MULTI-POPULATION GENETIC ALGORITHM FOR CRYSTAL STRUCTURES SOLUTION FROM X-RAY POWDER DIFFRACTION DATA[1];

Aleksandr Zaloga[1], Igor Yakimov[2], Petr Dubinin[3]

Siberian Federal University
e-mails: [1]zaloga@yandex.ru, [2]i-s-yakimov@yandex.ru, [3]dubinin-2005@yandex.ru
Russian Federation

Abstract: The crystal structures of over two hundred new substances are annually solved from powder diffraction data by methods of global optimization. A common problem of these methods is a deterioration in the convergence with an increase in the complexity of determined structures due to a non-linearly growing of problem's complication and stagnation in the numerous local minima of the R-factor hypersurface. The present paper describes an approach for automated crystal structure solution from powder diffraction data using the multi-population genetic algorithm (MPGA). The MPGA convergence charts and the atomic positions distribution maps of the MPGA populations for solving one of crystal structures are given.

Key words: genetic algorithms, X-ray powder diffraction, crystal structure analysis in direct space, full-profile analysis, Rietveld method.

1. INTRODUCTION

Information about crystal structure is necessary for the explanation and prediction of physical and chemical properties of materials. This information is accumulated in the Crystal Structures Databases (DB) [1, 2] and includes the coordinates of atoms in symmetrically independent part of the crystal cell and some additional parameters. The structure of polycrystalline materials is studied by the different methods of X-Ray powder diffraction [3]. Structural investigation includes the definition of an approximate model of the atomic crystal structure and its optimization. The initial data are the chemical formula, unit cell parameters, space group symmetry and the X-Ray powder diffraction pattern of the material. Typically, the optimization of crystal structure model is performed using the method of Rietveld [4], which consist in full-profile fitting of calculated diffraction pattern to experimental one with using a nonlinear least squares method. The main problem is the search of a model of the crystal structure.

[1] The paper has been presented on the International Conference "The Sixth International Workshop on Mathematical Models and their Applications", 13-15.11.2017, Krasnoyarsk, the Russian Federation, and it is not published in the Conference Proceedings.

With the computing power increasing, global optimization methods can be effective to achieve this objective, as evidenced by the increasing in the number of structures solved by these methods [5, 6]. The main options are the variations of simulated annealing method [7]. Their disadvantage is the sequential approach, based on a large number (~ 107-108) of small variations of the structural configuration that allows the efficient full-profile structure refinement with the Rietveld method at the final stage only. Methods of the crystal structure model search based on genetic algorithms [8] immediately generate a lot of trial structural models (population) and carry out its evolution through the crossover, mutation and selection focusing on the R-factor minimum (as fitness function), but the Rietveld refinement is used only at the final stage too. The disadvantage of global optimization methods, including GA, is the stagnation of convergence process in the local minima.

The first version of the parallel genetic algorithm for crystal structure solution was proposed in [9]. This version is based on the successfully used single-population GA [10], which is complemented with a direct exchange of random structural models among different GA populations. However, this approach hasn't been sufficiently developed so far. At the same time, GAs have two essential advantages. Firstly, they simultaneously execute the evolution of the whole set (population) of trial structural models, i.e. explore in parallel way a wide region of the structural parameters space. Secondly, the parallel GA is much better suited for implementation on supercomputing clusters than a single one. This creates the possibility to use the full power of parallel computing for structural analysis [9]. The multi-population approach could help to solve more complex crystal structures, but it hasn't almost been investigated so far.

We present a multi-population parallel GA, which implements co-evolution of independent GA processes on computational processes of multicore PC or cluster. Co-evolution is performed by accumulating the best trial structural models on the managing process and then selectively transmitting them into populations on the working processes. Such approach contributes to getting out from local minima of the R-factor hypersurface, accelerates the accumulation of correct atomic positions in the populations and increases the probability of GA convergence when more complex structures [6] are managed.

2. DESCRIPTION OF THE MPGA

The design and the current features of MPGA are described in more detail in [11]. Below we present a general description of its original features. The main ones are:

1) Appointment of physically conditioned penalties to a crystal structure.

2) Working with the molecular fragments.

3) Automatically putting the atoms on the symmetry elements if near to them.

4) Working with multi-phase samples.

