\text{best}}(\mathbf{s})$ и показатели достоверности $m(\mathbf{s})$ равенством

$$m(\mathbf{s}) \exp\left[i\,\varphi^{\text{best}}(\mathbf{s})\right] = \frac{1}{K} \sum_{k=1}^{K} \exp\left[i\,\varphi_k^{\text{mask}}(\mathbf{s})\right]. \tag{16}$$

Далее в качестве полученного приближения к значениям электронной плотности мы будем рассматривать два типа синтезов Фурье: "невзвешенный"

$$\rho_u(\mathbf{r}) = \frac{1}{V} \sum_{\mathbf{s}} F^{\text{obs}}(\mathbf{s}) \exp\left[i\,\varphi^{\text{best}}(\mathbf{s})\right] \exp\left[-2\pi i(\mathbf{s},\mathbf{r})\right] \tag{17}$$

и "взвешенный" синтез Фурье

$$\rho_w(\mathbf{r}) = \frac{1}{V} \sum_{\mathbf{s}} m(\mathbf{s}) F^{\text{obs}}(\mathbf{s}) \exp\left[i\,\varphi^{\text{best}}(\mathbf{s})\right] \exp\left[-2\pi i(\mathbf{s},\mathbf{r})\right]. \tag{18}$$

Как нетрудно видеть, взвешенный синтез Фурье есть не что иное, как среднее арифметическое синтезов, рассчитанных с использованием экспериментальных значений модулей и альтернативными вариантами наборов фаз. В нашем исследовании взаимно сравнивались две стратегии построения априорных распределений (14), используемых при генерации масок. Стратегия с одновременным использованием всех

рефлексов на протяжении всех тестовых расчетов проверялась в двух вариантах, с использованием двух разных типов синтезов Фурье, (17) и (18).

3. ТЕСТОВЫЙ ОБЪЕКТ И УСЛОВИЯ ПРОВЕДЕНИЯ ТЕСТОВ

Для проверки возможностей предложенного критерия отбора масок в данной работе в качестве тестового объекта использовалась известная структура мономерной формы фотосистемы II (*Thermosynechococcus elongatus* photosystem II, [43], PDB код 3KZI [44]), далее PS-II. Этот комплекс содержит 19 белковых цепей и множество кофакторов и содержит около 25 тысяч неводородных атомов. Представленный в PDB комплекс имеет молекулярный вес примерно 360 KDa. Эта величина использовалась далее для оценки удельного объема области молекулы. Внешние размеры мономера составляют примерно 100 Å. Общая структура комплекса показана на рисунке 3.

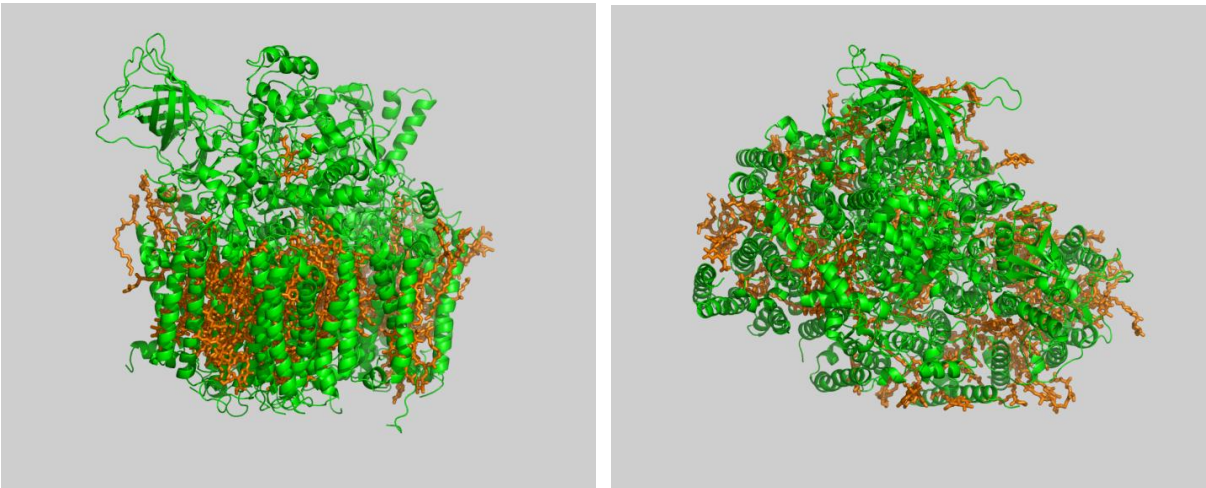

Рис. 3. Фотосистема II. Зеленой лентой показана белковая часть комплекса, оранжевыми линиями – кофакторы.

