*Математическая биология и биоинформатика*
*2020. Т. 15. № S. С. t1–t20. doi: 10.17537/2020.15.t1*

*Перевод оригинальной статьи, опубликованной на английском языке:*
*Лунин В., Лунина Н., Петрова Т. Математическая биология и биоинформатика. 2020;15(1):57–72.*
*doi: 10.17537/2020.15.57*

**========== ПЕРЕВОДЫ ОПУБЛИКОВАННЫХ СТАТЕЙ ==========**

# Восстановление модулей и расчет фаз для дифракционной картины изолированной частицы с использованием бинарных масок объекта

**Лунин В.Ю., Лунина Н.Л., Петрова Т.Е.**

*Институт математических проблем биологии РАН – филиал Института прикладной математики им. М.В. Келдыша Российской академии наук, Пущино, Россия*

***Аннотация.*** Развитие экспериментальной техники и, в частности, ввод в эксплуатацию рентгеновских лазеров на свободных электронах позволяют приблизиться к возможности регистрации рентгеновского рассеяния отдельной макромолекулярной частицей. Это открывает дорогу к определению методами рентгеновской дифракции структуры некристаллизованных макромолекулярных объектов. Возможность измерения интенсивностей не-Брэгговских рефлексов создает существенную избыточность экспериментальных данных, что существенно упрощает определение структуры объекта. Дискретизация непрерывной дифракционной картины на сетку с достаточно мелким шагом позволяет рассматривать проблему определения структуры как проблему определения структуры для "виртуального" кристалла с чрезвычайно большим относительным объемом растворителя в элементарной ячейке. В предположении, что область, занимаемая объектом в элементарной ячейке, известна, можно ожидать высокой эффективности решения фазовой проблемы итерационными методами, типа методов модификации электронной плотности. В то же время, итерационные методы чувствительны к точности задания области молекулы, неполноте экспериментальных данных и изначальной неединственности решения. Разработанный авторами метод предварительного решения фазовой проблемы осуществляет случайный поиск связных бинарных аппроксимаций распределения электронной плотности в объекте (масок области молекулы), воспроизводящих с достаточной точностью дифракционную картину, наблюдаемую в эксперименте. Выравнивание, в рамках группы эквивалентности решений фазовой проблемы, найденных масок с последующим усреднением позволяет получить приближенное решение фазовой проблемы. Помимо оценки неизвестных значений фаз структурных факторов, разработанный подход позволяет восстанавливать фрагменты дифракционной картины (значения модулей структурных факторов), потерянные в эксперименте. Примерами таких фрагментов могут служить нерегистрируемая центральная зона рентгенограммы или области "переэкспонированных" (ввиду ограниченности рабочего диапазона детектора) рефлексов.

***Ключевые слова:*** *биологические макромолекулы, одиночные частицы, рентгеновское рассеяние, рентгеновские лазеры на свободных электронах, фазовая проблема, восстановление дифракционных данных, эффективное разрешение.*

## 1. ВВЕДЕНИЕ

Среди биологов существует твердое убеждение, что в задаче определения структуры биологической макромолекулы главной проблемой является выращивание кристалла, состоящего из таких молекул. Специалисты, занимающиеся вычислительными методами биологической кристаллографии, делают все возможное, чтобы убедить биологов в справедливости этого утверждения, разрабатывая все более мощные и удобные для пользователя компьютерные программы, позволяющие расшифровывать структуру в почти автоматическом режиме [1–6]. Однако наличие таких инструментов не устраняет необходимости подготовки кристаллов. Еще одной постоянной проблемой является полнота набора собранных данных, что, опять же, определяется качеством имеющихся кристаллов. В последнее десятилетие развитие рентгеновского эксперимента, в частности, ввод в эксплуатацию рентгеновских лазеров на свободных электронах, создало предпосылки для разработки подходов, которые позволяют преодолеть эту проблему и определять трехмерные структуры некристаллических объектов [7–12]. В этой работе мы касаемся одной из проблем, возникающих при разработке таких подходов, а именно расшифровки структуры, в предположении, что набор экспериментальных данных собран. Основное содержание данной статьи было представлено в виде устного доклада на 32-ой Европейской кристаллографической конференции в Вене (Австрия) 18–23 августа 2019 года [13].

## 2. РЕНТГЕНОВСКИЙ ДИФРАКЦИОННЫЙ ЭКСПЕРИМЕНТ

В двух следующих разделах мы приводим базовые сведения о рентгеновском дифракционном эксперименте и особенностях картины дифракции, отсылая за более подробным описанием к работе [14]. В стандартном дифракционном эксперименте, как с одиночными частицами, так и с кристаллом, образец исследуемого вещества облучается первичным пучком рентгеновских лучей, и посредством какого-либо детектора регистрируются энергии новых рентгеновских пучков (электромагнитных волн), расходящихся во всевозможных направлениях от образца. Кинематическая теория рассеяния описывает рассеянные волны как результат суперпозиции вторичных сферических волн, излучаемых осциллирующими электронами образца, приведенными в движение воздействием первичной волны. Рассеянные волны традиционно называются отражениями или рефлексами. Эти термины восходят к первым экспериментам с кристаллическими образцами, где рассеянные волны интерпретировались, как результат отражения первичной волны в "кристаллических плоскостях" по законам геометрической оптики. В рамках кинематической теории рассеяния, результат эксперимента определяется распределением электронов в образце, описываемом функцией распределения электронной плотности $\rho(\mathbf{r}), \mathbf{r} \in \mathbf{R}^3$, а измеряемая в эксперименте энергия есть

$$E(\boldsymbol{\sigma}_0, \boldsymbol{\sigma}) = \varepsilon E_0 \left|\mathbf{F}(\mathbf{s})\right|^2 \quad . \tag{1}$$

Здесь $\boldsymbol{\sigma}_0$ и $\boldsymbol{\sigma}$ – векторы единичной длины, определяющие направления распространения первичной и рассеянной волн, $E_0$ – энергия первичной волны, константа ε является комбинацией физических констант и параметров эксперимента и не зависит от структуры исследуемого образца. Вектор **s**, именуемый вектором рассеяния, определяется как

$$\mathbf{s} = \frac{\boldsymbol{\sigma} - \boldsymbol{\sigma}_0}{\lambda} \quad , \tag{2}$$



Математическая биология и биоинформатика. 2020. Т. 15. № S. doi: 10.17537/2020.15.t1

где λ – длина волны первичной волны. Этот вектор играет важную роль в теории рассеяния и осуществляет связь между геометрией реального эксперимента с математическими объектами, используемыми для его описания. Трехмерное пространство, элементы которого рассматриваются как всевозможные вектора рассеяния (соответствующие различным комбинациям направлений $\boldsymbol{\sigma}_0$ и $\boldsymbol{\sigma}$), называется обратным кристаллографическим пространством. Потенциально, эксперимент позволяет измерить энергии рассеянных волн, отвечающих векторам рассеяния, заполняющим сферу $|\mathbf{s}| \le 2/\lambda$ в обратном пространстве. Однако, на практике, в зависимости от качества кристалла, чувствительности детектора и интенсивности облучения, набор рефлексов (и, соответственно, векторов рассеяния), для которых экспериментально измерена энергия, может быть существенно меньше. Отметим, что эта энергия зависит лишь от комбинации направлений $\boldsymbol{\sigma}_0$ и $\boldsymbol{\sigma}$, а не от направлений по отдельности. В связи с этим, термин "рефлекс" применяется не только к рассеянным волнам, но и к векторам рассеяния, им соответствующим.

Величины $I(\mathbf{s}) = |\mathbf{F}(\mathbf{s})|^2$, называемые интенсивностями рефлексов, являются квадратами комплексной трансформанты Фурье распределения ρ(**r**) электронной плотности:

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

где · означает скалярное произведение двух векторов. Интенсивности рефлексов зависят только от структуры образца. Они могут быть определены (в некой относительной шкале) непосредственно из эксперимента.

В обычных условиях эксперимента величина ε, входящая в равенство (1), чрезвычайно мала и может быть оценена как $10^{-24}$. Это делает экспериментальную регистрацию рассеяния чрезвычайно сложной задачей. До настоящего времени, единственным практическим путем измерения интенсивностей рефлексов являлось приготовление образца для эксперимента в виде монокристалла исследуемого объекта. В этом случае для некоторого дискретного набора рефлексов (Брэгговских рефлексов) интенсивности рассеянных лучей возрастают пропорционально квадрату числа элементарных ячеек в кристалле. В то же время, интенсивность прочих рефлексов чрезвычайно мала и теряется в эксперименте. Развитие техники рентгеновского эксперимента и, в частности, ввод в эксплуатацию рентгеновских лазеров на свободных электронах, открыло возможность регистрации непрерывной картины рассеяния для единичных (не кристаллических) объектов, хотя и в очень ограниченной зоне $|\mathbf{s}| \le s_{\min}$ обратного пространства.

