Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 17 July 2001
Accepted 21 September 2001

A guided simulated annealing method for crystallography

C. I. Chou[a]* and T. K. Lee[a,b]

[a]Institute of Physics, Academia Sinica, Nankang, Taipei, Taiwan, and [b]National Center for Theoretical Sciences, Hsinchu, Taiwan. Correspondence e-mail: cichou@phys.sinica.edu.tw

A new optimization algorithm, the guided simulated annealing method, for use in X-ray crystallographic studies is presented. In the traditional simulated annealing method, the search for the global minimum of a cost function is only determined by the ratio of energy change to the temperature. This method designs a new quality function to guide the search for a minimum. Using a multiresolution process, the method is much more efficient in finding the global minimum than the traditional method. Results for two large molecules, isoleucinomycin ($C_{60}H_{102}N_6O_{18}$) and an alkyl calix ($C_{72}H_{112}O_8 \cdot 4C_2H_6O$), with different space groups are reported.

1. Introduction

The direct method (Hauptman, 1986, 1995; Weeks & Miller, 1999; Xu *et al.*, 2000) has been used to solve the X-ray phase problem for more than 50 years with great success. Great strides have been made by the development of the Shake-and-Bake method (DeTitta *et al.*, 1994; Weeks *et al.*, 1994), which although mainly based upon the direct method also uses information in real space. It is of interest to develop alternative methods that might compliment the direct method. Recently, several groups (Karle, 1991; Su, 1995*a*,*b*; Liu & Su, 2000) have been pursuing a total real-space approach. The real-space approach so far has not been very successful. Whether this lack of success is due to the intrinsic difficulty with the approach, the lack of computer power or the deficiency of the algorithm is unclear. Here we will show a newly developed method that helps us to solve structures of molecules with about 100 non-hydrogen atoms. Although this size is not yet comparable to what Shake-and-Bake could solve, it is much improved over all the previous real-space approaches. We have reason to believe that this new method could be further improved to solve structures much larger than 100 non-hydrogen atoms.

In the usual real-space approaches, one tries to find the atomic positions by matching structure factors with the observed intensities. This becomes effectively an optimization problem with a large number of variables. The simulated annealing (SA) method (Kirkpatrick *et al.*, 1983; Wille, 1986; Su, 1995*a*,*b*) has often been used in these optimization problems. For systems with a large number of variables, the SA method usually only obtains solutions trapped in a local minimum instead of a global minimum unless exponentially large computing time is used.

In this paper, we report a new algorithm, the guided simulated annealing (GSA) method, that greatly improves the traditional SA method. We inject a quality or guiding function to guide the search for the global minimum instead of using only the value of the cost function as the determining factor during the annealing process. The quality function, which may be problem specific, is chosen to be the charge density for the X-ray crystallography problem discussed in this paper. In order to make sure that the quality function will not guide the search into deep minima, it is essential to employ a multigrid or multiresolution process. At the start, a large grid is chosen so that only a coarse-grained charge density or quality function is constructed and used. As the system begins to explore lower and lower energy configurations, the grid size is reduced and the charge density will have a gradually improved spatial resolution. The idea of introducing guiding or a quality function in the search for the global minimum is not new. However, to make it work efficiently, the multiresolution process we introduce below is essential. It not only decreases the probability of being trapped in a deep local minimum but also greatly reduces the computing time. Here we note that W. P. Su (Su, 1995*a*,*b*) has mentioned the idea of a multiresolution approach without using the guiding function.

Below we shall first present the GSA method for X-ray crystallography. Then the results obtained for two molecules with different symmetries are reported and the conclusion follows.

2. Methodology

The basic idea is quite similar to the least-squares-fitting method. If we can arrange the positions of non-H atoms in the unit cell to make the best fit of the observed structure-factor data, then we may have obtained the correct molecular structure. This is exactly a global optimization problem with the total number of variables equal to three times the number of non-H atoms.

Usually for X-ray crystallography, the energy function (or the cost function) is defined as

$$E = \sum_i [\lambda |F_c(\mathbf{k}_i)| - |F_o(\mathbf{k}_i)|]^2, \quad (1)$$

where $|F_o(\mathbf{k})|$ is the absolute value of the observed or measured structure factor. $|F_c(\mathbf{k})|$ is the calculated structure factor when we input the positions of all the non-H atoms in $F_c(\mathbf{k}) = \sum_j \exp(i\mathbf{k}\cdot\mathbf{r}_j) f_j$, where f_j is the atomic scattering factor. λ is a scale factor for the absolute intensity that, although usually not known from experiments, could easily be determined by Wilson's method (Ladd & Palmer, 1977). Hence, for a molecule with N atoms, there are $3N$ variables for this energy function E. To find the absolute minimum in this $3N$-dimensional space is obviously a nontrivial task. The SA method proposed by Kirkpatrick *et al.* (1983) is most often used for this kind of problem.

In the SA approach, the Metropolis Monte Carlo scheme (Metropolis *et al.*, 1953) is usually used. The simulation starts with a random atomic configuration. Then each of the N atoms is attempted to be moved to a new random position in succession. The change in the energy function ΔE due to the move is calculated. If $\Delta E < 0$ then the move is accepted and a new atom configuration results. If $\Delta E > 0$, the move is accepted with a probability $\exp(-\Delta E/T)$, where T is the effective temperature. During the annealing process, T is reduced gradually to lead the system to the atomic configuration with the lowest energy E. At high temperatures, the system moves between many configurations similar to a liquid state. As temperature is reduced, the system begins to sample only low-energy configurations. For large systems, there are usually a very large number of local minimum. The system could be easily trapped in such a minimum. Then one has to repeat this annealing process until the global minimum is located.

