Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 15 October 2003
Accepted 25 April 2004

© 2004 International Union of Crystallography
Printed in Great Britain – all rights reserved

Solving the Sayre equations for centrosymmetric structures with a genetic algorithm

Yi Zhou and Wu-Pei Su*

Department of Physics and Texas Center for Superconductivity and Advanced Materials, University of Houston, Houston, TX 77204, USA. Correspondence e-mail: wpsu@uh.edu

Sayre's equations give a set of relationships that exist among the structure factors of an equal-atom structure. In order to obtain the correct phases of the structure factors, a genetic algorithm is used to minimize a least-squares residual of Sayre's equations. In the genetic algorithm, a phase is treated as a gene and the whole set of phases is considered as a chromosome. Every chromosome is relaxed to a nearby local minimum by quenching after being produced from a previous generation. Trial calculations for a structure containing 92 non-H equal atoms with synthetic data and another structure containing 62 non-H equal atoms with real data are presented. Compared to simulated annealing, a genetic algorithm is a more efficient means of global optimization.

1. Introduction

X-ray diffraction is one of the most useful tools for crystal structure determination. If the amplitudes and phases of X-ray diffractions can be found, the electron density can be calculated through a Fourier transform directly. In practice, experiments give only the amplitudes without their phases. The recovery of the phases is termed the 'phase problem' and is a major research topic in crystallography. Pioneered by Hauptman & Karle (1953), direct methods determine the phases from the diffraction amplitudes. By formulating the phase problem as a global minimization problem, simulated annealing (Kirkpatrick *et al*., 1983) has been employed in a series of papers (Su, 1995; Chen & Su, 2000; Liu & Su, 2000) for structural determination. Among those methods, Chen & Su (2000) have used one that solves Sayre's equations (Sayre, 1952) by simulated annealing.

It is well known that a genetic algorithm (Goldberg, 1989) is in general a better technique of global optimization (Horst & Pardalos, 1995) than simulated annealing. First proposed by Holland (1975), the genetic algorithm belongs to a new generation of intelligent global optimization techniques. It has been applied to numerous problems (Lopez *et al*., 2000; Chu & Chu, 2001; Grigorenko *et al*., 2002) including the optimization of the atomic structures of small clusters (Deaven & Ho, 1995; Rata *et al*., 2000; Garzon *et al.*, 1998; Lemes *et al*., 2002; Zeiri, 1995). Its role in solving the phase problem has been largely unexplored. To date, there exist only two rather specialized applications. Landree *et al.* (1997) have used a multi-solution genetic algorithm to solve simple surface structures from very noisy and incomplete diffraction data. Webster & Hilgenfeld (2001) have used a different version of the genetic algorithm to reconstruct approximate envelopes of two protein structures. It would be desirable to extend the methodology to treat a general structure.

In this work, a least-squares residual of Sayre's equations is used to represent the fitness of a trial phase set with respect to the true phase set. For centrosymmetric structures, a phase is either 0 or π and can be easily represented by one bit (gene). A structure can be represented by encoding its phase set into a string of bits (genes) called a chromosome. A genetic algorithm is then used to find the fittest chromosome. After that, we use Fourier transform and real-space peak-picking to obtain the atomic coordinates of the structure.

An important ingredient of our algorithm is that we perform a relaxation immediately after every chromosome is created. This greatly improves the efficiency of the algorithm, since the relaxation significantly reduces the sample space and simplifies the landscape.

Compared to Chen & Su's (2000) work, the genetic algorithm gives better performance than simulated annealing in solving the Sayre equations. It can also handle real experimental data, as shown in our 62 non-H-atom structure example.

2. Methodology

2.1. Sayre's equations

A crystal structure that is composed of equal atoms has an electron-density function $\rho(\mathbf{x})$ composed of identical and resolved peaks. For such a structure, the function $\rho(\mathbf{x})$ and its square $\rho^2(\mathbf{x})$ are almost alike, except that the peaks have different shapes. If the atomic shape is known, the structure factor $F^{\rm sq}(\mathbf{h})$ of the squared structure can be expressed in terms of the structure factor $F(\mathbf{h})$ of the original structure as

$$F^{\rm sq}(\mathbf{h}) = S(\mathbf{h})F(\mathbf{h}), \qquad (1)$$

where $S(\mathbf{h})$ is the appropriate function to account for the change of atomic shape.