5) Providing built-in tools for the convergence process analysis: convergence chart for each process, 3D crystal structure, atoms distribution maps at each generation.

6) Being based on the well optimized FOX / ObjCryst++ library [12].

Figure 1 shows the flow-chart of MPGA. From this flow-chart the factors improving the MPGA convergence can be seen:

1) the execution of independent parallel processes with different settings on working processes;

2) the refinement of the best structures by Rietveld method;

3) the co-evolution involving accumulation and a managed exchange of the best solutions among the GA processes.

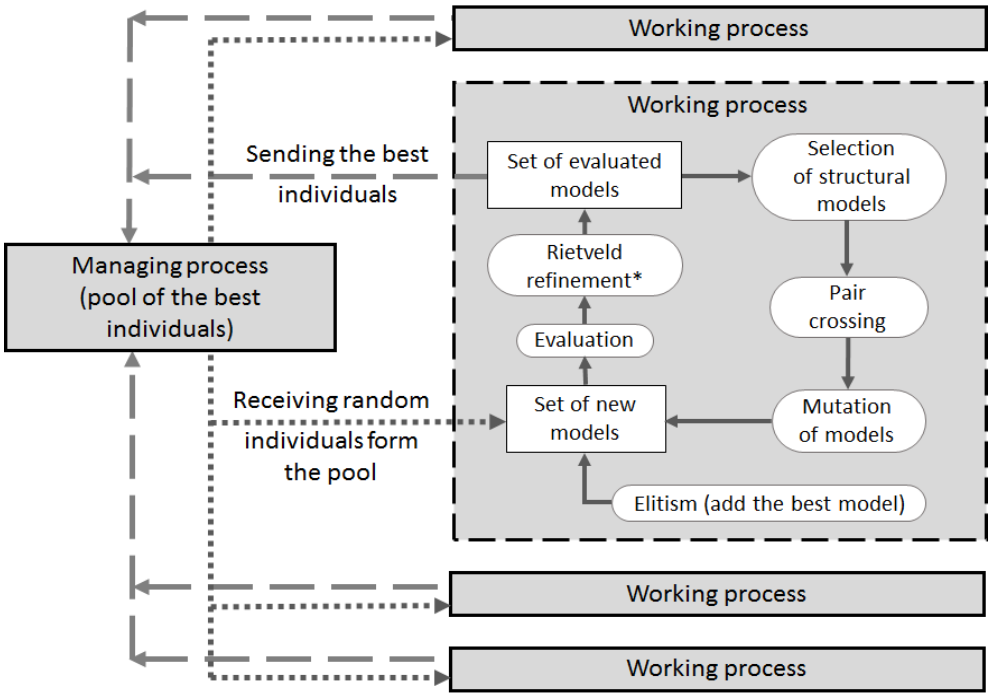

Fig. 1. The flowchart of MPGA
** the Rietveld refinement is performed once in several generations and only for the best individuals*

In previous versions of MPGA we used to use binary-coded chromosomes with 5 or 6 bits per each optimized parameter [11,13]. But now we use only real-coded chromosomes, because it is natural representation of crystal structure's parameters, and bit coding and decoding introduced distortions into the solutions found during the evolution process.

The evolution process uses the genetic operators described in the table 1.

The advantage of using co-evolution with an exchange of the best individuals compared with the using of the evolution with a single genetic algorithm without interpopulation exchange, is shown in our publication [14]. Here is should mentioned that an increasing of involved in MPGA computational processes significantly increases the probability of finding the right solution.

3. DETERMINATION OF THE CRYSTAL STRUCTURE OF $PbNaF_2NO_3$ USING MPGA

The synthesis of $PbNaF_2NO_3$ was described in [15]. The X-ray powder diffraction pattern was recorded on a Rigaku D2500/max diffractometer utilizing $CuKa_{1,2}$ radiation. The space group is Cmcm (No. 63). The unit cell parameters are: a = 5.5737(1) Å; b = 14.1136(2) Å; c = 5.6346(1) Å, V = 443.25(1) $Å^3$, Z=4.