Задача восстановления структуры объекта по картине рассеяния единичным экземпляром может быть сведена к стандартной задаче биологической кристаллографии. Введем в рассмотрение виртуальную элементарную ячейку *V*, понимая под этим параллелепипед достаточно больших размеров, чтобы вместить внутри себя исследуемую частицу, так что $\rho(\mathbf{r}) = 0$ вне элементарной ячейки. Обозначим {**a**, **b**, **c**} базис этой ячейки. Для простоты, мы будем далее предполагать, что ячейка является кубом со стороной $a_{cell}$. Внутри элементарной ячейки распределение электронной плотности может быть представлено в виде ряда Фурье

$$\rho(\mathbf{r}) = \frac{1}{|V|} \sum_{\mathbf{s}\in\mathfrak{R}'} \mathbf{F}_V(\mathbf{s}) \exp[-\mathrm{i}2\pi \mathbf{s}\cdot\mathbf{r}]\ ,\ \mathbf{r} \in V\ . \tag{4}$$



*Математическая биология и биоинформатика*. 2020. Т. 15. № S. doi: 10.17537/2020.15.t1

Здесь $|V|$ – объем элементарной ячейки и коэффициенты Фурье (структурные факторы) $\mathbf{F}_V(\mathbf{s})$ есть не что иное, как значения трансформанты Фурье (3). Множество $\mathfrak{R}'$ векторов рассеяния, по которым идет суммирование в (9), образует решетку в обратном пространстве, состоящую из векторов, удовлетворяющих условию

$$\mathbf{s}\cdot\mathbf{a}=h, \mathbf{s}\cdot\mathbf{b}=k, \mathbf{s}\cdot\mathbf{c}=l, \quad h,k,l-\text{целые числа}, \tag{5}$$

т.е. из векторов, имеющих целочисленные координаты в базисе $\{\mathbf{a}^*, \mathbf{b}^*, \mathbf{c}^*\}$, сопряженном к базису $\{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$. Мы будем называть соответствующие рефлексы Брэгговскими рефлексами. Равенство (4) не выполняется для точек $\mathbf{r}$, лежащих вне $V$. Введем в рассмотрение виртуальный кристалл, обладающий распределением электронной плотности $\rho^{cryst}(\mathbf{r})$, определённым равенством (4) для всех точек $\mathbf{r}\in\mathbf{R}^3$. Это распределение является периодическим с периодами $\{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$ и совпадает с распределением $\rho(\mathbf{r})$ в элементарной ячейке $V$. Задача расшифровки структуры исследуемого объекта является теперь стандартной задачей рентгеновской кристаллографии, а именно, задачей нахождения распределения электронной плотности в элементарной ячейке кристалла при наличии информации о величинах интенсивностей Брэгговских рефлексов или, что то же, о величинах модулей $|\mathbf{F}(\mathbf{s})|$ соответствующих структурных факторов. Эта задача может быть переформулирована как задача нахождения потерянных в эксперименте значений фаз $\varphi(\mathbf{s})$ структурных факторов, необходимых для расчета распределения плотности по формуле (4).

Аналогично, если непрерывная дифракционная картина $I(\mathbf{s}), \mathbf{s}\in\mathbf{R}^3$ дискретизирована на регулярную сетку в трехмерном пространстве, определяемую шагами дискретизации $\{\mathbf{a}^*, \mathbf{b}^*, \mathbf{c}^*\}$, то значения $I(\mathbf{s})$ в узлах этой сетки можно рассматривать как интенсивности Брэгговских рефлексов для виртуального кристалла с базисом $\{\mathbf{a}, \mathbf{b}, \mathbf{c}\}$, являющимся сопряженным к $\{\mathbf{a}^*, \mathbf{b}^*, \mathbf{c}^*\}$. Фундаментальное отличие от стандартной кристаллографической ситуации проявляется в том, что в случае одиночной частицы мы имеем существенную свободу в выборе размеров элементарной ячейки, или, что то же самое, в выборе шагов дискретизации обратного пространства $\{\mathbf{a}^*, \mathbf{b}^*, \mathbf{c}^*\}$. Необходимо лишь иметь размеры ячейки достаточно большими, чтобы вмещать внутри себя исследуемый объект.

Если значения фаз структурных факторов каким-то образом найдены, они могут быть использованы, совместно с экспериментально определенными значениями модулей, для приближенного расчета распределения электронной плотности по формуле (4). Результат такого расчета обычно именуется синтезом Фурье электронной плотности. Точность такой аппроксимации существенно зависит от количества членов ряда Фурье, включенных в расчет. Для того, чтобы численно охарактеризовать качество аппроксимации, полученной путем расчета частичного ряда Фурье, используется концепция разрешения. В биологической кристаллографии понятие разрешения определяется, вначале, для отдельного рефлекса $\mathbf{s}$, как величина периода соответствующий гармоники Фурье $\exp[-i2\pi\mathbf{s}\cdot\mathbf{r}]$ в разложении (4), рассматриваемой как функция $\mathbf{r}$. Этот период равен

$$d=\frac{1}{|\mathbf{s}|}=\frac{\lambda}{2\sin\theta}, \tag{6}$$

где $2\theta$ – угол между направлениями $\boldsymbol{\sigma}_0$ и $\boldsymbol{\sigma}$, $\lambda$ – длина волны используемого рентгеновского излучения. Принято говорить, что синтез (4) рассчитан с разрешением $d_{\min}$, если в расчет включены все (или "почти все") рефлексы с $|\mathbf{s}|\le s_{\max}=d_{\min}^{-1}$. Мы будем называть зоной разрешения сферическую область в обратном пространстве, заданную



Математическая биология и биоинформатика. 2020. Т. 15. № S. doi: 10.17537/2020.15.t1

условием $|\mathbf{s}| \le s_{\max}$, и оболочкой – сферический слой, определенный как $s_{\min} < |\mathbf{s}| \le s_{\max}$. Введенное понятие разрешения можно назвать "формальным" разрешением, поскольку учитывается только набор рефлексов, используемых для расчета, и не принимается во внимание точность определения величин модулей и фаз структурных факторов. Более сложные концепции разрешения [15–17] базируются на визуальном качестве "топографических" карт найденного приближения к распределению электронной плотности.

## 3. АНАЛИТИЧЕСКИЕ СВОЙСТВА ДИФРАКЦИОННОЙ КАРТИНЫ, ОТВЕЧАЮЩЕЙ ОДИНОЧНОЙ ЧАСТИЦЕ

Уменьшение шага дискретизации дифракционной картины приводит к увеличению количества экспериментальных данных, вовлеченных в работу в пределах заданной зоны разрешения. Однако, эти данные могут не являться независимыми. Если виртуальная ячейка достаточно велика, чтобы содержать внутри себя носитель функции $\rho(\mathbf{r})$ (т.е. все точки, в которых эта функция отлична от нуля), то для любой точки $\mathbf{u}$ значение трансформанты Фурье в этой точке является линейной комбинацией Брэгговских структурных факторов

$$\mathbf{F}(\mathbf{u}) = \sum_{\mathbf{s}\in\mathfrak{R}'} \mathbf{sinc}(\mathbf{u}-\mathbf{s})\mathbf{F}(\mathbf{s}),\ \mathbf{u}\in\mathbf{R}^3\ , \tag{7}$$

с предопределенными коэффициентами, не зависящими от функции $\rho(\mathbf{r})$:

$$\mathbf{sinc}\left(h\mathbf{a}^* + k\mathbf{b}^* + l\mathbf{c}^*\right) = \frac{\sin\pi h}{\pi h}\cdot\frac{\sin\pi k}{\pi k}\cdot\frac{\sin\pi l}{\pi l}. \tag{8}$$

При выводе последней формулы мы предполагаем, что начало координат находится в центре элементарной ячейки. Аналоги формулы (7) играют важную роль в теории передачи информации и, обычно, ассоциированы с именами В. Котельникова, К. Шеннона, Э. Уиттакера и Г. Найквиста, хотя ранние упоминания о ней могут быть найдены в работах Э. Бореля [18, 19]. Перспективы использования этой формулы в кристаллографии были отмечены Д. Сэйром [20], она использовалась в подходах, разработанных Ж. Бриконем [21, 22]. Из разложения (7) следует, что функции $\mathbf{F}(\mathbf{s})$ и $\rho(\mathbf{r})$ полностью определены во всех точках обратного и прямого пространств, если заданы значения структурных факторов для набора Брэгговских рефлексов, соответствующих минимально возможной элементарной ячейке, все еще вмещающей частицу.