According to Fourier theory, multiplication in real space is equivalent to convolution in Fourier space. Therefore, we have

$$F^{sq}(\mathbf{h}) = (1/v)\sum_{\mathbf{k}} F(\mathbf{k})F(\mathbf{h}-\mathbf{k}), \quad (2)$$

where v is the volume of the crystal unit cell in Å^3.

It follows that the structure factors satisfy the relations

$$vS(\mathbf{h})F(\mathbf{h}) = \sum_{\mathbf{k}} F(\mathbf{k})F(\mathbf{h}-\mathbf{k}). \quad (3)$$

This equation is true for all **h**, and is called a Sayre equation. For organic crystals, all atoms are alike except H atoms, which can be ignored for their small contribution to the electron density. A set of phases can be correct only if it satisfies all Sayre's equations for all **h**. A least-squares residual for Sayre's equations is

$$R = \sum_{\mathbf{h}} |vS(\mathbf{h})F(\mathbf{h}) - \sum_{\mathbf{k}} F(\mathbf{k})F(\mathbf{h}-\mathbf{k})|^2 \quad (4)$$

and the minimization of the residual yields the correct phase set.

2.2. Genetic algorithm

Within this formulation, the phase problem is converted into a global optimization problem. It is very complicated owing to high nonlinearity and the large number of degrees of freedom. Therefore, we use a well known global search method: the genetic algorithm (GA). What are GAs? GAs are search algorithms based on the mechanics of natural selection and natural genetics. First, an initial population of candidate solutions is introduced and the fitness of each individual of the population is calculated. Then, the population produces offspring to form a new population, according to the rule that fitter individuals have a better chance to produce offspring. Generally, the average fitness of the child generation is better than the parent generation. We repeat this procedure until a global minimum is found.

Coming back to our problem of centrosymmetric structures, a structure can be represented by encoding its phase set into a string of bits (genes) called a chromosome. Initial generation is composed of chromosomes with randomly generated phases and their fitness is calculated according to equation (4). The lower the residual, the fitter the chromosome is. It is important to point out that, in practice, we do not have the complete set of structure factors, so we must compensate for the omission of terms in equation (3) by multiplying the left-hand side by an empirical factor of the form $p - q|\mathbf{h}| - r|\mathbf{h}|^2$. It is found that the best values of p, q and r depend on the amount and type of data incompleteness, but are independent of the structure (Sayre, 1972).

Reproduction of a generation is performed mainly by three operations: direct copy, crossover and mutation. The parents are first selected with a probability function in terms of their fitness. A chromosome is first picked up randomly, and then its probability of being a parent is calculated as

$$P = \frac{R - R_{\rm thres}}{R_{\rm min} - R_{\rm thres}} \quad (P = 0 \text{ for } R > R_{\rm thres}), \quad (5)$$

where $R_{\rm thres} = 2R_{\rm ave} - R_{\rm min}$, $R_{\rm min}$ and $R_{\rm ave}$ are the minimum and average residual of the generation, respectively.

After a parent is selected, it is directly copied into a mating pool for further genetic operations. When we have enough parents, simple crossover is performed in two steps. First, two members are selected from the mating pool randomly. Second, a crossover position is selected uniformly at random, and the two parents interchange all the genes on one side of the position. After crossover, random sites are selected and the genes at those sites are flipped (*i.e.* $0 \rightarrow 1$, $1 \rightarrow 0$). This is called a mutation.

In addition to the above basic operations, we also use some advanced operations to make the algorithm more efficient. One problem with an ordinary genetic algorithm is that sometimes a rather good solution in the old generation is lost in the new generation. To prevent that from happening, we keep our best solutions by directly copying them into the next generation (elitism). Another problem is premature convergence, which means most of the population becomes similar before it finds its real optimum. We use a technique called niche to overcome this problem. Niche is a fitness sharing mechanism. Since we do not want the individuals in a population to resemble each other, we penalize highly similar individuals by artificially increasing their residuals.

