Computer Physics Communications 181 (2010) 1623–1632

Contents lists available at ScienceDirect

Computer Physics Communications

www.elsevier.com/locate/cpc

How to predict very large and complex crystal structures

Andriy O. Lyakhov [a,*], Artem R. Oganov [a,b], Mario Valle [c]

[a] Department of Geosciences and Department of Physics and Astronomy, and New York Center for Computational Sciences, Stony Brook University, Stony Brook, NY 11794-2100, USA
[b] Geology Department, Moscow State University, 119992, Moscow, Russia
[c] Data Analysis and Visualization Services, Swiss National Supercomputing Centre (CSCS), Via Cantonale, Manno 6928, Switzerland

ARTICLE INFO

Article history:
Received 29 January 2010
Received in revised form 28 May 2010
Accepted 5 June 2010
Available online 12 June 2010

Keywords:
Crystal structure prediction
Evolutionary algorithms
Genetic algorithms
Global optimization
Fingerprint function
Genetic drift
Order parameter

ABSTRACT

Evolutionary crystal structure prediction proved to be a powerful approach in discovering new materials. Certain limitations are encountered for systems with a large number of degrees of freedom ("large systems") and complex energy landscapes ("complex systems"). We explore the nature of these limitations and address them with a number of newly developed tools.

For large systems a major problem is the lack of diversity: any randomly produced population consists predominantly of high-energy disordered structures, offering virtually no routes toward the ordered ground state. We offer two solutions: first, modified variation operators that favor atoms with higher local order (a function we introduce here), and, second, construction of the first generation non-randomly, using pseudo-subcells with, in general, fractional atomic occupancies. This enhances order and diversity and improves energies of the structures. We introduce an additional variation operator, coordinate mutation, which applies preferentially to low-order ("badly placed") atoms. Biasing other variation operators by local order is also found to produce improved results. One promising version of coordinate mutation, explored here, displaces atoms along the eigenvector of the lowest-frequency vibrational mode. For complex energy landscapes, the key problem is the possible existence of several energy funnels – in this situation it is possible to get trapped in one funnel (not necessarily containing the ground state). To address this problem, we develop an algorithm incorporating the ideas of abstract "distance" between structures. These new ingredients improve the performance of the evolutionary algorithm USPEX, in terms of efficiency and reliability, for large and complex systems.

Published by Elsevier B.V.

1. Introduction

Evolutionary algorithms present an attractive approach to the crystal structure prediction problem [1–7]. Algorithm USPEX [1–3] is a particularly simple and efficient strategy for predicting the most stable crystal structure for a given compound without requiring any experimental input [8–11]. However, the use of evolutionary algorithms (and most other global optimization techniques) for very large and complex systems encounters fundamental problems.

Crystal structure prediction requires finding the global minimum in a (usually huge) search space with dimensionality

$$d = 3N + 3 \tag{1}$$

where N is the number of atoms in the unit cell. The complexity of this search increases exponentially with d. For small systems, up to about ten atoms in the unit cell, many different techniques were shown to be effective (see for example [12,13]). A prerequisite here is that a successful and efficient structure searching method must include local optimization, i.e. structure relaxation. It removes the usually large noise from the fitness function and decreases the intrinsic dimensionality of the search space by chemical constraints that appear during local optimization (i.e. the actual dimensionality can become much lower than d in Eq. (1), making search feasible [14,15]).

Evolutionary algorithms were successfully applied to very complex problems in other fields [16] and, in principle, it should be possible to predict efficiently structures with ∼300 degrees of freedom (i.e. ∼100 or more atoms in the unit cell for an unconstrained search, or more, if constraints related to structure or symmetry are used). For an evolutionary structure predicting algorithm to be effective, we need a balance between high diversity during the exploration of the search space (but especially in the initial population) and its learning power (enabled by selection and variation operators). Usually the diversity of the initial population is achieved by randomly choosing trial solutions from the search space. Surprisingly, random sampling for large systems leads to a very low diversity in the population of structures, giving it little chance to improve. In fact, diversity (as measured by σ^2 of the distribution of abstract distances between structures [14,17]) is *anti-*

* Corresponding author.
E-mail address: andriy.lyakhov@stonybrook.edu (A.O. Lyakhov).

0010-4655/$ – see front matter Published by Elsevier B.V.
doi:10.1016/j.cpc.2010.06.007

correlated with the intrinsic dimensionality d^* [15]. Therefore with increasing dimensionalities, the average distance between structures tends to zero (i.e. all structures become similar).

What this means is that if we randomly place atoms inside the unit cell, the vast majority of structures (even after relaxation) will be highly disordered and virtually identical. It can be visualized as taking a piece of a liquid or gas and trying to build a crystal by replicating it in all directions. Most of the thus generated structures have similar thermodynamic properties (in particular, high energies and configurational entropies), low degree of order (some of the definitions of which have been proposed in [14]), and chemically inferior arrangement of the atoms. This phenomenon will be illustrated later, see Fig. 1. Bad parents usually give bad offspring, and two nearly identical disordered parent structures produce virtually the same disordered structure. This makes it exceedingly difficult for the algorithm to find a good structure. Some researchers overcome this situation by seeding their initial population with structures that have the desired symmetry [18]. However, this becomes risky if we have zero prior knowledge about the symmetry of the solution. Here we present a trick to deal with this problem by initializing a sub-random first generation, which combines the desired unbiasedness of random sampling with a higher quality of the initial structures. In addition to constructing a higher-quality first generation, we also propose improved variation operators that promote local order (and, thus, more chemically reasonable and diverse structures).

Another problem that we encounter in crystal structure prediction is the complexity of the energy landscape – in this case, even a relatively small system may pose a considerable challenge for structure prediction, in terms of efficiency and success rate. This problem is typical for all evolutionary algorithms that operate in a complex search space: if there are many good local minima that correspond to very different crystal structures, there is always a risk to be trapped in the basin of attraction ("energy funnel") of one of them [19–22], and problems appear if that region of the energy landscape does not contain the global minimum. Various methods for escaping local minima have been developed in the context of other methods [23–29]; here we propose a simple strategy in the context of evolutionary structure prediction. Ideally, we want a method that avoids trapping in the local minima, rather than one for escaping them.

The remainder of this paper is organized as follows. In Section 2 we will introduce fundamental concepts of fingerprints and local order that allow us to build more efficient variation operators, described in Section 3. A cell splitting algorithm that generates the good initial population is described in Section 4. In Section 5 we will present numerical results to prove the effectiveness of the enhanced algorithm. Finally we summarize the paper with conclusions.

2. New tools and concepts: Fingerprint function and local order

The key of our method to deal with the problem of trapping in local minima is the ability to detect similar structures: if the same structure is allowed to proliferate in the population and comes to dominate it, it may become exceedingly difficult to produce any radically different solution. To prevent such proliferation (called "genetic drift" [30] or "cancer growth") identical structures must be detected and removed from the population. The following problems have to be overcome to perform this task. Obviously, one cannot directly compare atomic coordinates. Usually they are represented in the unit cell vectors basis and there are (in principle, infinitely) many equivalent ways of choosing the unit cell, i.e. the coordinate system. Thus, identical structures could have completely different atomic coordinates, as well as completely different structures could have the same coordinates in different unit cells. Correct comparison of crystal structures must be independent of the choice of the unit cell. For practical purposes, small numerical errors in the structure should not greatly influence the structure description. And last but not least, a good method of comparing structures should give a quantitative measure of the similarity between structures. Some algorithms use the energy difference as a measure of structural difference – however, this is a valid measure only for simple systems with just one energy funnel, and such systems present little challenge to the previous version of USPEX [1,2].

