ЯДЕРНАЯ ФИЗИКА И ИНЖИНИРИНГ, 2023, том 14, № 4, с. 408–414

МОДЕЛИРОВАНИЕ НАНОСТРУКТУР

УДК 538.97

РАЗВИТИЕ АЛГОРИТМОВ ДЛЯ АНАЛИЗА ДАННЫХ МАЛОУГЛОВОГО РЕНТГЕНОВСКОГО РАССЕЯНИЯ ОТ ПОЛИДИСПЕРСНЫХ И ЧАСТИЧНО УПОРЯДОЧЕННЫХ СИСТЕМ

© 2023 г. П. В. Конарев[a, b, *], В. В. Волков[a]

[a] *Институт кристаллографии им. А.В. Шубникова, ФНИЦ "Кристаллография и фотоника" РАН, Москва, 119333 Россия*

[b] *Национальный исследовательский центр "Курчатовский институт", Москва, 123182 Россия*

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

Поступила в редакцию 08.07.2022 г.
После доработки 14.07.2022 г.
Принята к публикации 15.07.2022 г.

Метод малоулового рентгеновского рассеяния позволяет исследовать структуру растворов белков, полимеров и металлических наночастиц в диапазоне 1–200 нм. Развитие новых и усовершенствование имеющихся алгоритмов для анализа экспериментальных данных малоуглового рассеяния представляет собой важную и актуальную задачу. В данной работе представлен ряд алгоритмов, позволяющих находить функции распределения рассеивающих неоднородностей по размерам, восстанавливать профили интенсивности отдельных компонент в белковых смесях, оценивать размеры области кристалличности и межплоскостные расстояния в частично упорядоченных системах. Ряд алгоритмов реализован в виде программ с использованием кроссплатформенной графической библиотеки Qt, что значительно расширяет круг их потенциальных пользователей. Эффективность работы алгоритмов продемонстрирована на ряде экспериментальных данных малоуглового рассеяния.

Ключевые слова: малоугловое рентгеновское рассеяние, эволюционный факторный анализ, гель-хроматография, функция распределения частиц по размерам, частично упорядоченные системы, размер области кристалличности, кроссплатформенная графическая библиотека Qt

DOI: 10.56304/S2079562922050244

ВВЕДЕНИЕ

Метод малоуглового рентгеновского рассеяния (МУРР) является эффективным структурным методом анализа растворов белков, полимеров и частично упорядоченных систем в наноразмерном диапазоне [1]. Одним из важных преимуществ метода МУРР является возможность изучения структуры биологических макромолекул в растворе в их естественных физиологических условиях и исследования реакции системы на изменения условий среды, таких как температура, pH, концентрация белка, состав буферного раствора.

Для эффективного определения трехмерной формы макромолекулы необходимо, чтобы исследуемая система была монодисперсной [2]. Метод использования гель-хроматографической колонки в сочетании с малоугловыми измерениями может обеспечить выполнение данного условия и во многих случаях разделить вклад отдельных компонентов, которые присутствуют в белковой смеси [3]. Однако, если временные профили выхода компонентов из хроматографической колонки перекрываются, прямое разделение компонент становится невозможным. В этом случае анализ данных требует специальной процедуры разложения для оценки количества компонентов и дальнейшего восстановления индивидуальных профилей рассеяния компонент по набору экспериментальных данных. С помощью ряда хемометрических методов (таких как метод альтернирующих наименьших квадратов или эволюционный факторный анализ) такое разложение можно провести. В работе [4] последний алгоритм был реализован в виде программы EFAMIX и успешно протестирован на ряде теоретических и экспериментальных наборов данных МУРР. Однако, эффективность алгоритма зависит как от относительного уровня шума в данных, так и от степени перекрытости хроматографических профилей выхода компонент. Поэтому, в ряде случаев автоматическое нахождение "временных" окон присутствия компонент оказывается неточным, и необходима "ручная" настройка параметров для работы алгоритма. Это удобно делать с помощью интерактивного графического меню, однако в первоначальной версии программы EFAMIX такой возможности не было.

Поэтому нами было проведена модернизация программы CHROMIXS [5], изначально предназначенной для интерактивной обработки набора малоугловых данных, в которую было добавлено специальное меню с заданием конфигурации вызова для программы EFAMIX. Новые возможности интерактивного моделирования данных МУРР будут показаны на примере олигомерной смеси белка альдолазы.