First, a sharing function $s(i, j)$ is defined as

$$s(i,j) = \begin{cases} 1 - d(i,j)/0.4 & \text{if } d(i,j) \le 0.4 \\ 0 & \text{if } d(i,j) > 0.4, \end{cases} \quad (6)$$

where $d(i, j)$, the distance, is the percentage of distinct phases between two individuals. By this definition, only neighbors within a distance of 0.4 have non-zero sharing function values. Then, for a given individual, the degree of sharing is obtained by summing the sharing function over all individuals. Finally, the individual's modified residual is calculated through

$$R_{\rm mod}(i) = \frac{R(i) - R_{\rm thres}}{\sum_j s(i,j)} + R_{\rm thres}, \quad (7)$$

where $R_{\rm mod}$ is the modified residual and the summation is over all the individuals. In this way, highly similar individuals have less chance of being reproduced, therefore premature convergence is prevented or at least slowed.

An important ingredient of our algorithm is relaxation. It is performed immediately after a chromosome is created. To understand the advantage of relaxation, we must introduce the concept of a schema first. A schema is a similarity template describing a set of genes at certain positions. In our case, a schema is a set of phase combinations. A GA is effective because it increases crucial schemata exponentially after they appear. Relaxation can produce such schemata quickly and effectively because it brings a chromosome into a nearby local minimum, which means some phase combinations are correct. This saves a lot of time for the algorithm to explore and find good schemata, and therefore makes the algorithm more efficient. In our problem, the large number of parameters (*i.e.* 926 parameters for the 62 atom case) makes the appearance of a correct schema very difficult in a typical GA. We have to

depend on those local minimum patterns produced in relaxation, otherwise the problem will be too difficult to solve. Relaxation is done by quenching, a simulated-annealing procedure (Press *et al.*, 1992) at zero annealing temperature $T = 0$. A random site is flipped and kept that way only if the residual decreases. Otherwise, it is flipped back. This process is continued until a local minimum is found and no more flips can decrease the residual.

3. Examples

In order to test the feasibility and efficiency of our approach, we choose two medium-sized molecules with centrosymmetric crystal structures. The first structure is tetraundecylpentacyclooctacosadodecaenoctol teraethanol solvate ($C_{72}H_{112}O_8 \cdot 4C_2H_6O$), with space group $P\overline{1}$, and unit-cell dimensions $a = 12.533$, $b = 12.649$, $c = 25.319$ Å, $\alpha = 84.79$, $\beta = 80.74$, $\gamma = 83.84^\circ$ (Hibbs *et al.*, 1998) (Fig. 1). This structure contains 92 non-H atoms. Synthetic data are fabricated by modeling every atom as a Gaussian ball with density $\exp[-(r/a_0)^2]$, where $a_0 = 0.5$ Å. The strongest 1291 independent reflections are selected from the synthetic data of 1 Å resolution. This data set contains only a small fraction of the total number of reflections. As mentioned previously, we have to compensate for the omission of terms by using an empirical factor of the form $p - q|\mathbf{h}| - r|\mathbf{h}|^2$. An artificial structure with similar data incompleteness and with the same unit-cell parameters and chemical formula is used to calculate p, q and r. In this example, they are 0.536, 0.281 and −0.073, respectively.

An initial population of 50 randomly generated chromosomes started the evolution procedure. The probabilities of crossover and mutation were set to 60% and 1%, respectively. After 12 trials, each taking an average of 4 h CPU on a Pentium 2 GHz machine, we obtained three correct structures. The success ratio is about 25%. Fig. 2 shows the average and minimum residuals of evolving generations of typical successful and unsuccessful trials. One can see that the successful process exhibits a sudden drop in the minimum residual curve before the end of the evolution, in contrast to the unsuccessful ones.