In our improved version of the algorithm we use discretized fingerprint function [14] as a crystal structure descriptor, a vector that we call fingerprint *FP*. Fingerprint function (defined by formula (3) in [14]) is related to the pair correlation function and diffraction spectra, and has all the desired properties listed above. It does not depend on absolute atomic coordinates, but only on interatomic distances. Small deviations in atomic positions will influence fingerprints only slightly. Contrary to methods that analyze only geometrical properties of the structure [28,31,32], fingerprint function allows us to distinguish structures where different atoms are swapped in their positions. One could also measure the similarity between structures by computing the distance between their fingerprints. In our algorithm we use the cosine distance

$$d_{ij} = 0.5\left(1 - \frac{FP_i FP_j}{\|FP_i\|\|FP_j\|}\right) \tag{2}$$

One could use other metrics as well, for example Euclidean distance or Minkowski norm. The advantages of cosine distances are that (1) they can only take values between 0 and 1, enabling universal distance criteria (e.g. 0.01 – very small distance, 0.20 – large, 0.5 – very large distance) and (2) cosine distances are less affected by the "distance concentration effects" that generally create trouble in multidimensional spaces [33].

Fingerprint function describes the correlations between atomic positions. Therefore it can be used to characterize the degree of order Π in the system. In [14] a dimensionless and scale-invariant definition was proposed:

$$\Pi = \frac{\Delta}{\lambda}|FP|^2 \tag{3}$$

Here Δ is a fingerprint function discretization step and λ is a characteristic length (for example a cubic root from volume per atom for a given system). In this paper we introduce the local degree of order Π_i for individual atoms $i = [1, \ldots, N]$ in the unit cell. To do this we define, as was done in [14], the fingerprint function of the individual atom A_i relative to all atoms of type B surrounding it as

$$F_{A_i B}(R) = \sum_{B_j} \frac{\delta(R - R_{ij})}{4\pi R_{ij}^2 (N_B/V)\Delta} - 1 \tag{4}$$

where the sum runs over all jth atoms of type B within some distance threshold $R_{\max}$, V is the unit cell volume, N_B is the number of atoms of the type B in the unit cell and R_{ij} is interatomic distance between atoms A_i and B_j. $\delta(R - R_{ij})$ is a Gaussian-smeared delta function, absorbing numerical errors and making $F(R)$ a smooth function. Then we discretize $F(R)$ over bins of width Δ to obtain the fingerprint vector $FP_{A_i B}$. The local order is defined as

$$\Pi_i = \sqrt{\sum_B \frac{N_B}{N} \frac{\Delta}{(V/N)^{1/3}} |F_{A_i B}|^2} \tag{5}$$

where N is the total number of atoms in the unit cell. Taking a square root in (5) leads to a more linear correlation between average order and energy of the structure.

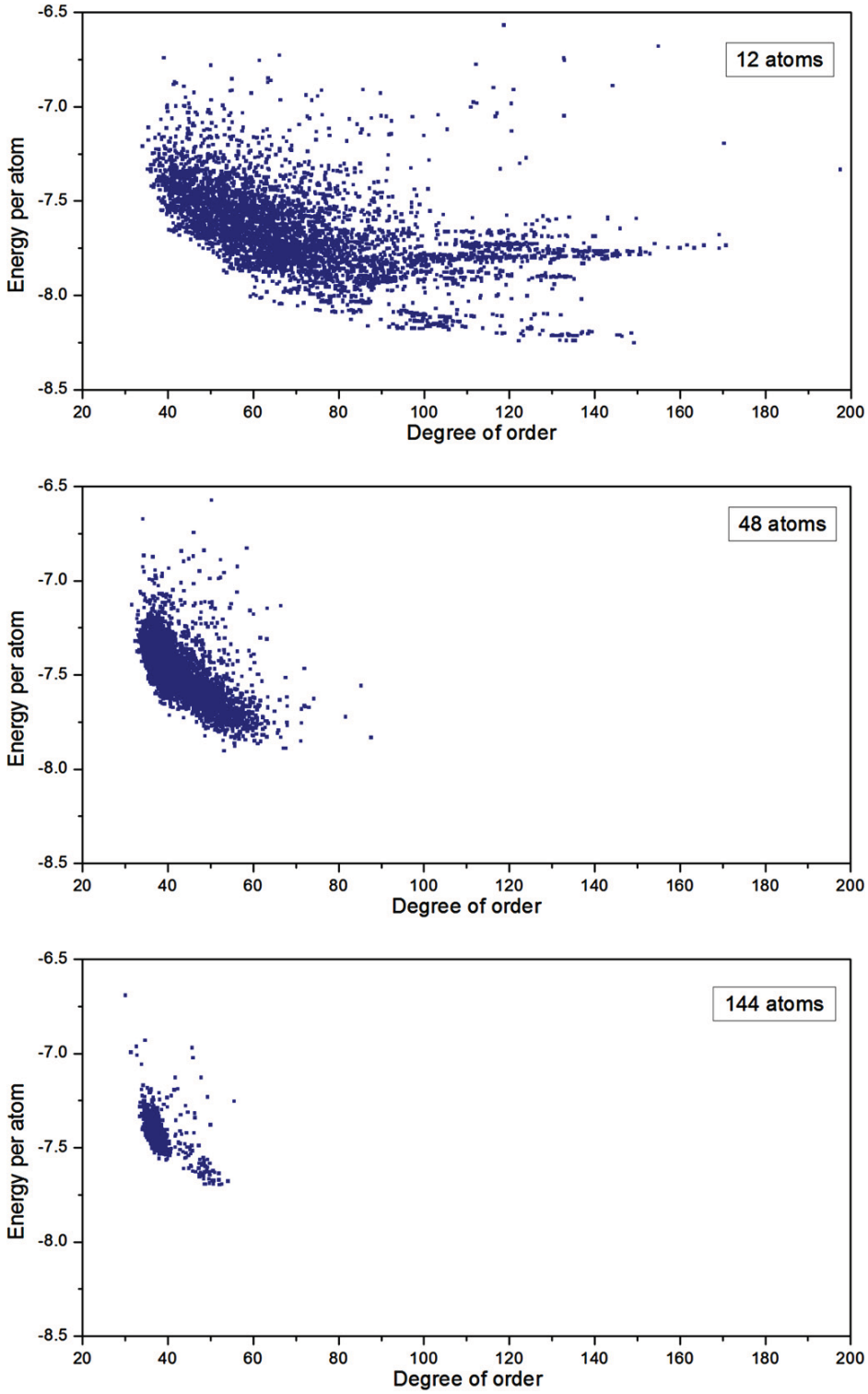

Fig. 1. Energy and degree of order Π (defined in [14]) of randomly generated structures for 12, 48 and 144 atoms in the unit cell. 5000 structures were generated and optimized for systems with 12 and 48 atoms in the unit cell and around 1000 structures were generated and optimized for system with 144 atoms in the unit cell.

