КРИСТАЛЛОГРАФИЯ, 2023, том 68, № 6, с. 941–945

ДИФРАКЦИЯ И РАССЕЯНИЕ ИОНИЗИРУЮЩИХ ИЗЛУЧЕНИЙ

УДК 539.219.1

ОПРЕДЕЛЕНИЕ ФОРМЫ СПИРАЛЬНОЙ ЧАСТИЦЫ ПО ДАННЫМ МАЛОУГЛОВОГО РЕНТГЕНОВСКОГО РАССЕЯНИЯ: МОДИФИКАЦИЯ АЛГОРИТМА “ИМИТАЦИИ ОТЖИГА”

© 2023 г. В. А. Григорьев[1,*], П. В. Конарев[1], В. В. Волков[1]

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

**E-mail: vasiliy.grigorev.1996@mail.ru*

Поступила в редакцию 09.06.2023 г.
После доработки 05.07.2023 г.
Принята к публикации 05.07.2023 г.

Модифицированный алгоритм “имитации отжига”, реализованный в программе DAMMINV, позволяет получать 10–15 различных моделей наночастицы, приближающих данные малоуглового рентгеновского рассеяния. Основой метода является режим перемежающихся весовых коэффициентов целевой функции, в котором балансируются вклады от штрафных коэффициентов, отвечающих за реалистичность модели, и вклад от невязки между экспериментальной и модельной кривыми рассеяния. Исследовано влияние шума на кривых рассеяния на качество восстановления трехмерной формы спирали, а также проведено сравнение результатов с полученными стандартными программами. Метод протестирован на модельных данных, свободных от шума и с наложением пуассоновского шума на примере спиральной частицы с толщиной витков, сопоставимой с характерным размером пространства между ними. Проведен сравнительный анализ восстановленных моделей с 3D-формами, полученными алгоритмами “имитации отжига” в стандартном режиме работы.

DOI: 10.31857/S0023476123600295, **EDN:** FZZKKB

ВВЕДЕНИЕ

Малоугловое рентгеновское рассеяние (**МУРР**) является мощным экспериментальным методом, позволяющим изучать структуру и динамику макромолекул, наночастиц, полимерных соединений, пористых материалов в масштабе от 1 до 200 нм. МУРР позволяет извлекать информацию о размере, форме, ориентации и взаимодействии структурных элементов наночастиц [1]. Поиск 3D-формы частицы в случае монодисперсной системы проводится по одномерным данным МУРР, измеренным в ограниченном угловом диапазоне. Отсюда вытекают такие проблемы, как неоднозначность определения формы и плохая обусловленность задачи поиска решения. В настоящий момент доступны различные пути решения данных проблем. В случае использования метода случайного поиска можно решать задачу десятки раз при одних и тех же начальных условиях и кластеризировать результаты по схожим признакам, как это делает, например, программа DAMCLUST [2].

Для поиска формы частицы распространенным является метод “имитации отжига” [3], который реализован в программе DAMMIN [4], а также в ее глубокой модификации DAMMIF [5], отличающейся тем, что максимальный размер частицы не ограничивается при поиске и трехмерная сетка модели является динамической величиной. Существуют и другие подходы, например в [6, 7] использована аппроксимация сферическими гармониками, которая в некоторых случаях может обеспечить однозначность восстановления формы [7], если число независимых параметров, описывающих форму, невелико (8–10) и больше числа шенноновских каналов N_{sh} ($N_{sh} = s_{max}D_{max}/\pi$, где s_{max} – максимальный модуль вектора рассеяния, D_{max} – максимальный размер частицы [8]). При этом число угловых отсчетов должно быть не меньше числа параметров. Однако в реальных данных из-за наличия шумов проблема однозначности возникает снова [7]. В [9] методом итеративного поиска структурных факторов вычислена электронная плотность различных биологических макромолекул. В [10] использовали метод машинного обучения на экспериментальных данных МУРР и автокодировщик 3D-моделей белков, обученный сжимать информацию о форме частицы в векторы 200-мерного пространства, которые оптимизируются с использованием генетических алгоритмов для построения 3D-моделей, приближающих данные МУРР.

