Journal of Siberian Federal University. Chemistry 4 (2014 7) 573-581
~ ~ ~

УДК 539.26:519.65:519.68

Multipopulation Genetic Algorithm for Simulation of the Crystal Structure from X-Ray Diffraction Data

Alexander N. Zaloga*, Sergey V. Burakov, Eugeniy S. Semenkin and Igor S. Yakimov
Siberian Federal University
79 Svobodny, Krasnoyarsk, 660041 Russia

Received 12.10.2014, received in revised form 18.11.2014, accepted 06.12.2014

A multipopulation genetic algorithm for a crystal structure solution from the X-ray powder diffraction data is proposed. Individual genetic algorithms are executed on different units of the computing cluster. The local optimization is performed periodically by the full-profile structure analysis (Rietveld method). The best trial structures are accumulated on the control unit for migration back to the routine compute units. The work of multi-population algorithm is discussed on 3 example of test structures with 9-10 independent atoms. The reliability of the structure search increases in a half order of magnitude more due to migration.

Keywords: X-ray powder diffraction, crystal structure analysis, Rietveld method, genetics algorithm.

Мультипопуляционный генетический алгоритм моделирования кристаллической структуры из рентгенодифракционных данных

А.Н. Залога, С.В. Бураков, Е.С. Семенкин, И.С. Якимов
Сибирский федеральный университет
Россия, 660041, Красноярск, пр. Свободный, 79

Предложен мультипопуляционный параллельный генетический алгоритм для моделирования атомной кристаллической структуры химических соединений из данных рентгеновской порошковой дифракции. Индивидуальные генетические алгоритмы выполняются на разных

* Corresponding author E-mail address: zaloga@yandex.ru

узлах вычислительного кластера. Лучшие структурные модели из всех узлов подвергаются локальной оптимизации по методу полнопрофильного структурного анализа и накапливаются на управляющем узле. Он управляет их выборочной миграцией обратно в популяции на вычислительных узлах. Работа мультипопуляционного алгоритма обсуждается на тестовых структурах с 9-10 независимыми атомами.

Ключевые слова: кристаллическая структура, рентгеновская порошковая дифракция, эволюционные генетические алгоритмы, полнопрофильный структурный анализ, метод Ритвельда.

Введение

Информация об атомной кристаллической структуре химических соединений включает координаты атомов в элементарной кристаллической ячейке и параметры их тепловых колебаний и накапливается в структурных базах данных [1, 2]. Основным методом для изучения структуры новых веществ, получаемых в поликристаллической форме, служит рентгеновская порошковая дифракция [3]. В структурное исследование входит определение приближенной модели атомной кристаллической структуры и ее оптимизация. Исходные данные для определения модели структуры – химическая формула, параметры осей ***a***, ***b***, ***c*** кристаллической ячейки, пространственная группа симметрии и полнопрофильная порошковая дифрактограмма вещества, которая может быть рассчитана из атомной кристаллической структуры [3]. Оптимизация модели кристаллической структуры осуществляется с помощью метода полнопрофильного анализа дифрактограмм Ритвельда [4]. Основным критерием служит минимум профильного R-фактора – невязки между расчетной и экспериментальной дифрактограммами, выраженной в относительных процентах. Основной проблемой дифракционного структурного анализа является поиск модели кристаллической структуры.

С ростом мощности компьютеров эффективными для этой цели становятся вычислительно емкие методы поиска структурных моделей в прямом пространстве [5, 6], основанные, подобно методу Ритвельда, на непосредственном моделировании экспериментальной дифрактограммы. Основными считаются варианты метода имитации отжига [7] и эволюционных генетических алгоритмов [8]. Последние имитируют биологическую эволюцию, т.е. осуществляют поэтапную стохастическую глобальную оптимизацию сразу целого множества (популяции) структурных моделей путем скрещиваний, мутаций и селекции моделей в новые поколения популяции по критерию минимума R-фактора. Преимущество этих методов – высокая степень автоматизации и относительная простота, что позволяет выполнять структурный анализ непосредственно в тех химических или материаловедческих лабораториях, где было синтезировано вещество. Основным недостатком признана частая преждевременная стагнация процесса поиска в локальных минимумах R-фактора – до сходимости одной из структурных моделей к истинной структуре вещества.

