*КРИСТАЛЛОГРАФИЯ, 2021, том 66, № 5, с. 793–801*

## СТРУКТУРА МАКРОМОЛЕКУЛЯРНЫХ СОЕДИНЕНИЙ

*УДК 539.219.1*

# О ТАКТИКЕ *AB INITIO*-ПОИСКА ФОРМЫ БЕЛКОВЫХ ЧАСТИЦ ПО ДАННЫМ МАЛОУГЛОВОГО РЕНТГЕНОВСКОГО РАССЕЯНИЯ

**© 2021 г. В. В. Волков[1,*]**

[1] *Институт кристаллографии им. А.В. Шубникова ФНИЦ "Кристаллография и фотоника" РАН, Москва, Россия*

*****E-mail: volkicras@mail.ru*

Поступила в редакцию 23.09.2020 г.
После доработки 02.01.2021 г.
Принята к публикации 12.03.2021 г.

Определение 3D-формы частиц по одномерным данным малоуглового рассеяния от растворов макромолекул неоднозначно. В силу плохой обусловленности обратной задачи, решение неустойчиво и зависит от параметров алгоритма поиска. Следовательно, на практике необходимо не только оценивать степень стабильности решения, но и подбирать параметры как метода решения, так и самой модели. Рассмотрена тактика поиска, заключающаяся в последовательном определении набора моделей формы частицы, представленной в виде структуры, состоящей из плотноупакованных шариков малого размера. Набор решений получают при варьировании относительного вклада членов целевой функции: критерия отклонения модельной кривой рассеяния от экспериментальной, штрафов за рыхлость и разрыв структуры тела, за отклонение среднего числа контактов шариков от заданного значения. Приведены примеры решения модельных задач и определения формы молекул по измерениям, выложенным в банке малоугловых данных и моделей SASBDB.

**DOI:** 10.31857/S0023476121050234

## ВВЕДЕНИЕ

К настоящему времени разработано несколько методов *ab initio*-восстановления формы с использованием модели малых объемных элементов [1–3]. Поиск 3D-формы частицы осуществляется минимизацией суммарной квадратичной невязки между экспериментальной кривой рассеяния и теоретической, рассчитываемой от модели структуры, представленной набором плотноупакованных элементов малого объема. В качестве элементов берут шарики или кубики такого размера, чтобы дифракция на их упаковке влияла на кривую рассеяния за пределами экспериментального диапазона. Кубические элементы позволяют более однородно заполнить пространство, однако и в случае шариков поправку на однородность легко учесть с помощью поправочного коэффициента. Поиск решения (пространственного размещения шариков в модели) ведут с помощью методов глобальной минимизации, например моделированием отжига (simulated annealing [4]) [1, 2] или с помощью генетических алгоритмов [3]. Алгоритм, основанный на поиске решения методом моделирования отжига, оказался исключительно эффективным в силу ряда алгоритмических решений, в частности из-за отказа от вариации на каждом шаге поиска всех параметров модели (число параметров равно числу узлов в прямом пространстве, в которых могут быть размещены шарики, заполняющие пространство и моделирующие электронную плотность модели). Вместо этого проводят варьирование только одного параметра, выбираемого случайным образом. При этом экспериментальные данные рассеяния приближают с наложением ограничений (прежде всего, плотность и неразрывность структуры) на решение.

## ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

Коротко основная идея моделирования, реализованная в программе DAMMIN [1], заключается в следующем. Область поиска (например, сфера с диаметром, равным максимальному размеру частицы $D_{max}$) разбивается на $N$ ($\sim 10^3$–$10^5$) узлов с плотнейшей упаковкой. $D_{max}$ определяют из функции парных расстояний $p(r)$, вычисленной из экспериментальных данных $I(s)$, например, с помощью программы GNOM [5] из пакета ATSAS [6] по формуле

$$p(r) = \frac{1}{2\pi^2}\int_{s=0}^{\infty} I(s)\frac{\sin(sr)}{sr}ds, \qquad (1)$$

где $r$ – длина отрезка в структуре, $s = 4\pi\sin\theta/\lambda$ – модуль вектора рассеяния, $2\theta$ – угол рассеяния, $\lambda$ – длина волны излучения. Каждому $j$-му узлу приписывается индекс $X_j$, обозначающий фазу,

которой принадлежит данный узел ($X_j = 0$ означает растворитель, $X_j = 1$ – частица). Таким образом, структура модели описывается конфигурационным вектором длиной $N$. В методе DAMMIN интенсивность рассеяния от такой модели быстро рассчитывается с помощью мультипольного разложения функции амплитуды рассеяния от объемного элемента. Тактика моделирования отжига заключается в старте с произвольной модели и ее последующих случайных модификациях до тех пор, пока теоретическое рассеяние от модели не станет приближать данные. В обычном методе Монте-Карло на каждом шаге используют две модели – лучшую и пробную, полученную случайной вариацией лучшей модели. Пробная модель отвергается, если она хуже, и принимается в качестве лучшей в противном случае, после чего цикл повторяют. Однако такой метод, несмотря на его стохастическую природу, на практике не способен найти решение за приемлемое время, застревая в широких локальных минимумах целевой функции. В методе отжига в качестве текущей структуры принимают модель, которая не обязательно лучше предыдущей. Случайную вариацию проводят относительно текущей модели и получают пробную. Вероятность принять худшую пробную модель в качестве новой текущей зависит от некоторого параметра, который называют в литературе "температурой" $T$. Большая температура означает высокую вероятность, вычисляемую как $\exp(-\Delta/T)$, принять в качестве новой текущей модели худшую, если $\Delta$ (пробное значение целевой функции минус текущее) больше 0 (решение хуже). Если пробная модель оказывается лучше, т.е. $\Delta < 0$, то она всегда принимается в качестве новой текущей. В начале поиска температуру выбирают достаточно высокой, чтобы частота принятия худших решений в качестве текущих превышала частоту обновления лучших моделей. Это заставляет программу "блуждать" по области поиска и в случае попадания модели в окрестность другого локального минимума, в конце концов, выбираться из текущего. По мере роста числа испытаний температуру периодически снижают. Чаще всего используют мультипликативный закон $T_{next} = T_{curr}F$, где $F$ – фактор отжига, равный в программе DAMMIN 0.9–0.98. Значение целевой функции уменьшается по аналогии с уменьшением внутренней энергии системы по мере снижения температуры, из-за чего метод и получил название "моделирование отжига" [4].