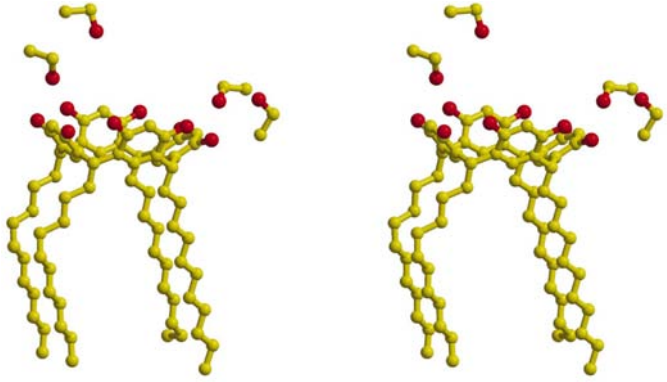

Figure 1
Stereodrawing of the molecular structure of the first example.

What is the origin of the sudden drops that distinguish successful trials from failed ones? Let us examine the solution hyperspace first. Our solution hyperspace resides in 1288 dimensions (three of the dimensions are pre-assigned to define the origin). Although in every dimension there are only two valid values (0 and 1), the whole configuration space is still astronomically large (2^{1288}). It is like finding a golf hole in a very large golf course without any sign or indication. A typical GA will give a rather smoothly evolving curve, because the GA is very good for exploration but not for refinement. We have found that our relaxation operation helps a typical GA to do the refining job; without the relaxation, the GA either converges very slowly or does not converge at all. In other words, a typical GA's job is to explore the golf course and find a good candidate neighborhood, and relaxation is used to find the lowest point in such a neighborhood. For a successful run, the sudden drop indicates the relaxation operation finding a rather deep hole. Whether this is the global minimum needs to be checked further. This can be done by performing a Fourier transform to obtain the electron-density map. In this case,

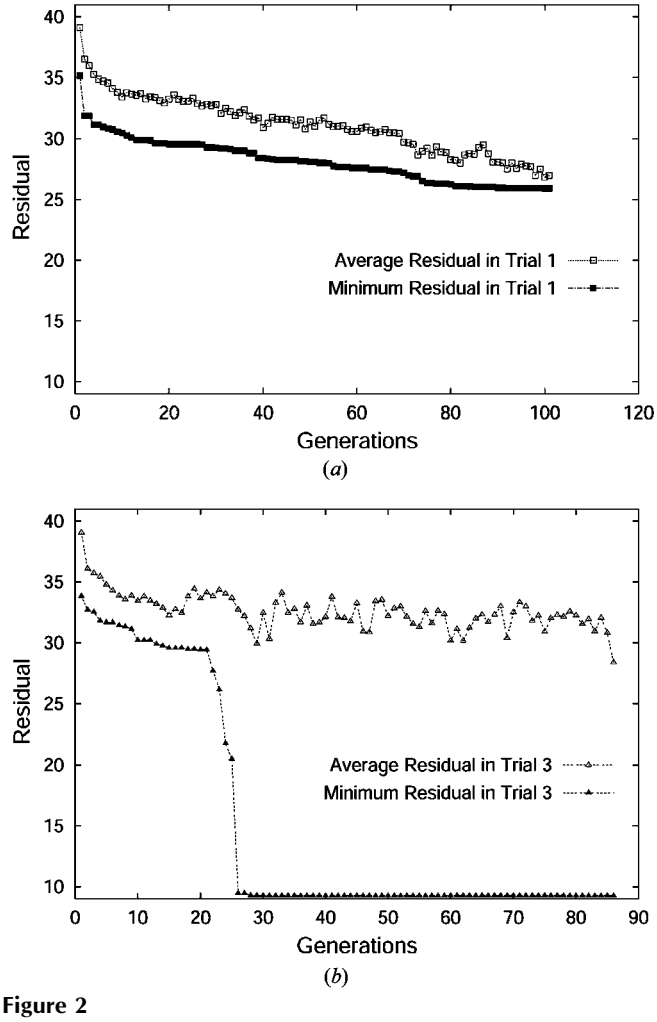

Figure 2
(*a*) A typical residual curve of an unsuccessful trial. The curves are smooth without a sudden drop. (*b*) A successful trial gives a sudden drop in the minimum residual curve. The smoothness of the average curve indicates the population has not converged.