Table 1. Description of used genetic operators and additional functions

Operator	***Application details***	***Settings with typical used values***
Selection	Tournament (can be different for different working processes)	Tourn. size = 3…5
Crossover	Single-point crossing or Clone. Crossover coef. = 0.75 means the that 75% of the selected individuals will be crossed and 25% will be cloned (can be different for different working processes).	Crossover coef. = 0.75
Mutation	Changing the mutated parameter to new random value within specified limits (can be different for different working processes)	Mutation coef. = 2%
Local optimization	Refinement of the best structural models by Rietveld method	1) Number of individuals for refinement = 3; 2) Run every 10 generations
Local elitism	Copying the best individual from previous generation to new generation	Number of generations to keep the best individual = 3…5
Sending from the managing process (from pool) to working processes	Each working process sends its best individual to the pool every generation. And the managing process dispatches some random individuals from accumulated ones to all working processes (only individuals with fitness less than the average accumulated fitness may be sent)	1) Send interval = 5, 7 or 10 generations; 2) Number individuals to send = 1 or 2.

The profile parameters of the diffraction pattern and the target value of the profile R-factor equal to 8.82% were determined using the Le Bail method. Restrictions on the interatomic distances were imposed according to the statistics of the interatomic distances distribution, which has been calculated by Diamond software [16] using structures having a similar composition. The required parameters for solving this structure by MPGA were the locations of Pb, F, Na, N atoms and two O atoms. The total number of degrees of freedom for this structure is 18.

The advantage of using co-evolution with an exchange of the best individuals compared with the using of the evolution with a single genetic algorithm without interpopulation exchange, is shown in our publication [14]. Here is should mentioned that an increasing of involved in MPGA computational processes significantly increases the probability of finding the right solution.

3. DETERMINATION OF THE CRYSTAL STRUCTURE OF $PbNaF_2NO_3$ USING MPGA

The synthesis of $PbNaF_2NO_3$ was described in [15]. The X-ray powder diffraction pattern was recorded on a Rigaku D2500/max diffractometer utilizing $CuKa_{1,2}$ radiation. The space group is Cmcm (No. 63). The unit cell parameters are: a = 5.5737(1) Å; b = 14.1136(2) Å; c = 5.6346(1) Å, V = 443.25(1) $Å^3$, Z=4.

The profile parameters of the diffraction pattern and the target value of the profile R-factor equal to 8.82% were determined using the Le Bail method. Restrictions on the interatomic distances were imposed according to the statistics of the interatomic distances distribution, which has been calculated by Diamond software [16] using structures having a similar composition. The required parameters for solving this structure by MPGA were the locations of Pb, F, Na, N atoms and two O atoms. The total number of degrees of freedom for this structure is 18.

The crystal structure determination was performed using the MPGA software on a computer equipped with an Intel Core i3-3110M processor having four processing threads. Three of them were involved in the evolutionary search of the crystal structure at independent populations and one thread was managing the accumulation of the best structural models from the generated populations and the exchange between populations [13]. The MPGA version with real-coded atomic coordinates and the periodical local optimization of the best structural models by local search (LS) by the Rietveld method [11] was used. The evolutionary search of the crystal structure was performed automatically after the MPGA launch, and the process was visually controlled with convergence charts, charts of distribution of atomic positions in the populations, and comparative charts between the experimental powder pattern and that one calculated from the best structural model in a current generation. It should be mentioned that the presetting of the search parameters for the MPGA operation is required because their quality affect the probability of GA to converge to the correct structure. This was achieved by an empirical selection from the parameters used in the MPGA launches for searching the testing crystalline structures.

4. THE MPGA CONVERGENCE DURING THE STRUCTURE DETERMINATION OF $PbNaF_2NO_3$

Figure 2 shows the graph of the MPGA convergence for the $PbNaF_2NO_3$ structure determination. The X-axis on the figure is the number of generations in the population (the number zero is the initial random generation of all populations); the Y-axis is the fitness function value (R-factor plus penalties for violation of the structural limitations) of the best structural model selected from the accumulated structures in a pool at the managing process. The solid line is the fitness value for the best found structural models. The dot-and-dash line is the best fitness value obtained after the refinement of the best structural model by the LS method (it has a stepped appearance when the LS execution mode sets to once every few generations). The dashed line is the contribution of penalties to the fitness function value. The combination of the reported diagrams allows the control and the analysis of the convergence process.