Поскольку для определения формы используются данные рассеяния, диапазон которых соответствует низкому пространственному разрешению, искомая модель также должна иметь более низкое разрешение по отношению к размеру объемного элемента. Поэтому на модель накладываются условия связности и плотности. Полная целевая функция, использованная в данной работе, имеет вид штрафной функции, зависящей от элементов конфигурационного вектора:

$$\Phi(\mathbf{X}) = w_R R + w_D P_D + w_L P_L + w_C P_C + \ldots. \qquad (2)$$

Здесь $R$ – критерий сходства модельной и экспериментальных кривых рассеяния

$$R = \left\{ \frac{\sum_{i=1}^{N} [(I_{\exp}(s_i) - \xi I_{\mathrm{mod}}(s_i))W(s_i)]^2}{\sum_{i=1}^{N} [I_{\exp}^2(s_i)W^2(s_i)]} \right\}, \qquad (3)$$

где $\xi = \dfrac{(\mathbf{I}_{\exp} \cdot \mathbf{I}_{\mathrm{mod}})}{\|\mathbf{I}_{\exp}\|^2}$ – МНК-множитель, совмещающий кривые рассеяния, умноженные на $W(s)$ – весовую функцию, которую назначают в виде

$$W(s) = \begin{cases} s^n, & n = 0,1,2,3,4 \\ \text{если} & s > s\big|_{\max[I_{\exp}(s)s^n]}, \\ 0.5\left\{ s^n + \dfrac{\max[I_{\exp}(s)s^n]}{I_{\exp}(s)} \right\} & \\ \text{если} & s \le s\big|_{\max[I_{\exp}(s)s^n]}. \end{cases} \qquad (4)$$

Умножение на $W(s)$ позволяет выравнивать вклад отклонений вдоль кривой рассеяния: при $n = 0$ расчет невязки (3) происходит на исходной шкале данных (что имеет смысл при спаде интенсивности не более одного порядка), с увеличением $n$ кривая интенсивности трансформируется в контур, проходящий через максимальное значение на некотором $s = s\big|_{\max[I_{\exp}(s)s^n]}$ и спадающий до некоторого значения, тем большего, чем больше $n$. Тем самым достигается возможность ослабить вклад в суммарную невязку начального участка, который наиболее подвержен искажающему влиянию рассеяния от агрегатов частиц, и одновременно увеличить вклад малоинтенсивного участка при больших $s$. Нижняя строка в (4) служит для того, чтобы ослабление начального участка было не слишком большим. На практике $n$ рекомендуется выбирать таким образом, чтобы величина $\dfrac{\max[I_{\exp}(s)s^n]}{\min[I_{\exp}(s)s^n]}$ была в диапазоне 10–50. При таком взвешивании обычное в МНК-методе деление невязок на стандартное отклонение экспериментальных шумов не применяется.

Для адекватного описания формы частицы и исключения влияния на кривую рассеяния упаковки шариков их диаметр должен быть достаточно мал. Для обеспечения этого число узлов в области поиска выбирают не менее 2000–5000 для компактных частиц и до 10000–30000 для анизометричных тел. При этом размер шариков оказывается в 5–10 раз меньше, чем пространственное

разрешение структурной модели, и для описания однородных областей структуры они должны находиться в плотноупакованном состоянии. Для обеспечения этого в целевую функцию добавлены штрафные члены, отражающие требования неразрывности структуры $w_D P_D$ и ее диффузности $w_L P_L$. Величину $P_D$ вычисляют как отношение общего числа шариков структуры к числу шариков, составляющих максимальный связный домен. $P_L$ (штраф за "рыхлость") вычисляют по формуле $P_L = 1 - \langle 1 - \exp(-N_e) + \exp(-12) \rangle_N$, где 12 – максимальное число контактов шарика с соседями, $N_e$ – фактическое число контактов, $\langle\ \rangle_N$ означает усреднение числа контактов, приведенное к одному шарику по структуре. Для весовых коэффициентов $w_D$ и $w_L$ рекомендованы значения 0.01–0.001 из соображения примерного равенства вкладов от $R$ и штрафов в точке минимума функционала (2). $P_C$ – штраф за относительное смещение центра тяжести частицы из центра области поиска. Этот член предотвращает "прилипание" модели к границе области и тем самым влияние границы на форму.

Намного более эффективный алгоритм поиска шариковых моделей реализован в программе DAMMIF [2]. Если DAMMIN осуществляет поиск модели в ограниченной области пространства и вхолостую просматривает много узлов, находящихся вдали от сформировавшегося структурного кластера, то DAMMIF всегда работает со связной структурой в неограниченном пространстве. Можно сказать, что DAMMIN "собирает" частицу, а DAMMIF ее "выращивает", работая при этом в 5–10 раз быстрее и делая значительно больше случайных модификаций. Тем не менее постоянная связность графа структуры может ограничивать ее изменчивость в ходе поиска, поэтому для определения разнообразия возможных моделей программу DAMMIF (как и DAMMIN) следует запускать несколько раз с разными параметрами: числом модификаций в цикле постоянной температуры, числом принимаемых модификаций до перехода к следующей температуре, начальной температурой и фактором ее снижения, балансом весов штрафов. Такая работа слишком затратная по времени и требует от пользователя большого опыта решения модельных задач.

