

## SASCUBE – ОБНОВЛЕННЫЙ МЕТОД КУБИКОВ ДЛЯ РАСЧЕТА ИНТЕНСИВНОСТИ РЕНТГЕНОВСКОГО РАССЕЯНИЯ БИОПОЛИМЕРОВ В РАСТВОРЕ

© 2019 г. Б.А. Федоров\*, А.В. Смирнов\*, В.В. Ярошенко\*, Ю.Б. Порозов\* \*\*

\*Санкт-Петербургский Национальный исследовательский университет информационных технологий, механики и оптики, 197101, Санкт-Петербург, Кронверкский просп. 49

\*\*Первый Московский государственный медицинский университет им. И.М. Сеченова, 119991, Москва, ул. Трубецкая, 8

E-mail: borfedorov1@gmail.com

Поступила в редакцию 03.07.18 г.

После доработки 11.10.18 г.

Принята к публикации 29.10.18 г.

Описан обновленный метод кубиков, который позволяет рассчитывать кривые SAS для биополимеров в растворе на основе координат их атомов без введения подгоночных параметров. Эффективность метода проверена на ряде модельных систем. Для восьми глобулярных белков проведено сравнение интенсивностей рассеяния, рассчитанных по программе SASCUBE (<https://sourceforge.net/projects/sascube>), с соответствующими экспериментальными кривыми. Рассмотрен дискуссионный вопрос о характере искажений электронной плотности воды вблизи поверхности белка. С помощью молекулярно-динамических симуляций показано, что распределение электронной плотности воды в приграничном слое белков толщиной 0,3 нм никак не может рассматриваться как однородное, а ее среднее значение может быть меньше средней электронной плотности воды. Проиллюстрировано заметное влияние конформационной подвижности белков на их кривую рассеяния. Обсуждается возможность совместного использования метода кубиков (программы SASCUBE) и молекулярной динамики для получения интенсивностей рассеяния белков с учетом всех факторов влияния на кривую рассеяния (конформационная подвижность белков, искажения структур воды и др.).

*Ключевые слова:* малоугловое рентгеновское рассеяние, молекулярная динамика, структура белков в растворе, структура воды.

**DOI:** 10.1134/S000630291901006X

Несколько десятилетий назад одним из авторов с сотрудниками [1–4] был разработан метод кубиков – строгий метод расчета интенсивности рассеяния глобулярными белками в растворе исходя из координат их неводородных атомов в кристалле. Были рассчитаны интенсивности для ряда белков в области малых и средних углов рассеяния [1–5], однако сравнение полученных кривых с экспериментом в те времена было весьма ограничено в силу отсутствия соответствующих экспериментальных кривых в широкой области углов рассеяния. Кроме того, расчет кривых рассеяния требовал большого «машинного» времени, что объясняется, разумеется, низкой производительностью вычислительных машин тех времен.

В настоящее время указанные ограничения фактически сняты: быстродействие современных компьютеров на много порядков превышает возможности вычислительных машин 70-х – 80-х годов прошлого века, а успехи в выделении и очистке белков в сочетании с прогрессом рентгеновской техники (прежде всего, с использованием синхротронного излучения) позволили создать обширную базу экспериментальных кривых рассеяния белков в растворе (см., например, базы данных BioI sis [6] и SASBDB [7]).

За истекшие годы были развиты новые методы расчета интенсивности рассеяния биополимеров в растворе [8–13]. И хотя эти методы дают, как правило, хорошее согласие с экспериментом, они содержат в своей основе некоторые принципиальные недостатки, которые, по-видимому, компенсируются введением по крайней мере двух подгоночных параметров.

Сокращение: МД – молекулярная динамика.

К сравнению теоретических основ этих методов и метода кубиков мы обратимся в разделе «Результаты и обсуждение».

Ниже мы рассмотрим обновленный метод кубиков. Хотя его основы остались практически неизменными по сравнению с нашими старыми работами, его программная реализация существенно изменилась, стала более обоснованной и «прозрачной» с точки зрения теории дифракции макромолекул в растворе. И, естественно, существенно изменилось время расчета интенсивностей рассеяния: если ранее для получения кривой рассеяния для глобулярного белка уходили часы, то сейчас при использовании многоядерных компьютеров с высокой производительностью требуются десятки секунд или минуты.

Программа SASCUBE, доступная по адресу <https://sourceforge.net/projects/sascube>, позволяет рассчитывать интенсивности рассеяния белков в растворе по координатам атомов, заданных в формате базы данных PDB.

В геометрической части программа позволяет:

- выбирать длину ребра кубика;
- выбирать радиус молекулы растворителя;
- выбирать один из вариантов радиусов атомов и атомных групп, входящих в состав белка;
- определять число кубиков молекулярной и доступной поверхностей;
- вычислять молекулярный объем и гидратированный объем, заключенный внутри доступной поверхности (объем, недоступный для центров молекул растворителя);
- рассчитывать и строить сечения белка и доступной поверхности поперек любой координатной оси с шагом, равным длине ребра кубика;

Вторая часть программы, выполняющая расчет интенсивности рассеяния на основе описанного ниже метода кубиков, включает:

- выбор интервала углов рассеяния и количество точек для расчета интенсивности;
- выбор числа точек в обратном пространстве для получения усредненной по ориентациям интенсивности рассеяния;
- численное представление интенсивности рассеяния и построение графика;
- расчет радиуса инерции белка в выбранном растворителе.

## МЕТОД РАСЧЕТА

Интенсивность рассеяния макромолекулы, погруженной в растворитель, существенно отличается от интенсивности рассеяния той же макромолекулы в вакуумном окружении. Это объясняется прежде всего уменьшением разности электронных плотностей между макромолекулой и ее окружением, а также неоднородным распределением электронной плотности в самой макромолекуле. Если предположить, что известны координаты всех неводородных атомов макромолекулы, а растворитель является однородной субстанцией с электронной плотностью  $\rho_0$ , то интенсивность рассеяния  $I(\mathbf{q})$  такой макромолекулы в точке  $\mathbf{q}$  обратного пространства может быть представлена в виде [14]:

$$I(\mathbf{q}) = |F(\mathbf{q}) - \rho_0(\mathbf{q})|^2, \quad (1)$$

где  $F(\mathbf{q})$  – амплитуда рассеяния от однородной частицы, имеющей объем и форму макромолекулы.

Интенсивность  $I(\mathbf{q})$ , усредненная в обратном пространстве по сфере радиуса  $q$  (что соответствует свободной ориентации частицы в реальном пространстве), равна

$$I(q) = \langle I(\mathbf{q}) \rangle, \quad (2)$$

где  $\mathbf{q}$  – вектор обратного пространства ( $|\mathbf{q}| = q = 4\pi \sin \theta / \lambda$ ,  $\lambda$  – длина волны рентгеновских лучей,  $2\theta$  – угол рассеяния),  $\rho_0$  – электронная плотность растворителя;

$$F(\mathbf{q}) = \sum_k f_k(q) \exp(i\mathbf{q}\mathbf{r}_k) \quad (3)$$

– амплитуда рассеяния макромолекулы в вакууме,  $\mathbf{r}_k$  – радиус-вектор ее  $k$ -ого атома,  $f_k(q)$  – атомный фактор рассеяния этого атома.