It can be seen from Figure 2 that on the 15-th generation of the evolution process the fitness was sharply decreased from the value of 52 to the value of 39% rel. On this stage the heavy atoms were reached their correct positions. However, the light atoms are still not in correct positions, and the interatomic distances were less than the limits (a penalty level for

this trial structure was about half of the fitness value). By the 76th generation the penalty had disappeared, all the interatomic distances had become correct and the fitness value had coincided with the R-factor value. Next, the R-factor was gradually reduced from 15.5 to 13.3% rel., mainly due to the refinement of light atoms positions, and by the 100th generation the convergence had been completed. It should be noted that a local optimization of the best models on computation processes occurred every 10 generations. But for this solution process it was useful only on 10th generation of the evolution; on the next generations declines of the solid line was not accompanied by declines of the dot-and-dash line.

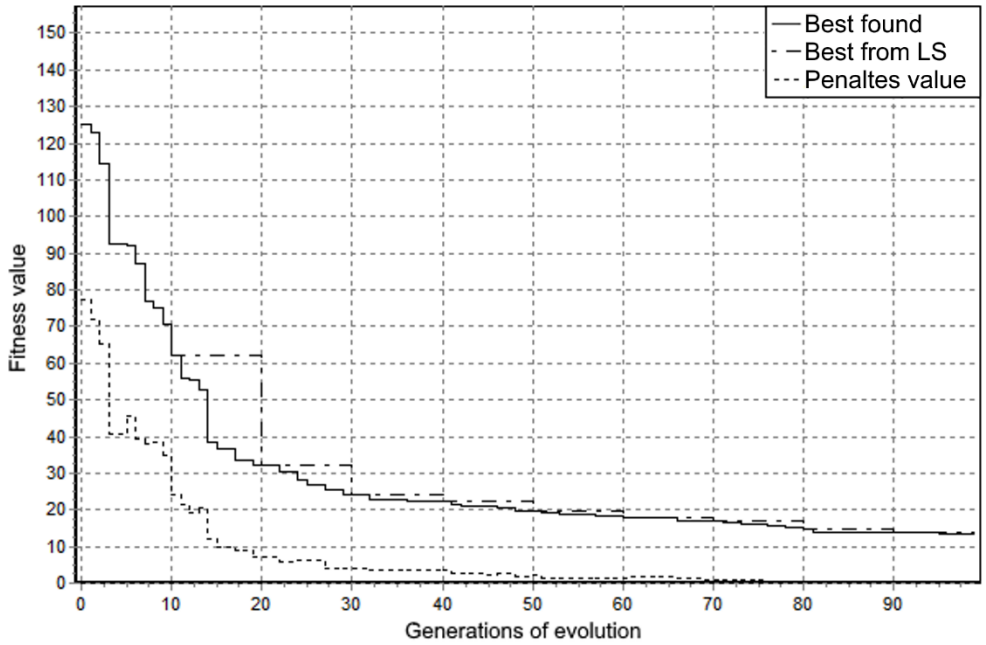

Fig. 2. MPGA convergence graph for the search of $PbNaF_2NO_3$ structure

5. RESULTS AND DISCUSSION

A new visualizer was developed to better understanding of GA convergence in populations at working processes. It allows to visualize a projection of atomic positions on chosen basic planes of the unit cell (e.g., *ac*) for all structural models in a population at a chosen generation. It is possible to preset the atomic coordinates of a known structure for comparison with them, and their positions will be marked with special shapes. Thus, the visualizer allows to scroll and compare in real time the distributions of atoms in different populations at different evolutionary generations and match them with the working processes convergence charts, etc. In particular, this tool is very useful to study the MPGA convergence on the test crystal structures.

To illustrate the MPGA convergence process, atomic positions distribution maps for the $PbNaF_2NO_3$ search are presented on Figure 3, Figure 4, and Figure 5. The projection of an independent part of the crystal cell on the *ab* plane is shown. Shapes with solid black fill indicate correct positions; big gray shapes indicate the positions of the best structure's atoms;

small gray shapes indicate the other atoms. Pb atoms are shown by circles, Na – by squares, F – by triangles, N and O – by diamonds. The pictures demonstrate how the atomic positions in chosen population of structure models evolve from randomly generated (Figure 3) to the correct positions (Figure 5).

The X-axis on the Figures 3-5 are the direction *b* (Å) in the unit cell; the Y-axis are the direction *a* (Å) in the unit cell.