В данной работе рассмотрен модифицированный алгоритм DAMMINV, автоматизирующий установку значений параметров поиска. Близкие результаты в некоторых случаях могут быть получены и с использованием DAMMIN и DAMMIF при соответствующей подготовке серии текстовых файлов заданий или повторном запуске программ в диалоговом режиме. Модифицированная версия DAMMINV ориентирована на последовательный поиск 10–15 решений в режиме варьирующейся величины весовых коэффициентов при штрафных членах ("перемежающийся режим", "alternating mode"). В DAMMINV введен дополнительный штраф за наличие шариков, не находящихся в контакте с основной структурой (наподобие алгоритма DAMMIF, но радиус контакта может быть увеличен для повышения степени варьируемости модели; подробнее ниже). Для моделирования диффузных структур введен также дополнительный штраф за отклонение среднего числа контактов шариков от заданного значения, который заменяет штраф за рыхлость (например, штраф за отклонение числа контактов от 2 форсирует поиск структуры в виде рыхлого клубка).

## ЭКСПЕРИМЕНТАЛЬНАЯ ЧАСТЬ

### *Предлагаемая последовательность шагов поиска формы частицы*

1. Выбрать правило взвешивания интенсивности рассеяния $W$ (формула (4)). Для этого можно воспользоваться программой SASPLOT из [6], в которой можно оценить площади левой и правой половин кривой рассеяния. В большинстве случаев степень $n$ в формуле (4) должна обеспечивать 3–10-кратное доминирование площади левой (начальной) половины взвешенной кривой, тем большее, чем шире угловой диапазон данных рассеяния. В DAMMINV выбор делается автоматически.

Важное замечание: при расчете функции парных расстояний при наличии небольшой агрегации частиц не следует искусственно уменьшать $D_{\max}$ во избежание артефактов в решении. Артефакты выражаются в виде распластывания структуры вблизи поверхности области поиска. Вместо этого лучше увеличивать начальный угол в данных, отбрасывая область, искаженную рассеянием от больших образований. В таком случае артефакты в моделях формы часто выглядят как легко различимые отдельные домены или "хвосты".

2. Для быстрого предварительного поиска модели можно установить больший радиус шарика, чем предлагается по умолчанию, из расчета 3–5 тысяч узлов в области поиска (или, например, удвоенный радиус в DAMMIF). Так как разрешение данных малоуглового рассеяния (**МУР**) невысокое, детали модели, меньшие чем 1/5–1/10 от максимального диаметра частицы, отражают только наличие малоразмерных внутренних неоднородностей, но не их фактическую структуру. Заметим также, что в программе DAMMIF можно увеличить число успешных модификаций модели (критерий перехода к следующей температуре), по умолчанию их доля составляет 0.1 от максимального числа модификаций при постоянной температуре и ее можно увеличить до 0.3–0.5. Далее будет показано, что в некоторых случаях такое

увеличение приводит к нахождению более адекватной структуры.

3. Стартовую температуру $T_{init}$ и множитель ее снижения $T_F$ при использовании программ DAMMIN и DAMMIF можно не варьировать. Однако слишком большая $T_{init}$ в DAMMIN приводит к большому числу начальных температурных циклов, в которых структура остается в виде случайного диффузного облака без существенного уменьшения невязки. В DAMMINV начальная температура устанавливается равной $T_{init} = 2.5\sigma_{\Phi(X)}$, где $\sigma_{\Phi(X)}$ – стандартное отклонение изменений целевой функции (2), вычисляемое с помощью таблицы разностей по значениям $\Phi(\mathbf{X})$, полученным путем 100–300 вариаций стартовой модели (обычно это случайным образом заполненная область поиска). Такой выбор $T_{init}$ обеспечивает начало снижения $\Phi(\mathbf{X})$ уже на втором–третьем температурном цикле в ходе поиска. В ходе поиска программа может дополнительно уменьшать температуру в случае, если критерий сходства формы экспериментальной и модельной кривых рассеяния (угол между векторами данных) становится больше 3°. Это предотвращает превращение сформированной к данному моменту модели в полностью случайное расположение шариков.

4. Осуществить поиск решения при $w_R = 1.0$ (уравнение (2)). В DAMMINV для повышения эффективности применен режим, названный **SBS** (Skip Bulk Solvent), при котором пропускаются модификации, размещающие новый шарик в узел, не контактирующий с шариками текущей модели, если относительное число оторванных шариков в модели становится меньше 3%. Это позволяет программе осуществить большее число вариаций связной структуры без увеличения их максимально допустимого числа и штрафа за наличие отдельно расположенных шариков. Этот режим почти аналогичен реализованному в алгоритме DAMMIF, но радиус контакта может превышать радиус упаковки шариков в сетке узлов (в этом случае штраф за отдельные шарики остается ненулевым). Для обеспечения большей варьируемости модели при поиске при включении SBS-режима температуру повышают в 1.5 раза.

5. Перед поиском следующего решения вес $w_R$ уменьшают в 10 раз, назначают $T_{init}$ в 1.5–2 раза меньше, чем на шаге 3. Вес штрафа за смещение центра тяжести частицы от центра области поиска $w_C$ увеличивают в 4 раза. Старт процедуры отжига проводят с полученной модели и получают решение с несколько худшим значением $R$, но лучшими значениями штрафов. Результат, полученный на шаге 4, обычно не рассматривают в качестве решения, если невязка $R$ значительно ухудшается. Цель этого шага – получение модели, в большей степени удовлетворяющей требованиям штрафов.

6. Веса всех штрафных членов возвращают в первоначальные значения, кроме штрафа за смещение центра тяжести частицы от центра области поиска, который уменьшают до значения 5% от первоначального (это безопасно, так как малая величина смещения уже получена на шаге 4). Устанавливают $T_{init} = 1.5\sigma_{\Phi(X)}$ или в 1.5–2 раза меньше первоначального значения, чтобы сохранить основные черты модели. Отметим, что минимизация штрафа за рыхлость тела нередко приводит к ориентированию частицы в области поиска так, что некоторые участки ее поверхности выстраиваются вдоль плоскостей упаковки шариков, и форма оказывается граненой. Величина штрафа $w_L P_L$ становится настолько малой, что полученная форма может сохраняться в следующих решениях. Чтобы этого избежать, на этом шаге в DAMMINV предусмотрен поворот тела относительно (фиксированной) области поиска на эйлеровы углы вокруг осей $X$ и $Z$ с соответствующим пересчетом координат шариков, чтобы направления граней не совпадали с плоскостями упаковки пространственных узлов. Так как новые координаты шариков теперь не совпадают с пространственной сеткой, структура частицы пересчитывается на ближайшие узлы. Переходят к шагу 3, на котором программа продолжает поиск решения, которое должно улучшить невязку, это основная задача данной фазы поиска. Итерации продолжают до достижения общего числа циклов поиска с фиксированной температурой 500–600.