Несмотря на разнообразие методов восстановления 3D-формы наночастиц по данным МУРР, проблему неоднозначности решений удается ограничить, а в некоторых случаях решить путем многократного запуска алгоритмов со случайными стартовыми параметрами модели. Работа [11] посвящена алгоритму поиска последовательного определения набора моделей формы частицы с одной стартовой случайной модели. В этом случае модифицированный режим работы алгоритма "имитации отжига" позволяет периодически варьировать параметры процедуры поиска таким образом, чтобы последовательно получить 10–15 потенциально различных моделей, соответствующих различным локальным минимумам целевой функции, включающей в себя невязку между экспериментальной и теоретической кривыми рассеяния и слагаемые, обеспечивающие компактность и связность модели. Находимые формы частиц нередко сильно зависят от баланса вкладов штрафных членов и, несмотря на близкие значения критерия качества подгонки к данным, могут в силу математической неоднозначности решения значительно различаться. Такое число моделей необходимо получать для того, чтобы была возможность выбирать из них те, которые наиболее соответствуют реальной форме частицы по другим косвенным данным. Однако из-за новизны метода имеется ряд требующих дополнительного исследования вопросов, таких как информация об оптимальном значении адаптированных параметров и о применении метода для поиска сложных форм частиц, например, имеющих высокую степень анизометрии с пропорциями 1 : 5–1 : 10. Отметим, что в [12] частично был решен вопрос нахождения оптимальных значений параметров модифицированного алгоритма "имитации отжига", таких как число шенноновских каналов и степень весовой функции, на которую взвешивается кривая рассеяния, для спиральных структур.

Настоящая работа посвящена сравнительному анализу модифицированного (DAMMINV [11]) и стандартного алгоритмов "имитации отжига" (DAMMIN и DAMMIF) на примере трехвитковой спирали с толщиной витков, сопоставимой с характерным размером пространства между ними, а именно расстоянием между витками и внутренним диаметром. Такая геометрия усложняет восстановление трехмерной формы, что делает ее привлекательной для исследования. Также рассмотрено влияние уровня статистического шума в данных МУРР на реконструкцию 3D-формы спирали.

МОДИФИЦИРОВАННЫЙ АЛГОРИТМ "ИМИТАЦИИ ОТЖИГА"

Форма частицы в программе DAMMIN и ее модификациях представляется в виде плотно упакованных однородных шариков, размещенных в узлах гексагональной упаковки. Теоретическая кривая рассеяния рассчитывается по формуле Дебая:

$$I_{\mathrm{mod}}(s) = n\Phi^2(sa)\left[1 + \frac{2}{n}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\frac{\sin sr_{ij}}{sr_{ij}}\right], \qquad (1)$$

где n, a – число и радиус шариков модели, r_{ij} – расстояние между центрами шариков, $\Phi^2(sa)$ – интенсивность рассеяния одним шариком [1]:

$$\Phi(t) = 3\frac{\sin t - t\cos t}{t^3}, \quad t = sa. \qquad (2)$$

Для ускорения расчетов расчет интенсивности рассеяния проводится с помощью математического аппарата сферических гармоник, в ряд по которым раскладывается амплитуда (2) до 20–50 номера гармоники [4].

Для стартовой модели рассчитывается целевая функция от вектора параметров **X**:

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

где w_i – весовые коэффициенты, P_D, P_L, P_C – штрафы, отвечающие соответственно за разрывность, рыхлость частицы и отклонение центра тяжести частицы от центра области поиска, R – значение невязки между экспериментальной и теоретической кривыми рассеяния:

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

где $\xi = (I_{\mathrm{exp}}I_{\mathrm{mod}})/\|I_{\mathrm{exp}}\|^2$ – шкалирующий множитель, $W(s)$ – весовая функция:

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

Целевая функция минимизируется с помощью алгоритма "имитации отжига" в режиме перемежающихся весовых коэффициентов, что и составляет основу модификации алгоритма (подробнее в [11]). В этом режиме после достижения минимума целевой функции весовые коэффициенты меняются так, чтобы на нечетных итерациях преобладала минимизация невязки (2), а на четных – минимизация вклада штрафных членов. Таким образом, после каждой температурной итерации рельеф целевой функции меняется, и текущее значение вектора **X** соответствует ненулевому градиенту целевой функции, что провоцирует

поиск нового минимума. К финальной итерации генерируется ансамбль из 10–15 моделей с наименьшими значениями целевой функции при некоторой комбинации весовых коэффициентов, каждую из которых можно рассматривать как искомую.