В [9] нами предложен эволюционный гибридный двухуровневый генетический алгоритм (ГА) для определения кристаллических структур в прямом пространстве. В данной работе сообщается о разработке на базе ГА экспериментального мультипопуляционного параллельного

генетического алгоритма (МПГА), предназначенного для работы на многоядерных ПК и вычислительном кластере СФУ, и обсуждаются результаты исследований по предотвращению стагнации и ускорению сходимости за счет обмена структурными моделями между разными популяциями.

Метод МПГА и экспериментальная оценка его эффективности

Схема МПГА изображена на рис. 1. Предварительно по встроенному в МПГА методу [10] выполняется декомпозиция дифрактограммы с определением профильных параметров, таких как форма и ширина дифракционных линий и т.п., которые затем фиксируются. Далее МПГА выполняет случайную генерацию *n* различных популяций структурных моделей вещества и на каждом из *n* вычислительных ядер ПК или кластера запускает индивидуальный однопопуляционный ГА. При изучении механизма сходимости ГА нами ранее было показано, что уточнение структурных моделей статистически обеспечивает установку отдельных атомов в истинные позиции в структуре с понижением R-фактора и является важным фактором ускорения сходимости. Поэтому лучшие в смысле R-фактора структурные модели с каждого эволюционного цикла ГА уточняются с помощью метода полнопрофильного анализа и добавляются в популяцию моделей следующего поколения. Небходимо отметить, что для ускорения и синхронизации межпопуляционной эволюции здесь исключен хорошо зарекомендовавший себя, но вычислительно емкий 2-й уровень ГА, оптимизирующий выбор хорошо уточняемых структурных параметров для полнопрофильного анализа. Вместо него включен менее эффективный для ГА, но более быстрый алгоритм из объектной библиотеки кристаллографических программ свободного доступа [11], используемый здесь для полнопрофильного уточнения структурной модели сразу по всем параметрам. Это позволило ускорить вычисления однопопуляционного ГА на порядок.

Основной операцией МПГА является случайная миграция структурных моделей между популяциями, эволюционирующими на разных вычислительных ядрах. Для этой цели определенное количество структурных моделей с относительно низким R-фактором, получаемых на каждом ядре после нескольких циклов эволюции, передаются и накапливаются на управляющем ядре (УЯ). Определенное их количество случайно отбирается и периодически рассылается с УЯ обратно в популяции на вычислительных ядрах или на те ядра, где замедляется темп снижения R-фактора.

При достижении одной из структурных моделей на УЯ целевого значения R-фактора (определенного в результате декомпозиции дифрактограммы) МПГА завершает работу, а модель подвергается заключительному полнопрофильному уточнению по всем профильным и структурным параметрам. После этого исследователь может провести кристаллохимическую верификацию найденной структуры, включающую оценку корректности межатомных расстояний, валентных углов и т.п.

Важными характеристиками любого стохастического метода структурного анализа являются скорость и надежность сходимости и сложность определяемых структур. Надежность выражает долю успешных пусков алгоритма, а сложность – количество независимо определяемых структурных параметров (степеней свободы структуры). Для статистической оценки этих характеристик был выполнен многократный поиск по МПГА хорошо известных тестовых