Dimensionless local degree of order defined by (5) turns out to be a generally very useful concept. It can characterize the quality of the environment and its symmetry for a given atomic position, see Figs. 2 and 3. Fig. 2 shows the degree of order for an atom that is moved in the horizontal mirror plane (perpendicular to a 4-fold axis) within the unit cell of a simple cubic lattice. One can see that the positions with higher symmetry correspond in general to higher degree of order. If we decrease the Gaussian smearing of the delta function in (5) then the peaks are sharper.

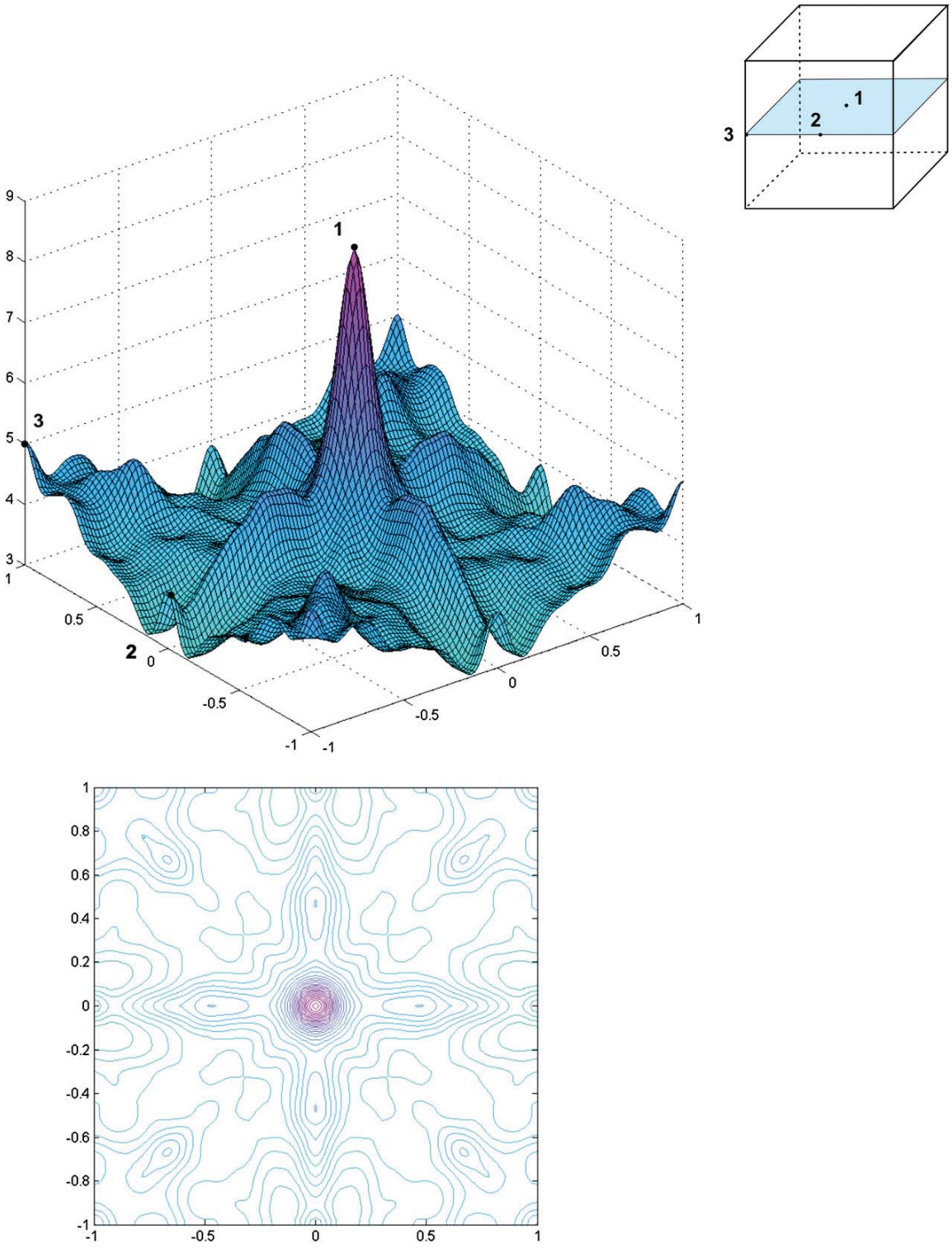

Fig. 2. The local degree of order measured in the horizontal mirror plane for a simple cubic lattice.

Fig. 3 shows one of the structures that we obtained for SiO_2 with 96 atoms in the unit cell (blue – low order, red – high order). One can see that local geometrical defects decrease the order of the surrounding atoms; compare, for example, pairs of atoms labeled as 1, 1′ and 2, 2′. Our tests show that local order is usually higher for atoms in more symmetric and "happy" environments and practically does not depend on atomic size or coordination number. We believe that this novel concept has many applications beyond those described in this paper.

If we average the local order of all atoms in the unit cell, the resulting average order is anticorrelated with the energy of the