Другим важным направлением анализа данных МУРР от полидисперсных систем является поиск функций распределения частиц по размерам. В настоящее время разработан ряд алгоритмов, различающихся между собой либо поиском гладкой функции распределения частиц по размерам произвольной формы (программы GNOM [6], GIFT [7], VOLDIS [8], McSAS [9]), либо путем ее задания в виде суперпозиции аналитических функций (программы POLYMIX [8], MIXTURE [10], SASFIT [11], SASVIEW [12]). Недавно был предложен новый подход [8], предусматривающий комбинированное использование как методов поиска, так и оптимизационных схем при варьировании параметров модели. Этот подход позволил существенно расширить область сходимости к правильному решению в многомерном пространстве стартовых параметров модели на примере систем сферических частиц, и, таким образом, улучшить надежность анализа сложных полидисперсных систем по данным малоуглового рассеяния. В данной работе нами будут описаны дополнительные возможности указанных выше программ (MIXTURE, POLYMIX, VOLDIS) для анализа полидисперсных систем с частицами более сложной формы (глобулярных частиц типа “ядро–оболочка”, частиц цилиндрического типа с одновременной полидисперсностью как по радиусу, так и по длине цилиндра, взаимодействующих сфер в приближении Кулоновского потенциала).

В случае частично упорядоченных систем кривые малоуглового рассеяния содержат уширенные дифракционные пики, по которым можно определить такие структурные параметры, как межплоскостное расстояние и средний размер кристаллитов. Ранее нами была разработана графическая программа Peak [10], однако поскольку она использовала графическую библиотеку QuickWin, то могла работать только на Windows-платформе, что в определенной мере ограничивало ее использование. Поэтому нами была создана новая кроссплатформенная версия программы Peak, которая использует графическую библиотеку Qt, а также улучшенную процедуру оптимизации структурных параметров, описывающих дифракционные пики.

Указанные выше алгоритмы позволяют получать важную структурную информацию о составе полидисперсных и частично упорядоченных систем по данным МУРР. Описание новых возможностей модернизированных программ будет представлено в следующих разделах данной работы.

ЭВОЛЮЦИОННЫЙ ФАКТОРНЫЙ АНАЛИЗ И ЕГО ПРИМЕНЕНИЕ К АНАЛИЗУ ДАННЫХ МУРР

Набор данных МУРР можно описать матрицей $\mathbf{A} = \{A_{ik}\} = \{I^{(k)}(s_i)\}$, ($i = 1,...,N$, $k = 1,..., K$, где N – количество экспериментальных точек на кривой рассеяния, K – количество временных кадров в наборе данных МУРР (например для измерений, проведенных с использованием гель-хроматографической колонки). Эта матрица может быть представлена как $\mathbf{A} = \mathbf{U} \times \mathbf{S} \times \mathbf{V}^{\mathbf{T}}$ (что представляет собой сингулярное разложение матрицы данных), где матрица $\mathbf{S}$ диагональная, а столбцы ортогональных матриц $\mathbf{U}$ и $\mathbf{V}$ являются собственными векторами матриц $\mathbf{A} \times \mathbf{A}^{\mathbf{T}}$ и $\mathbf{A}^{\mathbf{T}} \times \mathbf{A}$, соответственно. Метод эволюционного факторного анализа (ЭФА) [13] использует сингулярное разложение набора данных МУРР для нахождения начальной и конечной точек присутствия каждого компонента смеси в процессе эволюции системы (задающих границы “временных окон” компонентов). Каждый компонент вне своего “временного окна” отсутствует, и поэтому его концентрация равна нулю. Также предполагается, что компоненты выходят друг за другом по времени, то есть компонент, первым появившийся в системе, будет и первым, который исчезнет, второй компонент исчезнет следующим, и так далее.