В данной работе начальные значения весовых коэффициентов брали следующие: вес штрафа за разрывность w_D – 7.5×10^{-4}–1.5×10^{-2}, вес штрафа за рыхлость w_L – 6.0×10^{-3}–6.0×10^{-2}. Степень весовой функции n = 1.5, 2.0. Данные значения лежат в диапазоне оптимальных значений, который был получен опытным путем, а также является результатом исследования [12]. Остальные параметры процедуры поиска установлены по умолчанию, определенные опытным путем на начальном этапе модификации алгоритма. В качестве частицы со сложной формой выбрана круговая спираль из трех витков с закругленными концами. Параметры модели следующие: шаг – 55 Å, высота – 165 Å, внешний и внутренний диаметры – 100 и 40 Å, толщина витка спирали – 30 Å, радиус шариков – 2.2 Å. Интенсивность рассеяния от модели была рассчитана по формуле (1) и взята в качестве экспериментальной на угловом диапазоне, соответствующем числу шенноновских каналов N_{sh} = 28. Шум на интенсивность накладывали в виде комбинации гауссовского и пуассоновского шумов согласно алгоритмам [13, 14], что соответствует системам регистрации большинства используемых рентгеновских детекторов. Для гауссовского шума, моделирующего шумы системы регистрации, дисперсия была σ = 0.5% (отн.), для пуассоновского – 1, 3, 5%. При этом согласно критерию Дарбина–Ватсона [15, 16] (который оказался равным 1.9, 2.0, 1.95 соответственно) последовательность полученных значений шумов является случайной с вероятностью не менее 98%. Качество реконструкции модели оценивали с помощью критерия **NSD** (normalized spatial discrepancy) [17], количественно отражающего сходство двух структурных моделей, предварительно выравненных между собой по минимальному среднеквадратичному расстоянию (в идеальном случае NSD = 0, при полном несоответствии NSD $\gg$ 1). В данной работе сходство оценивали между теоретической и восстановленной спиралями. Среди семи решений, полученных с помощью программы DAMMINV, выбиралось то, которое имеет наилучшее значение NSD.

РЕЗУЛЬТАТЫ И ИХ ОБСУЖДЕНИЕ

Эталонное значение NSD не равно нулю, поскольку радиусы шариков теоретической и восстановленной спиралей различаются. Теоретическая кривая была рассчитана от модели, содержащей 7800 шариков радиусом 2.2 Å, тогда как при восстановлении формы спирали методом имитации отжига радиус шарика выбирался равным 3.5 Å. Это связано с тем, что использование слишком большого числа шариков малого радиуса приводит не только к существенному увеличению времени поиска, но и к увеличению числа локальных минимумов, т.е. к увеличению числа вариантов решений. В данной работе эталонное значение NSD между точными моделями спиралей (с радиусом шариков 2.2 и 3.5 Å) составляет 0.4504. На рис. 1 представлены графики зависимости значений NSD от уровня пуассоновского шума для спиралей, реконструированных при различных параметрах алгоритма: весе штрафов за разрывность и рыхлость, а также степенях весовой функции $W(s)$. Видно, что добавление шума при любом из заданных уровней повышает значение NSD (за исключением отдельных точек), т.е. решения ухудшаются. Однако при определенных значениях весов штрафов NSD достаточно близко к теоретическому (кривая *3* при уровне шума 1% на рис. 1а и кривые *4*, *1* при уровне шума 1 и 3% соответственно на рис. 1б). Такое неоднозначное следствие наложения шума объясняется тем, что шум, с одной стороны, может приводить к смещению математического ожидания решения, с другой – расширяет коридор приближения теоретической кривой рассеяния к экспериментальной, что, возможно, сглаживает профиль целевой функции и облегчает поиск глобального экстремума. При этом диапазон возможных решений покрывает истинное решение, к которому алгоритм все еще может приблизиться. На рис. 2 для сравнения показаны значения NSD моделей, полученных с помощью программ DAMMIN и DAMMIF с параметрами по умолчанию ($w_D = w_L$ = 0.003, $W \sim s^{-4}$ и w_D = 0, w_L = 0.01, $W \sim s^{-4}$ соответственно). Для DAMMINV показаны значения, усредненные по всем решениям, приведенным на рис. 1. Очевидно, что модифицированный алгоритм справляется с восстановлением данной спирали намного лучше, чем стандартные программы. На рис. 3 показаны наилучшие восстановленные модели среди полученных при разных уровнях шума с помощью программ DAMMIN и DAMMIF, а также наилучшие восстановления с помощью программы DAMMINV для каждого уровня шума. Соответствующие кривые рассеяния (теоретическая и модельные) отображены на рис. 4. Несмотря на хорошее соответствие кривых рассеяния, рассчитанных с помощью программ DAMMIN ($R_F = 3.61 \times 10^{-3}$), DAMMIF ($R_F = 2.2 \times 10^{-3}$) и DAMMINV ($R_F = 1.08 \times \times 10^{-3}$, $R_F = 1.15 \times 10^{-3}$, $R_F = 0.95 \times 10^{-3}$, $R_F = 1.58 \times \times 10^{-3}$ для кривых *3*–*6* на рис. 4), только последняя программа с достаточной точностью позволяет описать теоретическую модель, в то время как DAMMIN и DAMMIF позволяют восстанав-