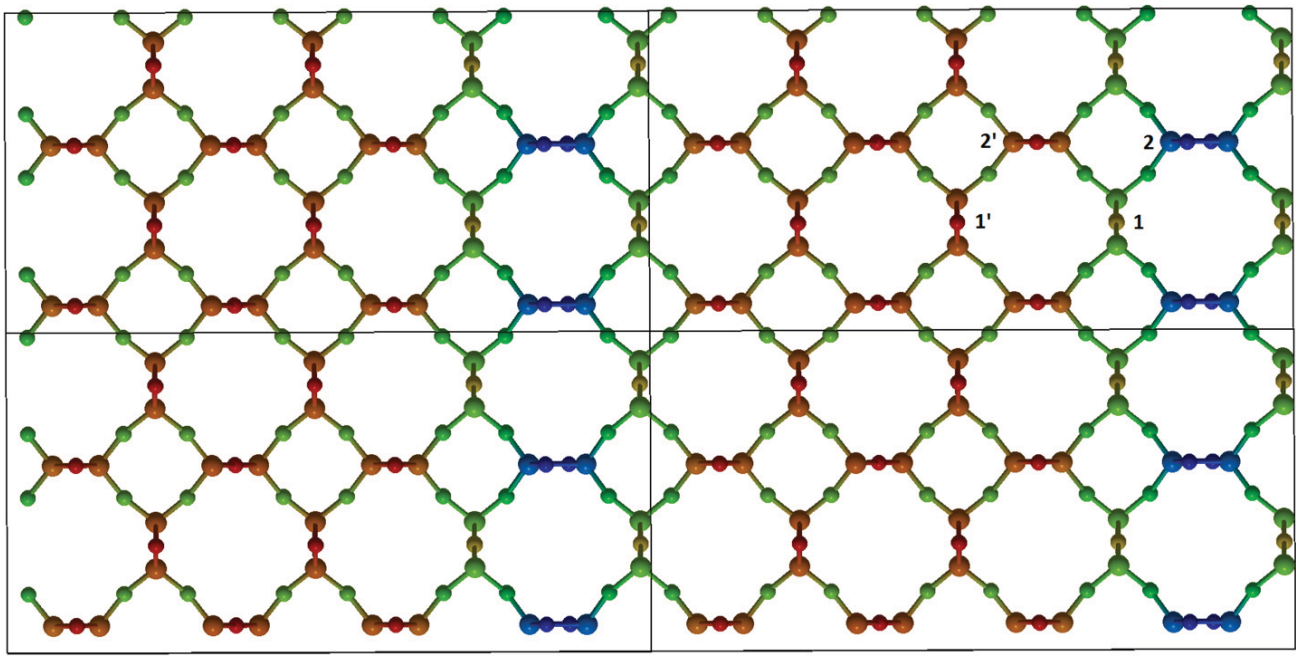

Fig. 3. One of the structures obtained for SiO_2 with 96 atoms in the unit cell. Atoms are colored according to their local degree of order (blue – low order, red – high order). Defective regions are clearly seen by low values of local order (blue atoms). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

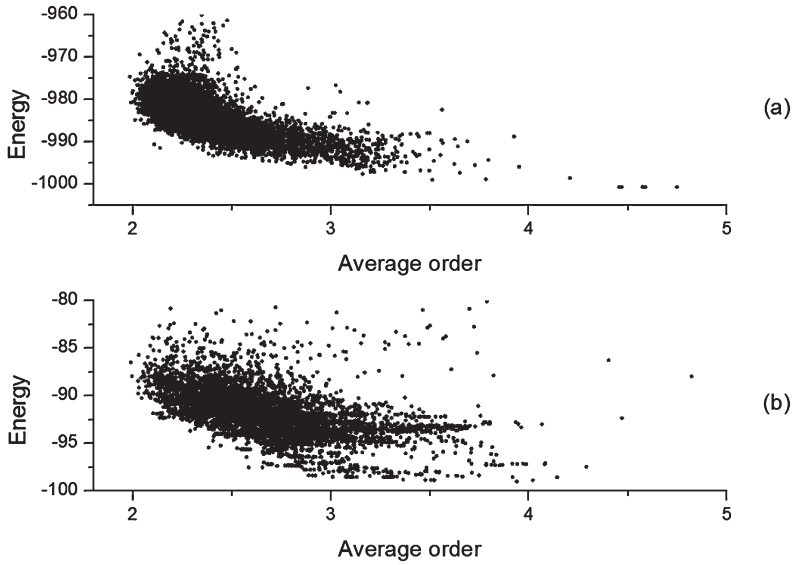

Fig. 4. Energy vs average local order for (a) 8800 structures of SiO_2 generated randomly and locally optimized, (b) 5000 binary Lennard-Jones structures with composition A_4B_8 (even in this frustrated system there is a clear anticorrelation; this anticorrelation is stronger in non-frustrated systems, like SiO_2 from panel (a)).

structure, see Fig. 4. Usually structures with lower energy have higher degrees of order.

3. Improved selection and evolutionary operators

Usually, evolutionary algorithm strives to achieve a balance between the diversity of the population and its quality. The first property is needed to preclude the 'cancer growth phenomena' described below. And the second property is necessary if we want to find a solution in a reasonable amount of time. We will show how the concepts of fingerprints and local order allow us to achieve such a balance in the new version of USPEX.

3.1. Overcoming 'cancer growth phenomena'

Even for chemically simple compounds, the energy landscape can have a complex topology with more than just one energy funnel. There is a risk to converge to the neighborhood of one solution, which tends to create its own replicas and overwhelm the population (we call it 'cancer growth phenomena'). This solution is not always the global minimum, and its domination precludes the exploration of other, possibly better, solutions, see Fig. 5. An extreme (fortunately, such extremes are rare) example of this is MgNH – even with 12 atoms in the unit cell it poses a significant problem for the standard algorithm (in two out of three attempted

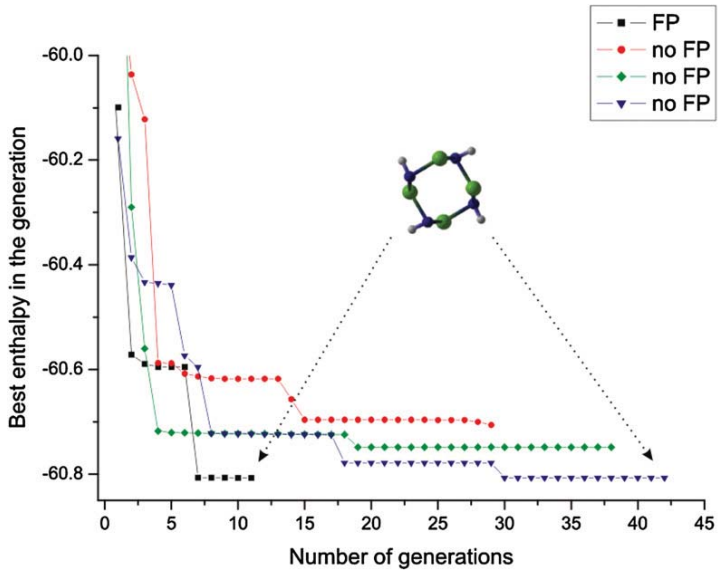

Fig. 5. Effectiveness of evolutionary algorithm with and without the use of fingerprint functions for MgNH.

runs a suboptimal solution was found and survived for a long time).

The degeneration of the population into clones of one structure can be visualized using the so called similarity matrix [17]. One could also visualize the dynamics of the diversity using the collective quasi-entropy, a quantity defined as:

$$S = \frac{-\sum_{i \neq j}(1 - d_{ij}) \log(1 - d_{ij})}{N^2 - N} \tag{6}$$

where the indices i and j run over all N structures that participate in the creation of the next generation (usually, the best 60–70% of the population). Quasi-entropy was defined in [14]; here we slightly modified its definition to make it more similar to the standard formula for the entropy expressed through occupation numbers.

Structure fingerprints provide a way to avoid trapping in a suboptimal energy funnel. We calculate the distance between fingerprints of different structures in a population and treat two structures as 'similar' when this distance is less than some small user-defined threshold $D_{\min}$. When the lowest-energy structures are selected to participate in producing the next generation, we ensure that all the selected structures are different. If our goal is to explore metastable states at the expense of convergence time, fingerprints allow us to use the *niche proportional population* [34,35] approach.

One has to mention that similarity is not transitive in our case. If structure A is similar to structure B and structure B is similar to structure C, this does not imply that structures A and C will be similar in our sense as well. This is a good property, since together with a proper choice of $D_{\min}$ it allows us to find a good balance between ability to find a local minimum and avoiding trapping in this minimum. On the one hand, we do not want to have too many structures from the same funnel, but on the other hand we need quite a few of them to efficiently explore that funnel.