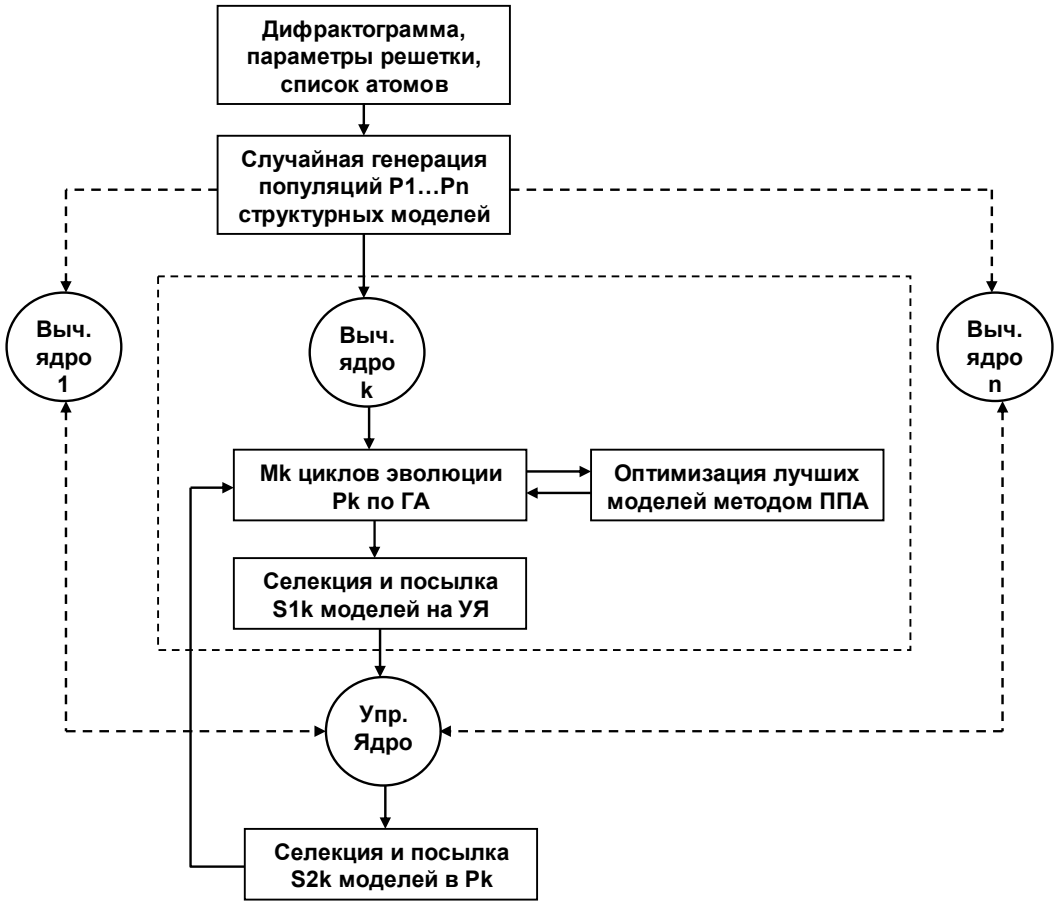

Рис. 1. Схема мультипопуляционного параллельного генетического алгоритма

структур трех химических соединений различной степени сложности, указанных в табл. 1. Выполнялся поиск координат атомов в общих позициях, температурные факторы были заданы априори, так как на практике они могут быть взяты из известных близких по составу и кристаллическому строению структур.

Для получения достаточно объективных характеристик по надежности и скорости сходимости МПГА для поиска каждой из структур было выполнено по 100 независимых пусков экспериментальной программы МПГА на 4-ядерном ПК. Для сравнительной оценки сходимости МПГА и однопопуляционного ГА в программе отключили режим пересылки структурных моделей с УЯ на вычислительные ядра и выполнили по 20 пусков поиска каждой из структур. При этом в каждом пуске программы на трех вычислительных ядрах выполнялся независимый однопопуляционный ГА, так что в 20 пусках было получено по 60 независимых результатов. Процесс поиска проводился в течение заданного количества поколений эволюции. Число структурных моделей, отбираемых для миграции, составляло 2 % от размера популяций. Параметры генетических операций – турнирного парного скрещивания, мутации атомных координат и селекции структурных моделей в очередные поколения эволюции по R-фактору – были идентичны. Кроме того, использовалась операция временного элитизма – сохранения в популяции