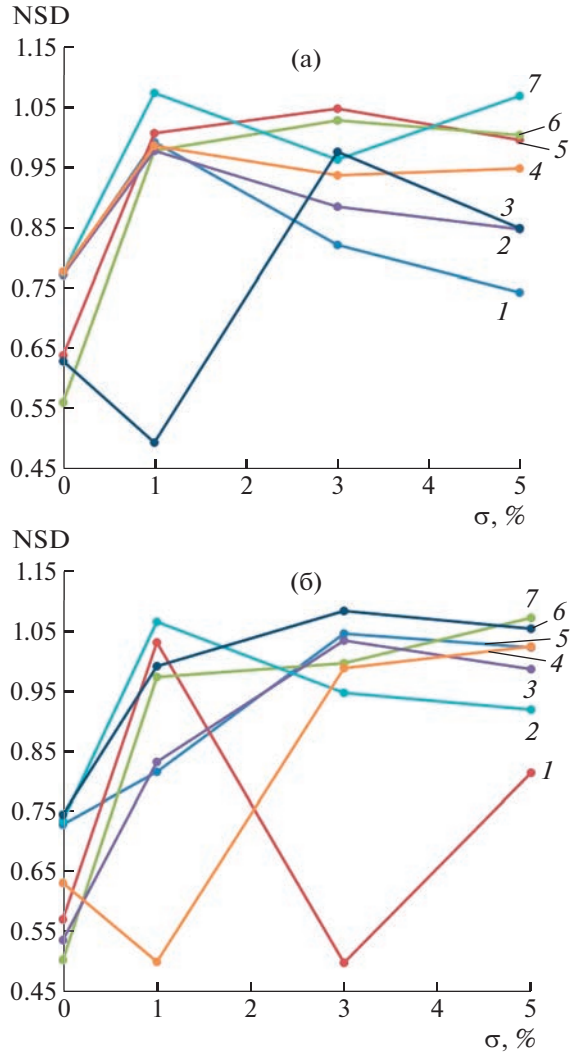

Рис. 1. Значения NSD для спиралей, восстановленных с помощью программы DAMMINV: а – весовая функция $W = s^{1.5}$, веса w_D, w_L: *1* – 0.005, 0.054, *2* – 0.003, 0.02, *3* – 7.5×10^{-4}, 0.03, *4* – 7.5×10^{-4}, 0.06, *5* – 0.005, 0.015, *6* – 0.003, 0.06, *7* – 0.0015, 0.012; б – $W = s^2$, веса w_D, w_L: *1* – 0.005, 0.048, *2* – 0.003, 0.036, *3* – 0.003, 0.042, *4* – 0.0015, 0.03, *5* – 0.015, 0.042, *6* – 0.0015, 0.02, *7* – 0.005, 0.036.

ливать не похожие на спираль структуры. Отметим, что для данных без шума и с пуассоновским шумом 1 и 3% спирали восстанавливаются практически идеально, тогда как для данных с шумом 5% на восстановленной модели (модель *7* на рис. 3) угадываются только некоторые характерные признаки спирали, а именно верхний и центральный витки, что указывает на неплохой результат, однако требующий повторного уточнения модели. Таким образом, наличие пуассоновского шума при числе точек на кривой рассеяния ~1000–2000 лишь незначительно влияет на хорошее качество восстановления структуры с помощью программы DAMMINV, которую не удается восстановить