Для проведения тестов была выбрана ячейка воображаемого кристалла с размерами $200 \times 200 \times 200$ Å, и по координатам атомов был рассчитан набор структурных факторов $\left\{F^{\text{obs}}(\mathbf{s}), \varphi^{\text{true}}(\mathbf{s})\right\}$ разрешения 16 Å (4108 независимых рефлексов). Далее в тестах модули этих структурных факторов $\left\{F^{\text{obs}}(\mathbf{s})\right\}$ рассматривались как экспериментально полученная информация об объекте, а значения фаз $\left\{\varphi^{\text{true}}(\mathbf{s})\right\}$ использовались лишь для контроля полученных результатов. На каждом шаге построение случайных масок велось до тех пор, пока не набиралось 100 масок, удовлетворяющих поставленным ограничениям. Значение контраста t в распределении (13) бралось равным 10^6.

Точность определения фаз характеризовалась коэффициентами корреляции синтезов Фурье, рассчитанных по разным зонам разрешения с экспериментальными значениями модулей и сравниваемыми наборами фаз [5]. В терминах значений структурных факторов эти коэффициенты корреляции могут быть представлены в виде

$$CP\left[d_{\max}, d_{\min}\right]\left(\left\{\varphi(\mathbf{s})\right\}\right) = \frac{\sum\limits_{\mathbf{s}\in\mathbf{S}} \left(F^{\text{obs}}(\mathbf{s})\right)^2 \cos\left(\varphi^{\text{true}}(\mathbf{s}) - \tilde{\varphi}(\mathbf{s})\right)}{\sum\limits_{\mathbf{s}\in\mathbf{S}} \left(F^{\text{obs}}(\mathbf{s})\right)^2}. \tag{19}$$

Здесь суммирование идет по структурным факторам с $\mathbf{s} \in \mathbf{S} = \left\{\mathbf{s} : 1/d_{\max} < s < 1/d_{\min}\right\}$, $\left\{\tilde{\varphi}(\mathbf{s})\right\}$ – набор фаз, полученный из оцениваемого набора $\left\{\varphi(\mathbf{s})\right\}$ его оптимальным

выравниванием по отношению к $\left\{\varphi^{\text{true}}(\mathbf{s})\right\}$. В тестах 2 и 3 выравнивание велось по всем рефлексам зоны ∞ – 16 Å, а в тесте 1 в зависимости от этапа работы (25, 20, 16Å соответственно).

Одним из основных параметров, определяющих описываемый метод, является размер генерируемых масок или, что то же самое, предполагаемый объем частицы. В данной работе мы характеризуем предполагаемый размер области либо количеством точек в маске, либо удельным объемом, определяемым как отношение объема области (в кубических ангстремах) к молекулярной массе частицы (в дальтонах). В кристаллографии при оценке процентного содержания растворителя в элементарной ячейке принято считать удельный объем области молекулы равным 1.23 $Å^3$/Da [45, 46]. Однако эта оценка не является оптимальной при работе в зоне низкого разрешения, когда границы молекулы существенно сглажены. В связи с этим, в данной работе удельный объем области молекулы принимался равным 1.33 $Å^3$/Da. Как результат, сетка и число точек в маске при работе при разрешении 25, 20 и 16 Å были определены как (24 × 24 × 24, 831), (30 × 30 × 30, 1623) и (40 × 40 × 40, 3847) соответственно.