В настоящей работе амплитуду  $f_k(q)$  для отдельных атомов рассчитывали из данных International Tables for Crystallography [15] путем их аппроксимации согласно формуле

$$f_k(q) = \sum_{m=0}^5 a_{k,m} q^{2m}. \quad (4)$$

Коэффициенты  $a_{k,m}$  были вычислены по методу наименьших квадратов. При этом относительное отклонение аппроксимации (4) от данных работы [15] в диапазоне значений  $q = 0 \dots 20 \text{ нм}^{-1}$  не превышает  $10^{-4}$ . Если атом связан с одним или несколькими водородными атомами, образуя атомную группу (например, группу  $\text{CH}_n$ ), амплитуду рассеяния этой группы рассчитывали по формуле

$$f_{\text{сн}_n}(q) = f_{\text{с}}(q) + n f_{\text{н}}(q). \quad (5)$$

В дальнейшем атомы и атомные группы рассматривали как сферы, центры которых совпадают с центрами неводородных атомов. Молекулу растворителя (в настоящей работе – водный растворитель) также рассматривали в виде сферы.

Хотя приводимый ниже анализ относится к любым макромолекулам, мы будем рассматривать глобулярные белки.

#### ПРОГРАММНАЯ РЕАЛИЗАЦИЯ МЕТОДА КУБИКОВ

**Геометрическая часть.** Для расчета амплитуды рассеяния ( $q$ ) необходимо прежде всего построить однородную частицу, имеющую объем и форму исследуемой макромолекулы (белка). В используемом нами методе заполнение объема белка производится мелкими кубиками, плотная упаковка которых обеспечивает однородную плотность внутри частицы.

Мысленно опишем вокруг каждого атома (атомной группы) «гидратированную» сферу, радиус которой равен сумме ван-дер-ваальсова радиуса атома (атомной группы) и радиуса молекулы воды. Система таких пересекающихся сфер образует «гидратированную» макромолекулу, причем ее поверхность («доступная поверхность») ограничивает объем, в котором не могут находиться центры окружающих молекул воды. Далее опишем вокруг этой «гидратированной» макромолекулы параллелепипед, разобъем его на мелкие плотно упакованные кубики и каждый кубик будем маркировать цифрами 0, 1, или 2 в зависимости от того, находится ли кубик внутри доступной поверхности (1), на поверхности (2) или вне ее (0). Если каждый кубик доступной поверхности окружить сферой радиуса молекулы воды и обратить все кубики внутри этой сферы в состояние 0, то можно перейти от объема «гидратированной» макромолекулы к ее реальному объему. Этот объем – сумма объемов всех оставшихся кубиков, маркированных цифрой (1). Те из этих кубиков, у которых хотя бы один из соседей маркирован цифрой 0, являются граничными кубиками и составляют *молекулярную поверхность* макромолекулы.

Рассмотрим работу алгоритма программы более подробно. Расчет начинается с построения таблицы, включающей название входящих в белки атомов (группы атомов), их ван-дер-ваальсовы радиусы и упомянутые выше коэффициенты аппроксимации атомных

факторов при различных  $q$ . Эта таблица является универсальной и будет использована при расчете интенсивности рассеяния для всех исследуемых белков.

Из PDB-файла извлекаются координаты всех атомов и групп атомов исследуемого белка и составляется вторая таблица, включающая в той же последовательности, как и в PDB-файле, названия атомов (группы атомов) и их координаты, также в эту таблицу переносятся из первой таблицы радиусы атомов (группы атомов) и коэффициенты аппроксимации.

Следующий этап работы – построение «гидратированного» белка и описание его системой плотно упакованных мелких кубиков. Для удобства последующих действий каждый неводородный атом слегка смещаем так, чтобы его центр совпал с центром близлежащего кубика. Это смещение составляет доли ангстрема и не вносит существенных искажений в дальнейшие расчеты, так как разрешение координат атомов в базе данных PDB составляет, как правило, более 0,15 нм. Для каждого атома (группы атомов) создается «шаблон» – шар из плотно упакованных кубиков с радиусом, равным сумме радиуса атома (группы атомов) и радиуса молекулы воды («кубики-1»). Далее, центр каждого атома (группы атомов) совмещается с центром своего шаблона, и все кубики, заполняющие этот «шаблон», маркируются цифрой 1. Таким образом, образуется объем «гидратированного» белка. Координаты этих «кубиков-1» записываются в таблицу, в которой, как нетрудно видеть, кубики могут повторяться, то есть иметь одинаковые координаты. Из повторяющихся кубиков исключаются все, кроме одного.

Далее следует анализ положения каждого кубика относительно его соседей. Если все шесть граней рассматриваемого кубика «закрываются», то есть соприкасаются с такими же «кубиками-1», его маркировка «1» остается прежней. Но если хотя бы одна из граней кубика «открыта», его маркировка переводится из состояния «1» в состояние «2» («кубики-2»). Эти кубики являются кубиками *доступной* поверхности. Эта поверхность ограничивает объем, в котором *не могут* находиться центры окружающих молекул воды.

Для молекулы воды также создается «шаблон» – шар из плотно упакованных кубиков с радиусом молекулы воды. Для перехода от «гидратированного» объема белка к его реальному объему центр каждого «кубика-2» совмещается с центром «шаблона» для воды, и все кубики внутри этого «шаблона» маркируются

цифрой 0, то есть выбывают из игры. Образно говоря, молекула воды, моделируемая состоящим из кубиков шаром, как бы катится по всей поверхности белка, обращая все кубики внутри этого катящегося шара в состояние 0. Оставшиеся «кубики-1» полностью повторяют очертание молекулы белка, создают однородную внутреннюю структуру, а их количество определяет объем молекулы.

Для удобства расчета функции  $I(\mathbf{q})$  все кубики, имеющие одинаковые координаты по двум координатным осям и плотно упакованные вдоль третьей оси, объединяются в параллелепипеды. Очевидно, что граничные кубики каждого параллелепипеда являются кубиками молекулярной поверхности. Отметим, что идея «катания» молекулы воды для определения молекулярной поверхности частицы впервые была предложена в работе [16].

**Расчет интенсивности рассеяния.** Будем считать, что составлена система из  $N$  параллелепипедов, имеющих направление, например, вдоль оси  $Y$ . Пусть  $j$ -й параллелепипед имеет длину  $b_j$  и радиус-вектор центра масс  $\mathbf{r}_j$ . Длина ребра составляющих его кубиков равна  $a$ . Тогда амплитуда рассеяния такой системой может быть рассчитана по следующей формуле [17]:

$$I(\mathbf{q}) = 8 \frac{\sin(q_x a / 2) \sin(q_z a / 2)}{q_x q_y q_z} \prod_{j=1}^N \sin(q_y b_j / 2) \exp(i\mathbf{q} \mathbf{r}_j), \quad (6)$$

причем амплитуда  $I(\mathbf{q})$  равна объему молекулы, то есть сумме объемов всех параллелепипедов, входящих в систему. Интенсивность рассеяния  $I(\mathbf{q})$  может быть рассчитана по формулам (1)–(5).