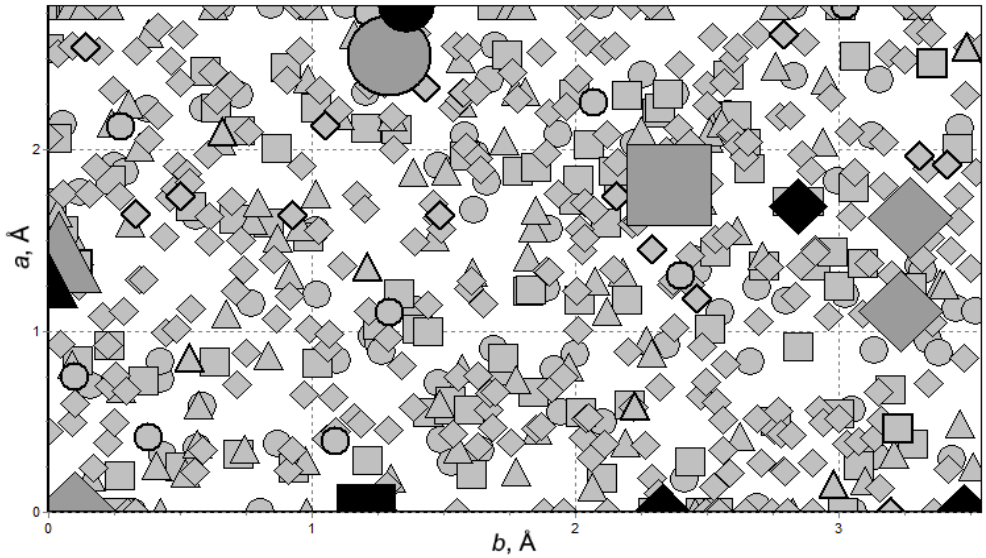

Fig. 3. 1st generation. Completely random structures. Fitness value of the best solution = 125.35%

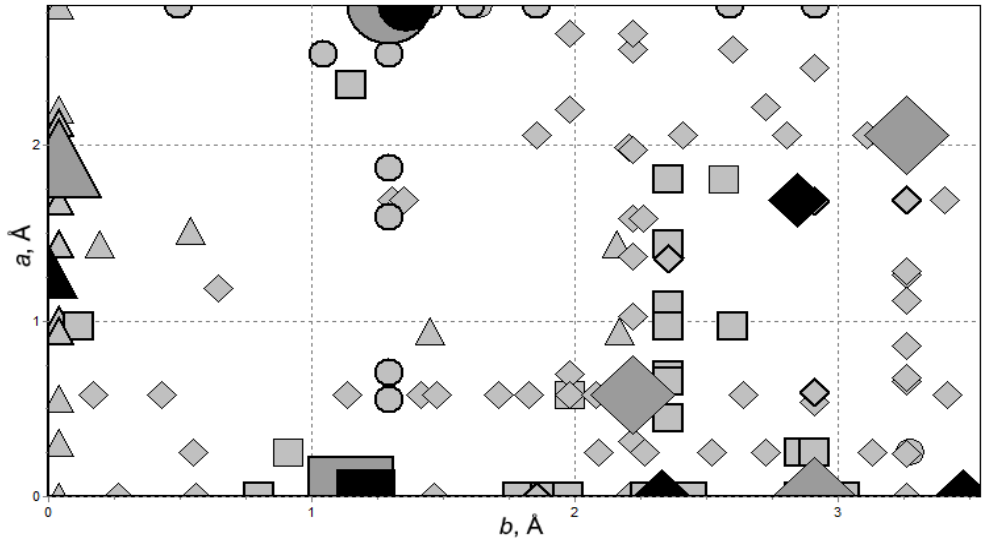

Fig. 4. 15st generation. Heavy atoms are close to the correct positions. Fitness value of the best solution = 38.49%