Основная идея ЭФА состоит в том, чтобы следить за изменением или эволюцией ранга матрицы $\mathbf{A}$ данных в зависимости от количества принятых во внимание временных кадров. Ранг матрицы определяется как число значимых сингулярных чисел в матрице $\mathbf{S}$. Эффективным альтернативным способом может быть подсчет нешумовых векторов левой сингулярной матрицы $\mathbf{U}$ по критериям автокорреляции их элементов. Для этого обычно делают ЭФА в направлении увеличения времени и в обратном направлении, чтобы определить, когда компонент появляется/исчезает из системы. Получив данную информацию, далее можно последовательно найти матрицу концентраций и матрицу профилей рассеяния компонентов.

Данный алгоритм был реализован в программе EFAMIX [4] и успешно протестирован на ряде теоретических и экспериментальных наборов данных МУРР. Определение концентрационных “окон присутствия” компонент проводится в автоматическом режиме, однако поскольку эффективность работы ЭФА алгоритма имеет определенные ограничения, зависящие от степени перекры-

вания концентрационных профилей компонент, а также степени их асимметрии, то в ряде случаев автоматический режим работы программы может приводить к смещенным оценкам, что уменьшает вероятность получения правильного решения. Поэтому важно предусмотреть возможность использования интерактивного моделирования для быстрого сканирования различных вариантов “окон присутствия” компонент и проверки результатов восстановления профилей рассеяния компонентов.

Для интерактивной обработки набора данных МУРР, полученных с использованием хроматографической колонки и содержащих только один компонент, ранее была разработана программа CHROMIXS [5], входящая в пакет ATSAS [14]. Анализ данных проводится путем усреднения временных кадров от образца и буфера и последующего вычитания буферного сигнала из кривой рассеяния образцом. Удобный графический интерфейс, использующий кроссплатформенную графическую библиотеку Qt, позволяет быстро проводить манипуляции с данными.

Для проведения интерактивного моделирования данных МУРР от белковых смесей, содержащих от двух до четырех компонент, программа CHROMIXS была нами модернизирована. В нее было добавлено специальное меню вызова программы EFAMIX, в котором можно задавать количество компонент, “временные” окна присутствия компонент и угловой диапазон используемых данных МУРР. После завершения работы программы EFAMIX восстановленные концентрационные профили и кривые рассеяния компонентов отображаются с помощью программы PrimusQt [15]. Графический интерфейс модернизированной программы CHROMIXS представлен на рис. 1.

На ней показаны результаты разложения олигомерной смеси белка альдолаза (гексамер–октамер) с помощью ЭФА алгоритма, которые хорошо согласуются с известными кристаллографическими структурами [4]. Таким образом, модернизация программы CHROMIXS позволяет проводить в интерактивном режиме обработку наборов данных МУРР с помощью ЭФА алгоритма.

НОВЫЕ ВОЗМОЖНОСТИ АНАЛИЗА ДАННЫХ МУРР ОТ ПОЛИДИСПЕРСНЫХ СИСТЕМ ПРИ ЗАДАНИИ ФУНКЦИЙ РАСПРЕДЕЛЕНИЯ ПО РАЗМЕРАМ В ПАРАМЕТРИЧЕСКОМ ВИДЕ

Интенсивность рассеяния многокомпонентной системы (под компонентом имеются ввиду частицы одной формы с мономодальным распределением по размерам) можно представить в следующем виде [16]:

$$I(s) = \sum_k \nu_k I_k(s), \quad (1)$$

где суммирование проводится по разным компонентам, v_k – относительная объемная доля k-ого компонента, $I_k(s)$ – интенсивность k-ого компонента, модуль вектора рассеяния $s = (4\pi/\lambda)\sin(\theta)$, 2θ – угол рассеяния, λ – длина волны.

Для полидисперсной системы взаимодействующих частиц, интенсивность каждого компонента можно представить в следующем виде:

$$I_k(s) = T_k(s)\int D_k(R)V_k(R)\left[\Delta\rho_k(R)\right]^2 i_{0k}(s,R)dR, \quad (2)$$

где R – размер частицы, k – индекс компонента, $D_k(R)$ – объемное распределение частиц по размеру для k-ого компонента, $V_k(R)$ – объем частицы в к-ом компоненте, $\Delta\rho_k(R)$ – контраст электронной плотности, $i_{0k}(s,R)$ – нормированный интенсивность рассеяния от частицы, $T_k(s)$ – структурный фактор для данного компонента, ответственный за межчастичную интерференцию в рассеянии.