Таблица 1. Список тестовых кристаллических структур для МПГА

Химическая формула соединения	Параметры решетки; простр. группа симметрии; число формульных единиц Z в ячейке	Число атомов в независимой части ячейки	Число степеней свободы атомных координат
K_4SnO_4	a=6,48Å, b=6,51Å, c=9,70Å, α=71,82°, β=99,89°, γ=113,13°; P -1; Z=2	9	27
K_2PbO_2	a=10,9Å, b=7,6Å, c=7,32Å, α=119,3°, β=88,4°, γ=117,7°; P -1; Z=2	10	30
$Ca_2Al_3O_6F$	a=7,3205Å, c=6,9988Å; R -3; Z=12	9	25

Таблица 2. Усредненные характеристики поиска структур по МПГА и ГА

Тестовая структура	Размер популяции; количество поколений эволюции	Параметры миграции		Процент сходимости поиска, %	Среднее время сходимости, мин	Средний R-фактор и σ(R)при сходимости
		S1	S2			
МПГА						
K_4SnO_4	50; 200	1	1	76	1,72	5,53 (0,01)
K_2PbO_2	100; 200	2	2	64	2,77	7,76 (0,03)
$Ca_2Al_3O_6F$	100; 300	2	2	27	4,21	7,09 (0,01)
Однопопуляционный ГА						
K_4SnO_4	50; 200	1	0	50	3,08	5,53 (0,001)
K_2PbO_2	100; 200	2	0	27	5,20	7,75 (0,001)
$Ca_2Al_3O_6F$	100; 300	2	0	6,7	4,50	7,09 (0,006)

лучшей структуры на протяжении трех поколений. Усредненные характеристики результатов поиска структур сведены в табл. 2.

При поиске структур по МПГА на суперкомпьютерном кластере СФУ с использованием 24 вычислительных узлов структура K_4SnO_4 определяется со 100%-ной надежностью, а более сложная тестовая структура $Er_{10}W_2O_{21}$ (пр. гр. P b c n) с 54 степенями свободы атомных координат была определена в 3 случаях из 5 (23 популяции размера 500, число поколений эволюции 500).

Обсуждение результатов

Вариант МПГА с миграцией структурных моделей между популяциями был впервые предложен в [12]. Особенностью этого варианта стало создание МПГА на базе успешно действующего однопопуляционного ГА [13], дополненного средствами случайного обмена структурными моделями непосредственно между разными популяциями. Структурные модели для миграции выбирают с вероятностью, обратно пропорциональной их R-фактору. На примере многократного определения тестовой структуры органического соединения [14], включающего поиск 13 структурных координат, показано, что 4-популяционный МПГА с размером популяций из 100 структурных моделей обеспечил сходимость 42 % (в 10 случаях из 24 пусков), а однопопуляционный ГА – 18 % (в 3 случаях из 17 пусков) при одинаковых вычислительных ре-

сурсах (на 20-ядерном вычислительном кластере). Другими словами, было получено ускорение сходимости в ~ 2,4 раза относительно однопопуляционного ГА, что показало эффективность межпопуляционной миграции. Однако данный подход не получил достаточного развития, возможно, потому, что для структуры с 13 степенями свободы сходимость недостаточна и заметно уступает широко используемым методам имитации отжига, особенно при их работе на суперкомпьютерном кластере [15]. Тем не менее во многих работах [5-7, 16] отмечается перспективность развития именно генетических алгоритмов для структурного анализа.

Из данных табл. 2 видно, что наш вариант МПГА на 4-ядерном ПК при поиске структур с 25–30 степенями свободы структурных переменных демонстрирует среднюю сходимость 56 % при повышении сходимости МПГА относительно однопопуляционного ГА в 1,5–4 раза (при одинаковых вычислительных ресурсах). Главной особенностью нашего подхода является постоянное накопление на управляющем ядре МПГА лучших (в смысле R-фактора) структурных моделей из всех популяций и обратная миграция моделей, случайно выбираемых для этого среди лучшей половины моделей УЯ. Такая обратная миграция обеспечивает постоянное присутствие во всех популяциях структурных моделей с R-фактором выше среднего. Для предотвращения преждевременного разрушения лучших моделей, мигрировавших с УЯ или собственных, вследствие неудачной мутации или скрещивания используется операция временного элитизма (на 3-4 поколения). Вместе с тем для предотвращения чрезмерного распространения одной из них в популяциях и обеспечения разнообразия структурных моделей необходимо применять небольшой коэффициент миграции с УЯ (S2 ~ 1-2 % от размера популяции). Размеры популяций зависят от сложности искомой структуры и должны превышать количество степеней свободы структурных параметров в несколько раз. Число заданных поколений эволюции должно быть выше размера популяций.