Fig. 5 shows a comparison of performance of the USPEX algorithm with and without the use of fingerprints in the selection process. Calculations were done within the generalized gradient approximation [36] using the VASP code [37]. For complex structures the higher diversity of the population helps to find the global minimum faster. And while the original algorithm showed that it can get trapped in a local minimum, the new improved version successfully found the global minimum every time it was used. Only one out of three runs with the old algorithm for MgNH with 12 atoms in the unit cell found the correct solution (even then, only after 41 generations). At the same time all four runs with the new algorithm found the solution in not more than 20 generations (namely in 7, 7, 13 and 17 generations).

In general, our tests with different systems – N, GaAs, SiO_2, various Lennard-Jones systems – show that the new version of USPEX is clearly superior. However, sometimes, especially for very simple systems, the use of fingerprints may slightly slow down global optimization.

3.2. Improved 'survival of the fittest' and discarding too different parents

Many good evolutionary algorithms let a few best structures from a preceding generation survive into the next generation unchanged. This enhances the learning power and ensures that good solutions are not lost. This has an additional benefit of finding low-energy metastable states in addition to the ground state. However, as mentioned in the previous section, it is important to avoid trapping in a local minimum, which implies that the surviving structures should be very different, not merely non-identical. This is achieved by adding an additional 'similarity' threshold $D_{\text{best\_min}} > D_{\min}$, so that only one (lowest-energy) of the structures with distances less than this threshold will survive for the next generation.

Making many structures survive is beneficial, but has a danger if some of the surviving structures are energetically and chemically poor. Therefore, we propose a dynamical survival using the following formula:

$$N_{\text{keep}} = \min(N_{\text{best}}, N_{\alpha}) \tag{7}$$

Here N_{α} is determined as the highest position in ranking that satisfies the condition: the variance of fitness function values among surviving structures should be less than variance among the rest of the population times some coefficient α:

$$\text{Var}(E_1, \ldots, E_{N_{\alpha}}) < \alpha \cdot \text{Var}(E_{N_{\alpha}+1}, \ldots, E_N) \tag{8}$$

We also fix the maximum value for the number of surviving structures by setting the parameter N_{best}. Dynamical N_{keep} and threshold $D_{\text{best\_min}}$ improve our initial algorithm and also

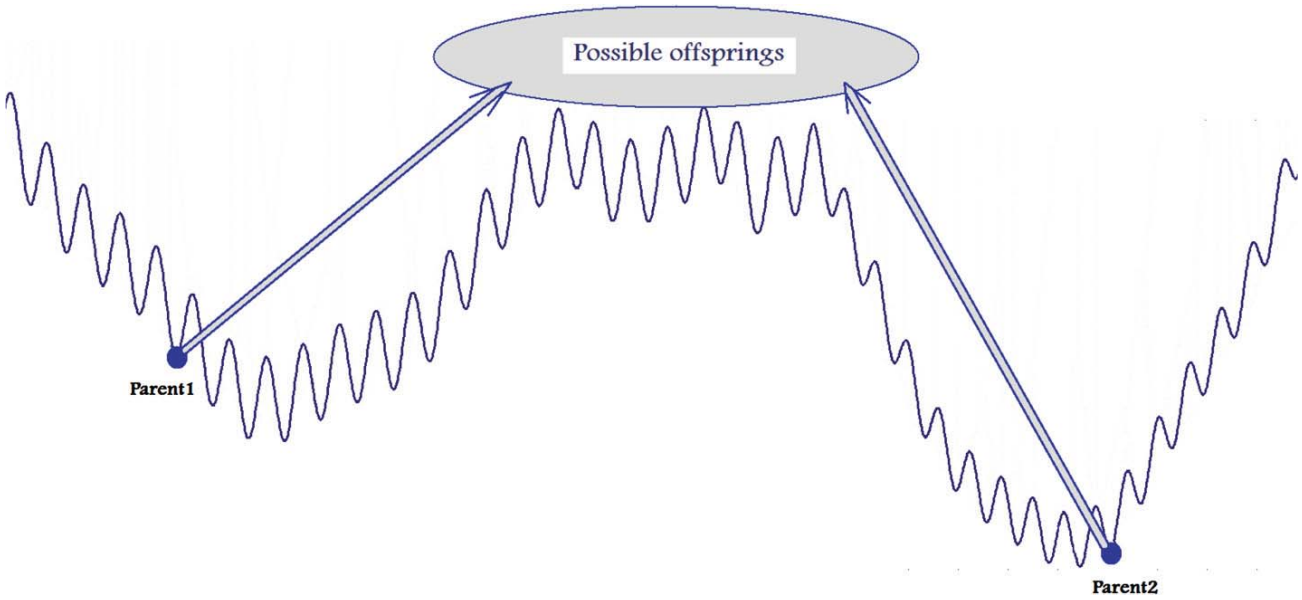

Fig. 6. Two parents from different funnels that have quite different structures usually produce a poor offspring in the high-energy areas between the funnels.

keep good suboptimal solutions for search of metastable states, if needed.

Another situation, where fingerprint functions allow us to reduce the number of poor offspring structures, is when heredity operator is applied to two very different parents. Usually, if we take two good parents from different funnels, their offspring will be extremely bad, see Fig. 6. This is largely avoided by enforcing a maximum distance between structures that are allowed to mate. This trick is similar to '*niching*' [38] used in cluster structure prediction calculations (and is more universal and less empirical). Such a small change can have a large impact on studies of complex systems. When there are good structures belonging to different funnels, the probability of choosing two very different good structures as parents is quite high, while the chances of a good offspring are expected to be low.

3.3. Improved heredity and mutation operators using local order

The ability to distinguish nicely located atoms using the concept of local order allows us to enhance the spatial heredity operator [2] and introduce a new mutation operator, the coordinate mutation. When the offspring structure is produced by heredity operator there is a corrector step adjusting the number of atoms of each kind so that chemical composition of the offspring is the same as for parents. Atoms in excess are removed randomly and atoms in shortage are added randomly from one of the parents until the desired number of atoms of each kind is reached. We can improve this step by relating the probability of an atom being removed or added to its order. Atoms with higher order have higher probability to be added and lower probability to be removed. Our tests show that this increases the general efficiency of the algorithm. We also improved the heredity operator by creating from each parent structure a few random slices of the same thickness instead of a single slice [2] and then using the one with the highest average order to produce a child structure.

As for mutations, so far USPEX used only lattice mutation and atomic identity swaps (permutation). Coordinate mutation was found [2] to be ineffective, because "blind" displacement of the atoms is much more likely to decrease the quality of a structure than to increase it. However, the local degree of order allows us to introduce a non-blind coordinate mutation (where poorly located atoms with low order are mutated more vigorously), which actually improves the algorithm. In this new variation operator, all atoms in the unit cell are moved in random direction, the distance for this movement for atom i is picked from a normal Gaussian distribution with sigma defined as

$$\sigma_i = \sigma_{\max} \frac{\Pi_{\max} - \Pi_i}{\Pi_{\max} - \Pi_{\min}} \tag{9}$$

Thus atoms with higher order are perturbed less than atoms with low order. Such operator is useful for fast exploration of the funnel and is reminiscent of annealing – upon heating, weaker bound atoms make larger displacements and can even diffuse to new positions, which results in a new structure upon cooling (or local optimization), while clusters of nicely ordered atoms are more likely to be preserved by this operator – small atomic displacements are removed after local optimization. In the tests described here we use $\sigma_{\max}$ of the order of a typical bond length (e.g. $\sigma_{\max} = 2$ Å for SiO_2). Our numerical results show that new enhanced variation operators greatly increase the efficiency of the algorithm.