В программе MIXTURE [10], входящей в пакет ATSAS, есть возможность задавать до 10 различных форм частиц в виде простых геометрических тел и учитывать межчастичное взаимодействие. В качестве геометрических тел можно использовать сферы, цилиндры и эллипсоиды вращения. Для сфер и цилиндров можно независимо задавать плотности ядра и оболочки, а также учитывать полидисперсность по внешнему радиусу частицы (в то время как длина цилиндров задается фиксированной). Для расчета структурного фактора используется приближение Перкуса–Йовика для потенциала притягивающихся жестких сфер.

Для расширения возможностей моделирования полидисперсных систем нами были созданы модифицированные варианты программы MIXTURE со следующими дополнительными опциями:

1) расчет полидисперсных цилиндров с одновременным учетом полидисперсности как по радиусу, так и по длине цилиндра;

2) расчет полидисперсных сфер типа “ядро–оболочка” с одновременным учетом полидисперсности как радиуса ядра, так и толщины оболочки в предположении, что они подчиняются распределению Шульца [17]. Аналитические выражения для расчета интенсивностей были выведены в работе [18];

3) расчет структурного фактора для взаимодействующих частиц в усредненном сферическом приближении (MSA) для Кулоновского потенциала согласно работе [19].

Данные опции позволяют проводить моделирование полидисперсных систем с большим числом степеней свободы и не налагать на форму частиц и их взаимодействие ограничения, в случаях, когда у нас нет какой-либо априорной информации