На рис. 2 представлены графики сходимости МПГА на УЯ при поиске структуры $Ca_2Al_3O_6F$. На оси абсцисс указан номер поколения эволюции (нулевой номер – начальная случайная генерация всех популяций), на оси ординат – соответствующие ему значения R-фактора.

Для повышения скорости вычислений вначале рассчитывают приближенные значения R-фактора (красная линия), а затем точные (черная линия) для лучшей из структурных моделей, попавших на УЯ к данному поколению эволюции. Монотонно снижающаяся ступенчатая форма этого графика отражает постепенное накопление на УЯ глобально лучших структурных моделей, формируемых в популяциях на вычислительных ядрах. Фиолетовая линия показывает размер штрафа, наложенного на R-фактор за чрезмерное (не физичное) сближение атомов в лучшей модели для того, чтобы она со временем отсеялась из популяции. Синяя линия изображает среднее значение R-фактора для лучшей половины структурных моделей на УЯ. Ее сближение с черной до достижения последней целевого значения R-фактора служит индикатором стагнации процесса эволюции. Момент сходимости МПГА к целевому значению R-фактора, а соответствующей структурной модели – к истинной структуре отмечен стрелкой (на 89-м поколении). Зеленая линия показывает худшее значение R-фактора для всех лучших моделей на всех вычислительных ядрах на данном поколении. Ее скачки вверх иллюстрируют разрушение лучшей модели в какой-либо из популяций, а скачки вниз – включение одной из лучших моделей во все популяции. Из-за случайного характера миграции сближение зеленой линии с черной почти всегда запаздывает, а расстояние между ними характеризует скорость