98% of the 1288 reflections have the correct phases and the density map is rather accurate and makes chemical sense, as shown in Fig. 3. Further peak-picking and refinement can be done routinely to yield the accurate atomic coordinates and is not discussed here.

To illustrate the evolution progress, we show in Fig. 4(*a*) the detailed population distribution of a successful trial in every generation. Compared to a trial without the niche option, as shown in Fig. 4(*b*), we see that premature convergence is successfully avoided.

In order to further test the robustness of our algorithm, we use real diffraction data in the second trial calculation. The structure is 1,1,1-tris(4-hydroxyphenyl)ethane-1,4,8,11-tetraazacyclotetradecane–methanol $(C_{20}H_{18}O_3)_2\cdot C_{10}H_{24}N_4\cdot CH_4O$, with space group $P\overline{1}$ and unit-cell dimensions $a = 8.221$, $b = 16.245$, $c = 17.337$ Å, $\alpha = 81.694$, $\beta = 89.656$, $\gamma = 86.468^\circ$ (Ferguson *et al.*, 1998) (Fig. 5). This structure contains 62 non-H atoms. The strongest 926 reflections were selected from the 8108 experimental ones, for which h ranges from −10 to 10, k from 0 to 10 and l from −10 to 22. p, q, r are calculated to be 0.557, −0.048 and 0.358, respectively. As in the previous example, the p, q, r are obtained from a random structure with identical lattice parameters and chemical formula. In addition, the atomic scattering factors listed in *International Tables for X-ray Crystallography* (Cromer & Waber, 1974) are used since real diffraction data are considered here. Using the same setting for parameters as in the first example, we obtained two correct structures after ten trials, each taking an average of 3 h CPU. The success ratio is thus about 20%. The evolution curves for successful and unsuccessful trials show similar properties to the previous example. Since the real experimental data contains more errors, the average time required to find a correct answer for this smaller structure is about the same as that of the previous example.