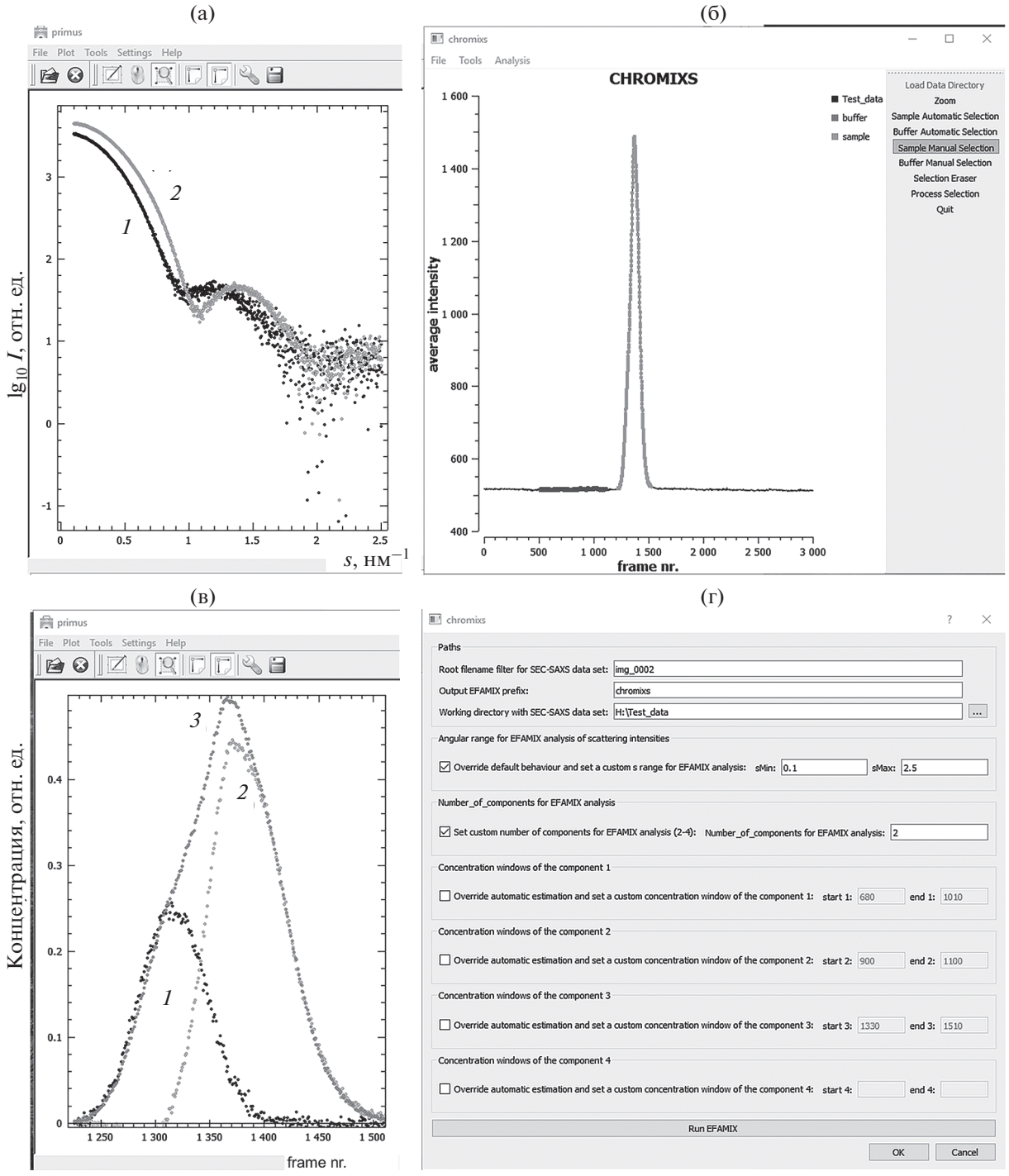

Рис. 1. Графический интерфейс вызова программы EFAMIX с помощью программы CHROMIXS, позволяющий проводить моделирование в интерактивном режиме. Данные МУРР от белка альдолаза, представляющего собой в растворе смесь гексамеров и октамеров [4]. Представлены: (а) восстановленные программой EFAMIX профили рассеяния октамеров (кривая (*1*)) и гексамеров (кривая (*2*)) белка альдолазы; (б) хроматографический (“временной”) профиль выхода компонент, область фонового (“буферного”) сигнала соответствует временным кадрам 500–1000, область присутсвия образца – временным кадрам 1200–1500; (в) найденные с помощью эволюционного факторного анализа концентрационные профили компонент (октамеры – *1*, гексамеры – *2*) и их суммарный профиль – *3*; (г) меню вызова программы EFAMIX.

об исследуемой системе. В то же время неоднозначность решения при этом может возрастать, поэтому для получения более надежного и устойчивого решения рекомендуется использовать комбинированный подход, с использованием программ GNOM, VOLDIS, POLYMIX/MIXTURE, подробно описанный в работе [8]. В частности, программа прямого непараметрического поиска гистограммы распределения VOLDIS позволяет достаточно надежно оценить стартовые величины параметров компонентов распределений. Ее основные особенности следующие: 1) поиск минимума невязки осуществляется быстрым методом Левенберга–Марквардта. 2) Программа напрямую меняет значения распределения в каждом узле гистограммы. Однако это распределение будет состоять из большого числа узких пиков из-за сильной корреляции форм кривых интенсивности рассеяния соседних (т.е., близких) по размеру частиц. 3) Чтобы преодолеть эту трудность, интенсивность модельного рассеяния рассчитывается по сглаженному контуру распределения. Это дополнительно улучшает обусловленность задачи. 4) Проводится поиск решений при нескольких (5–10) возрастающих степенях сглаживания распределения. 5) Из полученного набора решений выбирают максимально гладкое при приемлемом критерии качества приближения экспериментальных интенсивностей рассеяния, вычисляемом как χ^2.

ОЦЕНКА СТРУКТУРНЫХ ПАРАМЕТРОВ ДИФРАКЦИОННЫХ ПИКОВ ОТ ЧАСТИЧНО УПОРЯДОЧЕННЫХ СИСТЕМ ПО ДАННЫМ МУРР

Структурные характеристики частично упорядоченных систем можно оценить по максимумам профилей рассеяния с использованием программы PEAK [10]. На первом шаге проводится интерактивный выбор одного или несколько пиков на кривой рассеяния, затем выполняется приближение выбранной области пика с помощью Гауссовой функции и двухпараметрического вычитания фона. К оцениваемым структурным параметрам относятся: межплоскостное расстояние, размер области кристалличности и степень разупорядочения, рассчитанные по положению пика и его ширине с использованием стандартных уравнений дифракционной теории [20].