Аналогично, интерполяционная функция может быть выписана для непрерывного распределения интенсивностей рефлексов $I(\mathbf{s}) = |\mathbf{F}(\mathbf{s})|^2$ в предположении, что виртуальная ячейка достаточно велика, чтобы содержать носитель автокорреляционной функции. Автокорреляционная функция вводится как

$$A(\mathbf{r}) = \rho * \breve{\rho} = \int_{\mathbf{R}^3} \rho(\mathbf{u})\rho(\mathbf{u}-\mathbf{r})\,dV_{\mathbf{u}},\ \mathbf{r}\in\mathbf{R}^3\ , \tag{9}$$

где $\breve{\rho}(\mathbf{r}) = \rho(-\mathbf{r})$ – энантиомер функции $\rho(\mathbf{r})$. Предполагая, что $\operatorname{supp} A(\mathbf{r}) \subset V$, мы получаем равенство

$$I(\mathbf{u}) = \sum_{\mathbf{s}\in\mathfrak{R}'} \mathbf{sinc}(\mathbf{u}-\mathbf{s})I(\mathbf{s}),\ \mathbf{u}\in\mathbf{R}^3\ . \tag{10}$$

Достаточным условием справедливости этого равенства является выбор размеров виртуальной ячейки вдвое больше диаметра исследуемой частицы. Это отвечает выбору



Математическая биология и биоинформатика. 2020. Т. 15. № S. doi: 10.17537/2020.15.t1

шага дискретизации данных, который был бы вдвое меньше величины, обратной диаметру частицы. Такой выбор шага часто называется пределом Найквиста.

Из равенства (9) следует, что непрерывная дифракционная картина $I(\mathbf{s})$ полностью определена значениями интенсивностей Брэгговских рефлексов, отвечающих минимально возможному выбору виртуальной ячейки, все еще вмещающей носитель автокорреляционной функции. Может создаться впечатление, что дальнейшее уменьшение шага дискретизации не привносит в работу новой информации. Это не вполне так. Формулы (7) и (10) предполагают суммирование по бесконечному набору рефлексов. В случаях, когда для части Брэгговских рефлексов интенсивности неизвестны или содержат ошибки, измерение интенсивности любого не-Брэгговского рефлекса порождает уравнение, ограничивающее значения неизвестных Брэгговских интенсивностей [23].

Важность значений, отвечающих промежуточным точкам обратного пространства, подчеркивается фундаментальной теоремой Шварца – Пэлли – Винера. Эта теорема гласит, что преобразование Фурье функции с компактным носителем является целой голоморфной функцией. Термин "целой" означает, в частности, что ряд Тейлора с центром в произвольной точке $\mathbf{s}_0$ обратного пространства сходится в любой другой точке этого пространства $\mathbf{s}$, определяя, тем самым, значение $I(\mathbf{s})$. Коэффициентами разложения в ряд Тейлора являются частные производные функции $I(\mathbf{s})$, вычисленные в точке $\mathbf{s}_0$. Для того, чтобы определить эти производные, достаточно знать значения функции $I(\mathbf{s})$ в сколь угодно малой окрестности точки $\mathbf{s}_0$. Таким образом, мы приходим к удивительному следствию: при исследовании единичной частицы для того, чтобы знать значения интенсивностей рефлексов во всех точках обратного пространства, достаточно знать все значения интенсивностей из малой окрестности какой-либо точки $\mathbf{s}_0$. Это, потенциально, означает возможность вычисления значений интенсивностей для рефлексов любого разрешения, не ограничиваясь даже теоретически пределом $\lambda/2$. К сожалению, описанная процедура восстановления неизвестных значений интенсивностей имеет чисто теоретический характер и не реализуема на практике. Задача разработки алгоритмов, позволяющих практически осуществлять расширение набора известных интенсивностей, является вызовом для методов вычислительной математики.

## 4. МЕТОДЫ МОДИФИКАЦИИ ЭЛЕКТРОННОЙ ПЛОТНОСТИ И ПРОЕКЦИОННЫЕ МЕТОДЫ

В кристаллах биологических макромолекул существенная часть объема (в среднем 50 %) занята растворителем. При исследовании одиночных частиц аналогом области растворителя является часть виртуальной ячейки, не занятая частицей. Свобода в выборе виртуальной элементарной ячейки позволяет существенно увеличивать эту область. К примеру, выбор для сферической частицы размеров ячейки, вдвое превышающих диаметр частицы ("предел Найквиста"), приводит к тому, что примерно в 93 % объема виртуальной ячейки электронная плотность оказывается равной (или близкой) нулю. Присутствие большого объема растворителя в ячейке создает предпосылки для эффективного применения метода "сглаживания растворителя" (solvent flattening) для решения фазовой проблемы. Этот метод принадлежит к широкому классу "методов модификации электронной плотности", которые применяются в биологической кристаллографии, начиная с 70-х годов прошлого столетия [21, 22, 24–26]. Метод сглаживания растворителя [21, 27] (разработанный независимо в оптике под именем "phase retrieval algorithm") кратко может быть изложен следующим образом. Предположим, что нам известна область в элементарной ячейке, занятая изучаемым объектом (область молекулы) и, тем самым, дополнительная область – область

растворителя. Пусть для некоторого набора рефлексов (с известными модулями структурных факторов) заданы некоторые стартовые значения фаз. Это позволяет рассчитать синтез Фурье (4). В полученном синтезе значения электронной плотности в области растворителя, вообще говоря, не равны нулю. Поэтому мы модифицируем распределение электронной плотности, принудительно делая равными нулю значения в области растворителя. Теперь мы получаем функцию, равную нулю вне области молекулы, но модули структурных факторов, рассчитанных по этой функции, не совпадают с экспериментальными значениями. Поэтому от рассчитанных структурных факторов мы берем только значения фаз, рассматривая их как следующее приближение к решению фазовой проблемы. Это шаг процедуры далее итерационно повторяется. Такой метод уточнения значений фаз структурных факторов является частью многих комплексов кристаллографических программ.

С математической точки зрения, это подход принадлежит к классу проекционных методов и может быть сформулирован следующим образом. Рассмотрим конфигурационное пространство всех распределений электронной плотности. В этом пространстве выделим два класса функций (два многообразия). Первое многообразие образовано функциями, обращающимися в ноль в области растворителя. Второе многообразие – функции, имеющие заданные(экспериментальные) величины модулей структурных факторов. Наша задача заключается в поиске распределения, принадлежащего одновременно обоим многообразиям. Итерационная процедура состоит в последовательном переходе с одного многообразия на другое. При этом каждый раз переход осуществляется в ближайшую точку альтернативного многообразия. Такой переход называется операцией проектирования на соответствующее многообразие. Сходимость такой процедуры может быть медленной, и для ее ускорения может применяться следующий прием. На очередном шаге процедуры после того, как точка была спроектирована на альтернативное многообразие, движение точки не прекращается, а продолжается в том же направлении на такой же шаг. Этот переход можно интерпретировать как "отражение" точки относительно многообразия. Такие операции называются рефлекторами. Формально, рефлектор связан с соответствующим проектором равенством $R = 2P - I$, где $I$ – тождественный оператор. Используя это равенство, нетрудно видеть, что рефлектор, соответствующий обнулению плотности в области растворителя, это хорошо известный "перескок плотности" (density flip) [31, 32]. При этой модификации значения плотности в области молекулы не меняются, а в области растворителя меняют знак ($\rho \to -\rho$). Аналогично, рефлектор, соответствующий операции подмены модулей, есть не что иное, как вычисление широко применяющегося комбинированного "2-1" синтеза Фурье с коэффициентами $\left(2F^{obs} - F^{calc}, \varphi^{calc}\right)$. Использование рефлекторов может существенно ускорить сходимость, но делает процесс менее устойчивым. Для ускорения сходимости итераций с сохранением устойчивости предложены многочисленные комбинации этих четырех базовых операторов, включаемых с различными весами. Перечень наиболее часто применяемых схем может быть найден в [29]. Эти схемы, основанные на четырех базисных операциях, а именно зануление плотности в области растворителя и перескок плотности, и расчет $\left(F^{obs}, \varphi^{calc}\right)$ либо $\left(2F^{obs} - F^{calc}, \varphi^{calc}\right)$ синтеза Фурье, реализованы в интерактивной программе Hawk [33]. Эта программа обеспечивает интерактивный контроль параметров метода и визуализацию промежуточных результатов. Тестовые расчеты с использованием этой программы, проводимые на объектах с известной структурой, приводят к удивительно хорошим результатам. После серии пробных расчетов удается подобрать параметры, приводящие к нужному решению. Ситуация становится более сложной в случае данных, содержащих ошибки, и неизвестном ответе.



Математическая биология и биоинформатика. 2020. Т. 15. № S. doi: 10.17537/2020.15.t1

Это становится особенно существенно при работе с данными низкого разрешения, когда нет четкого понимания, как должен выглядеть "правильный" синтез Фурье.

Дополнительные сложности привносит принципиальная неединственность решения фазовой проблемы. Произвольный сдвиг в пространстве распределения электронной плотности, как целого, или переход к энантиомеру приводит к функции, имеющей те же самые значения модулей структурных факторов, но другие значения их фаз, т.е. к новому решению фазовой проблемы. Для более полного исследования конфигурационного пространства и получения начального приближения могут использоваться методы глобального поиска, учитывающие группу неединственности решения.