Следующий этап работы состоит в переходе к интенсивности рассеяния  $I(q)$  путем усреднения  $I(\mathbf{q})$  по всем точкам обратного пространства, расположенным на расстоянии  $q$  от его начала. С этой целью вся поверхность сферы радиуса  $q$  разбивается на малые участки с одинаковой площадью, для каждого участка рассчитывается интенсивность  $I(\mathbf{q})$ , которая затем усредняется по всем участкам разбиения. Так как элемент поверхности сферы  $q^2 \sin \theta d\theta d\phi$  (и соответственно азимутальный и полярный углы), то при равном выборе углов равномерность разбиения сферы на равные площади сводится к утверждению  $\sin \theta = \text{const}$ . Легко показать, что при этом проекция  $Z$  дуги  $q$  на ось  $Z$  есть величина постоянная. Таким образом, для нахождения

полярного угла искомого малого участка поверхности необходимо разделить ось  $Z$  внутри сферы на  $n$  равных участков, найти расстояние  $Z_i$ -го участка от центра сферы ( $Z_i = Zi, 1 < i < n$ ) и рассчитать  $\theta_i = \arccos(Z_i/q)$ . Так как  $d\theta = 2/m$  ( $m$  – число разбиений в азимутальной плоскости), то  $d\phi = 2\pi/m$  ( $1 < j < m$ ) и

$$q_x = q \sin \theta_i \cos \phi_j, \quad q_y = q \sin \theta_i \sin \phi_j, \quad q_z = q \cos \theta_i. \quad (7)$$

Рассчитав интенсивность рассеяния  $I(\mathbf{q})$  в каждой выбранной точке поверхности сферы и усреднив ее по всем точкам  $i$  и  $j$ , можно определить искоемую интенсивность рассеяния:

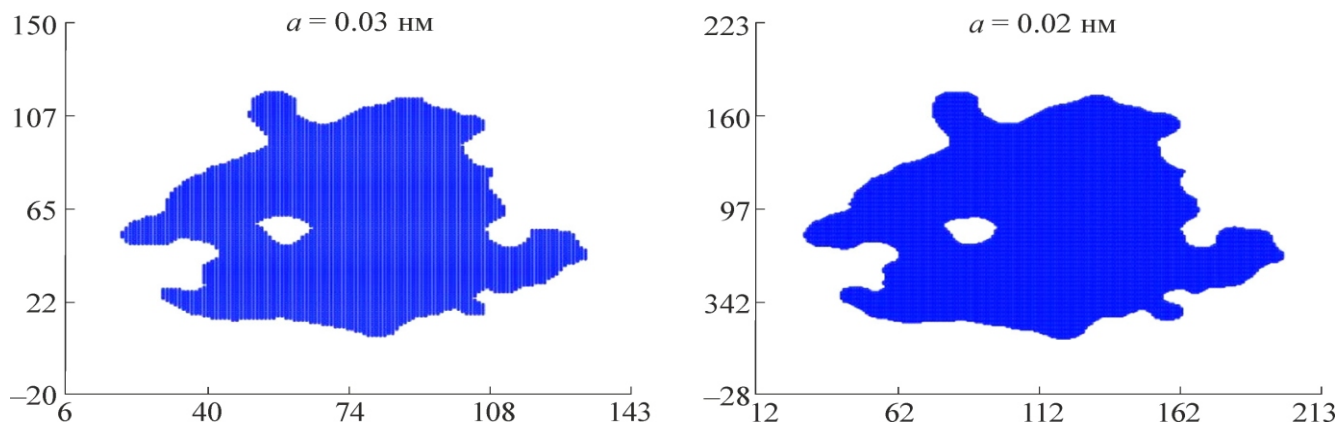
$$I(q) = \frac{1}{mn} \sum_{i=1}^n \sum_{j=1}^m I(q_x, q_y, q_z). \quad (8)$$

Как показывает опыт, для несильно вытянутых частиц выбор значений  $n = m = 40$  уже обеспечивает высокую точность интенсивности рассеяния в области углов до  $q = 10 \text{ нм}^{-1}$ .

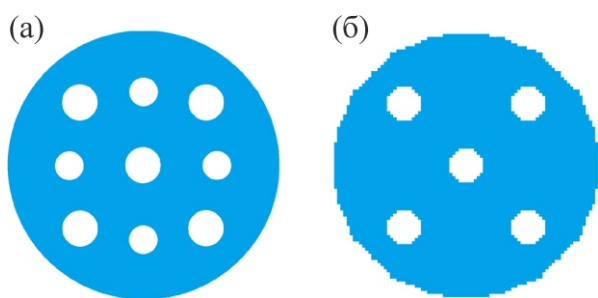
## ПРОВЕРКА ПРОГРАММЫ

**Выбор размера кубиков.** В качестве основы была выбрана длина ребра кубика  $a = 0,03 \text{ нм}$ . Эта величина достаточно мала по сравнению с размерами всех входящих в состав белка неводородных атомов, а также молекулы воды. Если существенно изменить длину ребра кубика, выбрав, например,  $a = 0,2 \text{ нм}$ , то изменения основных геометрических параметров (на примере лизоцима 6LYZ) окажутся незначительными: молекулярный объем, доступная и молекулярная поверхность изменятся менее чем на 2%, радиус инерции белка, заполненного кубиками, – на доли процента. На рис. 1 представлено одно из сечений указанного белка при его заполнении кубиками  $a = 0,03 \text{ нм}$  (рис. 1, слева) и кубиками  $a = 0,02 \text{ нм}$  (рис. 1, справа). Из сравнения этих сечений видно, что их очертания и площадь практически полностью повторяют друг друга.

**Поворот молекулы.** Еще одним способом проверки программы является оценка инвариантности основных рассчитываемых параметров к повороту молекулы относительно координатных осей. Были выбраны повороты молекулы на  $45^\circ$  и на  $60^\circ$ . Как оказалось, при обоих поворотах изменения доступной и молекулярной поверхности, а также молекулярного объема и радиуса инерции белка нигде не превышают долей процента. Эта оценка относится как к кубикам с  $a = 0,03 \text{ нм}$ , так и к кубикам с  $a = 0,02 \text{ нм}$ .



**Рис. 1.** Одно из сечений молекулы лизоцима (6LYZ), объем которой заполнен кубиками с разной длиной ребра  $a$ . По координатным осям отложены относительные номера кубиков.



**Рис 2.** (а) – Сечение модельной системы, состоящей из большого равномерно заполненного шара и мелких разноразмерных шаровых полостей; (б) – та же система описана с помощью кубиков. Подробности в тексте.

**Оценка чувствительности программы к наличию внутренних полостей.** Рассмотрен однородный шар диаметром 2,65 нм, в котором заключены неперекрывающиеся шарообразные полости. Часть полостей имеет диаметр 0,33 нм, другая часть – диаметр 0,27 нм. Диаметр молекулы воды 0,30 нм. При погружении этой системы в водную среду молекулы воды, проникая в большой шар, не могут занимать объем полостей малого диаметра, и, следовательно, объем таких полостей следует присоединить к объему большого шара. В то же время молекулы воды могут размещаться в полостях большего диаметра, и объем этих полостей исключается из объема большого шара, образуя тем самым внутреннюю полость. На рис. 2 представлены сечение большого шара с большими и малыми полостями в плоскости двух координатных осей и сечение в той же плоскости этой системы, описанной с помощью геометрической части программы SASCUBE с длиной ребра 0,03 нм. Как видно из рисунка, полости меньшего диаметра исчезли, а их объем слился с объемом большого шара.