Рассмотренная тактика направлена на получение решений, отличающихся друг от друга, и вместе с тем на повышение вероятности найти более глубокий минимум целевой функции. По разбросу полученных структур (обычно 6–10 моделей) с помощью процедур DAMAVER или DAMCLUST можно оценить стабильность решения, как это сделано, например, в [7].

Еще одним отличием программы DAMMINV является учет рассеяния от шариков модели, которое дает существенный вклад в области больших углов. Форм-фактор шариков заменен на рассеяние от эллипсоидов [8] с пропорциями осей 0.71 : 1.0 : 1.4, соответствующими типовым тепловым колебаниям атомов. Это уменьшает влияние на форму кривой рассеяния дифракции от пространственной упаковки элементов, что позволяет более адекватно определять структуры, модели которых состоят из небольшого числа шариков (менее 50–100), без необходимости добавлять или вычитать константу из данных рассеяния. Если такая необходимость существует (например, при неадекватном вычитании рассеяния от кюветы с растворителем), в программе предусмотрена соответствующая опция.

Однозначность определения структуры может быть выражена в терминах дисперсии решений

для конкретной задачи. Информацию о степени разброса моделей дают программы DAMAVER и DAMCLUST. Из-за отсутствия линейной связи между параметрами структуры (координатами шариков) и входными данными рассеяния выражение для дисперсии решений аналитически получить невозможно. Тем не менее можно дать ответ на вопрос о степени однозначности (или устойчивости), проводя анализ величин парных корреляций моделей в серии решений, которые получены методом случайного поиска и, вследствие этого, независимы. Конечно, такой ответ всегда является неполным, и для получения дополнительной информации об однозначности в пакете ATSAS предусмотрены дополнительные процедуры (программа AMBIMETER [6, 9]).

В общем случае сложно различить неоднозначность решения задачи определения структурной модели и его неустойчивость. И то, и другое приводит к тому, что находимые модели частиц отличаются друг от друга, если варьировать угловой диапазон данных рассеяния, проводить поиск с разных стартовых приближений, изменять параметры самого метода минимизации. Для получения окончательного ответа необходимо проводить серию численных экспериментов с варьированием условий поиска, анализировать отдельные решения с точки зрения "физического смысла", отбирать группы похожих решений (например, по $R$-фактору в прямом пространстве) и т.п. В итоге анализ данных МУР оказывается довольно сложной экспертной задачей, хотя нередко и удается получить окончательный ответ без проведения большого количества численных экспериментов. В данной публикации не рассматриваются все влияющие на решение факторы, так как их последовательный анализ может быть предметом не одной статьи.

Примеры экспериментальных данных брали из базы моделей белковых частиц, определенных по данным МУР [https://www.sasbdb.org/data]. Образцы отбирали по линейности графика Гинье и не слишком большому уровню шума.

*Пример неоднозначности модели частицы.* Факт неоднозначности отмечали многие исследователи, однако в литературе до сих пор не было примеров двух разных по структуре тел, МУР от которых было бы идентично. Идентичность кривых рассеяния означает, что телам должны соответствовать одинаковые функции парных расстояний $p(r)$ (1), из чего следует идентичность максимальных размеров $D_{max}$, объемов $V$ и радиусов инерции $R_g$ [8]. Это ограничивает число возможных форм частиц. В литературе существует пример расположения пяти точек на плоскости двумя разными способами при идентичных наборах парных расстояний между вершинами графов и дополнительно идентичности наборов площадей

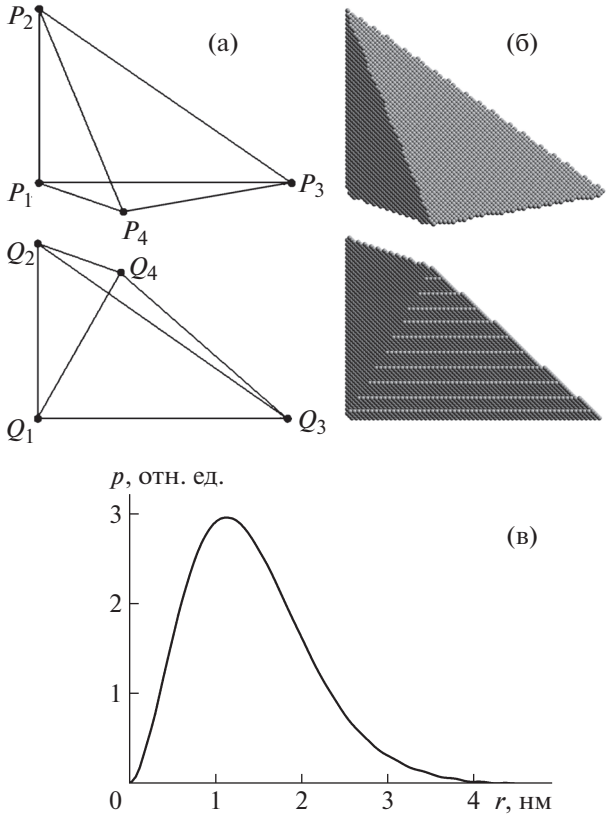


**Рис. 1.** Пример двух пирамидальных тел одинаковой высоты относительно плоскости рисунка (точки $P_4$ и $Q_4$ вынесены из плоскости), обладающих идентичными функциями парных расстояний $p(r)$: а – опубликованная плоская модель [10]; б – объемные шариковые модели, составленные примерно из 20 000 шариков; в – вид функций парных расстояний, рассчитанных по шариковым моделям (совпадают с точностью до толщины линии на графике).