Использование метода сглаживания растворителя предполагает, что положение в пространстве и форма области молекулы заранее известна. Определение этой области часто модифицируется итерационно путем фильтрации зашумленного синтеза Фурье, построенного на текущем этапе работы. Одним из путей такого определения области молекулы является двухшаговая фильтрация, предложенная в работах [27, 34–36]. На первом шаге этой процедуры каждая точка в элементарной ячейке получает вес, отражающий степень уверенности в ее принадлежности области молекулы. В простейшем случае этот вес может устанавливаться равным единице для точек с текущим значением плотности, превышающим заданный уровень, и нулевым в противном случае. В более продвинутых схемах единичный вес может устанавливаться как для точек с самыми высокими, так и с самыми низкими значениями плотности. На втором шаге значения назначенных весов заменяются результатом их усреднения в окрестности данной точки. Полученная сглаженная функция используется для выделения области молекулы, как области, содержащей максимальные значения этой функции.

## 5. РЕШЕНИЕ ФАЗОВОЙ ПРОБЛЕМЫ С ИСПОЛЬЗОВАНИЕМ СВЯЗНЫХ МАСОК

Предложенный ранее подход к использованию связных масок области молекулы [37, 38] в задачах биологической кристаллографии может рассматриваться с двух точек зрения. Во-первых, он является методом решения фазовой проблемы биологической кристаллографии при низком разрешении и позволяет получить приближенные значения фаз структурных факторов и значения индивидуальных показателей достоверности найденного значения той или иной фазы. С другой стороны, он является методом поиска маски области молекулы, совместной с экспериментальными данными. Найденная маска может быть использована далее в процедурах модификации электронной плотности для уточнения значений фаз. Требование связности искомой маски является существенным ограничением на допустимые наборы значений фаз и может быть само по себе использовано для уточнения значений фаз. Основой подхода является генерация случайным образом большого количества гипотетических связных масок области молекулы. Сгенерированная маска считается допустимой, если рассчитанные по ней модули структурных факторов достаточно точно воспроизводят их экспериментальные значения. Допустимые маски запоминаются для дальнейшего использования. Процесс генерации продолжается до тех пор, пока не набрано оговоренное количество (100 в наших тестах) допустимых масок. Наборы структурных факторов, отвечающих отобранным маскам, выравниваются и усредняются. Найденные значения фаз могут быть использованы для расчета синтеза Фурье или в качестве входных данных для программ уточнения значений фаз.

Мы называем маской бинарную функцию, определенную, обычно, на некоторой сетке в элементарной ячейке. Мы говорим, что маска связна, если любые две точки маски могут быть соединены путем, в котором каждые две следующие друг за другом точки являются соседними. (В наших тестах для каждой точки соседями считались шесть

ближайших к ней точек сетки). Рисунок 1 иллюстрирует процесс построения случайно маски. Важной характеристикой маски является ее объем (или количество точек сетки, включенных в маску). Мы характеризуем размер маски удельным объемом, который равен объему в маске, приходящемуся на один дальтон молекулярной массы объекта. В кристаллографических подходах, при оценке содержания растворителя в ячейке принято считать удельный объем области молекулы равным 1.23 $Å^3Da^{-1}$ [39, 40].

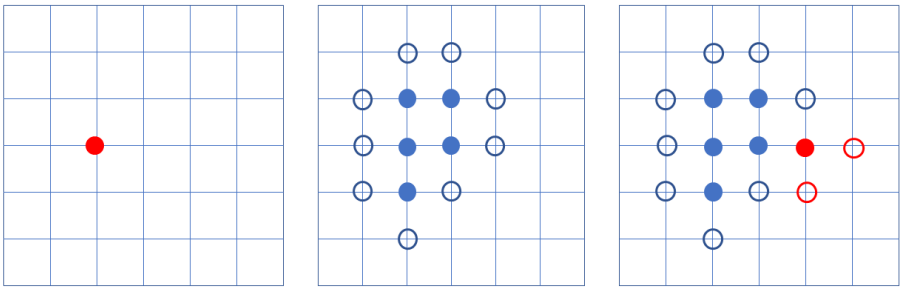



**Рис. 1**. Процедура генерации случайной маски. Маска строится точка за точкой. **a**) Начальная точка выбирается случайно. **b**) Для построенной частичной маски (показана закрашенными кругами) определяются граничные точки (показаны окружностями). **c**) Случайно выбранная граничная точка добавляется к маске. Корректируется состав множества граничных точек.

Процесс отбора допустимых масок регулируется критериями, определяющими, какие из сгенерированных масок считаются допустимыми. Простейший пример такого критерия – нецентрированная корреляция рассчитанных по маске модулей структурных факторов с экспериментально определенными величинами

$$CM_{(d_{\max},d_{\min})} = \frac{\sum_{\mathbf{s}} F^{obs}(\mathbf{s}) F^{mask}(\mathbf{s})}{\sqrt{\sum_{\mathbf{s}} \left(F^{obs}(\mathbf{s})\right)^2 \sum_{\mathbf{s}} \left(F^{mask}(\mathbf{s})\right)^2}}, \qquad (11)$$

где суммы рассчитываются по рефлексам, удовлетворяющим условию $d_{\max}^{-1} < |\mathbf{s}| \le d_{\min}^{-1}$. Сгенерированная маска считается допустимой, если эта величина превосходит заданный уровень $CM^{crit}$. Могут применяться и другие критерии отбора, например, статистическое правдоподобие маски.

При тестировании методов на объектах с известной структурой (т.е. с известными величинами фаз структурных факторов) можно непосредственно сравнить полученные значения фаз с их точными величинами. Популярной мерой, используемой при таком сравнении, является коэффициент корреляции двух распределений электронной плотности, рассчитанных с одними и теми же (экспериментальными) модулями структурных факторов и разными наборами фаз [43]. Этот показатель (Map Correlation Coefficient, *MCC*) может быть рассчитан, также, в терминах структурных факторов, как взвешенная сумма косинусов фазовых ошибок:

$$MCC_{(d_{\max},d_{\min})} = \frac{\sum_{\mathbf{s}} \left(F^{obs}(\mathbf{s})\right)^2 \cos\left(\varphi^{true}(\mathbf{s}) - \varphi^{calc}(\mathbf{s})\right)}{\sum_{\mathbf{s}} \left(F^{obs}(\mathbf{s})\right)^2}, \qquad (12)$$

где суммы рассчитываются по рефлексам, удовлетворяющим условию $d_{\max}^{-1} < |\mathbf{s}| \le d_{\min}^{-1}$.

Следует иметь в виду, что, в силу принципиальной неоднозначности решения фазовой проблемы, в процессе отбора случайно сгенерированных масок мы можем

получать допустимые маски, выглядящие как сильно различающиеся, но становящиеся похожими, если применить к одной из масок сдвиг и (или) смену энантиомера. Поэтому прежде, чем сравнивать или усреднять соответствующие отобранным маскам наборы фаз, эти наборы должны быть выровнены посредством преобразований тривиальной группы неоднозначности. В нашем случае эта группа включает все сдвиги начала координат и переход к энантиомеру. Помимо тривиальной неоднозначности могут существовать более сложные случаи неоднозначности – "гомометрические решения". Их наличие может быть обнаружено посредством анализа дендрограммы, соответствующей процедуре кластерного анализа отобранных вариантов. После того, как наборы структурных факторов, соответствующих отобранным маскам, выровнены, для каждого рефлекса могут быть определены "наилучшая" фаза $\varphi^{best}(\mathbf{s})$ и показатель достоверности $m(\mathbf{s})$

$$m(\mathbf{s})\exp\left[i\varphi^{best}(\mathbf{s})\right]=\sum_{j=1}^{M}\exp\left[i\varphi_j(\mathbf{s})\right]. \qquad (13)$$

Здесь $\varphi_j(\mathbf{s})$ – фазы структурных факторов, соответствующие $j$-той из $M$ отобранных масок.

Рисунок 2 демонстрирует качество наборов фаз, полученных после выравнивания и усреднения 100 наборов фаз, соответствующим маскам, отобранным на основе величины корреляции модулей (11) в тестах [37] с известной структурой фотосистемы II [44]. Разные кривые соответствуют разным пороговым значениям для коэффициента корреляции, использованным при отборе. Точность определения фаз растет с повышением жесткости отбора масок. Следует отметить, что даже при отсутствии отбора масок (нижняя кривая на графике), усреднение фаз, отвечающих случайным связным маскам правильного объема, приводит к разумным значениям фаз. Эти фазы, будучи использованы вместе с правильными значениями модулей, приводят к довольно большому значению коэффициента корреляции полученного синтеза Фурье с точным. Варьирование величины заданного удельного объема маски показывает, что при работе с данными низкого разрешения оптимальное значение этого параметра может несколько превосходить стандартную оценку 1.23 $Å^3Da^{-1}$.