## РЕЗУЛЬТАТЫ И ОБСУЖДЕНИЕ

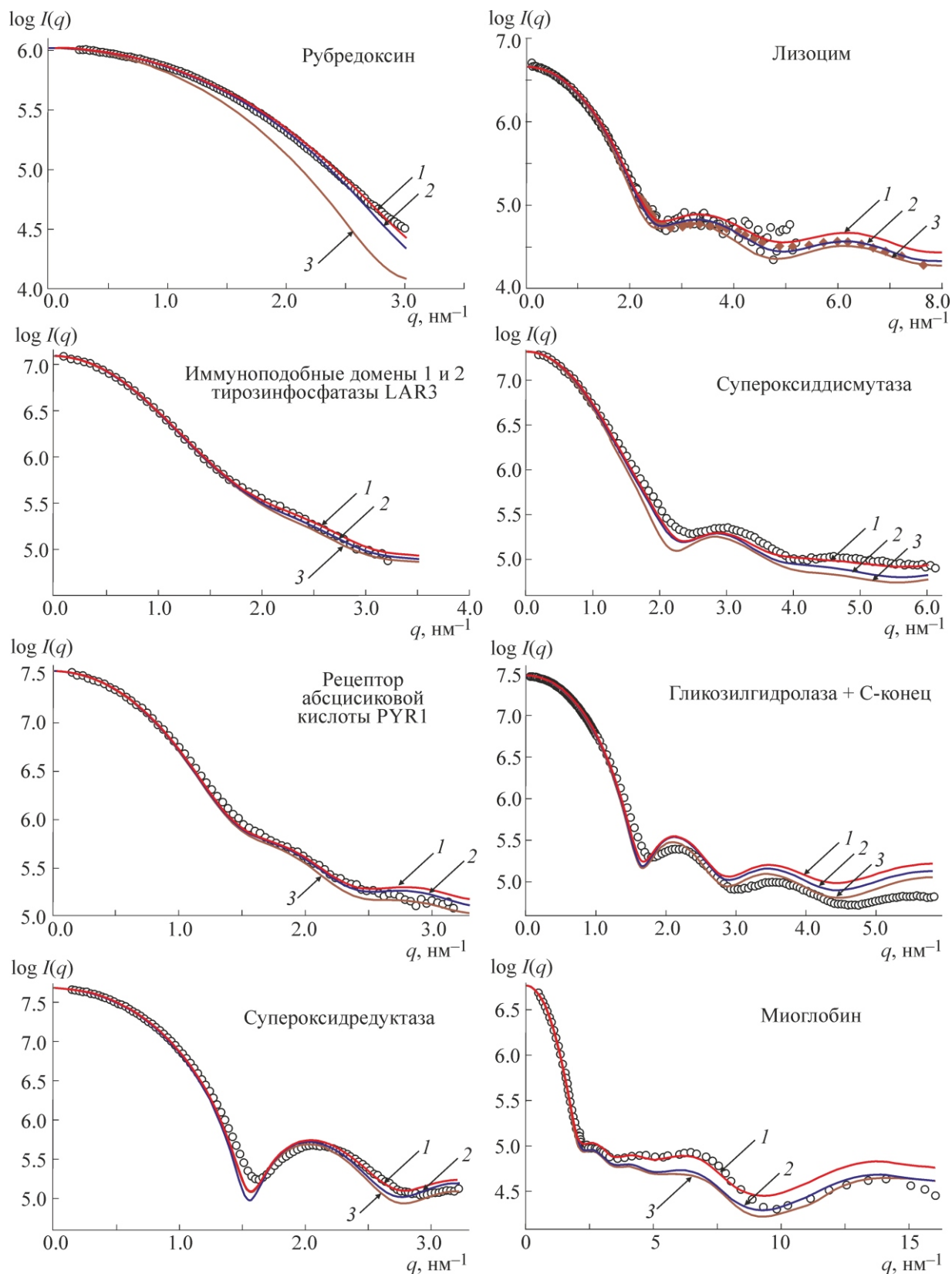
### Сравнение расчетных интенсивностей рассеяния с экспериментальными кривыми.

Для восьми глобулярных белков было проведено сравнение интенсивностей рассеяния, рассчитанных с помощью программы SASCUBE, с соответствующими экспериментальными индикатрисами рассеяния, выбранными в основном из базы данных BIOISIS. Координаты атомов белков взяты из базы данных PDB. В расчетах использованы радиусы, приводимые в работе [18] (радиусы I) и работе [16] (радиусы II), причем радиусы II в целом меньше радиусов I. Результаты сравнения представлены на рис. 3.

Важно отметить, что наиболее информативной с точки зрения анализа структуры белков «средних» размеров (с диаметром  $\sim 5$ – $7$  нм) является область углов рассеяния от  $q \sim 1$  нм $^{-1}$  до  $q \sim 10$  нм $^{-1}$ . Для более крупных белков «интересная» область расширяется в сторону меньших углов. В то же время при  $q > 10$  нм $^{-1}$  кривая рассеяния мало специфична к структуре белка, а диффузный максимум при  $\sim 14$  нм $^{-1}$ , наблюдающийся на кривых рассеяния всех глобулярных белков, связан, как известно [19], со средним ван-дер-ваальсовым расстоянием между атомами при плотной упаковке полипептидной цепи.

Были рассмотрены следующие белки (рис. 3):

- 1) Рубредоксин: 1BQ9(PDB), 1RBDGP (BIOISIS-ID) [6].
- 2) Лизоцим: 6LYZ(PDB) [8,20].
- 3) Иммуноподобные домены 1 и 2 тирозинфосфатазы LAR3: 3PXJ (PDB), LAR12P (BIOISIS-ID) [21].



**Рис. 3.** Экспериментальные и расчетные кривые рассеяния для восьми белков (см. текст). Кружки – экспериментальные кривые; для лизоцима приведены две экспериментальные кривые: кружки – из источника [8], ромбы – из источника [20]. Расчетные кривые (SASCUBE): красная – на основе радиусов I без учета фиксированных молекул воды, синяя – на основе радиусов II без учета фиксированных молекул воды, коричневая – на основе радиусов II с учетом фиксированных молекул воды.

4) Супероксиддисмутаза: 1HL5 (PDB), APSOD (BIOISIS-ID) [22].

5) Рецептор абсцисиковой кислоты PYR1: 3K3K (PDB), 1PYR1P (BIOISIS-ID) [23].

6) Гликозилгидролаза + С-конец: 1EDG (PDB), AT5GHP (BIOISIS-ID) [24].

7) Супероксидредуктаза: 1DQK (PDB), 1SPXGP (BIOISIS-ID) [6].

8) Миоглобин: 1MLA (PDB) [20].

Общий обзор представленных интенсивностей свидетельствует о том, что рассчитанные кривые рассеяния дают качественное согласие с соответствующими экспериментальными индикатрисами рассеяния. Однако во многих случаях наблюдаются заметные расхождения между теорией и экспериментом, и это представляется вполне естественным. Мы бы выделили ряд факторов, которые в большей или в меньшей степени могут влиять на ход расчетных кривых рассеяния.

**Влияние фиксированных (по PDB-данным) молекул воды на интенсивность рассеяния белков в растворе.** Как описано выше, в программе SASCUBE вода, окружающая молекулу белка, рассматривается как континуум с равномерно распределенной электронной плотностью. Такое приближение является весьма грубым. Вода может активно взаимодействовать с поверхностными атомами молекулы белка и, как следует из ряда экспериментальных исследований [25–28], ее структура в примыкающей к белку области может заметно искажаться по сравнению с далекими от белка областями. Наличие этих искажений было отмечено и в работах [29–31], а также в работах [32–35] и других. К вопросу влияния этих искажений на интенсивность рассеяния мы еще вернемся ниже.

На графиках на рис. 3 в качестве иллюстрации показано влияние на кривую рассеяния *лишь фиксированных* молекул воды, координаты которых представлены в базе данных PDB наряду с координатами атомов белков в кристаллической структуре. Расчет состоял в следующем. Представим себе, что молекулы белка без искажения своей структуры переходят из кристалла в водный раствор, увлекая за собой также без искажения все фиксированные молекулы воды. Будем считать, что весь комплекс «белок + фиксированная вода» представляет новую частицу, погруженную в воду однородной плотности, и для этой частицы по программе SASCUBE рассчитывается интенсивность рассеяния. Как видно, почти на всех графиках на рис. 3 наличие фиксированных молекул воды лишь слабо влияет на общий ход