треугольников [10]. Эксперимент с опубликованным примером показал, что, если для обеих структур вывести центральную точку из плоскости на одинаковые расстояния (что обеспечивает равенство объемов полученных пирамид) и, полагая тела однородными по плотности, численно рассчитать функции расстояний $p(r)$ с уменьшающимся радиусом шариков, можно увидеть, что кривые распределений стремятся к совпадению. Формы этих моделей показаны на рис. 1.

*Влияние параметра взвешивания интенсивности рассеяния.* На рис. 2 представлены кривые рассеяния от частицы рис. 3 (*1*), взвешенные согласно формуле (4). Модель на рис. 3 (*2*) получена с помощью программы DAMMIN с установками по умолчанию с весом $W = s^2$. Тот же вес для DAMMINV, несмотря на перемежающийся режим, не привел к удовлетворительному восстановлению формы (рис. 3 (*3*)), тогда как $W = s^1$ (рис. 2 (*2*)) позволяет найти близкую к точной мо-

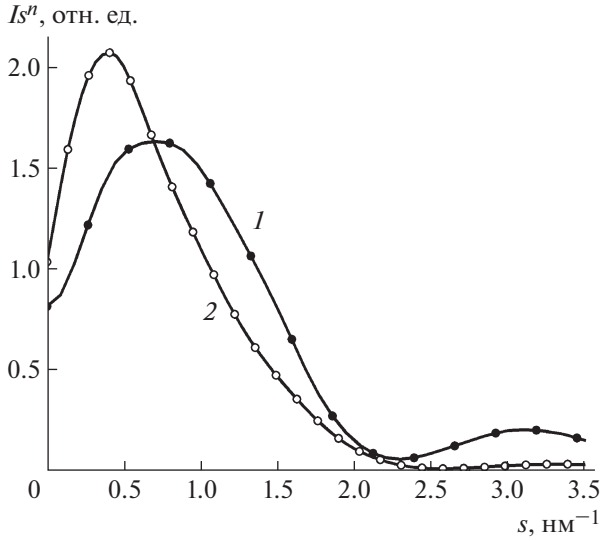


**Рис. 2.** Взвешенные интенсивности рассеяния, используемые для расчета невязки $R$. Точки – теоретическое рассеяние от частицы, представленной на рис. 3 (*1*), сплошная линия – типичная кривая рассеяния от найденных моделей, показанных на рис. 3 (*2*–*4*). На графике: *1* – весовая функция $W = s^2$, *2* – $W = s^1$.

дели форму (рис. 3 (*4*)). Значения критерия подгонки $R$ рассчитаны для невзвешенных данных.

*Поиск формы частиц в перемежающемся режиме.* Другой пример, демонстрирующий работу поиска с периодическим изменением относительных вкладов членов целевой функции, представлен на рис. 4, 5. В случае программы DAMMINV применяли вес $W = s^1$. Остальные параметры поиска оставляли по умолчанию. Во всех случаях значение $R$ составляло $(1 \pm 0.2) \times 10^{-4}$. Видно, что одиночные поиски DAMMIN и DAMMIF с установками по умолчанию и DAMMINV (формы *4*–*6* на рис. 5) не приводят к сколько-нибудь приемлемому результату, несмотря на хорошее соответствие кривых рассеяния (рис. 4). Последовательный поиск в перемежающемся режиме на пятой итерации дает приемлемое восстановление формы, сохраняющееся на следующих итерациях. Как отмечено ранее, увеличение числа принятых при фиксированной температуре модификаций относительно полного числа испытаний до 0.5 (100000 вместо 20000) в данном случае привело к нахождению адекватной модели во всех трех случаях.

*Формы молекулы бычьего альбумина как пример анализа данных с частичной агрегацией белка.* Приведем результаты анализа типичного сложного случая изучения белка в растворе. На рис. 6а показаны кривые МУР от раствора альбумина. Расчет функции парных расстояний показал, что распределение имеет протяженное плечо в обла-

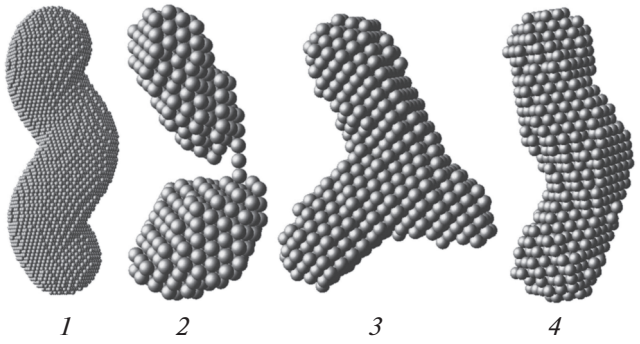


**Рис. 3.** Типичные результаты восстановления формы частицы (*1*) по теоретическим данным рассеяния, показанным на рис. 2: *2* – программа DAMMIN ($R = 1.3 \times 10^{-3}$, $W = s^2$), *3* – DAMMINV ($R = 0.73 \times 10^{-4}$, $W = s^2$, перемежающийся режим), *4* – DAMMINV ($R = 0.71 \times 10^{-4}$, $W = s^1$, перемежающийся режим).