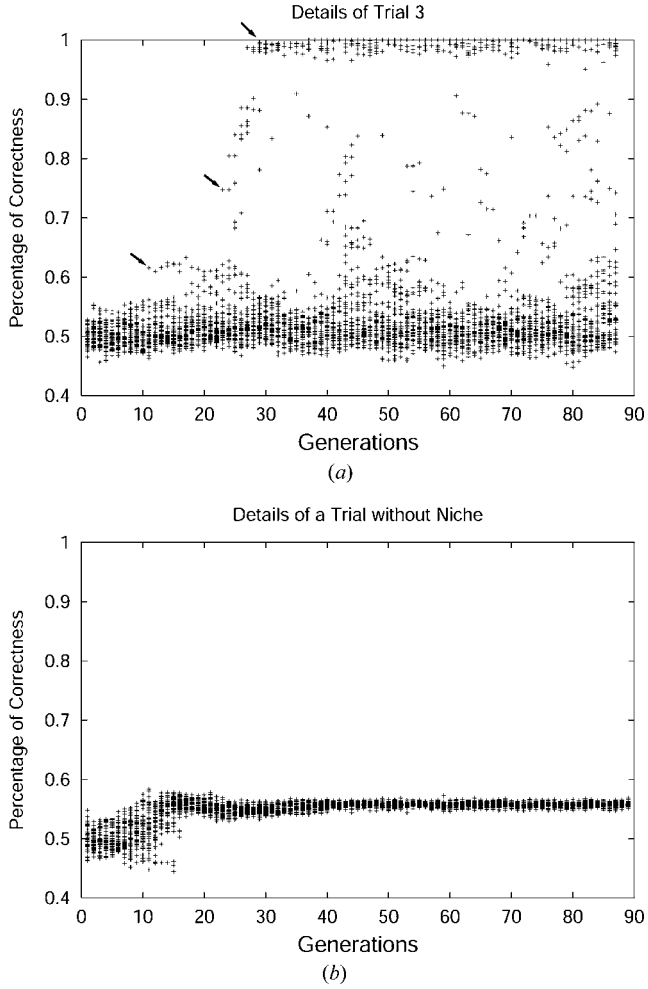

Figure 4
(*a*) The detailed population distribution of every generation in the successful trial. In this figure, at generation 11, an individual becomes obviously better and begins to attract more individuals to its neighborhood. Thanks to the niche method, this neighborhood is not overcrowded. The group of individuals continues to search its nearby areas for several generations and finds a much better configuration at generation 23. Finally, the group reaches the global optimum at generation 29. Convergence is successfully prevented in this process. (*b*) The detailed population distribution in a trial without the niche option. The population converges prematurely after only 15 generations.

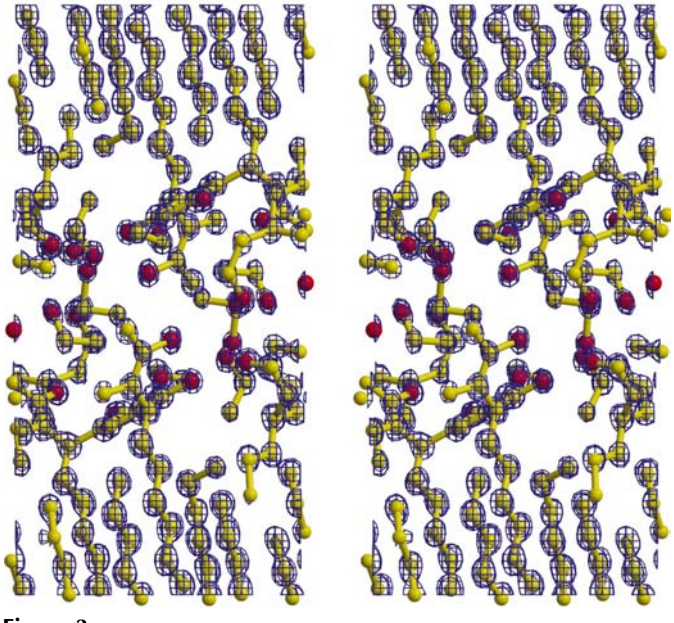

Figure 3
Stereodrawing of the calculated density map superimposed on the true structure in one unit cell. They match very well.

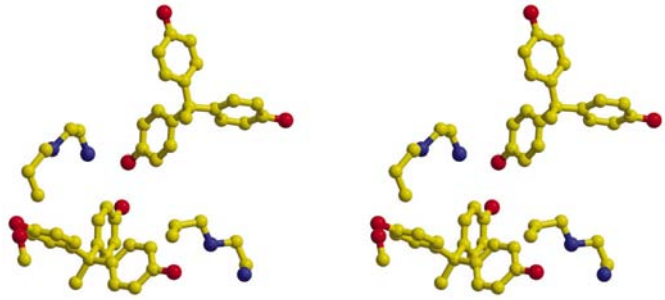

Figure 5
Stereodrawing of the molecular structure of the second example.

4. Discussion

In summary, we have developed a new method to solve the phase problem. Although we only present the results for centrosymmetric structures, it is not too hard to extend the method to other symmetry groups. In this method, the phase problem is formulated as a global minimization problem related to Sayre's equations, and the genetic algorithm is employed to find the global minimum. It is easy to understand in concept and very straightforward to implement. In this genetic algorithm, niche, a fitness sharing mechanism, is used to prevent premature convergence. Relaxation, which is not typical in a genetic algorithm, is used in our algorithm for its ability to significantly reduce the configuration space. Compared to simulated annealing, our method is much more efficient. It has a success ratio of about 20–25%, compared to 2–3% in the simulated-annealing case. It also requires much less computer time, less than 20% of the time required for a simulated-annealing procedure. Compared to modern direct methods such as the Shake-and-Bake method (Weeks & Miller, 1999) and the *SHELXS* program (Sheldrick, 1990), which can solve most structures with under 100 non-H atoms in minutes, the GA approach is still two orders of magnitude slower. However, as a new methodology, the speed might improve with time and it might solve structures not solvable by other means. Following Sayre's (1974) work, our method might be useful for refining high-resolution protein structures.