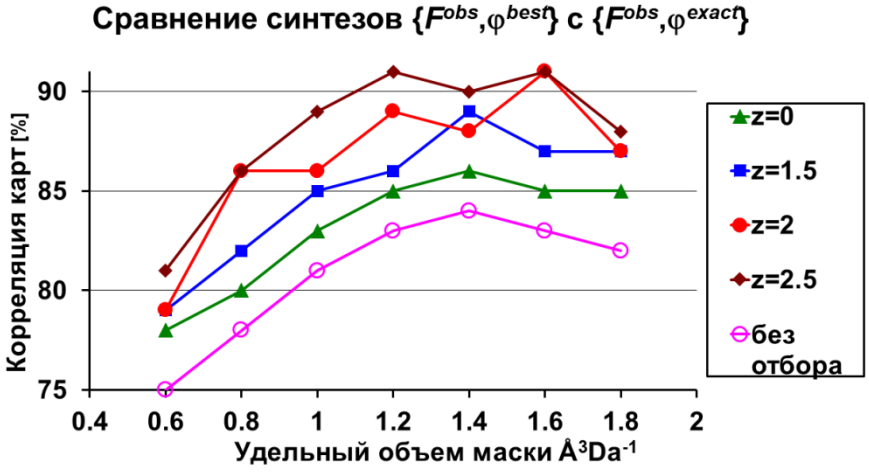
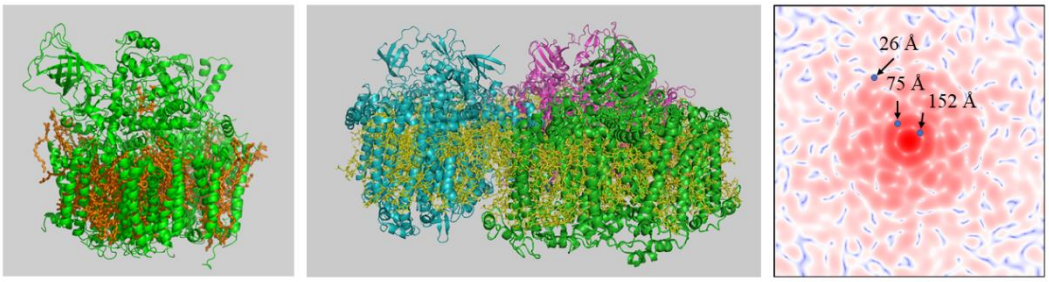



**Рис. 2**. Результаты тестового определения значений фаз для смоделированных данных для мономера фотосистемы II (рис. 3). Коэффициент корреляции карт для найденных значений фаз и точных значений показан как функция удельного объема использованных в работе масок. При отборе допустимых масок нижний уровень приемлемой корреляции модулей структурных факторов определялся в виде $CM^{crti}$ = MEAN($CM$) + z·RMSD($CM$), где среднее и RMSD величины для величины CM определялись по всем сгенерированным маскам. (См. [37]).



*Математическая биология и биоинформатика*. 2020. Т. 15. № S. doi: 10.17537/2020.15.t1




**a**) **b**) **c**)

**Рис. 3.** Объекты, использованные в тестах: **a**) Мономер фотоситемы II ([44], PDB-код 3kzi, *MW* = 300 kDa). **b**) Тример фотосистемы I ([45], PDB-код 1jbo, *MW* = 1068 kDa). **c**) Одно из сечений трехмерного набора смоделированных данных для тримера фотосистемы I. Значения интенсивностей $I(\mathbf{s})$ представлены в логарифмической шкале.

Найденные значения фаз могут далее быть уточнены, а набор включенных в работу рефлексов расширен [37, 41, 46]. При генерации случайной маски построение маски идет точка за точкой (рис. 1). Выбор новой точки идет из множества граничных точек для уже построенной частичной маски. Этот выбор может осуществляться либо с равной вероятностью для всех точек, либо с учетом предварительно заданного трехмерного набора значений вероятностей того, что данная точка принадлежит области молекулы. Априорное распределение вероятностей может быть построено на базе полученного на предыдущем шаге синтеза Фурье электронной плотности $\rho(\mathbf{r})$ в виде

$$P(\mathbf{r}) = C\exp\left[\kappa\rho(\mathbf{r})\right], \qquad (14)$$

где $\kappa$ – параметр, определяющий "остроту" пиков распределения вероятностей. Основу процедуры итерационного уточнения набора фаз и расширения количества вовлеченных в работу рефлексов составляет определение на очередном шаге фаз априорных распределений вероятностей. Эти фазы строятся на основе синтеза Фурье электронной плотности, вычисленного по результатам предыдущего шага.

На рисунке 4 показан прогресс в определении значений фаз в результате 20 шагов уточнения их значений. Для наборов фаз, полученных на разных этапах уточнения, показана их корреляция с точными значениями, рассчитанная как функция разрешения рефлексов.

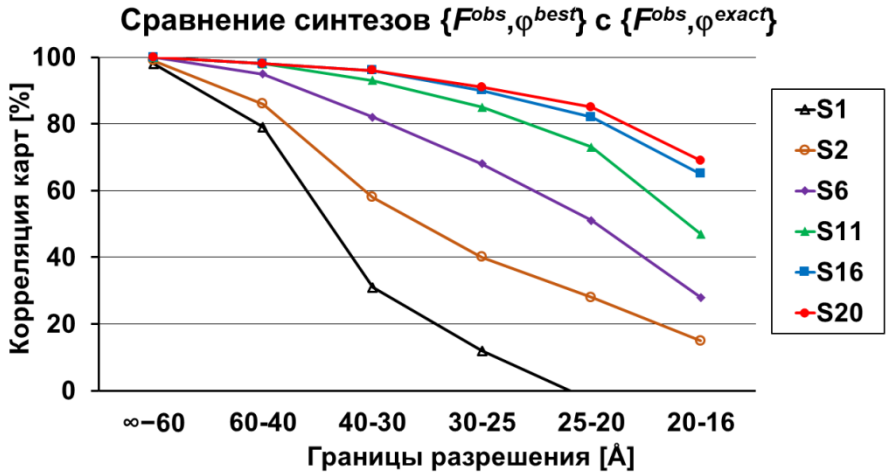
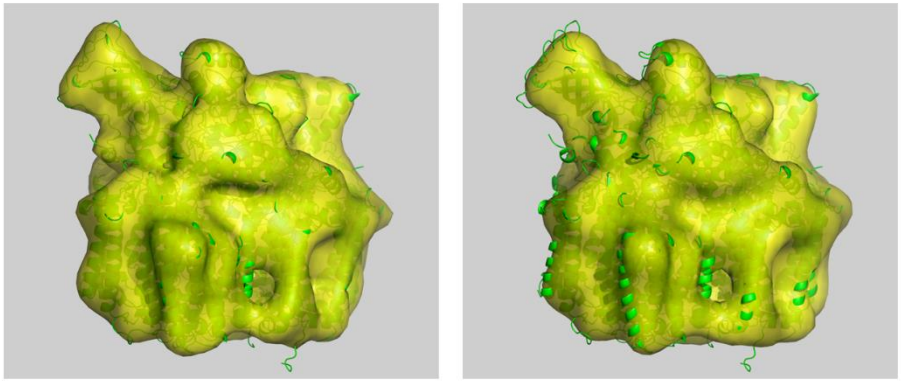



**Рис. 4**. Результат уточнения значений фаз и расширения набора рефлексов при тестировании метода на смоделированных данных для мономера фотосистемы II (рис. 3). Корреляция найденных значений фаз с точными как функция разрешения показана для разных шагов работы [41].

Рисунок 4 демонстрирует прогресс по ходу 20 шагов работы по уточнению значений фаз и расширению набора структурных факторов с 25 до 16 Å. Графики показывают корреляцию найденных значений фаз с точными для последовательных сферических оболочек в обратном пространстве. На рисунке 5 показано изображение частицы, построенное на основе точного синтеза разрешения 16 Å и синтеза, рассчитанного с использованием точных значений модулей и найденных на последнем шаге значений фаз.



**a**) **b**)

**Рис. 5**. Изображения частицы, построенные на основе синтезов Фурье разрешения 16 Å, рассчитанных с точными значениями модулей, но разными значениями фаз структурных факторов. На обоих рисунках поверхность ограничивает область, имеющую удельный объем 1.23 $Å^3Da^{-1}$. Показана модель, соответствующая белковой части комплекса. **a**) Точные значения фаз. **b**) Фазы получены на шаге S20 процедуры определения фаз [41].