We have also successfully explored a special coordinate mutation operator (that we call 'softmutation') to improve the structures, especially those in the first generation. The idea is to perform concerted mutation of atomic coordinates instead of a random one. In this case, atoms are moved along the eigenvector of the softest mode (both positive and negative directions need to be tried). The amplitude of displacement is a user-defined parameter, and if initial positions of the atoms after displacement get too close to each other, this amplitude is increased until constraints are satisfied. If a structure has already been softmutated, the next lowest-frequency non-degenerate mode is used. To check the effectiveness of the new operator we did calculations for 100 structures of garnet pyrope ($Mg_3Al_2(SiO_4)_3$) with 160(!) atoms in the unit cell. They were generated randomly and then each structure was mutated along the softest mode. Each mutant structure was relaxed. Out of 100 attempts, 79 led to lower energies, 2 remained unchanged and 19 become worse. The success rate of 79% is extremely high for system with that many atoms in the unit cell.

For the usability of softmutation, the crucial step is the efficient construction of the dynamical matrix (ab initio construction

of the dynamical matrix would be extremely expensive). To calculate the softest modes we construct the dynamical matrix [39] from bond hardness coefficients [40] (disregarding differences in atomic masses).

$$D_{\alpha\beta}(a,b) = \sum_m \frac{\partial^2}{\partial\alpha_a^0 \partial\beta_b^m}\left(\frac{1}{2}\sum_{i,j,l,n} H_{i,j}^{l,n}\left[\left(x_i^l - x_j^n\right)^2 + \left(y_i^l - y_j^n\right)^2 + \left(z_i^l - z_j^n\right)^2\right]\right) \quad (10)$$

Here coefficients α, β denote coordinates (x, y, z); coefficients a, b, i, j describe the atom in the unit cell; coefficients l, m, n describe the unit cell number. Therefore x_i^l is, for example, the x-coordinate of atom i in the unit cell l. $H_{i,j}^{l,n}$ is bond hardness coefficient between the atom i in the unit cell l and atom j in the unit cell n. For more details see [39] and [40].

Eigenvectors of $D_{\alpha\beta}(a,b)$ corresponding to lowest non-zero eigenvalues determine the direction of softmutation. More accurate (and expensive) ways of constructing the dynamical matrix could be used. However this is hardly needed, given the results of our approach.

Dynamical matrix can also be used for another coordinate mutation operator when atoms are moved in random directions with the amplitudes determined by their thermal ellipsoids (that can be calculated using $D_{\alpha\beta}(a,b)$). We are continuing tests of these variation operators to gather more comprehensive statistics. Preliminary results show us that using softmutation as a variation operator we can enhance the effectiveness of the algorithm.

4. Dealing with large systems

As we mentioned in the Introduction, large randomly created systems resemble a glassy disordered state. This is illustrated in Fig. 1, which shows very low degrees of order for such structures, in comparison with smaller-cell structures. Such structures are energetically poor and extremely similar to each other (i.e. increasing the population size does little to improve it). Therefore random sampling has to be improved to increase the diversity of the population. We propose a very simple way to do it: unit cell splitting. The idea is to split the cell into smaller subcells, generate one subcell using random sampling and then replicate it to fill the cell. Thus we obtain more ordered structures, with are usually much better from the chemical and energetic points of view. Such structures have a higher translational symmetry, leading to potential risks that the final solution may be biased toward higher-symmetry solutions. Such risks are minimized when the subcells are not too small ($>$15–20 atoms/subcell) and several different splittings are applied in the population concurrently. The higher translational symmetry of parent structures is broken in the offspring by heredity and permutation operators, while the lattice mutation operator preserves translational symmetry (see [2] for details on variation operators).

The main advantage of this method is a higher-quality initial generation compared to the normal random sampling, see Fig. 7. This can speed up the global optimization. In fact, for a large system this (or seeding with promising non-random structures) is a key to finding any good solutions at all. Even for small systems this can increase the efficiency of the algorithm, but certain caution is needed.

When the number of atoms in the unit cell is not good for splitting into identical subcells (e.g. prime numbers), the algorithm creates random vacancies to keep the correct number of atoms in the whole unit cell. This situation is illustrated in Fig. 8. Such structures are well known in non-stoichiometric compounds and even in the elements – a similar (though not identical) situation exists in the high-pressure orthorhombic phases Cs-III and Rb-III with 84 and 52 atoms in the unit cell, respectively, which can be represented as supercells of the body-centered cubic structure with additionally inserted atoms [41]. Pseudo-subcells are a very general case, they lead to non-trivial solutions that do not incorporate extra symmetry, but do have an increased degree of order. We also expect them to be an effective tool to improve the conventional random sampling methods [42,43] when dealing with large systems. For our test SiO_2 system, see next section, it increased the success rate for random search by almost an order of magnitude. The success rate for random sampling without unit cell splitting was 0.04%: 7 out of 18 821 randomly generated structures gave the ground state after local optimization. With unit cell splitting it had 0.29% success rate: 61 ground states were found in 21 020 attempts.

5. Numerical results

We have tested the new ideas on a relatively large and complex system – SiO_2 with 24 atoms in the unit cell (75 degrees of freedom). This system possesses at the same time a huge search space (almost unlimited number of tetrahedral frameworks, in addition to non-tetrahedral and non-framework structures), propensity for disorder, and a complex energy landscape, while being computationally convenient (enabling reasonable statistics to be collected in a reasonable time). Structure relaxation and energy calculations were performed using GULP code [44] and a Buckingham potential based on the one by Sanders et al. [45]. The results are shown in Table 1. All calculations were limited to 50 generations, each containing 40 structures. Those calculations where the global minimum was found ('guessed') in the first generation were discarded from statistics. Algorithm #1 is the original version of USPEX; all newer versions compared in Table 1 show improved performance (except when the only change was to simply kill strictly identical structures, while letting many lowest-energy structures to survive, as proposed in [31] – in this case on average 22.35 generations were needed to find the global minimum). All enhanced versions of the algorithm use fingerprints to improve selection. Algorithm #2 does not use cell splitting and variation operators improved by order. Algorithm #3 uses cell splitting and improved heredity, but does not use improved variation operators. Algorithms #4–7 have different ratios of offspring created by different variation operators. Algorithm #4 doesn't use coordinate mutation. Algorithms #5 and #7 use mostly coordinate mutation and only rarely lattice mutation, the difference is in the value of $\sigma_{\max}$ (2 Å for algorithm #5 and 1 Å for algorithm #7). Algorithm #6 does not use heredity and different mutation operators were applied with similar probability.

One can see that the combination of the ideas described above significantly speeds up global optimization – the average number of generations required to reach the global minimum is reduced by at least 30% and the success rate is substantially increased.

We have tested (but without collecting such comprehensive statistics) that the new algorithm also works well for much larger systems – e.g. SiO_2 with 48 atoms/cell and 96 atoms/cell.