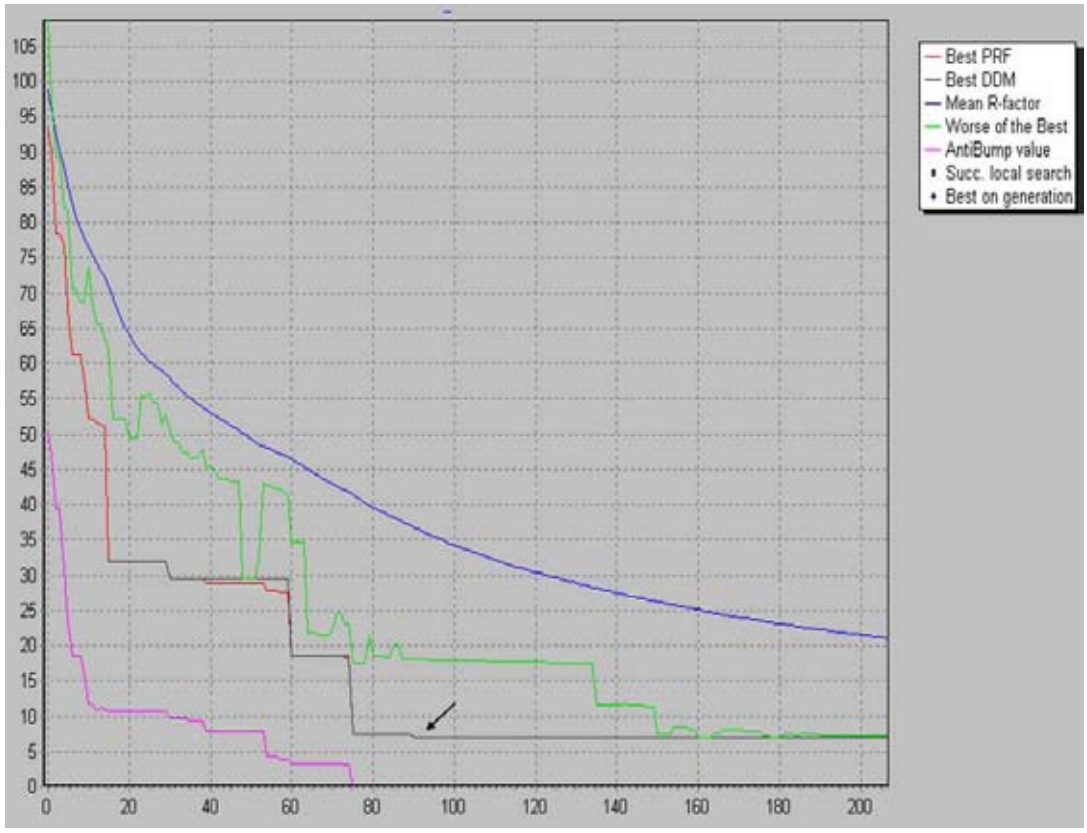

Рис. 2. Графики сходимости МПГА при поиске структуры $Ca_2Al_3O_6F$; ось абсцисс – номер поколения эволюции, ось ординат – значения R-фактора

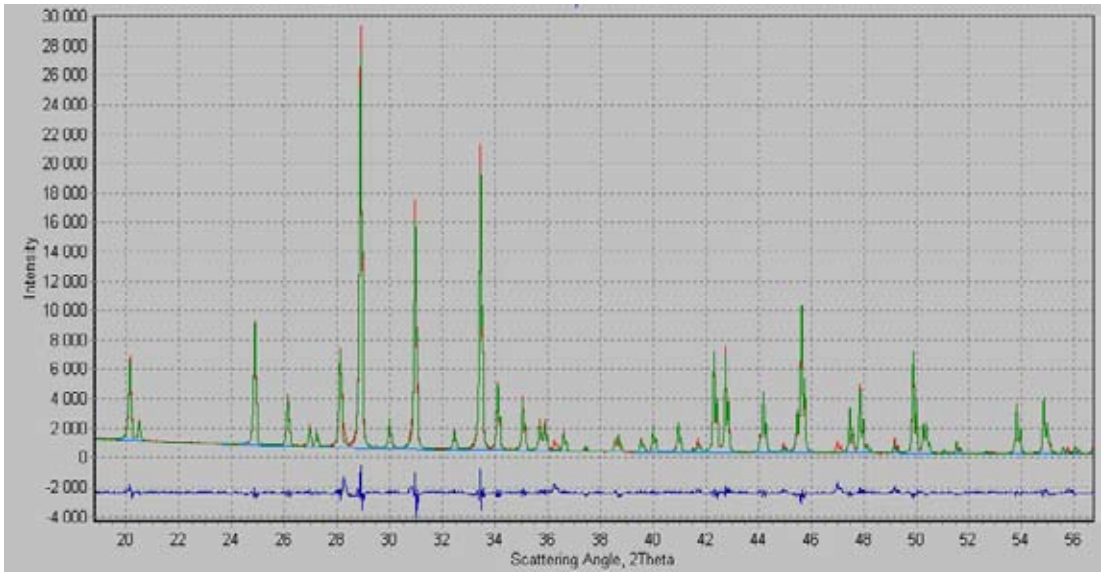

Рис. 3. Расчетная (зеленая) и экспериментальная (красная) дифрактограммы и их разность (синяя) для найденной по МПГА структуры $Ca_2Al_3O_6F$; Rwp-фактор 7,09 % отн.

распространения лучших структурных моделей с УЯ по популяциям (может регулироваться коэффициентами миграции S1 и S2). Таким образом, анализ динамики изменения этих графиков позволяет корректировать процесс сходимости МПГА. По завершении работы МПГА для лучшей структурной модели строится график (рис. 3), показывающий соответствие расчетной и экспериментальной дифрактограмм.