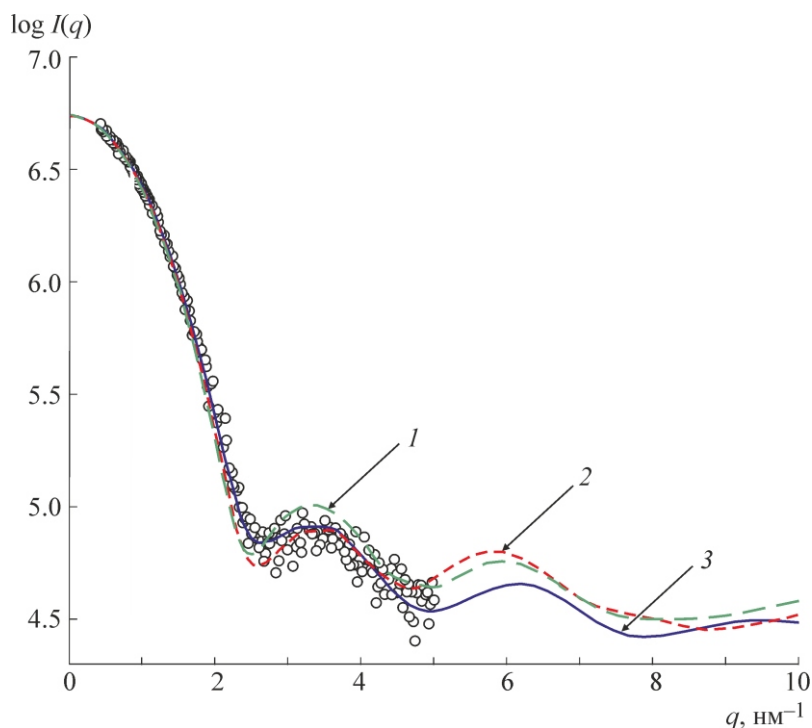
кривой рассеяния. Возможно, это связано с тем, что число таких молекул относительно невелико в сравнении с количеством атомов белка. Исключением является лишь белок рубредоксин, в котором по кристаллическим данным количество фиксированных молекул воды составляет около 40% от количества атомов в белке. Отметим, что та же слабая зависимости интенсивности рассеяния от учета фиксированных молекул воды была отмечена еще в наших старых работах [1–4].

**Влияние конформационной подвижности белков в растворе на их интенсивность рассеяния.** При анализе водных растворов белков следует учитывать не только перестройки молекул воды вблизи поверхности белка, но и динамический характер конформационных изменений в самой структуре белка. С этой целью в работе был использован метод молекулярной динамики (МД). Симуляцию проводили в программной среде Desmond [36–38] при температуре 310 К и давлении 1,01325 бар.

В качестве иллюстрации того, что структура белков в растворе подвержена существенным динамическим изменениям, на рис. 4 проведен расчет интенсивности рассеяния от двух конформаций лизоцима 6LYZ, выбранных на основе МД-симуляций через 50 нс (кривая 1) и через 300 нс (кривая 2) после помещения белка в водный растворитель (изменения структуры воды при этом не рассматривались). Кривая 3 – интенсивность рассеяния от «базисной» конформации, взятой из базы данных PDB. Как видно, кривые рассеяния для выбранных двух конформаций заметно отличаются как между собой, так и от «базисной» конформации.

**Влияние выбора радиусов атомов и групп атомов белков на их интенсивность рассеяния.** Как следует из графиков на рис. 3, в большинстве случаев теоретические кривые рассеяния слабо зависят от выбора радиусов I или II. Из выбранных восьми белков исключение составляет лишь график для белка супероксидредуктазы, на котором наблюдаются весьма заметные различия между кривыми, относящимися к разным группам радиусов.

На графиках, представленных на рис. 3, мы не приводим кривые рассеяния, рассчитанные по программе CRY SOL и программам, близким к CRY SOL, так как вводимые авторами этих программ два подгоночных параметра (о чем см. ниже) обеспечивают практически идеальное совпадение расчетных кривых с экспериментом.



**Рис. 4.** Кривые рассеяния лизоцима 6LYS в растворе для двух конформаций белка, выбранных на основе МД-симуляций, и кривая рассеяния от «базисной» конформации. Кружки – экспериментальная кривая рассеяния [8]. Подробности в тексте.

#### СРАВНИТЕЛЬНЫЙ АНАЛИЗ МЕТОДА КУБИКОВ SASCUBE И ДРУГИХ МЕТОДОВ РАСЧЕТА ИНТЕНСИВНОСТИ БЕЛКОВ В РАСТВОРЕ

В представленной в настоящей работе программе SASCUBE нет подгоночных параметров, и в этом состоит ее главное отличие от большинства других программ по расчету интенсивностей рассеяния макромолекул в растворе по известным координатам атомов.

**Замечания по поводу программы CRY SOL.** В этом аспекте остановимся подробнее на развитой в работе [8] программе CRY SOL. В программе использован метод разложения экспоненциальных функций по сферическим гармоникам, что является удачным подходом при расчете сферически симметричной картины рассеяния свободно ориентированными частицами. Эта работа является пионерской, она позволила получать очень хорошее совпадение расчетных кривых с экспериментом и, что немаловажно, при минимальном времени расчета. Этот метод с рядом модификаций стал широко использоваться в мировой литературе [10,11,25,39].

В то же время следует отметить, что программа CRY SOL включает ряд допущений, со-

мнительность которых отмечают и сами авторы.

Одно из ключевых допущений, используемых в программе CRY SOL, состоит в том, что вдоль всей поверхности молекулы белка существует приграничный слой воды одинаковой толщины с равномерно распределенной электронной плотностью, отличной от средних электронных плотностей как белка, так и воды. Это допущение широко используется многими авторами при расчете интенсивностей рассеяния частиц в растворе, так как оно, при подборе соответствующих параметров этого слоя (наряду с другим подгоночным параметром), существенно улучшает совпадение расчетной кривой с экспериментальной. Однако трудно представить наличие такого однородного слоя для глобулярных белков с сильно развитой поверхностью, да еще, вероятно, имеющих внутренние полости (см. рис. 1).

Отметим еще ряд мало оправданных допущений, используемых в программе CRY SOL. Способ введения функции, описывающей поверхность частицы, может быть применен, строго говоря, лишь для выпуклых частиц. В ином случае лучи, исходящие из центра частицы, могут несколько раз пересекать ее границу, и «не замечать» возможную сложную форму ее поверхности (например, большие или малые



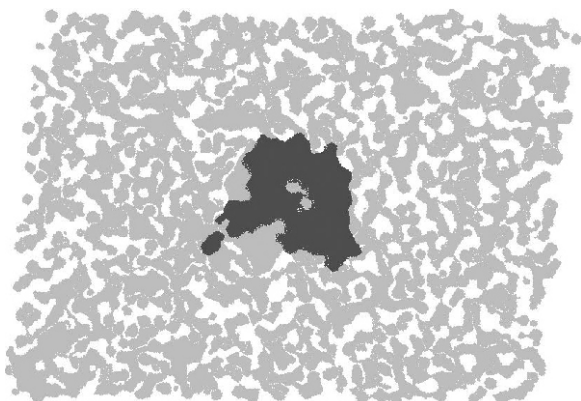


Рис. 5. Сечение объема молекулы лизоцима 6LYZ и окружающих ее водных кластеров по данным МД-симуляции. Подробности в тексте.

«заливы»), а также внутренние полости. Как следствие, форма и объем частицы могут существенно искажаться. Кроме того, метод не обеспечивает пространственную однородность вытесненного белком объема, что существенно при расчете амплитуды рассеяния. Этот же недостаток присущ и другим методам, в которых объем белка заполняется полыми шарами. В программе CRY SOL вводится еще один параметр, регулирующий радиусы шаров, замещающих атомы (dummy atoms) и группы атомов белка (то есть регулирующий объем белка), который по сути является еще одним подгоночным параметром.

При взаимодействии молекул воды с поверхностными молекулами белка, как уже отмечалось, безусловно, происходит перестройка молекул воды и, следовательно, происходят локальные изменения электронной плотности близ белковой границы. Вопросы состоят лишь в том, насколько эти изменения существенны, равномерно ли распределены электронные плотности по всему объему приграничного слоя и можно ли считать толщину этого слоя одинаковой вдоль всей поверхности белка.