4. РЕЗУЛЬТАТЫ ТЕСТИРОВАНИЯ

4.1. Тест 1

Традиционно в биологической кристаллографии определение структуры объекта проходит через несколько стадий с постепенным увеличением количества экспериментальных данных, включенных в работу [47]. Следуя этому правилу, тестовое восстановление значений фаз структурных факторов было разбито на три стадии, с удвоением количества используемых данных на каждом этапе. Работа начиналась с использования набора рефлексов разрешения 25 Å (1054 рефлексов, шаги 1–5). На следующих стадиях разрешение набора было повышено до 20 Å (2082 рефлексов, шаги 6–10) и 16 Å (4108 рефлексов, шаги 11–16). На первом шаге генерация масок шла с равномерным априорным распределением, на следующих шагах использовалось априорное распределение вида (13) со значением контраста $t = 10^6$ и невзвешенным синтезом (16), построенным по результатам предыдущего шага. На рисунке 4 показано изменение по ходу работы величины фазовой корреляции с точными фазами для различных оболочек в обратном пространстве. В таблице 1 приведены значения корреляции финального фазового набора с точными фазами по расширяющимся зонам и оболочкам в обратном пространстве.

Таблица 1. Качество финальных наборов фаз, полученных в тестах. Приведена фазовая корреляция (19) (в процентах)

	Границы зон разрешения $d_{\max} - d_{\min}$ (Å)										
	∞–60	∞–40	∞–30	∞–25	∞–20	∞–16	60–40	40–30	30–25	25–20	20–16
Тест 1	100	100	99	99	99	98	97	95	92	85	71
Тест 2	100	100	99	99	98	96	97	94	89	72	54
Тест 3	100	100	99	99	98	97	98	93	87	83	66
Число рефл. в зоне	85	255	618	1054	2082	4108	170	363	436	1028	2026