Однако программа PEAK имела существенные ограничения, поскольку использовала графическую библиотеку QuickWin и могла работать только на Windows-платформе, что сильно ограничивало ее использование. Поэтому нами была создана принципиально новая кроссплатформенная версия программы PEAK с использованием графи-

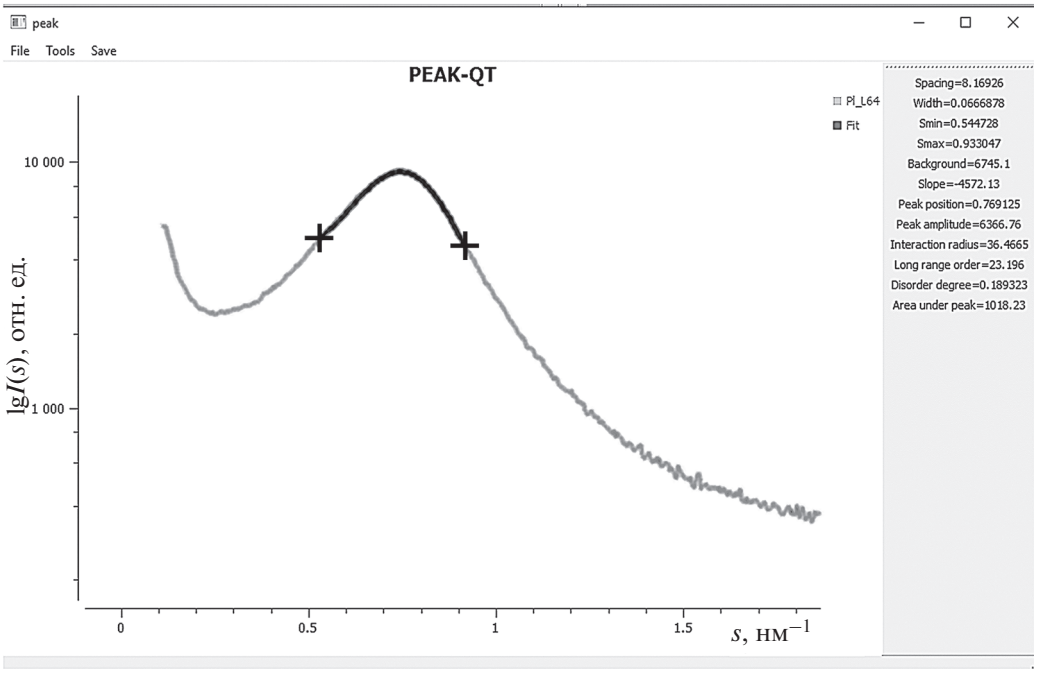

Рис. 2. Графический интерфейс кроссплатформенной версии программы PEAK. Светло-серым цветом обозначены экспериментальные данные МУРР (композитный твердый электролит на основе блок сополимера Плюроник PL-L64 и SiO_2/Al [21]), темно-серым цветом – приближение программы PEAK. Крестики показывают выбранную область данных для моделирования. Справа на панели указаны вычисленные структурные параметры (межлоскостное расстояние (spacing), размер кристаллита (long range order), степень разупорядочения структуры (disorder degree) и др.).

ческой библиотеки Qt. На рис. 2 показан графический интерфейс новой версии программы с данными МУРР от композитного твердого электролита на основе блок сополимера Плюроник PL-L64 и SiO_2/Al [21].

Помимо улучшенной графики, удобного масштабирования графика с помощью компьютерной мыши, программа получила более удобный пользовательский интерфейс, результаты моделирования отображаются непосредственно на рабочей панели программы. Кроме того, в программе для оптимизации структурных параметров используется современный минимизационный пакет MINPACK [22], что также повысило точность получаемых результатов.

ЗАКЛЮЧЕНИЕ

В рамках работы разработан новый графический интерфейс в программе CHROMIXS для вызова программы EFAMIX, восстанавливающей профили рассеяния отдельных компонентов в белковых смесях по данным МУРР с использованием хроматографической колонки. Добавлены опции в программы MIXTURE и VOLDIS, расширяющие возможности моделирования полидисперсных систем. Создана новая кроссплатформенная версия программы PEAK для анализа данных МУРР от частично упорядоченных систем.

БЛАГОДАРНОСТИ

Работа выполнена при поддержке Министерства науки и высшего образования Российской Федерации в рамках Федеральной научно-технической программы развития синхротронных и нейтронных исследований и исследовательской инфраструктуры на 2019–2027 гг. (соглашение №. 075-15-2021-1355 от 12.10.2021 г.).