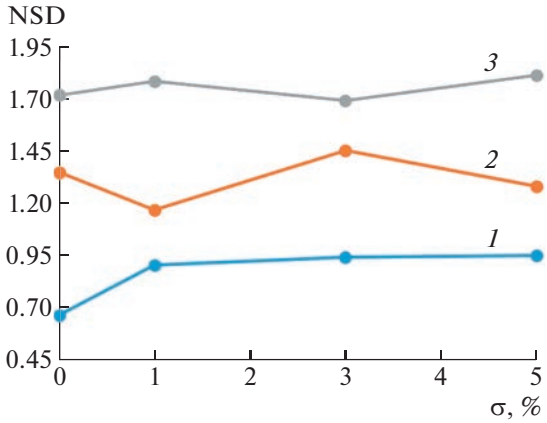

Рис. 2. Значения NSD по результатам расчетов с помощью программ DAMMINV (*1*), DAMMIN (*2*), DAMMIF (*3*). Для DAMMINV показаны значения, усредненные по данным на рис. 1.

программами DAMMIN и DAMMIF со значениями весов штрафов, задаваемыми по умолчанию. Несмотря на то что наличие умеренного уровня шума должно усложнять процесс восстановления

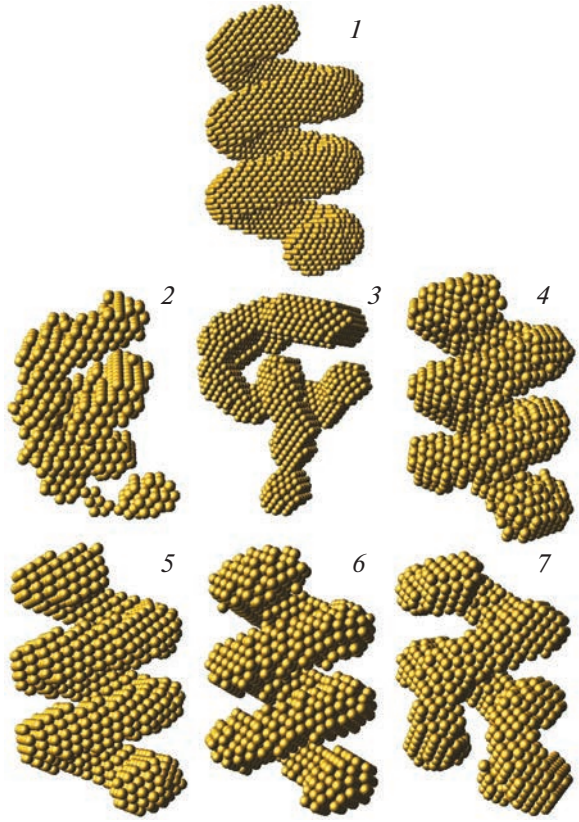

Рис. 3. Теоретическая (*1*) и восстановленные спирали с помощью программ DAMMIN (*2*), DAMMIF (*3*), DAMMINV (наилучшие решения): *4* – без шума, $W = s^2$, $w_D = 0.005$, $w_L = 0.036$, NSD = 0.50; *5* – σ = 1%, $W = s^{1.5}$, $w_D = 7.5 \times 10^{-4}$, $w_L = 0.03$, NSD = 0.49; *6* – σ = 3%, $W = s^2$, $w_D = 0.005$, $w_L = 0.048$, NSD = 0.50; *7* – σ = 5%, $W = s^{1.5}$, $w_D = 0.005$, $w_L = 0.054$, NSD = 0.74.

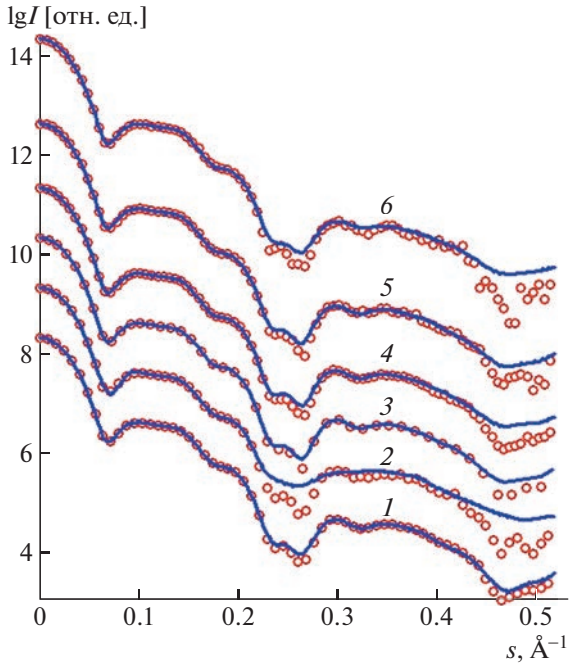

Рис. 4. Теоретические (кружки) и модельные (линии) кривые рассеяния: модельные кривые *1*–*6* (R_F = $= 3.61 \times 10^{-3}$, 2.2×10^{-3}, 1.08×10^{-3}, 1.15×10^{-3}, 0.95×10^{-3}, 1.58×10^{-3}) соответствуют моделям *2*–*7* на рис. 3 соответственно.