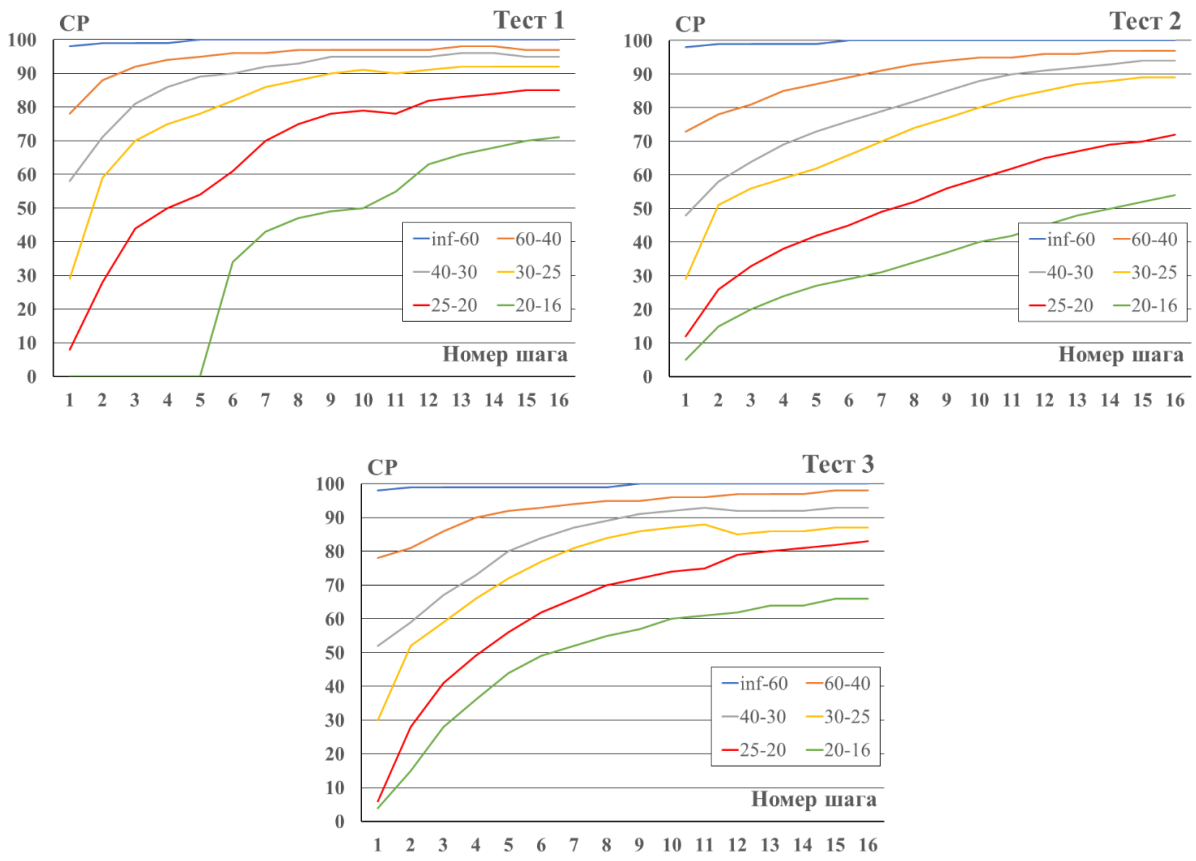

Рис. 4. Изменение величины корреляции текущих значений фаз с точными фазами в процессе решения фазовой проблемы. Показаны значения, вычисленные по формуле (19) для разных интервалов $d_{\max} - d_{\min}$.

4.2. Тест 2

Во втором тесте работа осуществлялась со всеми рефлексами зоны разрешения 16 Å сразу (4108 рефлексов). Рисунок 4 показывает изменение величины фазовой корреляции для различных оболочек в обратном пространстве на разных шагах теста. Соответствующие значения корреляции финального фазового набора с точными фазами по расширяющимся зонам и оболочкам в обратном пространстве приведены в таблице 1.

4.3. Тест 3

В третьем тесте использовалась промежуточная стратегия. Формально, в работу были включены все рефлексы, но для построения априорного распределения использовался взвешенный синтез Фурье (17). Поскольку на начальных шагах работы показатели достоверности (15) для оболочек высокого разрешения малы, то это, фактически, выводило рефлексы высокого разрешения из построения априорных распределений. В то же время при расчете критериев отбора использовались все рефлексы зоны 16 Å. Рисунок 4 и таблица 1 представляют результаты расчетов, аналогично предыдущим двум тестам.