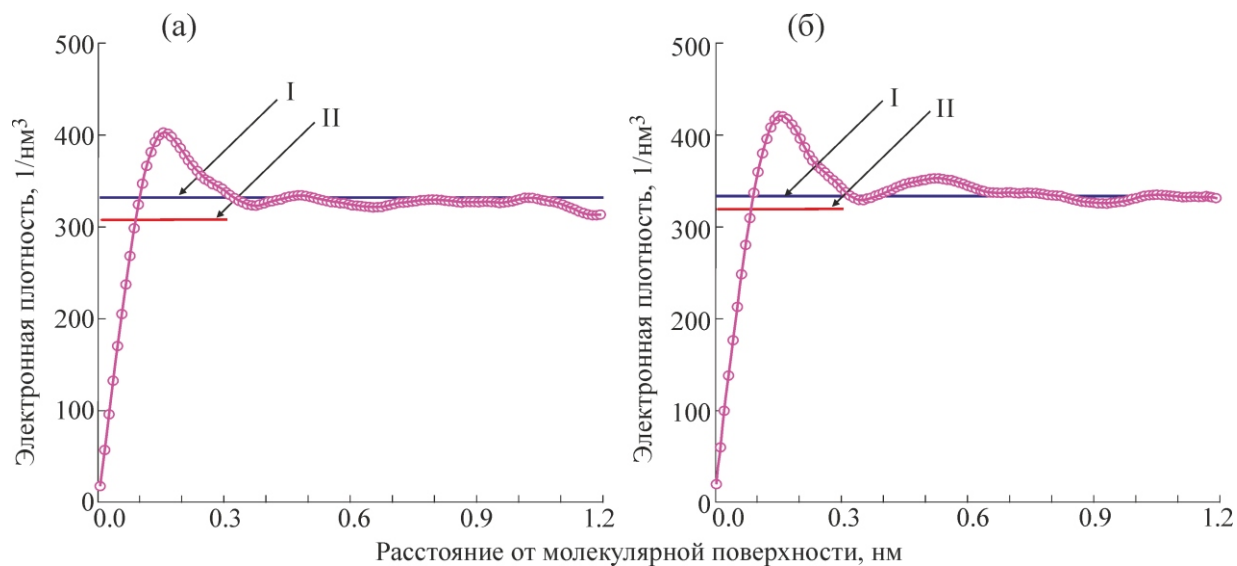
Для ответа на указанные вопросы мы провели следующие два рассмотрения.

**Распределение кластеров воды вокруг молекулы белка по данным молекулярно-динамической симуляции.** С помощью МД-симуляций с использованием ячейки  $10,40 \times 9,55 \times 9,26 \text{ нм}^3$  были определены координаты атомов лизоцима 6LYZ (PDB6LYS) и всех молекул воды через  $\sim 100 \text{ нс}$  после начала взаимодействия молекул воды как между собой, так и с молекулой белка. Затем, с использованием мгновенные значения координат всех атомов в ячейке, по программе SAS CUBE

(геометрическая часть) были получены объемные очертания молекулы белка, а также всех кластеров молекул воды, окружающих белок. На рис. 5 представлено одно из сечений этого объема. Темным цветом обозначено сечение молекулы белка, светлым – кластеров воды. Как видно из рисунка, ничто не свидетельствует о том, что у границы белка имеется приграничный слой одинаковой толщины и одинаковой плотности.

**Зависимость средней электронной плотности воды от расстояния до молекулярной поверхности белка по данным молекулярно-динамической симуляции.** Следующее рассмотрение, также использующее МД-симуляцию, касается расчета зависимости средней электронной плотности ( $\rho$ ), создаваемой молекулами воды, от расстояния  $r$  до молекулярной поверхности белка. Для рассмотрения выбраны два белка разной молекулярной массы: лизоцим 6LYZ (1001 неводородный атом) и супероксиддисмутаза APSOD (2231 неводородный атом). Исходные данные для лизоцима: нормальное атмосферное давление  $p_0 = 101,3 \text{ кПа}$ , температура  $T = 310 \text{ К}$ ; при этих условиях плотность воды  $\rho_w = 992,6 \text{ кг/м}^3$ , электронная плотность воды  $\rho_0 = 331,8 \text{ нм}^{-3}$ . Исходные данные для супероксиддисмутазы: нормальное атмосферное давление  $p_0 = 101,3 \text{ кПа}$ , температура  $T = 300 \text{ К}$ ; при этих условиях электронная плотность воды  $\rho_w = 996,5 \text{ кг/м}^3$ , средняя электронная плотность воды  $333,1 \text{ нм}^{-3}$ .

Для расчета указанной зависимости ( $\rho$ ) использовали следующую процедуру. Рассчитывали молекулярную поверхность белка, затем все пространство вокруг этой поверхности разбивали на слои толщиной  $0,01 \text{ нм}$  и рассчитывали объем каждого слоя. Расчет молекулярной поверхности и объемов слоев проводили с помощью метода кубиков. Методом МД-симуляции были определены через достаточно большой промежуток времени (время от начала симуляции для обоих белков составляло  $100 \text{ нс}$ ) координаты атомов белковых молекул, а также окружающих их молекул воды. Для каждой молекулы воды определяли расстояние  $L$  от ее кислородного атома до ближайшего атома молекулы белка. Если ван-дер-ваальсов радиус этого атома (или группы атомов) обозначить через  $R$ , то расстояние  $r$  от молекулы воды до молекулярной поверхности рассчитывали по формуле  $r = L - R$ . Электронную плотность ( $\rho$ ) каждого слоя определяли как отношение числа электронов в слое к объему этого слоя. Молекулы воды считали равномерно за-



**Рис. 6.** Распределение электронной плотности в приграничной зоне молекул лизоцима (а) и супероксиддисмутазы (б). Конформация белков и координаты окружающих их молекул воды определены с помощью МД-симуляции. Синяя линия – электронная плотность невозмущенного растворителя, красная линия – среднее значение электронной плотности растворителя в приграничном слое в интервале  $0 < r < 0,3$  нм. Подробности в тексте.

ряженными шариками. Радиус шарика выбирали равным 0,15 нм. При вычислении количества электронов в слое учитывали части объемов шариков, попадающие в каждый слой.

Графики распределения электронной плотности для двух рассмотренных белков представлены на рис. 6. Как видно из рисунка, в случае исследуемых белков приграничный слой воды, «возмущенный» присутствием белка, простирается до значений  $r \sim 0,5-0,6$  нм (в работе [20] этот слой определен как  $r \sim 0,7$  нм), а в области  $r < 0,5$  нм наблюдается существенно неравномерное распределение электронной плотности, максимум которой для обоих белков приходится на расстояние  $r \sim 0,15$  нм. Отметим, что среднее значение в интервале  $0 < r < 0,3$  нм составляет  $\rho = 308,3$  эл/нм<sup>3</sup> для лизоцима и  $\rho = 319,0$  эл/нм<sup>3</sup> для супероксиддисмутазы (на обоих графиках – линия II), что заметно меньше электронной плотности воды в этих условиях. Таким образом, если ввести пограничный слой толщиной 0,3 нм (как в работе [6]) и даже допустить равномерное распределение электронной плотности в этом слое (что, как видно из того же рис. 6, безусловно, неверно), определяемая внутри этого слоя средняя электронная плотность окажется ниже, а не выше электронной плотности воды, что противоречит общепринятому мнению. При расчетах радиусы атомов и групп атомов взяты из работы [16], однако если их взять из работы [18], графики на рис. 6 останутся практически теми же. Разумеется,

мы рассмотрели частные случаи, но они свидетельствуют о том, что по крайней мере для части белков подгоночные параметры не имеют ясно выраженного физического смысла.