СПИСОК ЛИТЕРАТУРЫ/REFERENCES

1. *Svergun D.I. et al.* Small Angle X-ray and Neutron Scattering from Solutions of Biological Macromolecules. 2013. Oxford: Oxford Univ. Press.
2. *Jeffries C.M. et al.* // Nat. Prot. 2016. V. 11. P. 2122.
3. *Mathew E., Mirza A., Menhart N.* // J. Synchr. Rad. 2004. V. 11. P. 314.
4. *Konarev P.V. et al.* // Protein Sci. 2022. V. 31. P. 269.
5. *Panjkovich A., Svergun D.I.* // Bioinformatics. 2018. V. 34. P. 1944.
6. *Svergun D.I.* // J. Appl. Cryst. 1992. V. 25 (4). P. 495.
7. *Glatter O.* // J. Appl. Cryst. 1980. V. 13 (1). P. 7.
8. *Volkov V.V., Konarev P.V., Kryukova A.E.* // JETP Lett. 2020. V. 112 (9). P. 591.
9. *Bressler I., Pauw B.R., Thünemann A.F.* // J. Appl. Cryst. 2015. V. 48 (3). P. 962.
10. *Konarev P.V. et al.* // J. Appl. Cryst. 2003. V. 36 (5). P. 1277.
11. *Breßler I., Kohlbrecher J., Thünemann A.F.* // J. Appl. Cryst. 2015. V. 48 (5). P. 1587.
12. *Alina G. et al.* // SasView for Small-Angle Scattering Analysis. http://www.sasview.org/.
13. *Keller H.R., Massart D.L.* // Chemom. Intell. Lab. Syst. 1992. V. 12 (3). P. 209.
14. *Manalastas-Cantos K. et al.* // J. Appl. Cryst. 2021. V. 54 (1). P. 343.
15. *Petoukhov M.V. et al.* // J. Appl. Cryst. 2012. V. 45 (2). P. 342.
16. *Svergun D.I. et al.* // J. Chem. Phys. 2000. V. 113 (4). P. 1651.
17. *Schulz G.V.* // Z. Phys. Chem. B. 1935. V. 30. P. 379.
18. *Wagner J.* // J. Appl. Cryst. 2012. V. 45 (3). P. 513.
19. *Hansen J.-P., Hayter J.B.* // Mol. Phys. 1982. V. 46. P. 651.
20. *Vainshtein B.K.* Diffraction of X-rays by Chain Molecules. 1966. Amsterdam: Elsevier.
21. *Bronstein L.M., Karlinsey R.L., Yi Z. et al.* // Chem. Mater. 2007. V. 19 (25). P. 6258.
22. MINPACK: Numerical Library for Function Minimization and Least-Squares Solutions. https://www.math.utah.edu/software/minpack.html.

Development of Algorithms for Analysis of Small-Angle X-Ray Scattering Data from Polydisperse and Partially Ordered Systems

P. V. Konarev[1, 2, *] and V. V. Volkov[1]

[1]*Shubnikov Institute of Crystallography, FSRC "Crystallography and Photonics", Russian Academy of Sciences, Moscow, 119333 Russia*

[2]*National Research Centre "Kurchatov Institute", Moscow, 123182 Russia*

**e-mail: peter_konarev@mail.ru*

Received July 8, 2022; revised July 14, 2022; accepted July 15, 2022

Abstract—The small-angle X-ray scattering method allows studying the structure of solutions of proteins, polymers and metal nanoparticles in the range of 1–200 nm. The development of new and improvement of the available algorithms for the analysis of experimental data of small-angle X-ray scattering data is an important task. This study presents a number of algorithms that allow one to find the particle size distribution functions, restore the intensity profiles of individual components in protein mixtures, and estimate the size of the crystallinity region and spacing distances in partially ordered systems. A number of algorithms are im-

plemented in computer programs using the cross-platform graphics library Qt, which significantly expands the number of the potential users. The efficiency of the algorithms has been demonstrated on a number of theoretical and experimental small-angle X-ray scattering data.

Keywords: small-angle X-ray scattering, evolving factor analysis, size exclusion chromatography, particle size distribution, partially ordered systems, crystallite size, cross-platform graphics library Qt