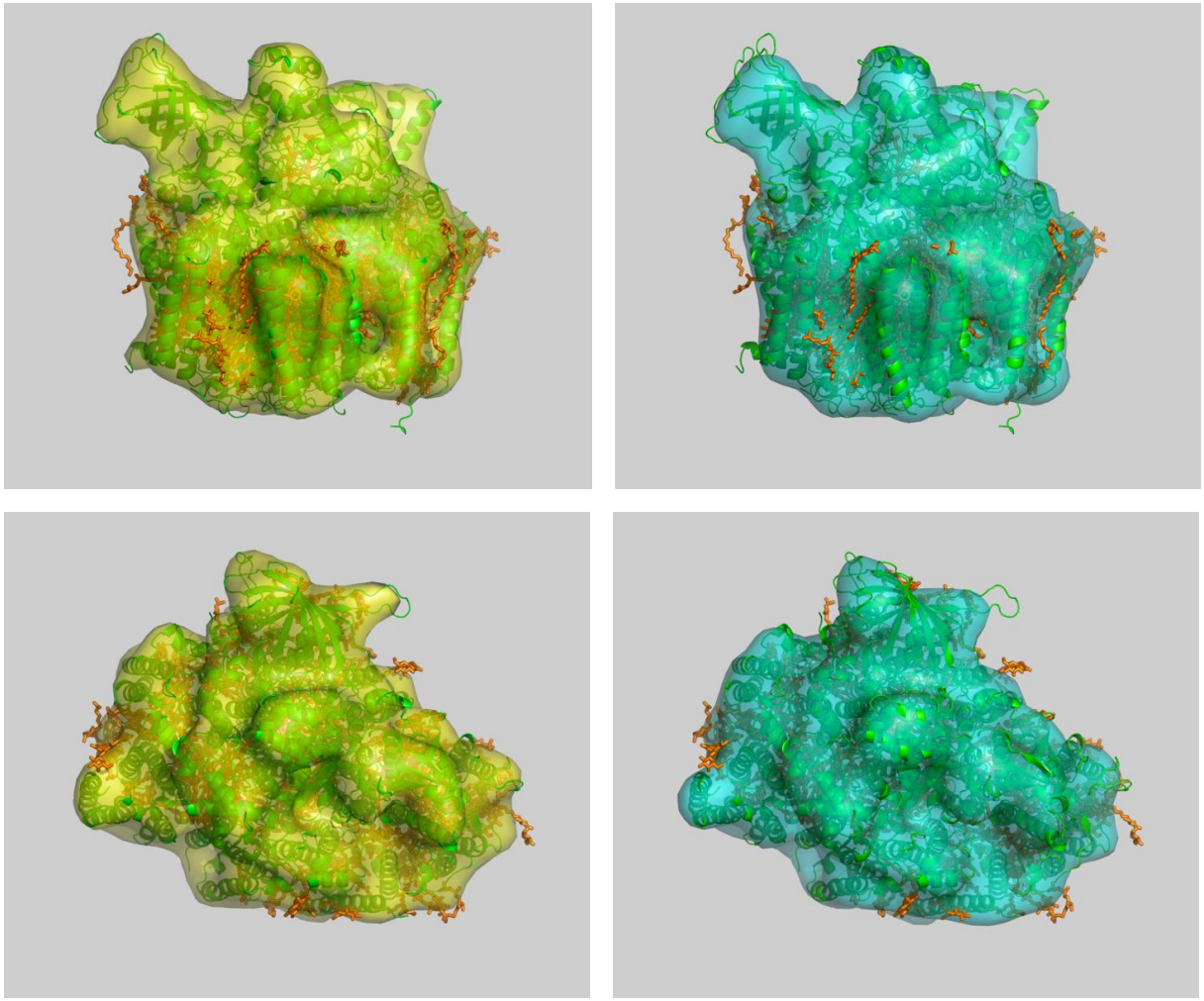

Рис. 5. "Идеальный" синтез Фурье разрешения 16 Å (слева) и синтез, построенный с использованием значений фаз, полученных на 16 шаге теста 3 (справа), совмещенные с моделью комплекса. Удельный объем показанной области объекта 1.23 Å3/Da.

На рисунке 5 показаны финальный синтез Фурье, полученный в третьем тесте, и "идеальный" синтез Фурье разрешения 16 Å, рассчитанный с точными значениями фаз.

5. ВЫВОДЫ

В представленной работе показано, что предложенный новый тип критерия отбора позволяет решать фазовую проблему при исследовании изолированных частиц методом рентгеновской дифракции. При этом использование нового критерия позволило вдвое увеличить число определяемых значений фаз рефлексов по сравнению с тестами, проведенными в предыдущих работах на основе прежнего критерия отбора – коэффициента корреляции модулей [20, 37].

Постепенное повышения разрешения, как в априорных распределениях, так и при вычислении критериев отбора, приводит к лучшим результатам, нежели одновременное использование всех рефлексов сразу.

Проведенное исследование показало, что использование взвешенных синтезов при построении априорного распределения дает для рефлексов наиболее высокого разрешения сходимость к решению более быструю, нежели при работе с невзвешенными синтезами. Другое преимущество этого подхода состоит в том, что он подключает рефлексы высшего разрешения к работе автоматически, согласно увеличивающейся надёжности соответствующей фазовой информации, и не требует ручного вмешательства и экспертной оценки.

Работа была поддержана грантом РФФИ 16-04-01037а. Расчеты производились на многопроцессорном вычислительном комплексе K100 Института прикладной математики им. М.В.Келдыша РАН. Работа А.Г. Уржумцева была поддержана ресурсами FRISBI (Французской инфраструктуры Интегральной структурной биологии, ANR-10-INBS-05) и Instruct-ERIC.