Интересно отметить, что при  $r = 0$  электронная плотность не имеет нулевого значения. Это можно объяснить некорректностью представления молекул воду в виде шариков, а также тем, что некоторые молекулы вода проникают внутрь поверхности белка, образуя, например, водородные связи.

**Замечания к ряду других работ по расчету интенсивности рассеяния белков в растворе.** Рассмотрим некоторые другие работы, в которых используются те же два основных допущения работы [8]: введение однородного и одинаковой толщины поверхностного слоя с отличной от белка и воды электронной плотностью и параметра, регулирующим исключенный объем молекулы.

В работе [39] также используются указанные допущения, а для расчета интенсивности рассеяния используется формула Дебая, в которой амплитуда рассеяния учитывает, по мнению авторов, электронную плотность фиксированного по толщине поверхностного слоя и регулятор объема белка. Более того, с целью сокращения времени расчета путем дальнейших огрублений авторы сводят двойную сумму в формуле Дебая к одинарной сумме.

Авторы работы [10] пытаются обеспечить пространственную однородность вытесненного белком объема путем описания этого объе-

ма системой плотно упакованных разноразмерных кубиков вплоть до кубиков с длиной ребра 0,05 нм. Однако такой подход, похоже, не в состоянии учесть глубокие узкие «заливы» на поверхности белка, а также его внутренние полости.

С целью существенного упрощения и ускорения расчета в ряде работ были рассмотрены «крупнозернистые» модели белковых молекул для прогнозирования их сворачивания и предсказания новых структур [40]. Так, в работах [41,42] положение каждой аминокислоты просто представлено ее С-атомом, в который помещаются все электроны этой аминокислоты. В работах [43–45] уже включены в рассмотрение атомы основной цепи, а «крупным зерном» является боковой остаток, также связанный с положением С-атома, в который помещаются электроны боковой цепи. Наконец, в работе [46] боковые цепи белков также представлены С-атомами с соответствующими электронами, но в этой модели в рассмотрение включены в явном виде молекулы воды, находящиеся вблизи молекулы белка. Для крупных субъединичных белков, исследование которых имеет целью описание взаимного расположения составляющих их блоков, «крупнозернистый» подход, возможно, разумен, так как именно расположения блоков в наибольшей степени сказывается на самой малоугловой части кривой рассеяния, а детали структуры субъединиц в этой области углов не «видны». Но для средних и небольших белков такой подход, по-видимому, малопригоден.

**Заключительные замечания по методам расчета интенсивности рассеяния белков в растворе.** В программе CRY SOL, а также в других программах, близких к CRY SOL, вводятся два физически мало аргументированных подгоночных параметра. Вариации этих параметров обеспечивают хорошее согласие между расчетными и экспериментальными кривыми рассеяния, так как при этом компенсируются недостатки метода, а также неучтенные объективные факторы, влияющие на экспериментальные интенсивности рассеяния.

Создается впечатление, что основная задача, которую ставит перед собой ряд авторов, – добиться наилучшего согласия между теорией, основанной на данных PDB, и экспериментом и при этом максимально сократить время расчета. Нам кажется, это неправильная постановка задачи. Допустим, мы добились, умело используя подгоночные параметры, идеального согласия с экспериментом. Ну и что? Можно ли утверждать, что полученное согласие свидетельствует о том, что белок имеет именно

структуру, описанную в базе данных PDB, и именно то расположение молекул воды вблизи белка, к которому приводят подгоночные параметры? Ведь экспериментальные кривые – это интенсивности рассеяния белков *в растворе* (за вычетом, разумеется, рассеяния от растворителя), в то время как практически все методы расчета (в том числе и метод SASCUBE) используют структуру белковой макромолекулы в кристалле. При переходе белка в раствор, как уже отмечалось, могут происходить значительные структурные перестройки в самой макромолекуле, влияющие на кривую рассеяния [47], может меняться расположение молекул воды в окружающей белок области, могут меняться эффективные радиусы атомов и атомных групп. Кроме того, проявляются объемные эффекты, которые даже при небольших концентрациях белка влияют на малоугловую часть кривой рассеяния [48]. Все это приводит к тому, что экспериментальная кривая может существенно отличаться от расчетной кривой, основанной на PDB данных. Есть ли смысл гнаться за точным совпадением расчетной кривой с экспериментальной?

Существенным шагом вперед следует считать совместное использование методов малоуглового рассеяния и метода МД. Метод МД позволяет не только следить во времени за изменением структуры белка, но и за смещением молекул воды как внутри белка, так и в его окружении. В этом аспекте одной из пионерских является работа [20], в которой авторы – *не рассматривая изменения структуры белка* – учитывали лишь смещения молекул воды, связанные как с взаимодействием между собой, так и с молекулой белка. Но даже при таком ограниченном подходе было получено хорошее совпадение теоретических кривых рассеяния с экспериментальными кривыми для двух белков.

## ЗАКЛЮЧЕНИЕ

Сочетание МД и малоуглового рассеяния представляется наиболее перспективным направлением при исследовании структуры белков в растворе. Одна из возможностей таких исследований описана в работе [20], в которой формула (26) дает строгое выражение для интенсивности рассеяния белков при динамическом описании изменений их конформаций в растворе.

Другой строгий метод может быть основан на использовании МД и программы SASCUBE. Как показано в той же работе [20], глубина неравномерного распределения молекул воды в

приграничной области белков распространяется не далее 0,7 нм от поверхности белка. Пусть, исходя из данных МД, нам известны координаты всех атомов  $i$ -й конформации белка и всех молекул воды в приграничной 0,7 нм зоне. Если при  $i$ -й конформации белка взять за основу комплекс «белок + приграничный слой воды» (то есть рассмотреть этот комплекс как новый «супербелок») то на основе формулы (1) можно с помощью программы SASCUBE строго рассчитать интенсивность рассеяния всего комплекса. В этом случае первый член формулы (1) – амплитуда рассеяния от всех атомов, входящих в комплекс, в том числе и атомов воды, а во втором члене ( $\mathbf{q}$ ) – амплитуда рассеяния от «полого» комплекса, то есть всего комплекса, заполненного плотно упакованными кубиками. Далее, разумеется, следует усреднение полученной интенсивности по максимально возможному числу таких комплексов, создаваемых с помощью МД-симуляции.

Оба этих подхода в принципе эквивалентны, и вопрос в том, какой из них выбирать. Ответ зависит прежде всего от трудностей в составлении кодовых программ и времени расчета. В случае, если эта задача будет решена и мы сможем получать без каких-либо подгоночных параметров хорошее согласие расчетной кривой с экспериментом для белков в широкой области углов рассеяния, можно перейти к следующему исследованию. Как следует из предварительных данных по МД-симуляции, некоторые белки в растворе структурно можно разделить на ряд отличающихся друг от друга конформеров, находящихся между собой в состоянии динамического равновесия, причем структуры белков, относящихся к одному конформеру, весьма близки. Если для данного белка имеется всего несколько таких конформеров, то, зная интенсивность рассеяния всего белка в растворе и рассчитав интенсивность рассеяния для каждого конформера, можно определить относительную долю каждого из них в растворе.

Авторы приносят сердечную благодарность магистранту Университета ИТМО А. Суворовой и аспиранту Университета ИТМО М. Гурееву за помощь в работе.

Работа была выполнена при частичной финансовой поддержке Правительства Российской Федерации (грант 08-08) и Российского академического проекта «5-100» (Университет им. И.М. Сеченова).