## 6. ВОССТАНОВЛЕНИЕ ПОТЕРЯННЫХ РЕФЛЕКСОВ

При работе с данными, полученными в дифракционном эксперименте, часто встречается ситуация, когда не удается измерить интенсивности рефлексов самого низкого разрешения (центральной зоны обратного пространства), а часть наиболее сильных рефлексов оказывается переэкспонированной. Значения модулей структурных факторов для таких рефлексов (равно как и соответствующие фазы) могут быть восстановлены в рамках предложенной процедуры и использованы далее для расчета синтезов Фурье электронной плотности. В процессе работы при вычислении оценки соответствия маски экспериментальным данным, естественно, могут быть использованы только зарегистрированные в эксперименте значения. Однако, при наличии маски могут быть рассчитаны значения и модулей, и фаз структурных факторов для всех рефлексов рассматриваемой зоны разрешения. Поэтому, модули потерянных в эксперименте структурных факторов могут быть оценены средними значениями величин модулей, рассчитанных по отобранным допустимым маскам. Пример такой оценки показан на рисунке 6. В этом тесте с модельными данными для тримера фотосистемы I величины модулей структурных факторов для 182 рефлексов центральной зоны (136 Å) считались неизвестными. Помимо этого, неизвестными считались и значения модулей для 66 сильнейших рефлексов в оставшейся части обратного пространства. На диаграмме приведены точные и восстановленные значения модулей для списка из 248 рефлексов, упорядоченных по разрешению. Значение стандартного R-фактора между точными и восстановленными значениями составило 7 %. Средняя фазовая ошибка для восстановленных структурных факторов составила 12 градусов.



*Математическая биология и биоинформатика*. 2020. Т. 15. № S. doi: 10.17537/2020.15.t1

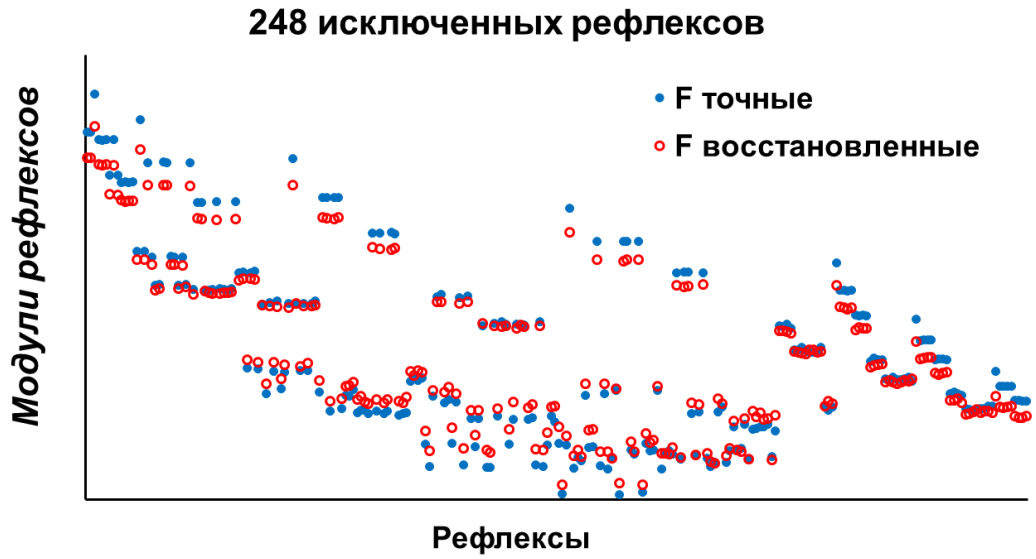



**Рис. 6**. Результат восстановления неизвестных значений модулей структурных факторов в тесте со смоделированными данными для тримера фотосистемы I (рис.3). Точные и восстановленные значения модулей показаны для 248 рефлексов, объявленных неизвестными. Рефлексы показаны в порядке повышения разрешения (возрастания величины $|\boldsymbol{s}|$).

## 7. ФУНКЦИЯ ЗАХВАТА МОДЕЛИ КАК ОЦЕНКА КАЧЕСТВА СИНТЕЗА ФУРЬЕ

Существует распространенная (хотя и серьёзно критикуемая [47, 48]) практика рассматривать значения фаз для некоторой оболочки обратного пространства как надежные, если коэффициент корреляции карт в этой оболочке превышает 50 %. В Таблице 1 для трех решений, полученных на различных шагах процедуры восстановления значений фаз для фотосистемы II, приведены коэффициенты корреляции с точными фазами в последовательных оболочках обратного пространства. Основываясь на этих коэффициентах, разрешение синтеза Фурье, полученного на шагах S6 и S11, может быть оценено как 20 Å, а для шага S20 как 16 Å. Однако оценка разрешения будет другой, если мы попробуем подойти к оценке разрешения с позиций полезности синтеза при построении атомной модели структуры [41].

**Таблица 1.** Коэффициент корреляции карт в сферических оболочках обратного пространства и оценка разрешения финальных карт распределения электронной плотности

| Шаг | Границы оболочки (Å) | | | | | | Оценка разрешения (Å) | |
|---|---|---|---|---|---|---|---|---|
| | ∞–60 | 60–40 | 40–30 | 30–25 | 25–20 | 20–16 | по *MCC** | по *MTF*** |
| S6 | 100 | 95 | 82 | 68 | 51 | 28 | 20 | 40 |
| S11 | 100 | 98 | 93 | 85 | 73 | 47 | 20 | 30 |
| S20 | 100 | 98 | 96 | 91 | 85 | 69 | 16 | 25 |
| Число рефлексов | 85 | 170 | 363 | 436 | 1028 | 2026 | | |

* Map Correlation Coefficient (12).
** Model Trapping Function (15)

Синтез Фурье удобен для построения модели, если область молекулы (определённая как область наиболее высоких значений синтеза) содержит позиции атомов объекта, но не включает области пространства, занятые растворителем. Такая ситуация имеет место

для синтезов Фурье высокого разрешения, но недостижима при низком разрешении. На рисунке 7 показаны области $\{\mathbf{r}:\rho(\mathbf{r})>\rho_{crit}\}$ одинакового объема на картах, построенных по точным синтезам Фурье разного разрешения. Можно видеть, что, например, при разрешении 3 Å такая область хорошо локализует положения атомов. Если разрешение синтеза ухудшается, то при сохранении объема области позиции некоторых атомов начинают вылезать из области. Количество таких атомов возрастает с понижением разрешения. Мы можем оценить качество такой области количеством атомов, захваченных ею. Изменение значения уровня $\rho_{crit}$ меняет размер области $\{\mathbf{r}:\rho(\mathbf{r})>\rho_{crit}\}$ и количество захваченных ею атомов. Поэтому мы будем характеризовать качество синтеза Фурье $\rho(\mathbf{r})$ функцией захвата модели (Model Trapping Function, *MTF*), вычисляемой как функция удельного объема κ области $\Omega_\kappa=\{\mathbf{r}:\rho(\mathbf{r})>\rho_{crit}\}$ и показывающей долю атомов модели, захваченных этой областью

$$MTF(\kappa)=\frac{\text{число атомов в области }\Omega_\kappa}{\text{общее число атомов в модели}}. \qquad (15)$$

При построении области $\Omega_\kappa$ критический уровень $\rho_{crit}$ выбирается таким, чтобы иметь объем области равным κ. Естественно, такая функция может быть рассчитана только при тестировании метода, когда атомная модель объекта известна. Функция захвата модели показывает, как доля захваченных областью атомов растет с ростом удельного объема области. На рисунке 8 показаны кривые захвата модели для точных синтезов Фурье разного разрешения и модели комплекса фотосистемы I.

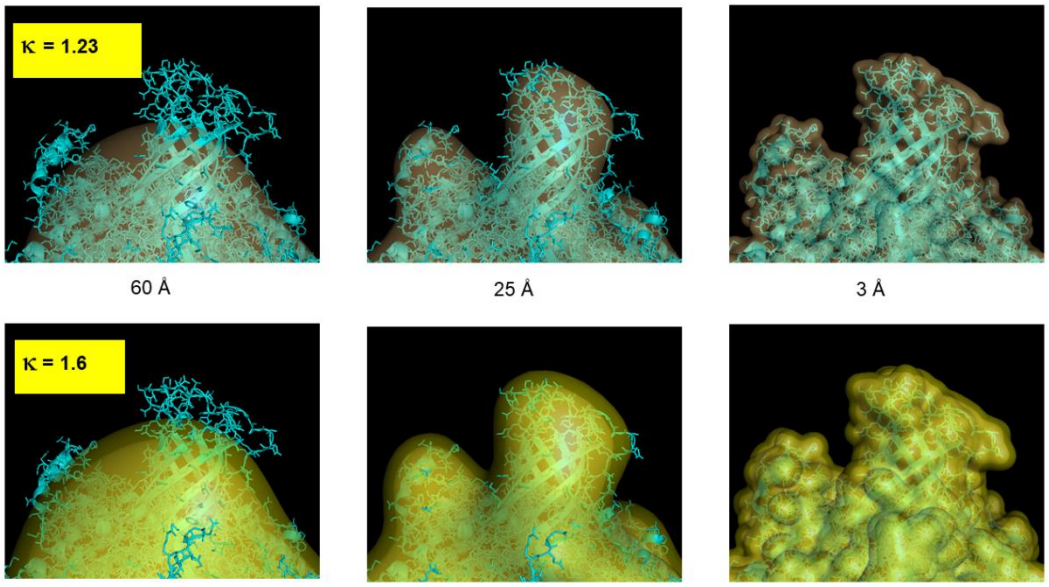



**Рис. 7**. Область молекулы, построенная с использованием точных синтезов Фурье возрастающего (слева направо) разрешения. Показаны области, отвечающие двум различным значениям удельного объема κ.