Заключение

Генетические алгоритмы обеспечивают ab initio поиск атомной кристаллической структуры химических соединений по порошковой дифрактограмме, хорошо автоматизируются и имеют существенный потенциал развития, в частности за счет параллельной организации для работы на многоядерных ПК и суперкомпьютерных кластерах. Повышение эффективности поиска структур за счет миграции в мультипопуляционных алгоритмах можно объяснить следующим. Стохастические процессы ГА на разных ядрах попадают в разные локальные минимумы на гиперповерхности R-фактора. Периодическое и умеренное «разбавление» плохо сходящихся популяций лучшими структурными моделями из других популяций приводит к «конкуренции локальных минимумов». Это обеспечивает выходы из них и улучшает общую сходимость эволюционного процесса в глобальный минимум. Нами планируется дальнейшее развитие этого подхода для определения более сложных кристаллических структур.

Работа выполнена при поддержке гранта ГФ-3 в рамках государственного задания МОН РФ Сибирскому федеральному университету на 2014 г.

Список литературы

1. Inorganic Crystal Structure Database. FIZ Karlsruhe. http://www.fiz-karlsruhe.de/icsd.html

2. Cambridge Structural Database. Cambridge Crystallographic Data Centre. http://www.ccdc.cam.ac.uk/products/csd/

3. Powder Diffraction Theory and Practice, ed. R.E. Dinnebier and S.J.L. Billinge / Royal Society of Chemistry, 2008. 507P.

4. Young R.A. The Rietveld Method / Oxford University Press. 1995. 298P.

5. David W. I. F., Shankland K.. Structure determination from powder diffraction data // Acta Cryst. (2008). A64, 52–64.

6. Harris K. D. M. Powder Diffraction Crystallography of Molecular Solids // Top. Curr. Chem. 2012,315, 133–177.

7. Radovan Cerny, Vincent Favre-Nicolin. Direct space methods of structure determination from powder diffraction: principles, guidelines, perspectives // Z. Kristallogr. 222 (2007) 105-113.

8. Kenneth D.M. Harris. Fundamentals and applications of genetic algorithms for structure solution from powder X-ray diffraction data // Computational Materials Science. V. 45. Issue 1. 2009. P. 16–20.

9. Yakimov Y. I., Semenkin E. S., Yakimov I. S. Two-level genetic algorithm for a fullprofile fitting of X-ray powder patterns // Z. Kristallogr. Suppl. 30 (2009) 21-26.

10. A. Le Bail. The profile of a Bragg reflection for extracting intensities / Chapter 5 in: Powder Diffraction: Theory and Practice, Ed. R.E. Dinnebier & S.J.L. Billinge, Royal Society of Chemistry, Cambridge (2008) 134-165.

11. Favre-Nicolin V., Cerny R. FOX, free objects for crystallography: a modular approach to ab initio structure determination from powder diffraction // J. Appl. Cryst. (2002). 35, pp.734–743.

12. Habershon, S.; Harris, K. D. M.; Johnston, R. L. Development of a Multipopulation Parallel Genetic Algorithm for Structure Solution from Powder Diffraction Data. // J. Comput. Chem. 2003, V.24, No.14, pp. 1766 - 1774.

13. Albesa-Jové D.; Kariuki B. M.; Kitchin S. J.; Grice L.; Cheung E. Y.; Harris K. D. M. Challenges in Direct-Space Structure Determination from Powder Diffraction Data: A Molecular Material with Four Independent Molecules in the Asymmetric Unit. ChemPhysChem. 2004, 5, 414–418.

14. Cheung, E. Y.; McCabe, E. E.; Harris, K. D. M.; Johnston, R. L.; Tedesco, E.; Raja, K. M. P.; Balaram, P. // Angewandte Chemie, Int. Ed. 2002, 41, 494.

15. Thomas A. N. Griffin, Kenneth Shankland, at al. GDASH: a grid-enabled program for structure solution from powder diffraction data. // J. Appl. Cryst. (2009). 42, 356–359.

16. Bryce Meredig, C. Wolverton. A hybrid computational–experimental approach for automated crystal structure solution // Nature Materials, 12, 123–127 (2013).