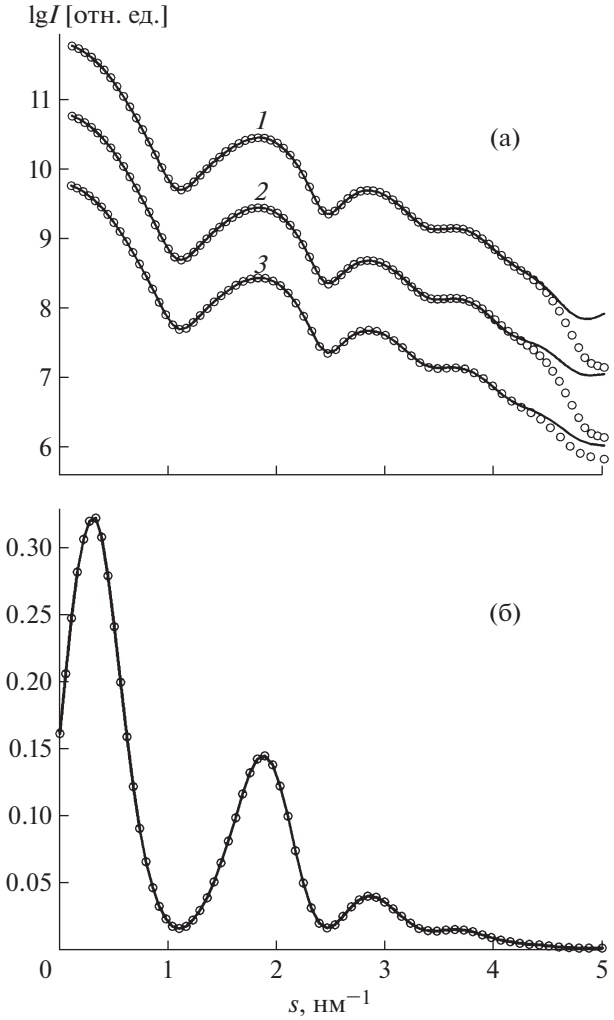


**Рис. 4.** Результаты подгонки модельной интенсивности рассеяния (точки), рассчитанной от структуры *1* на рис. 5. Сплошными линиями показано рассеяние от найденных структур: *1* – DAMMIN, *2* – DAMMIF, *3* – DAMMINV (структура *6* на рис. 5) (а). Взвешенная нормированная интенсивность рассеяния для расчета невязки $R$ (б).

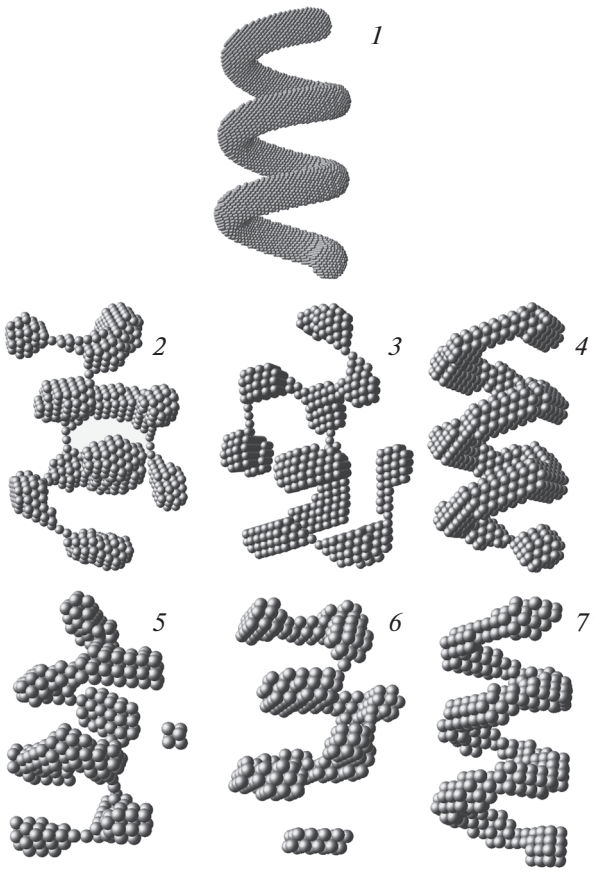


**Рис. 5.** Результаты моделирования исходной структуры (*1*): программа DAMMIN (*2*); DAMMIF с установками по умолчанию (*3*), DAMMIF с увеличенным до 100000 числом принятых при фиксированной температуре модификаций (*4*); DAMMINV – *5* – после первой итерации, *6* – после третьей, *7* – после пятой. Нечетные номера итераций соответствуют шагу 4 перемежающейся моды, который соответствует исходному (большому) относительному вкладу члена $R$ в целевой функции.

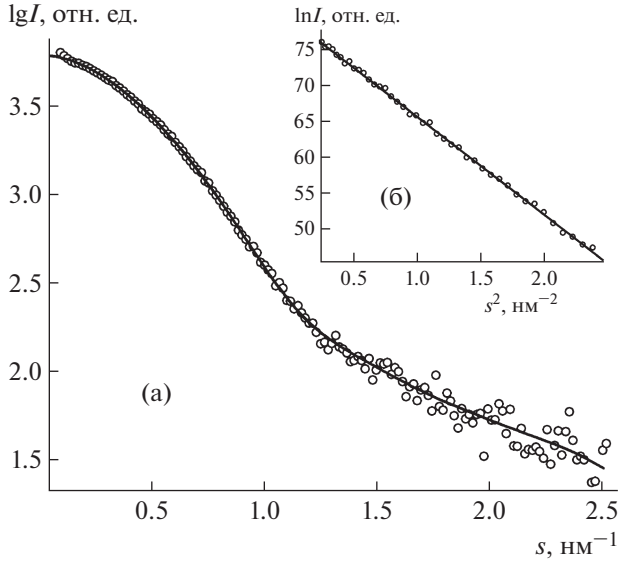


**Рис. 6.** Экспериментальная интенсивность рассеяния от раствора бычьего альбумина (точки) и типичная модельная кривая (линия) от структур, приведенных на рис. 8 (а); график Гинье для оценки радиуса инерции $R_g$ (б).