Введенная функция захвата модели может быть использована для оценки разрешения наборов фаз, полученных на разных шагах процедуры восстановления значений фаз при работе с тестовыми данными для фотосистемы I. Для каждого набора фаз можно рассчитать соответствующий синтез Фурье и кривую захвата модели этим синтезом. Полученная кривая может быть сравнена с набором эталонных кривых, отвечающих точным синтезам Фурье разного разрешения. На рисунке 9 показано, что, например, кривая, отвечающая синтезу Фурье, рассчитанному на шаге S6, близка к кривой,

отвечающей точному синтезу разрешения 40 Å. Поэтому результат шага S6 может быть оценен разрешением 40 Å с точки зрения пригодности этого синтеза для построения атомной модели. Аналогично, разрешение наборов фаз, полученных на шагах S11 и S20, может быть оценено как 30 и 25 Å соответственно (Таблица 1). Эти оценки далеки от традиционных оценок, полученных на основе правила 50 % и, скорее, склоняют к использованию правила 90 % корреляции для оценки качества набора фаз.

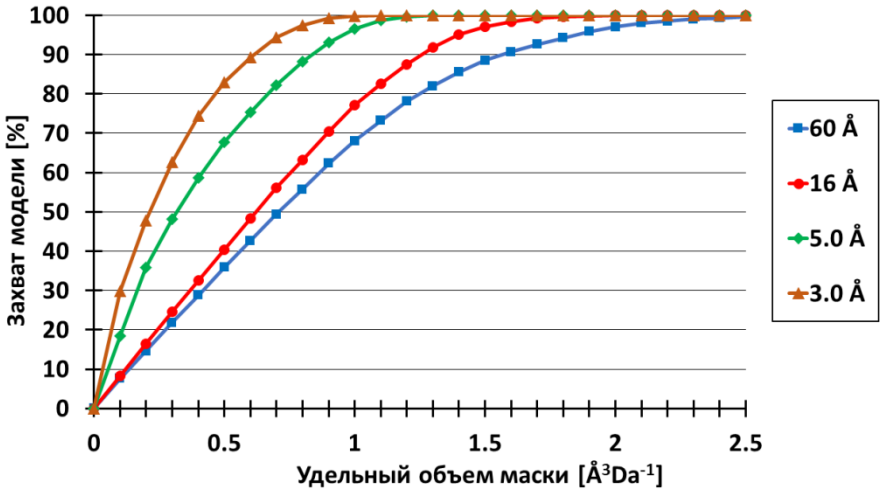



**Рис. 8**. Кривые захвата модели, отвечающие точным синтезам Фурье разного разрешения, рассчитанные с модельными данными для фотосистемы I [41].

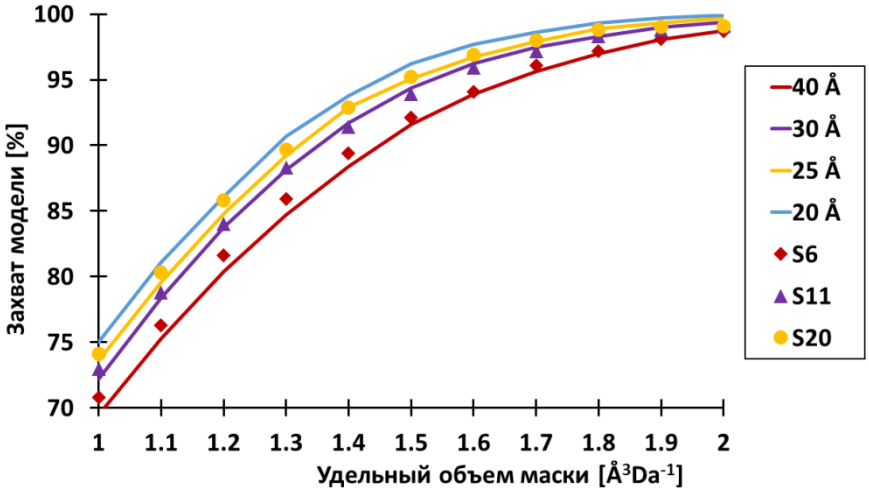



**Fig. 9**. Оценка разрешения для результатов тестового определения фаз для фотосистемы I на основе функции захвата модели. Кривые захвата модели для различных шагов процедуры определения фаз показаны маркерами. Эталонные кривые захвата модели для точных синтезов Фурье разного разрешения показаны сплошной линией. [41].

## 8. ЗАКЛЮЧЕНИЕ

В целом, решение фазовой проблемы при исследовании одиночных частиц осуществляется теми же методами, что и в биологической кристаллографии. Основное различие заключается в том, что возможность регистрация непрерывной картины рассеяния при исследовании одиночных частиц приводит к существенной избыточности экспериментальных данных, что, в свою очередь, повышает эффективность методов

решения фазовой проблемы и открывает возможность восстановления частей дифракционной картины, не зарегистрированных в конкретном эксперименте. Избыточность данных может быть использована различными путями, например, посредством методов модификации электронной плотности, интерполяционной формулы Шеннона и ее модификаций, а также на основе использования связных бинарных аппроксимаций исследуемого объекта.

## REFERENCES


1. Adams P.D., Afonine P.V., Bunkóczi G., Chen V.B., Davis I.W., Echols N., Headd J.J., Hung L.-W., Kapral G.J., Grosse-Kunstleve R.W. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica D*. 2010. V. 66. P. 213–221. doi: 10.1107/S0907444909052925.
2. Winn M.D., Ballard C.C., Cowtan K.D., Dodson E.J., Emsley P., Evans P.R., Keegan R.M., Krissinel E.B., Leslie A.G.W., McCoy A. et al. Overview of the CCP4 suite and current developments. *Acta Crystallographica D*. 2011. V. 67. P. 235–242. doi: 10.1107/S0907444910045749.
3. Sheldrick G.M. A short history of SHELX. *Acta Crystallographica A*. 2008. V. 64. P. 112–122. doi: 10.1107/S0108767307043930.
4. Bricogne G., Vonrhein C., Flensburg C., Schiltz M., Paciorek W. Generation, representation and flow of phase information in structure determination: recent developments in and around SHARP 2.0. *Acta Crystallographica D*. 2003. V. 59. P. 2023–2030. doi: 10.1107/S0907444903017694.
5. Blanc E., Roversi P., Vonrhein C., Flensburg C., Lea S.M., Bricogne G. Refinement of severely incomplete structures with maximum likelihood in BUSTER-TNT. *Acta Crystallographica D*. 2004. V. 60. P. 2210–2221. doi: 10.1107/S0907444904016427.
6. Minor W., Cymborowski M., Otwinowski Z., Chruszcz M. HKL-3000: the integration of data reduction and structure solution - from diffraction images to an initial model in minutes. *Acta Crystallographica D*. 2006. V. 62. P. 859–866. doi: 10.1107/S0907444906019949.
7. Spence J.C.H. XFELs for structure and dynamics in biology. *IUCrJ*. 2017. V. 4. P. 322–339. doi: 10.1107/S2052252517005760.
8. Standfuss J., Spence J. Serial crystallography at synchrotrons and X-ray lasers. *IUCrJ*. 2017. V. 4. P. 100–101. doi: 10.1107/S2052252517001877.
9. Aquila A., Barty A., Bostedt C., Boutet S., Carini G., dePonte D., Drell P., Doniach S., Downing K.H., Earnest T. et al. The linac coherent light source single particle imaging road map. *Structural Dynamics*. 2015. V. 2. doi: 10.1063/1.4918726.
10. Ayyer K., Geloni G., Kocharyan V., Saldin E., Serkez S., Yefanov O., Zagorodnov I. Perspectives for imaging single protein molecules with the present design of the European XFEL. *Structural Dynamics*. 2015. V. 2. doi: 10.1063/1.4919301.
11. Daurer B.J., Okamoto K., Bielecki J., Maia F.R.N.C., Muhlig K., Seibert M.M., Hantke M.F., Nettelblad C., Benner W.H., Svenda M. et al. Experimental strategies for imaging bioparticles with femtosecond hard X-ray pulses. *IUCrJ*. 2017. V. 4. P. 251–262. doi: 10.1107/S2052252517003591.
12. Lunin V.Y., Lunina N.L., Petrova T.E. The biological crystallography without crystals. *Mathematical Biology and Bioinformatics*. 2017. V. 12. No. 1. P. 55–72. doi: 10.17537/2017.12.55.
13. Lunin V.Y. Mask-based approach to restoring and phasing single-particle diffraction data. In: *32nd European Crystallographic Meeting, Vienna, Austria, August 18-23*: Abstract Booklet. 2019. P. 138.