6. Conclusions

We have discussed the ways to improve evolutionary algorithms for crystal structure prediction. The major problems that arise, especially when dealing with large and complex multi-funnel systems, are described. We have shown that one can increase both the efficiency and the success rate of the algorithm, as well as the range of its applicability, by using fingerprint functions, local degrees of order, and cell splitting. The fingerprint function improves the selection rules (through removing clones) and heredity operator (through niching), thus enabling studies of more complex

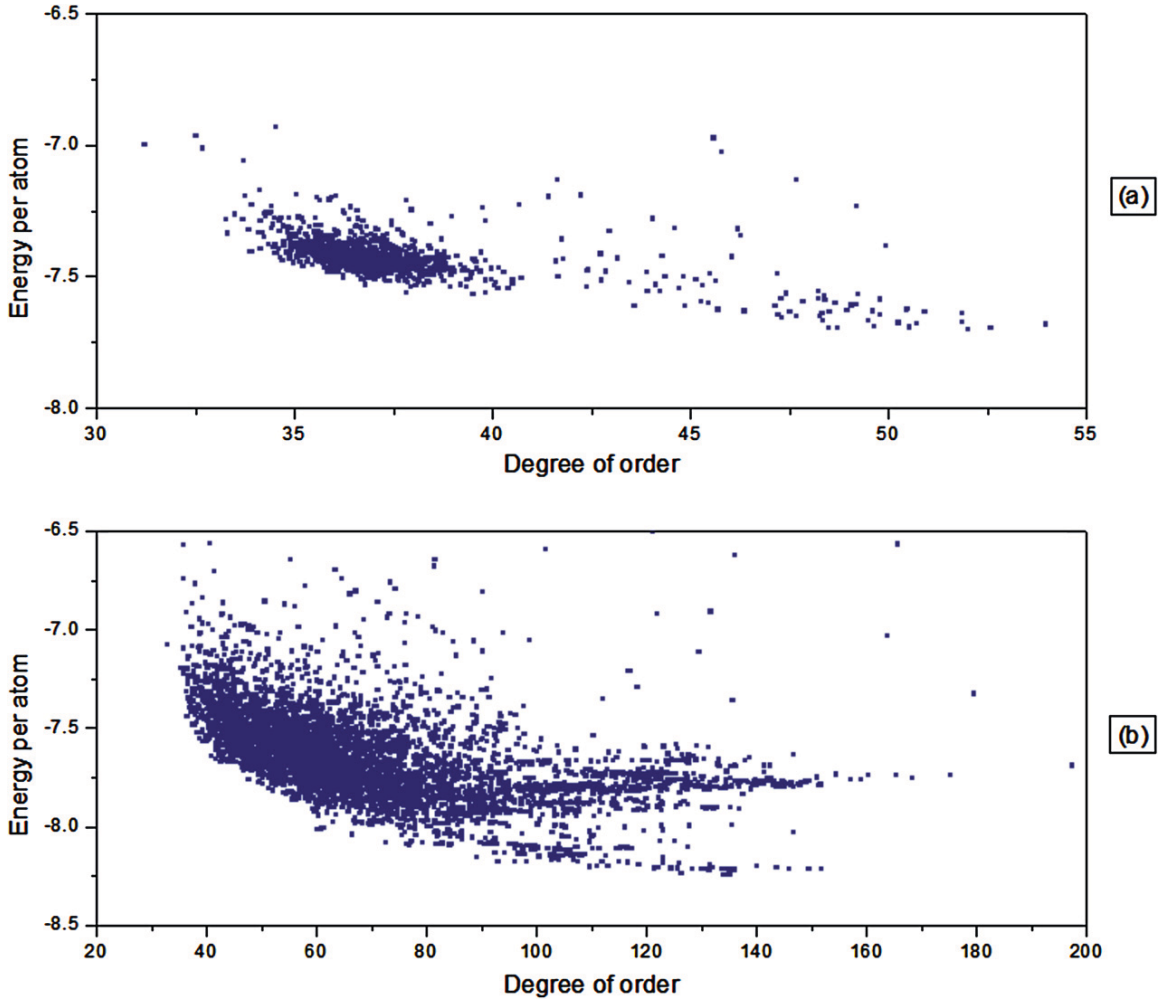

Fig. 7. Energy and degree of order of randomly generated structures of a binary Lennard-Jones $A_{48}B_{96}$ system (a) without the unit cell splitting and (b) with splitting into 12 smaller subcells.

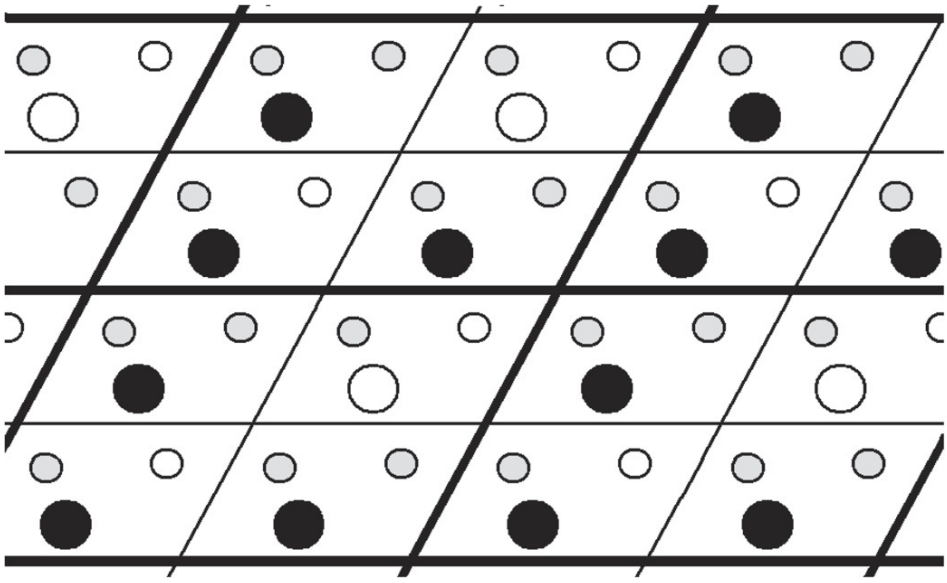

Fig. 8. Illustration of the pseudo-subcell algorithm for composition A_3B_6 (atoms A – large black circles, B – small gray circles, empty circles – vacancies). Thick lines show the true unit cell, split into 4 pseudo-subcells (thin lines).

Table 1
Average number of generations needed to find the global minimum for SiO_2 with 24 atoms in the unit cell.

	Algorithm						
	#1	#2	#3	#4	#5	#6	#7
Improved selection	–	+	+	+	+	+	+
Cell splitting	–	–	+	+	+	+	+
Niching	–	+	+	+	+	+	+
Variation operators improved by order	–	–	–	+	+	+	+
Average number of generations	**20.35**	**17.83**	**14.08**	**14.13**	**12.2**	**9.21**	**13.72**
Standard deviation	13.77	10.57	10.67	12.15	7.8	5.88	11.55
Success rate (%)	59.65	85.71	92.5	83.67	87.5	100	96.7
Number of calculations participating in statistics	57	21	39	49	56	24	91

systems. Cell splitting results in a better initial generation, and variation operators modified using local order drive search toward more chemically reasonable structures. Introduction of non-blind coordinate mutation is also seen as an important improvement. Our tests show that new ingredients clearly improve the efficiency and success rate of the algorithm.