СПИСОК ЛИТЕРАТУРЫ

1. Barends T. R. M., Foucar L., Botha S., Doak R. B., Shoeman R. L.,Nass K., Koglin J.E.,Williams G. J., Boutet S.,Messerschmidt M., Schlichting I. De novo protein crystal structure determination from X-ray free-electron laser data. *Nature*. 2014. V. 505. P. 244–247. doi: 10.1038/nature12773
2. Chapman H.N., Fromme P., Barty A., White T.A., Kirian R.A., Aquila A., Hunter M.S, Schulz J., DePonte D.P, Weierstall U., *et al*. Femtosecond X-ray protein nanocrystallography. *Nature*. 2011. V. 470. P. 73–77. doi: 10.1038/nature09750
3. Boutet S., Lomb L., Williams G.J., Barends T.R., Aquila A., Doak R.B., Weierstall U., DePonte D.P., Steinbrener J., Shoeman R.L., *et al*. High-resolution protein structure determination by serial femtosecond crystallography. *Science*. 2012. V. 337. P. 362–364. doi: 10.1126/science.1217737
4. Redecke L., Nass K., DePonte D.P., White T.A., Rehders D., Barty A., Stellato F., Liang M., Barends T.R.M., Boutet S., *et al*. Natively Inhibited Trypanosoma brucei Cathepsin B Structure Determined by Using an X-ray Laser. *Science*. 2013. V. 339. P. 227–230. doi: 10.1126/science.1229663
5. Lomb L., Barends T.R.M., Kassemeyer S., Aquila A., Epp S.W., Erk B., Foucar L., Hartmann R., Rudek B., Rolles D., *et al*. Radiation damage in protein serial femtosecond crystallography using an x-ray free-electron laser. *Physical Review B*. 2011. V. 84. 214111. doi: 10.1103/PhysRevB.84.214111
6. Johansson L.C., Arnlund D., White T.A., Katona G., DePonte D.P., Weierstall U., Doak R.B., Shoeman R.L., Lomb L., Malmerberg E., *et al*. Lipidic phase membrane protein serial femtosecond crystallography. *Nature Methods*. 2012. V. 9. P. 263–265. doi: 10.1038/nmeth.1867
7. Kern J., Alonso-Mori R., Hellmich J., Tran R., Hattne J., Laksmono H., Glöckner C., Echols N., Sierra R.G., Sellberg J., *et al*. Room temperature femtosecond X-ray diffraction of photosystem II microcrystals. *Proceedings of the National Academy of Sciences of the USA*. 2012. V. 109. P. 9721–9726. doi: 10.1073/pnas.1204598109
8. Лунин В.Ю., Лунина Н.Л., Петрова Т.Е. Биологическая кристаллография без кристаллов. *Математическая биология и биоинформатика*. 2017. Т. 12. № 1. С. 55–72. doi: 10.17537/2017.12.55
9. Thibault P., Elser V., Jacobsen C., Shapiro D., Sayre D. Reconstruction of a yeast cell from X-ray diffraction data. *Acta Crystallographica Section A: Foundations of Crystallography*. 2006. V. 62. P. 248–261. doi: 10.1107/S0108767306016515
10. Rodriguez J.A., Xu R., Chen C.C., Huang Z., Jiang H., Chen A.L., Raines K.S., Pryor A. Jr, Nam D., Wiegart L., Song C., Madsen A., Chushkin Y., Zontone F., Bradley P.J., Miao J. Three-dimensional coherent X-ray diffractive imaging of whole frozen-hydrated cells. *IUCr Journal*. 2015. V. 2. P. 575–583. doi: 10.1107/S205225251501235X
11. Takayama Y., Yonekura K. Cryogenic coherent X-ray diffraction imaging of biological samples at SACLA: a correlative approach with cryo-electron and light microscopy. *Acta Crystallographica Section A: Foundations and Advances*. 2016. V. 72. P. 179–189. doi: 10.1107/S2053273315023980
12. Munke A., Andreasson J, Aquila A., Awel S., Ayyer K., Barty A., Bean R.J., Berntsen P., Bielecki J., Boutet S. *et al*. Coherent diffraction of single Rice Dwarf virus particles using hard X-rays at the Linac Coherent Light Source. *Sci. Data*. 2016. V. 3. Article No. 160064. doi: 10.1038/sdata.2016.64