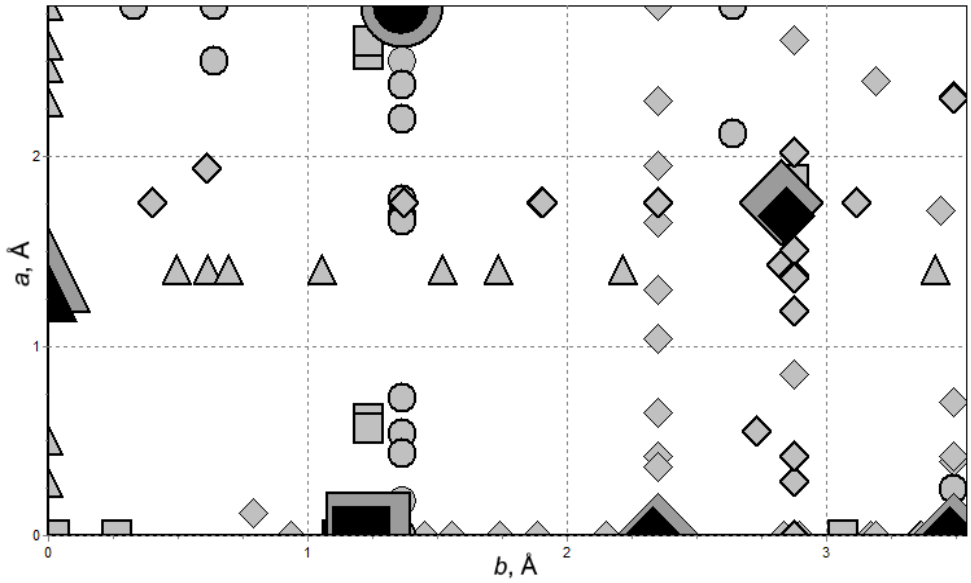

Fig. 5. 100th generation. Fitness value of the best solution = 13.39%

The probability that the atomic coordinate will be randomly generated in neighborhood of its correct position (no more than ~0.03 of the length of the unit cell's axis from the correct position; on the figures 3–5 it is the size of small shapes) is about 1/33. It is experimentally shown in [17] that a structural model having atoms quite precisely localized can be effectively refined by the LS method (at least, when the lengths of the unit cell's axes are up to 10–15 Å). The initial population of even 50 structural models contains in average one or two such coordinates for each atom (in different models). The MPGA convergence is produced by an evolutionary accumulation of "good" coordinates in a population (it goes step by step from heavy atoms to light atoms) due to successful crossings and the LS minimization. It is because the R-factor has statistical sensitivity to determining the correct set of the heavy atom atomic coordinates while the other lighter atoms are distributed randomly [17]. In addition, it is complemented by the exception of chemically incorrect structural models (by adding a penalty to the R-factor value).

The final refinement of the found structure $PbNaF_2NO_3$ was made using JANA2006 software [15]. The atomic coordinates are shown in Table 2.

Table 2. Fractional atomic coordinates (given in fractions of the corresponding sides of the crystal cell) for crystal structure of $PbNaF_2NO_3$

Atom	***x/a***	***y/b***	***z/c***
Pb	0.5	0.09578(4)	0.75
Na	0.0	0.0828(2)	0.25
F	0.2396(6)	0.0	0.5
N	0.0	0.2491(5)	0.0
O1	0.0	0.1631(4)	0.25
O2	0.3185(9)	0.2064(3)	0.25

6. CONCLUSION

The MPGA software, implementing a multi-population genetic algorithm and using parallel computing on multi-core PCs and supercomputer clusters, was developed. A further development of MPGA with using for solving more complex structures has been planned.

It was shown the application of new visualization tool to watch atomic positions convergence in all the population. The mechanism of atomic coordinates convergence was discussed.

The crystal structure of the complex compound $PbNaF_2NO_3$ was determined by multi-population genetic algorithm and then refined using JANA2006 software.

ACKNOWLEDGMENT

The reported study was funded by Russian Foundation for Basic Research, Government of Krasnoyarsk Territory, Krasnoyarsk Region Science and Technology Support Fund to the research project №16-43-243049.

REFERENCES

[1] Hellenbrandt, M. The Inorganic Crystal Structure Database (ICSD) – Present and Future. *Crystallography Reviews*, **10**, 2004, pp. 17-22.

[2] Allen, F.H. The Cambridge Structural Database: A quarter of a million crystal structures and rising. *Acta Cryst.*, **B58**, 2002, pp. 380-388.

[3] Dinnebier, R. E., Billinge, S. J. L. *Powder Diffraction Theory and Practice*. Royal Society of Chemistry, UK, 2008.