## СПИСОК ЛИТЕРАТУРЫ

1. М. Ю. Павлов и Б. А. Федоров, *Биофизика* **27** (4), 609 (1982).
2. Б. А. Федоров, О. Б. Птицын и Л. А. Воронин, *Молекуляр. биология* **8** (5), 693 (1974)
3. M. Yu. Pavlov and B. A. Fedorov, *Biopolymers* **22**, 1507 (1983).
4. B. A. Fedorov and A. I. Denesyuk, *J. Appl. Cryst.* **11**, 473 (1978).
5. M. Yu. Pavlov, M. A. Sinev, A. A. Timchenko, and O. B. Pitsyn, *Biopolymers* **25**, 1385 (1986).
6. G. L. Hura, A. L. Menon, M. Hammel, et al., *Nature Methods* **6**, 606 (2009).
7. E. Valentini, A. G. Kikhney, G. Previtali, et al., *Nucl. Acids Res.* **43**, D357 (2015).
8. D. Svergun, C. Barberato, and M. H. J. Koch, *J. Appl. Cryst.* **28**, 768 (1995).
9. D. Schneidman-Duhovny, M. Hammel, J. A. Tainer, and A. Sali, *Biophys. J.* **105**, 962 (2013).
10. J. Bardhan, S. Park, and L. Makowski, *J. Appl. Cryst.* **42**, 932 (2009).
11. H. Liu, A. Hexemer, and P. H. Zwart, *J. Appl. Cryst.* **45**, 587 (2012).
12. A. Grishaev, L. Guo, T. Irving, and A. Bax, *J. Am. Chem. Soc.* **132**, 15484 (2010).
13. S. Grudinina, M. Garkavenko, and A. Kazennov, *Acta Cryst.* **D73**, 449 (2017).
14. A. Hyman and P. A. Vaughan, *Small-Angle X-Ray Scattering* (Gordon and Breach, New York – London – Paris, 1967).
15. *International Tables for Crystallography*, vol. C., 3rd ed., Ed. by E. Prince (Kluwer Academic Publishers, Dordrecht – Boston – London, 2004).
16. B. Lee and F. M. Richards, *J. Mol. Biol.* **55**, 374 (1971).
17. Б. К. Вайнштейн, *Дифракция рентгеновых лучей на цепных молекулах* (Изд-во АН СССР, М., 1963).
18. A. Bondi, *Molecular Crystals, Liquid and Glasses* (John Wiley, New York, 1968).
19. Б. А. Федоров и О. Б. Птицын, *ДАН СССР* **233**, 716 (1977).
20. S. Park, J. P. Bardhan, B. Roux, and L. Makowski, *J. Chem. Phys.* **130**, 134014 (2009).
21. B. H. Biersmith, M. Hammel, E. R. Geisbrecht, and S. Bouyain, *J. Mol. Biol.* **408** (4), 616 (2011).
22. R. W. Strange, S. Antonyuk, M. A. Hough, et al., *J. Mol. Biol.* **328**, 877 (2003).
23. N. Nishimura, K. Hitomi, A. S. Arvai, et al., *Science* **326**, 1373 (2009).
24. M. Hammel, H.-P. Fierobe, M. Czjzek, et al., *J. Biol. Chem.* **280**, 38562 (2005).
25. D. I. Svergun, S. Richard, M. H. J. Koch, et al., *Proc. Natl. Acad. Sci. USA* **95**, 2267 (1998).
26. F. T. Burling, W. I. Weis, K. M. Flaherty, and A. T. Brünger, *Science* **271**, 72 (1996).
27. S. J. Perkins, *Eur. J. Biochem.* **157**, 169 (1986).

28. M. Gerstein, C. Chothia, Proc. Natl. Acad. Sci. USA **93**, 10167 (1996).
29. J. L. Finney, in *Water: A Comprehensive Treatise*, Ed. by F. Franks (Plenum, New York, 1979), V. 6, p. 47.
30. F. M. Richards, Annu. Rev. Biophys. Bioeng. **6**, 151 (1977).
31. S. Hubbard, K. Hodgson, and S. Doniach, J. Biol. Chem. **263**, 4151 (1988).
32. X. Cheng and B. P. Schoenborn, Acta Cryst. **B46**, 195 (1990).
33. J. Budger, Biophys. J. **65**, 1656 (1993).
34. F. Merzel and J. C. Smith, Proc. Natl. Acad. Sci. USA **99**, 5378 (2002).
35. M. Koizumi, H. Hirai, T. Onai, et al., J. Appl. Cryst. **40**, s175 (2007).
36. K. J. Bowers, D. E. Chow, H. Xu, et al., in *SC 2006 Conference, Proc. of the ACM/IEEE* (IEEE, 2006), p. 43.
37. A. A. Koshevoy, E. O. Stepanov, and Yu. B. Porozov, Biophysics **59**, 28 (2014).
38. G. S. Tamazian, J. H. Chang, S. N. Knyazev, et al., Proteins: Structure, Function and Bioinformatics **83**, 2217 (2015).
39. D. Schneidman-Duhovny, M. Hammel, J. A. Tainer, and A. Sali, Nucl. Acids Res. **44**, W424 (2016).
40. S. Yang, S. Park, L. Makowski, and B. Roux, Biophys. J. **96**, 4449 (2009).
41. W. Zheng and S. Doniach, J. Mol. Biol. **316**, 173 (2002).
42. Y. Wu, X. Tian, M. Lu, et al., Structure **13**, 1587 (2005).
43. D. Y. Guo, R. H. Blessing, D. A. Langs, and G. D. Smith, Acta Crystallogr. D, Biol. Crystallogr. **55**, 230 (1999).
44. D. Harker, Acta Crystallogr. **6**, 731 (1953).
45. D. Y. Guo, G. D. Smith, J. F. Griffin, and D. A. Langs, Acta Crystallogr. A **51**, 945 (1995).
46. D. I. Svergun, M. V. Petoukhov, and M. H. J. Koch, Biophys. J. **80**, 2946 (2001).
47. L. Makowski, D. Gore, S. Mandava, et al., Biopolymers **95**, 531 (2011).
48. A. V. Smirnov, I. N. Deryabin, and B. A. Fedorov, J. Appl. Cryst. **48**, 1089 (2015).

## **SASCUBE: an Updated Method of Cubes for Calculation of Intensity of X-ray Scattering by Biopolymers in Solution**

**B.A. Fedorov\*, A.V. Smirnov\*, V.V. Yaroshenko\*, and Yu.B. Porozov\* \*\***

*\*St. Petersburg National Research University of Information Technologies, Mechanics and Optics, Kronverskii prosp. 49, St. Petersburg, 197101 Russia*

*\*\*Sechenov First Moscow State Medical University, ul. Trubetskaya 8, Moscow, 119991 Russia*

The work describes an updated method of cubes, which allows calculating the SAS curves for biopolymers in solution on the basis of coordinates of their atoms without introducing any fitting parameters. The efficiency of the method was checked on a number of model systems. For eight globular proteins, the scattering intensities, calculated using SASCUBE software (<https://sourceforge.net/projects/sascube>), were compared with the corresponding experimental curves. This work addresses a contentious issue on the nature of distortions of the electron density of water near the protein surface. Molecular dynamic simulations were used to show that the electron density distribution of water in a 0.3 nm-thick near-border protein layer could not be considered uniform, while its average could be less than the average electron density of water. It was shown that the conformational mobility significantly influenced the scattering curve of proteins. The possibility of using the method of cubes (SASCUBE software) in combination with molecular dynamics is discussed in order to find the intensities of protein scattering with all influences on the scattering curve (conformational mobility of proteins, distortions of water structures etc.) factored in.

*Keywords: small-angle X-ray solution scattering, molecular dynamics, protein structure in solution, water structure*