13. Ekeberg T., Svenda M., Abergel C., Maia F.R.N.C., Seltzer V., Claverie J.M., Hantke M., Jönsson O., Nettelblad C., van der Schot G. *et al*. Three-Dimensional Reconstruction of the Giant Mimivirus Particle with an X-Ray Free-Electron Laser. *Physical Review Letters*. 2015. V. 114. Article No. 098102. doi: 10.1103/PhysRevLett.114.098102.
14. Song C., Jiang H., Mancuso A., Amirbekian B., Peng L., Sun R., Shah S.S., Zhou Z.H., Ishikawa T., Miao J. Quantitative Imaging of Single, Unstained Viruses with Coherent X Rays. *Physical Review Letters*. 2008. V. 101. Article: 158101. doi: 10.1103/PhysRevLett.101.158101
15. Seibert M.M., Ekeberg T, Maia F.R.N.C., Svenda M., Andreasson J.,Jönsson O., Odic D., Iwan B., Rocker A., Westphal1 D. Single mimivirus particles intercepted and imaged with an X-ray laser. *Nature*. 2011. V. 470. P. 78–82. doi: 10.1038/nature09748
16. Van der Schot G., Svenda M., Maia F.R.N.C., Hantke M., DePonte D., Seibert M.M., Aquila A., Schulz J., Kirian R., Liang M. *et al*. Imaging single cells in a beam of live cyanobacteria with an X-ray laser. *Nature Communication*. 2015. V. 6. Article number: 5704. doi: 10.1038/ncomms6704
17. Fienup J.R. Reconstruction of an object from the modulus of its Fourier transform. *Optics Letters*. 1978. V. 3. N. 1. P. 27–29. doi: 10.1364/OL.3.000027
18. Urzhumtseva L., Klaholz B., Urzhumtsev A. On effective and optical resolutions of diffraction data sets. *Acta Crystallographica Section D: Biological Crystallography*. 2013. V. 69. P. 625–634. doi: 10.1107/S0907444913016673
19. Sayre D. Some implications of a theorem due to Shannon. *Acta Crystallographica*. 1952. V. 5. P. 843. doi: 10.1107/S0365110X52002276
20. Lunin V.Y., Lunina N.L., Petrova T.E., Baumstark M.W., Urzhumtsev A.G. Mask-based approach to phasing of single-particle diffraction data. *Acta Crystallographica Section D: Structural Biology*. 2016. V. 72. P. 147–157. doi: 10.1107/S2059798315022652
21. Marchesini S., He H., Chapman H.N., Hau-Riege S.P., Noy A., Howells M.R., Weierstall U., Spence J.H.C. X-ray image reconstruction from a diffraction pattern alone. *Phis. Rev. B*. 2003. V. 68. Article No. 140101(R). doi: 10.1103/PhysRevB.68.140101
22. Maia F.R.N.C., Ekeberg T., Spoel D., Hajdu J. Hawk: the image reconstruction package for coherent X-ray diffractive imaging. *J. Applied Crystallography*. 2010. V. 43. P. 1535–1539. doi: 10.1107/S0021889810036083
23. Bricogne G. Geometric sources of redundancy in intensity data and their use for phase determination. *Acta Crystallographica Section A: Foundations of Crystallography*. 1974. V. 30. P. 349–405. doi: 10.1107/S0567739474010722
24. Rodriguez J.A., Xu R.,Chen C.-C., Zou Y., Miao J. Oversampling smoothness: an effective algorithm for phase retrieval of noisy diffraction intensities. J. Applied Crystallography. 2013. V. 46. P. 312–318. doi: 10.1107/S0021889813002471
25. Miao J., Kirz J., Sayre D. The oversampling phasing method. *Acta Crystallographica Section D: Biological Crystallography.* 2000. V. 56. P. 1312–1315. doi: 10.1107/S0907444900008970
26. He H., Su, W.P. Direct phasing of protein crystals with high solvent content. *Acta Crystallographica Section A: Foundations of Crystallography*. 2015. V. 71. P. 92–98. doi: 10.1107/S2053273314024097
27. 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. doi: 10.1107/S0567739476001691
28. Marchesini S. A unified evaluation of iterative projection algorithms for phase retrieval. *Rev. Sci. Instrum*. 2007. V. 78. Article No. 011301. doi: 10.1063/1.2403783