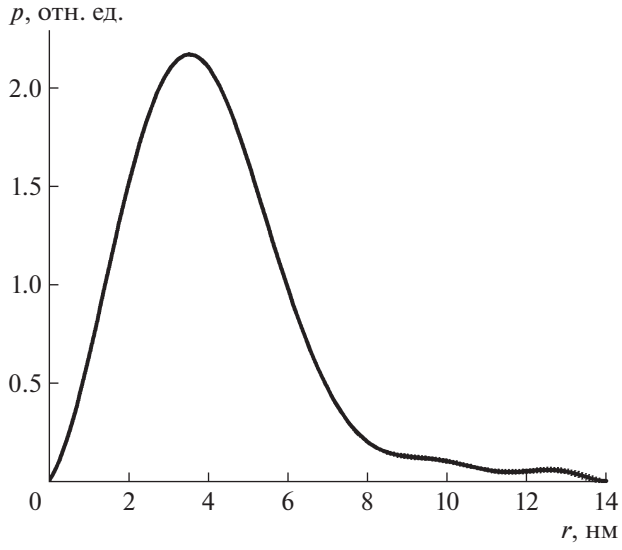


**Рис. 7.** Функция парных расстояний, рассчитанная по данным рис. 6а.

сти больших, более 9 нм, длин отрезков (рис. 7), что свидетельствует о частичной агрегации белка в растворе, хотя область Гинье прямолинейна (рис. 6б) и об этом не свидетельствует. Соответствующие радиусы инерции, найденные по области Гинье и по функции парных расстояний, равны соответственно $3.25 \pm 0.02$ и $3.26 \pm 0.04$ нм. Их близость также не свидетельствует об агрегации, хотя радиус инерции, рассчитанный в программе CRYSOL [11] по кристаллической структуре белка, даже с учетом гидратной оболочки меньше: 2.84–2.89 нм. Действительно, типичные модели формы молекулы (рис. 8), найденные в перемежающемся режиме поиска программой DAMMINV, демонстрируют наличие артефактов, искусственно увеличивающих длину максимальной хорды в структуре. Формы молекул определяли с использованием взвешивания $W = s^1$, так как взвешивание $W = s^2$ слишком сильно увеличивает интенсивность рассеяния на больших углах и результат оказывается несколько хуже. Так как артефакты формы возникают в решениях в разных местах структуры, усредненная с помощью DAMAVER структура не показывает хорошего совпадения с экспериментом ($\chi^2 = 10.4$) и здесь не приводится. Основной результат данного исследования – варианты формы макромолекулы после отбрасыва-

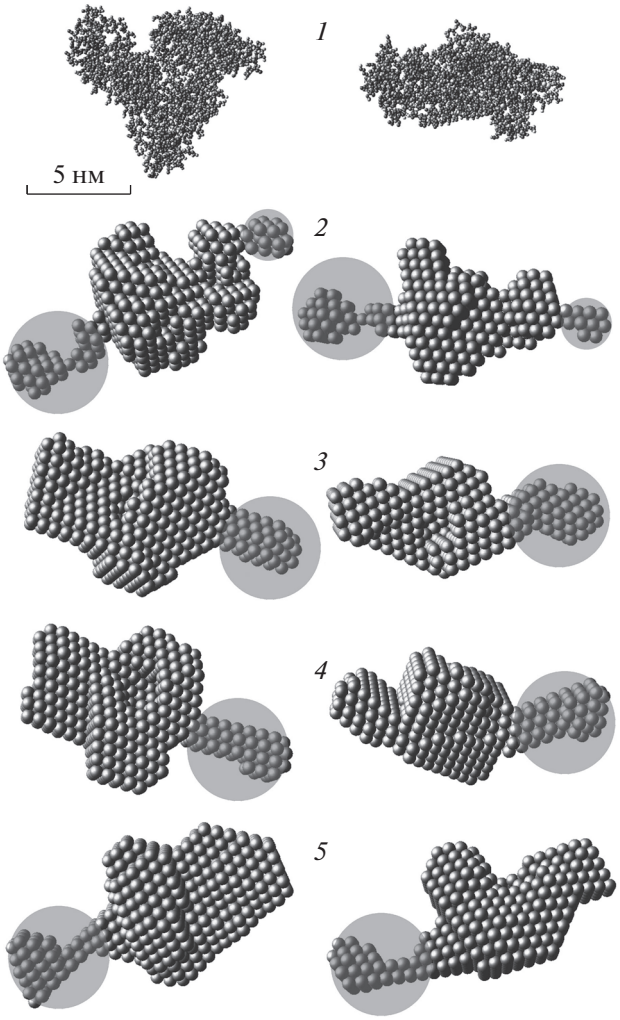


**Рис. 8.** Кристаллическая структура молекулы бычьего альбумина из белковой базы данных 3V03 (*1*) и последовательные решения DAMMINV после шага 3 (*2*–*5*). Кругами помечены артефакты из-за агрегации (предположительно – частичной агрегации молекул в растворе).

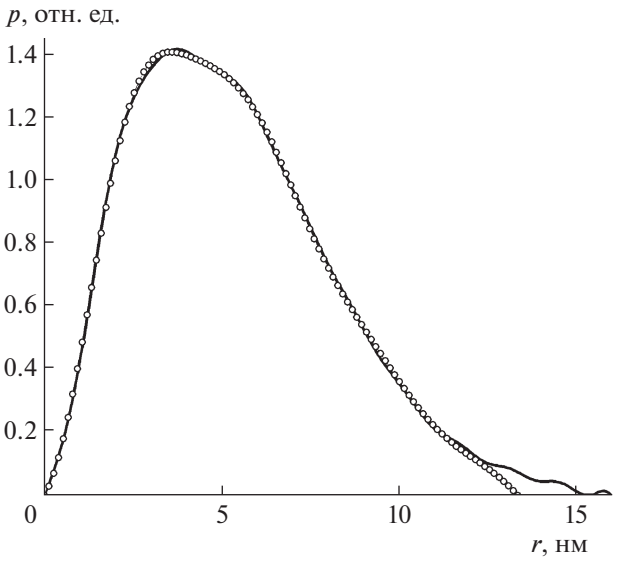


**Рис. 9.** Функции парных расстояний для образца FRET биосенсора, код данных в базе SASDAF5. Точки – опубликованные данные, линия – рассчитанное распределение, которое было использовано в качестве входных данных для программы DAMMINV.

ния артефактов агрегации, помеченных на рис. 8 серыми кругами. Эти варианты в силу предлагаемой тактики поиска должны отличаться друг от друга, охватывая круг возможных форм.