14. Lunin V.Y., Lunina N.L., Petrova T.E. Single particle study by X-ray diffraction: Crystallographic approach. *Mathematical Biology and Bioinformatics*. 2019. V. 14. No. 2. P. 500–516. doi: 10.17537/2019.14.500.
15. Urzhumtseva L., Klaholz B., Urzhumtsev A. On effective and optical resolutions of diffraction data sets. *Acta Crystallographica D*. 2013. V. 69. P. 1921–1934. doi: 10.1107/S0907444913016673.
16. Kucukelbir A., Sigworth F.J., Tagare H.D. Quantifying the local resolution of cryo-EM density maps. *Nature Methods*. 2014. V. 11. P. 63–65. doi: 10.1038/nmeth.2727.
17. Afonine P.V., Klaholz B.P., Moriarty N.W., Poon B.K., Sobolev O.V., Terwilliger T.C., Adams P.D., Urzhumtsev A. *Acta Crystallographica D*. 2018. V. 74. P. 814–840. doi: 10.1107/S2059798318009324.
18. Meijering E. A chronology of interpolation: from ancient astronomy to modern signal and image processing. *Proceedings of the IEEE*. 2002. V. 90. P. 319–342. doi: 10.1109/5.993400.
19. Kotel'nikov V.A. On the transmission capacity of 'ether' and wire in electric communications. *Physics-Uspekhi*. 2006. V. 49. № 7. P. 736–744. doi: 10.1070/PU2006v049n07ABEH006160.
20. Sayre D. Some implications of a theorem due to Shannon. *Acta Crystallographica*. 1952. V. 5. P. 843. doi: 10.1107/S0365110X52002276.
21. Bricogne G. Geometric sources of redundancy in intensity data and their use for phase determination. *Acta Crystallographica A*. 1974. V. 30. P. 395–405. doi: 10.1107/S0567739474010722.
22. Bricogne G. Methods and programs for direct-space exploitation of geometric redundancies. *Acta Crystallographica A*. 1976. V. 32. P. 832–847. doi: 10.1107/S0567739476001691
23. Lunin V.Y., Lunina N.L. Repairing of the diffraction pattern in the X-ray free electron laser study of biological particles. *Advanced Mathematical Models & Applications*. 2018. V. 3. P. 117–127.
24. Lunin V.Y. Use of the fast differentiation algorithm for phase refinement in protein crystallography. *Acta Crystallographica. A*. 1985. V. 41. P. 551–556. doi: 10.1107/S0108767385001209.
25. Podjarny A.D., Rees B., Urzhumtsev A.G. Density modification in X-ray crystallography. In: *Methods in Molecular Biology, Crystallographic Methods and Protocols*. Eds. Jones C., Milloy B, Sanderson M.R. Totowa, New Jersey: Humana Press, 1996. P. 205-226. (Methods in Molecular Biology, Vol. 56.). doi: 10.1385/0-89603-259-0:205.
26. Zhang K.Y.J., Cowtan K.D., Main P. Phase improvement by iterative density modification. In: *International Tables for Crystallography*. Vol. F. Eds. Arnold E., Himmel D.M., Rossmann M.G. Chichester: John Wiley and Sons, 2012. P. 385–400. doi: 10.1107/97809553602060000847.
27. Wang B.C. Resolution of phase ambiguity in macromolecular crystallography. *Methods in Enzymology*. 1985. V. 115. P. 90–111. doi: 10.1016/0076-6879(85)15009-3.
28. 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.
29. 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.
30. Millane R., Lo V.L. Iterative projection algorithms in protein crystallography. I. Theory. *Acta Crystallographica A*. 2013. V. 69. P. 517–527. doi: 10.1107/S0108767313015249.
31. Abrahams J.P. Bias reduction in phase refinement by modified interference functions: introducing the γ-correction. *Acta Crystallographica D*. 1997. V. 53. P. 371–376. doi: 10.1107/S0907444996015272.

32. Oslányi G., Sütő A. *Ab initio* structure solution by charge flipping. *Acta Crystallographica A*. 2004. V. 60. P. 134–141. doi: 10.1107/S0108767303027569.
33. 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.
34. Urzhumtsev A.G. *The use of local averaging in analysis of macromolecule images at electron density distribution maps*: Preprint. Pushchino, 1985 (in Russ.).
35. Urzhumtsev A.G., Lunin V.Y., Luzyanina T.B. Bounding a Molecule in a Noisy Synthesis. *Acta Crystallographica A*. 1989. V. 45. P. 34–39. doi: 10.1107/s0108767388008955.
36. 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.
37. 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 D*. 2016. V. 72. P. 147–157. doi: 10.1107/S2059798315022652.
38. Lunin V.Y., Lunina N.L., Petrova T.E. The use of connected masks for reconstructing the single particle image from X-ray diffraction data. *Mathematical Biology and Bioinformatics*. 2014. V. 10. № Suppl. P. t1–t19. doi: 10.17537/2015.10.t1.
39. Matthews B.W. Solvent Content of Protein Crystals. *Journal of Molecular Biology*. 1968. V. 33. P. 491–497. doi: 10.1016/0022-2836(68)90205-2.
40. Weichenberger C.X., Afonine P.V., Kantardjieff K., Rupp B. *Acta Crystallographica D*. 2015. V. 71. P. 1023–1038. doi: 10.1107/S1399004715006045.
41. 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. II. Likelihood-based selection criteria. *Acta Crystallographica D*. 2019. V. 75. P. 79–89. doi: 10.1107/S2059798318016959.
42. Lunina N.L., Petrova T.E., Urzhumtsev A.G., Lunin V.Y. The Use of Connected Masks for Reconstructing the Single Particle Image from X-Ray Diffraction Data. III. Maximum-Likelihood Based Strategies to Select Solution of the Phase Problem. *Mathematical Biology and Bioinformatics*. 2018. V. 13. № Supl. P. t70–t83. doi: 10.17537/2018.13.t70.
43. Lunin V.Y., Woolfson M.M. Mean Phase Error and the Map Correlation Coefficient. *Acta Crystallographica D*. 1993. V. 49. P. 530–533. doi: 10.1107/S0907444993005852.
44. 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.
45. Jordan P., Fromme P., Witt H.T., Klukas O., Saenger W., Krauß N. Three-dimensional structure of cyanobacterial photosystem I at 2.5 Å resolution. *Nature*. 2001. V. 411. P. 909–917. doi: 10.1038/35082000.
46. Lunina N.L., Petrova T.E., Urzhumtsev A.G., Lunin V.Y. The use of connected masks for reconstructing the single particle image from X-ray diffraction data. II. The dependence of the accuracy of the solution on the sampling step of experimental data. *Mathematical Biology and Bioinformatics*. 2015. V. 10. № Suppl. P. t56–t72. doi: 10.17537/2015.10.t56.
47. Van Heel M., Schatz M. Fourier shell correlation threshold criteria. *J. Struct. Biol*. 2005. V. 151. P. 250–262. doi: 10.1016/j.jsb.2005.05.009.
48. Van Heel M., Schatz M. Reassessing the Revolution's Resolutions. *bioRxiv*. 2017. Article No. 224402. doi: 10.1101/224402.


Опубликовано 05.06.2020.

***Ttranslation into Russian of the original article published in English:***
***Lunin V., Lunina N., Petrova T. Mathematical Biology and Bioinformatics. 2020;15(1):57–72.***
***doi: 10.17537/2020.15.57***

# Mask-based approach in phasing and restoring of single-particle diffraction data

**Lunin V.Y., Lunina N.L., Petrova T.E.**

*Institute of Mathematical Problems of Biology RAS, Keldysh Institute of Applied Mathematics of Russian Academy of Sciences, Pushchino, Moscow Region, 142290 Russia*

***Abstract.*** The development of experimental techniques, in particular the emergence of the X-ray free-electron lasers, allows one to register the scattering from an isolated particle and, thereby, opens a door to the study of a fine three-dimensional structure of non-crystalline biological objects by X-ray diffraction methods. The possibility to measure non-Bragg reflections makes experimental data mutually dependent and essentially simplifies the structure solution. The sampling of experimental scattering data to a sufficiently fine grid makes the structure determination equivalent to phasing of structure factor magnitudes for a 'virtual' crystal with extremely large solvent content. This makes density modification phasing methods especially powerful supposing the object envelope is known. At the same time, such methods may be sensitive to the accuracy of the predefined envelope and completeness of experimental data and may suffer from non-uniqueness of the solution of the phase problem. The mask-based approach is a preliminary phasing method that performs random search for connected object envelopes possessing of the structure factor magnitudes close to the values observed in X-ray experiment. The alignment and averaging of the phase sets corresponding to selected putative envelopes produce an approximate solution of the phase problem. Beside the estimation of unknown phase values this approach allows one to estimate the values of structure factor magnitudes lost in the experiment, e.g. those corresponding to beam-stop shade zone or to oversaturated reflections.

***Key words:*** *biological macromolecules, single-particles, X-ray scattering, X-ray free electron lasers, phase problem, magnitude retrieval, effective resolution.*