Acknowledgements

We thank two anonymous referees for useful comments. Funding from DARPA (Young Faculty Award), the Research Foundation of Stony Brook University, Intel Corp., and Rosnauka (Russia, contract 02.740.11.5102) is gratefully acknowledged, as well as access to the Blue Gene L/P complex of the New York Center for Computational Sciences, to the Skif MSU supercomputer (Moscow State University) and to the Joint Supercomputer Centre of the Russian Academy of Sciences. We thank the National Science Foundation of China for the Research Fellowship for International Young Scientists under grant No. 10910263.

References

[1] A.R. Oganov, C.W. Glass, J. Chem. Phys. 124 (2006) 244704.
[2] C.W. Glass, A.R. Oganov, N. Hansen, Comp. Phys. Comm. 175 (2006) 713–720.
[3] A.R. Oganov, Y. Ma, C.W. Glass, M. Valle, Psi–k Newslett. 84 (2007) 142–171.
[4] S.M. Woodley, Struct. Bonding 110 (2004) 95–132.
[5] T.S. Bush, C.R.A. Catlow, P.D. Battle, J. Mater. Chem. 5 (1995) 1269–1272.
[6] A.R. Oganov, Y. Ma, A.O. Lyakhov, M. Valle, C. Gatti, Rev. Mineral. Geochem. 71 (2010) 271–298.
[7] A.O. Lyakhov, A.R. Oganov, Y. Wang, Y. Ma, in: A.R. Oganov (Ed.), Crystal Structure Prediction, Wiley–VCH, 2010, submitted for publication.
[8] A.R. Oganov, C.W. Glass, S. Ono, Earth Planet. Sci. Lett. 241 (2006) 95–103.
[9] A.R. Oganov, J. Chen, C. Gatti, Y.-Z. Ma, Y. Ma, C.W. Glass, Z. Liu, T. Yu, O.O. Kurakevych, V.L. Solozhenko, Nature 457 (2009) 863–867.
[10] Y. Ma, M.I. Eremets, A.R. Oganov, Y. Xie, I. Trojan, S. Medvedev, A.O. Lyakhov, M. Valle, V. Prakapenka, Nature 458 (2009) 182–185.
[11] M. Martinez-Canales, A.R. Oganov, A.O. Lyakhov, Y. Ma, A. Bergara, Phys. Rev. Lett. 102 (2009) 087005.
[12] R. Martonak, A. Laio, M. Parrinello, Phys. Rev. Lett. 90 (2003) 075503.
[13] C.J. Pickard, R.J. Needs, Phys. Rev. Lett. 97 (2006) 045504.
[14] A.R. Oganov, M. Valle, J. Chem. Phys. 130 (2009) 104504.
[15] M. Valle, A.R. Oganov, Crystal fingerprints space. A novel paradigm to study crystal structures sets, Acta Cryst. A, in press.
[16] H.K. Tsai, J.M. Yang, Y.F. Tsai, C.Y. Kao, IEEE Trans. Syst. Man. Cybern. B Cybern. 34 (2004) 1718–1729.
[17] M. Valle, A.R. Oganov, in: Proc. IEEE VAST, 2008, pp. 11–18.
[18] M.D. Wolf, U.J. Landman, J. Phys. Chem. A 102 (1998) 6129.
[19] M. Eigen, Naturwiss. 55 (1971) 465–522.
[20] D.E. Goldberg, J. Richardson, in: Proceedings of the Second International Conference on Genetic Algorithms and Their Applications, Lawrence Erlbaum, New Jersey, 1987, pp. 41–49.
[21] S.W. Mahfoud, in: L.D. Whitley, M.D. Vose (Eds.), Proc. Foundations Genetic Algorithms, San Francisco, 1995, pp. 185–223.
[22] K. Deb, D.E. Goldberg, in: J.D. Schaffer (Ed.), Proc. 3rd Intl. Conf. Genetic Algorithms, San Mateo, 1989, pp. 42–50.
[23] S. Goedecker, J. Chem. Phys. 120 (2004) 9911–9917.
[24] A. Laio, M. Parrinello, Proc. Natl. Acad. Sci. USA 99 (2002) 12562–12566.
[25] J. Pannetier, J. Bassas-Alsina, J. Rodriguez-Carvajal, V. Caignaert, Nature 346 (1990) 343–345.
[26] J.C. Schön, M. Jansen, Angew. Chem. Int. Ed. Engl. 35 (1996) 1287.
[27] D.J. Wales, J.P.K. Doye, J. Phys. Chem. A 101 (1997) 5111.
[28] G. Rossi, R. Ferrando, Chem. Phys. Lett. 423 (2006) 17–22.
[29] Z. Li, H.A. Scheraga, Proc. Natl. Acad. Sci. USA 84 (1987) 6611–6615.
[30] D.E. Goldberg, P. Segrest, in: Proceedings of the Second International Conference on Genetic Algorithms and Their Applications, Lawrence Erlbaum, New Jersey, 1987, pp. 1–8.
[31] Y. Yao, J.S. Tse, D.D. Klug, J. Sun, Y. Le Page, Phys. Rev. B 78 (2008) 054506.
[32] N.L. Abraham, M.I.J. Probert, Phys. Rev. B 77 (2008) 134117.
[33] K. Beyer, J. Goldstein, R. Ramakrishnan, U. Shaft, in: ICDT '99 Proceedings of the 7th International Conference on Database Theory, in: Lecture Notes in Computer Science, vol. 1540, Springer-Verlag, London, 1999.
[34] B.L. Miller, M.J. Shaw, in: Proc. IEEE Int. Conf. Evolutionary Computation, 1996, pp. 786–791.
[35] A. Della Cioppa, C. De Stefano, A. Marcelli, IEEE Trans. Evol. Comput. 8 (2004) 580–592.
[36] J.P. Perdew, K. Burke, M. Ernzerhof, Phys. Rev. Lett. 77 (1996) 3685.
[37] G. Kresse, J. Furthmüller, Comp. Mater. Sci. 6 (1996) 15.
[38] B. Hartke, J. Comput. Chem. 20 (1999) 1752.
[39] M. Born, K. Huang, Dynamical Theory of Crystal Lattices, Oxford University Press, Oxford, 1954.
[40] K. Li, X. Wang, F. Zhang, D. Xue, Phys. Rev. Lett. 100 (2008) 235504.
[41] V.F. Degtyareva, Phys. Usp. 49 (2006) 369.
[42] C.M. Freeman, J.M. Newsam, S.M. Levine, C.R.A. Catlow, J. Mater. Chem. 3 (1993) 531–535.
[43] L.D. Lloyd, R.L. Johnston, Chem. Phys. 236 (1998) 107.
[44] J.D. Gale, Z. Kristallogr. 220 (2005) 552.
[45] M.J. Sanders, M. Leslie, C.R.A. Catlow, J. Chem. Soc. Chem. Commun. (1984) 1271.