*Определение формы молекул по опубликованным данным*. Данные для анализа взяты из базы данных МУР от белковых растворов SASBDB [12]. Данные рассеяния и ссылки на публикации не приведены, так как их можно найти в базе по коду образца. По данным рассеяния рассчитаны распределения по расстояниям $p(r)$ без принудительного уменьшения максимального расстояния $D_{max}$. Теоретические кривые рассеяния от полученных моделей формы описывают данные с точностью не хуже $\chi^2 = 1.12$ и не представляют интереса для обсуждения. Поэтому приводим только наборы полученных структур.

На рис. 9 показаны опубликованная и рассчитанная кривые распределения по расстояниям для структуры SASDAF5 (FRET биосенсора), рис. 10 представляет опубликованные и рассчитанные структуры. Видно, что расчет с использованием расширенной $p(r)$ предоставляет структуры, более близкие по своему характеру к молекулярной модели, что может служить ее подтверждением. Пустоты малого размера в найденных моделях, как было сказано ранее, не являются деталями структуры, а отражают только попытку описать разреженность структуры в прилегающей области. Анализ структур с помощью DAMCLUST показал небольшой разброс моделей, и на рис. 10 представлены наиболее различающиеся.

## ЗАКЛЮЧЕНИЕ

Представлены некоторые результаты численных экспериментов по определению формы частиц из данных МУР от разбавленных изотропных монодисперсных систем и предложена схема последовательных расчетов с помощью программы моделирования шариковыми структурами. Весь спектр возможных случаев охватить в одной работе невозможно, но главным результатом проведенного исследования служит утверждение: предварительное моделирование с использованием искусственных структур и последовательный поиск решений с варьированием параметров алгоритма поиска могут помочь ответить на во-

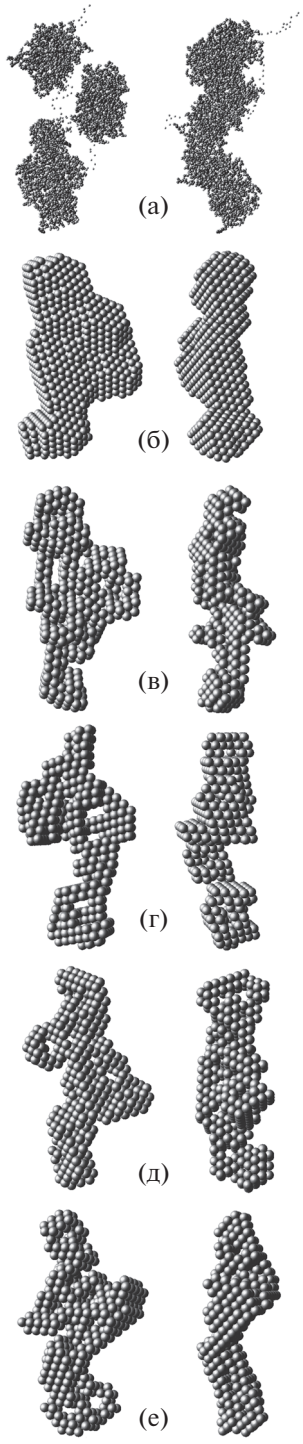


**Рис. 10.** Модели частицы FRET биосенсора (код SASDAF5): а, б – атомарная и шариковая опубликованные модели; в–е – найденные с помощью DAMMINV по данным рис. 9. Правый столбец – структуры повернуты на 90° относительно вертикальной оси.

прос о надежности полученных решений в исследовании белкового объекта.

Развитие методологии анализа данных МУР выполняется в тесном сотрудничестве с группой Д.И. Свергуна BIOSAXS [https://www.embl-hamburg.de/biosaxs/] в Европейской лаборатории молекулярной биологии (EMBL c/o DESY, Hamburg), которому автор выражает глубокую благодарность.

Работа выполнена при финансовой поддержке Российского научного фонда (грант № 19-14-00244) и Министерства науки и высшего образования в рамках выполнения работ по Государственному заданию ФНИЦ "Кристаллография и фотоника" РАН в части развития библиотеки подпрограмм общего назначения.

## СПИСОК ЛИТЕРАТУРЫ

1. *Svergun D.I.* // Biophys. J. 1999. V. 76. P. 2879. https://doi.org/10.1016/S0006-3495(99)77443-6
2. *Franke D., Svergun D.I.* // J. Appl. Cryst. 2009. V. 42. P. 342. https://doi.org/10.1107/S0021889809000338
3. *Chacon P., Moran F., Díaz J.F. et al.* // Biophys. J. 1998. V. 74. № 6. P. 2760. https://doi.org/10.1016/S0006-3495(98)77984-6
4. *Kirkpatrick S., Gelatt C.D., Vecchi M.P.* // Science. 1983. V. 220. P. 671. https://doi.org/10.1126/science.220.4598.671
5. *Semenyuk A.V., Svergun D.I.* // J. Appl. Cryst. 1991. V. 24. P. 537. https://doi.org/10.1107/S002188989100081X
6. *Manalastas-Cantos K., Konarev P.V., Hajizadeh N.R. et al.* // J. Appl. Cryst. 2021. V. 54. P. 1. https://doi.org/10.1107/S1600576720015368
7. *Volkov V.V., Svergun D.I.* // J. Appl. Cryst. 2003. V. 36. P. 860. https://doi.org/10.1107/S0021889803000268
8. *Свергун Д.И., Фейгин Л.А.* Рентгеновское и нейтронное малоугловое рассеяние. М.: Наука, 1986. 280 с.
9. *Petoukhov M.V., Svergun D.I.* // Acta Cryst. D. 2015. V. 71. P. 1051. https://doi.org/10.1107/S1399004715002576
10. *Boutin M., Kemper G.* // arXiv:math/0304192 [math.AC]. 2003. P. 1. https://arxiv.org/abs/math/0304192
11. *Svergun D.I., Barberato C., Koch M.H.J.* // J. Appl. Cryst. 1995. V. 28. P. 768. https://doi.org/10.1107/S0021889895007047
12. Small Angle Scattering Biological Data Bank SASBDB // https://www.sasbdb.org/