модели, модифицированный алгоритм “имитации отжига” эффективно работает и в таких условиях.

ЗАКЛЮЧЕНИЕ

Проанализировано восстановление спиральной структуры с помощью модифицированного алгоритма “имитации отжига” (DAMMINV), основанного на режиме перемежающихся штрафных весов целевой функции, и проведено сравнение с результатами, полученными программами DAMMIN и DAMMIF при стандартных параметрах поиска. Трехмерная форма спирали была восстановлена по модельным данным рассеяния без шума, а также с наложением пуассоновского шума с относительным уровнем 1, 3 и 5% и дополнительного гауссовского шума с уровнем 0.5%. Показано, что DAMMIN и DAMMIF при использовании параметров по умолчанию недостаточно эффективно восстанавливают пространственную форму спирали, тогда как DAMMINV позволяет получать решения, близкие к истинной модели для каждого уровня шума, рассмотренного в работе. Однако вопрос выбора оптимальных параметров для того или иного уровня шума, определяющих целевую функцию, остается открытым и является предметом для дальнейших исследований.

Работа выполнена при поддержке Министерства науки и высшего образования РФ в рамках выполнения работ по Государственному заданию ФНИЦ “Кристаллография и фотоника” РАН.

СПИСОК ЛИТЕРАТУРЫ

1. *Свергун Д.И., Фейгин Л.А.* Рентгеновское и малоугловое рассеяние. М.: Наука, 1986. 280 с.
2. *Petoukhov M.V., Franke D., Shkumatov A.V. et al.* // J. Appl. Cryst. 2012. V. 45. P. 342. https://doi.org/10.1107/S0021889812007662
3. *Kirkpatrick S., Gelatt C.D., Vecchi M.P.* // Science. 1983. V. 220. P. 671. https://doi.org/10.1126/science.220.4598.671
4. *Svergun D.I.* // Biophys J. 1999. V. 78. P. 2879. https://doi.org/10.1016/S0006-3495(99)77443-6
5. *Franke D., Svergun D.I.* // J. Appl. Cryst. 2009. V. 42. P. 342. https://doi.org/10.1107/S0021889809000338
6. *Svergun D.I., Stuhrmann H.B.* // Acta Cryst. A. 1991. V. 47. P. 736. https://doi.org/10.1107/S0108767391006414
7. *Svergun D.I., Volkov V.V., Kozin M.B. et al.* // Acta Cryst. A. 1996. V. 52. P. 419. https://doi.org/10.1107/S0108767396000177
8. *Shannon C.E., Weaver W.* The Mathematical Theory of Communication. University of Illinois Press, 1949. 125 p.
9. *Grant T.D.* // Nature Methods. 2018. V. 15. P. 191. https://doi.org/10.1038/nmeth.4581
10. *He H., Liu C., Liu H.* // iScience. 2020. V. 23. 100906.
11. *Волков В.В.* // Кристаллография. 2021. Т. 66. С. 793. https://doi.org/10.31857/S0023476121050234
12. *Григорьев В.А., Конарев П.В., Волков В.В.* // Успехи в химии и химической технологии. 2022. Т. 36. С. 53.
13. *Marsaglia G., Tsang W.W.* // SIAM J. Sci. Stat. Comput. 1984. V. 5. P. 349. https://doi.org/10.1137/0905026
14. *Devroye L.* // Computing. 1981. V. 26. P. 197. https://doi.org/10.1007/BF02243478
15. *Durbin J., Watson G.S.* // Biometrika. 1950. V. 37. P. 409. https://doi.org/10.1093/biomet/37.3-4.409
16. *Durbin J., Watson G.S.* // Biometrika. 1951. V. 38. P. 159. https://doi.org/10.2307/2332325
17. *Kozin M., Svergun D.* // J. Appl. Cryst. 2001. V. 34. P. 33. https://doi.org/10.1107/S0021889800014126