[4] Young, R. A. *The Rietveld Method*. Oxford University Press, UK, 1993.

[5] David, W. I. F., Shankland, K. Structure determination from powder diffraction data. *Acta Cryst*. **A64**, 2008, pp. 52-64.

[6] Le Bail, A., Cranswick, L. M. D. Third structure determination by powder diffractometry round robin (SDPDRR-3). *Powder Diffr.*, **24**, 2009, 254-262.

[7] Černý, R., Favre-Nicolin, V. Direct space methods of structure determination from powder diffraction: principles, guidelines, perspectives. *Z. Kristallogr.*, **222**, 2007, pp. 105-113.

[8] Harris, K. D. M. Fundamentals and applications of genetic algorithms for structure solution from powder X-ray diffraction data. *Computational Materials Science*, **45**, 2009, pp. 16-20.

[9] Habershon, S. et. al. Development of a Multipopulation Parallel Genetic Algorithm for Structure Solution from Powder Diffraction Data. *J. Comput. Chem.*, **24**, 2003, pp. 1766-1774.

[10] Albesa-Jové, D. et. al. Challenges in Direct-Space Structure Determination from Powder Diffraction Data: A Molecular Material with Four Independent Molecules in the Asymmetric Unit. *ChemPhysChem.*, 5, 2004, pp. 414-418.

[11] Burakov, S. V. et. al. Research on convergence of multipopulation binary- and real-coded genetic algorithms for solution of crystal structure from X-Ray powder diffraction data. *Crystal Research and Technology*, **50**, 2015, pp. 724–728.

[12] Favre-Nicolin, V., Černý, R. FOX, 'Free Objects for crystallography': a modular approach to ab initio structure determination from powder diffraction. *J. Appl. Crystallogr.*, **35**, 2002, pp. 734–743.

[13] Zaloga, A. N. et. al. Multi-population genetic algorithm for determination the crystal structure of substances from X-ray diffraction data. *Journal of Siberian Federal University. Chemistry*, **7**, 2014, pp. 572-580.

[14] Zaloga, A. N. et. al. Determination of the $[Pt(NH_3)_5Cl]Br_3$ crystal structure from X-ray powder diffraction data using multi-population genetic algorithm. *Powder Diffraction*, **32**, 2017, pp. 110-117.

[15] Charkin, D. O. et. al. One disorder out of two orders: Synthesis and crystal structures of cation-ordered $PbNaF_2NO_3$, anion-ordered Pb_2OFNO_3, and continuous disordered $(Pb,Na)_2(O,F)_2$-dNO_3 solid solution with Sillen-derived structures. *Solid State Sciences*, **71**, 2017, pp. 111-116.

[16] Pennington, W. T. DIAMOND – Visual Crystal Structure Information System. *J. Appl. Crystallogr.*, **32**, 1999, 1028-1029.

[17] Yakimov, Y. I. The evolutionary method of modelling the crystal structure of substances from powder diffraction data. *Journal of Siberian Federal University. Chemistry*, **6**, 2013, pp. 180-191.

[18] Brester, C. et. al. Restart Operator for Multi-Objective Genetic Algorithms: Implementation, Choice of Control Parameters and Ways of Improvement. *International Journal on Information Technologies and Security*, **4** (vol. 9), 2017, pp. 25-36.

[19] Sopov, E., Semenkin, E. Automated Synthesis of Selection Operators in Genetic Algorithms using Genetic Programming. *International Journal on Information Technologies and Security*, **4** (vol. 9), 2017, pp. 13-24.

Information about the authors:

Aleksandr Zaloga is research engineer at the Siberian Federal University (Krasnoyarsk, Russia). He is PhD in Physical Science. His research interests include powder diffraction, crystal structure determination, Rietveld method, evolutionary computation.

Igor Yakimov is professor at the Siberian Federal University (Krasnoyarsk, Russia). He is Dr. Sc. in Physical Science. His research interests include powder diffraction, quantitative phase analysis, Rietveld method, evolutionary computation.

Petr Dubinin is research engineer at the Siberian Federal University (Krasnoyarsk, Russia). He is PhD in Technical Science. His research interests include powder diffraction, quantitative phase analysis, Rietveld method.

Manuscript received on 28 January 2018