29. Zhang K.Y.J., Cowtan K.D., Main P. Phase improvement by iterative density modification. In: *International Tables for Crystallography*. V. F. Second Edition. Eds. Arnold E., Himmel D.M., Rossmann M.G. Chichester: John Wiley and Sons, 2012. P. 385–400. doi: 10.1107/97809553602060000847
30. Millane R., Lo V.L. Iterative projection algorithms in protein crystallography. I. Theory. *Acta Crystallographica Section A: Foundations of Crystallography*. 2013. V. 69. P. 517–527. doi: 10.1107/S0108767313015249
31. Wang B.C. Resolution of phase ambiguity in macromolecular crystallography. *Methods in Enzymology*. 1985. V. 115. P. 90–112.
32. Loh N.-T.D., Elser V. Reconstruction algorithm for single-particle diffraction imaging experiments. *Physical Review E*. 2009. V. 80. Article No. 026705. doi: 10.1103/PhysRevE.80.026705
33. Elser V. Solution of the crystallographic phase problem by iterated projections. *Acta Crystallographica Section A: Foundations of Crystallography*. 2003. V. 59. P. 201–209. doi: 10.1107/S0108767303002812
34. 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. doi: 10.1107/S0108767390003464
35. 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 A: Foundations of Crystallography*. 2000. V. 56. P. 375–382. doi: 10.1107/S0108767300004633
36. Petrova T.E., Lunin V.Y., Podjarny A.D. Ab initio low-resolution phasing in crystallography of macromolecules by maximization of likelihood. *Acta Crystallographica Section D: Biological Crystallography*. 2000. V. 56. P. 1245–1252. doi: 10.1107/S0907444900009343
37. Лунин В.Ю., Лунина Н.Л., Петрова Т.Е. Использование связных масок в задаче восстановления изображения изолированной частицы по данным рентгеновского рассеяния. *Математическая биология и биоинформатика*. 2014. Т. 9. № 2. С. 543–562. doi: 10.17537/2014.9.543
38. 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. doi: 10.1107/S0907444992008801
39. Lunin V.Y., Urzhumtsev A.G. Improvement of protein phases by coarse model modification. *Acta Crystallographica Section A: Foundations of Crystallography*. 1984. V. 40. P. 269–277. doi: 10.1107/S0108767384000544
40. Lunin V.Y., Skovoroda T.P. R-free likelihood-based estimates of errors for phases calculated from atomic models. *Acta Crystallographica Section A: Foundations of Crystallography*. 1995. V. 51. P. 880–887. doi: 10.1107/S010876739500688X
41. Urzhumtsev A.G., Skovoroda T.P., Lunin V.Y. A procedure compatible with X-PLOR for the calculation of electron-density maps weighted using an R-free-likelihood approach. J. Applied Crystallography. 1996. V. 29. P. 741–744. doi: 10.1107/S0021889896007194
42. Read R.J. Improved Fourier Coefficients for Maps Using Phases from Partial Structures with Errors. *Acta Crystallographica Section A: Foundations of Crystallography*. 1986. V. 42. P. 140-190. doi: 10.1107/S0108767386099622
43. Broser M., Gabdulkhakov A., Kern J., Guskov A., Müh F., Saenger W., Zouni A. Crystal structure of monomeric Photosystem II from Thermosynechococcus elongatus at 3.6 Å resolution. *J. Biol. Chem*. 2010. V. 285. P. 26255–26262. doi: 10.1074/jbc.M110.127589

44. Berman H.M., Westbrook J., Feng Z., Gilliland G., Bhat T.N., Weissig H., Shindyalov I.N., Bourne P.E. The Protein Data Bank. *Nucleic Acids Research*. 2000. V. 28. P. 235–242. doi: 10.1093/nar/28.1.235
45. Matthews B.M. Solvent content of protein crystals. *Journal of Molecular Biology*. 1968. V. 33. P. 491–497. doi: 10.1016/0022-2836(68)90205-2
46. Weichenberger C.X., Rupp B. Ten years of probabilistic estimates of biocrystal solvent content: new insights via nonparametric kernel density estimate. *Acta Crystallographica Section D: Biological Crystallography*. 2014. V. 70. P. 1579–1588. doi: 10.1107/S1399004714005550
47. Urzhumtsev A., Afonine P.V., Adams P.D. On the use of logarithmic scales for analysis of diffraction data. *Acta Crystallographica Section D: Biological Crystallography*. 2009. V. 65. P. 1283–1291. doi: 10.1107/S0907444909039638

Рукопись поступила в редакцию 29.11.2017.
Дата опубликования 13.12.2017.