To further improve our algorithm, various modifications may be applied to solve larger structures and to achieve a better success ratio. For instance, selected initial population may be used instead of a random one. This may save us some time because some schemata may already exist in the selected initial population. This can be done by using a real-space approach first to find an approximate density map and then performing a Fourier transform to get an initial phase set. Further premature convergence prevention techniques can also be tried to improve the success ratio.

This work was partially supported by the Robert A. Welch Foundation (under grant No. E-1070) and the Texas Center for Superconductivity and Advanced Materials. We thank Yan Chen for useful conversations.

References

Chen, Y. & Su, W.-P. (2000). *Acta Cryst.* A**56**, 127–131.
Chu, X. & Chu, S.-I. (2001). *Phys. Rev. A*, **64**, 021493.
Cromer, D. T. & Waber, J. T. (1974). *International Tables for X-ray Crystallography*, Vol. IV, p. 100. Birmingham: Kynoch Press.
Deaven, D. M. & Ho, K. M. (1995). *Phys. Rev. Lett.* **75**, 288–291.
Ferguson, G., Glidewell, C., Gregson, R. M. & Meehan, P. R. (1998). *Acta Cryst.* B**54**, 139–150.
Garzon, I. L., Michaelian, K., Beltran, M. R., Posada-amarillas, A., Ordejon, P., Artacho, E., Sanchez-Portal, D. & Soler, J. M. (1998). *Phys. Rev. Lett.* **81**, 1600–1603.
Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning.* New York: Addison-Wesley.
Grigorenko, I., Speer, O. & Garcia, M. E. (2002). *Phys. Rev. B*, **65**, 235309.
Hauptman, H. & Karle, J. (1953). *Solution of the Phase Problem. I. The Centrosymmetric Crystal.* American Crystallographic Association Monograph, No. 3. Ann Arbor, USA: Edwards.
Hibbs, D. E., Hursthouse, M. B., Malik, K. M. A., Adams, H., Stirling, C. J. M. & Davis, F. (1998). *Acta Cryst.* C**54**, 987–992.
Holland, J. H. (1975). *Adaptation in Natural and Artificial Systems.* University of Michigan, USA.
Horst, R. & Pardalos, P. M. (1995). *Handbook of Global Optimization.* Dordrecht: Kluwer Academic Publishers.
Kirkpatrick, S., Gelatt, C. D. Jr & Vecchi, M. P. (1983). *Science*, **220**, 671–680.
Landree, E., Collazo-Davila, C. & Marks, L. D. (1997). *Acta Cryst.* B**53**, 916–922.
Lemes, M. R., Marim, L. R. & Pino, A. D. Jr (2002). *Phys. Rev. A*, **66**, 023203.
Liu, X. & Su, W.-P. (2000). *Acta Cryst.* A**56**, 525–528.
Lopez, C., Alvarez, A. & Hernandez-Garcia, E. (2000). *Phys. Rev. Lett.* **85**, 2300–2303.
Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1992). *Numerical Recipes in C: the Art of Scientific Computing*, 2nd ed., pp. 444–455. Cambridge University Press.
Rata, I., Shvartsburg, A. A., Horoi, M., Frauenheim, T., Siu, K. W. M. & Jackson, K. A. (2000). *Phys. Rev. Lett.* **85**, 546–549.
Sayre, D. (1952). *Acta Cryst.* **5**, 60–65.
Sayre, D. (1972). *Acta Cryst.* A**28**, 210–212.
Sayre, D. (1974). *Acta Cryst.* A**30**, 180–184.
Sheldrick, G. M. (1990). *Acta Cryst.* A**46**, 467–473.
Su, W.-P. (1995). *Acta Cryst.* A**51**, 845–849.
Webster, G. & Hilgenfeld, R. (2001). *Acta Cryst.* A**57**, 351–358.
Weeks, C. M. & Miller, R. (1999). *J. Appl. Cryst.* **32**, 120–124.
Zeiri, Y. (1995). *Phys. Rev. E*, **51**, 2769–2772.