It is not difficult to guess the SA approach described above will most likely fail for the energy function E in (1) if we consider large molecules without heavy atoms. So far we are not aware of any report of success by using this approach to obtain the correct structure for molecules with more than 60 or 70 non-H atoms unless there are a number of heavy atoms present.

There are two difficulties associated with the simple SA method described above. During the annealing process, the system essentially moves randomly in a $3N$-dimensional space besides the preference for lower-energy configurations. This is clearly a very inefficient approach. In addition, it leads to the second difficulty of trapping in a local minimum far from the configuration with the absolute minimum in E. If there is a way to guide the search path toward the vicinity of the global minimum then both difficulties would be reduced. The GSA method developed by us is exactly aimed at providing such a guiding function.

Before we start to describe the GSA method in detail, we shall first cast the energy function in a different form. Notice that the observed intensities $|F_o(\mathbf{k})|^2$ are usually larger for small $|\mathbf{k}|$. There could be orders of magnitude differences for different $|\mathbf{k}|$. Small $|\mathbf{k}|$ structure factors will only provide a low-resolution structure. Structure factors for large $|\mathbf{k}|$ usually have smaller amplitude but they are more sensitive to the positions of the atoms. However, to group all of them together in the single energy function as in equation (1) is inappropriate. The configurations chosen are heavily influenced by making $[|F_c(\mathbf{k})| - |F_o(\mathbf{k})|]^2$ very small only for those $|\mathbf{k}|$ with very large $|F_o(\mathbf{k})|^2$. For other $|\mathbf{k}|$ with very weak intensities, even if the calculated amplitude is several times larger than the observed value, they do not make any significant contribution in (1). In other words, a small error in large $|F_o(\mathbf{k})|^2$ will be more important than having large errors in small $|F_o(\mathbf{k})|^2$. Hence it is better to rewrite the cost function such that the effect of weak intensity $|F_o(\mathbf{k})|^2$ is not overlooked.

In this work, structure factors are grouped into subsets according to the magnitude of their observed intensities. Our energy function is defined as:

$$\begin{aligned} E &= E_1 + E_2 + E_3 + \ldots, \\ E_1 &= \sum_{i_1=1}^{N_1} \lambda_1 [|F_o(\mathbf{k}_{i_1})| - |F_c(\mathbf{k}_{i_1})|]^2, \\ E_2 &= \sum_{i_2=1}^{N_2} \lambda_2 [|F_o(\mathbf{k}_{i_2})| - |F_c(\mathbf{k}_{i_2})|]^2, \\ &\vdots \\ E_m &= \sum_{i_m=1}^{N_m} \lambda_m [|F_o(\mathbf{k}_{i_m})| - |F_c(\mathbf{k}_{i_m})|]^2. \end{aligned} \quad (2)$$

Here we require observed intensities in E_1 greater than E_2, than E_3 *etc*. The scale factor λ_i are chosen to make each subset about the same weight in the total energy function E. We set

$$\lambda_1 \sum_{i_1=1}^{N_1} |F_o(\mathbf{k}_{i_1})| = \lambda_2 \sum_{i_2=1}^{N_2} |F_o(\mathbf{k}_{i_2})| = \ldots = \lambda_m \sum_{i_m=1}^{N_m} |F_o(\mathbf{k}_{i_m})|$$

with $\lambda_1 < \lambda_2 < \lambda_3 \ldots$.

Since in the energy function E_1 most $|\mathbf{k}|$ are small, they will only provide a low-resolution or a coarse-grained image of the charge density. When E_2, E_3, ... are considered, we will obtain better and better resolution of the structure. Thus in our approach we shall first consider E_1 only and then include E_2 and E_3 in succession.

Once the energy function is decided, we start the annealing process. Just like the SA method, we begin with a random configuration of atoms and then move these atoms to find new configurations with smaller and smaller values of E_1 according to the Monte Carlo rules. Unlike the traditional SA method, we are not interested in finding the very low temperature result but only to obtain a group of configurations having reasonably small E_1. When the acceptance rate for Monte Carlo moves is getting too small because the system might be trapped in a local minimum, we stop this round of simulation. Carrying out these annealing processes many times starting with different random configurations, we obtain many low E_1 configurations. Rotation and/or translation operations allowed by the space group are used on these configurations to fix the origin problem.

An average coarse-grained charge-density distribution ρ can be obtained from these low E_1 configurations by first broadening the δ-function-like atomic charge density of each

configuration into a Gaussian function with a width σ. Then we make a weighted average of each configuration's coarse-grained charge-density distribution. Configurations with lower energy are given a larger weight. For simplicity, we assign the weight of each configuration according to its energy in a linear function. This charge-density distribution ρ does not give us very accurate positions of atoms but only the regions in the unit cell that atoms prefer most. This fact makes the function ρ a good guiding function to search for a lower minimum in the next round of Monte Carlo simulation.

An alternate way to construct the guiding function ρ is to divide the unit cell into many grids. When an atom is moved into a grid, its charge density is uniformly distributed in this grid. Then ρ is just a weighted average of the histograms of the distribution of atoms in the unit cell.

Once we have a guiding function ρ, we can start the SA process by adopting different selection rules. Atoms at positions with small values of ρ are given a larger probability to be selected to change their positions. For simplicity we use $1/\rho(r)$ as the selection probability for the atom at position r. The new position the atom is to be moved to is not randomly chosen. The atom is moved to regions with large values of ρ. To avoid overpacking the atoms in a small region, we impose the rule that two atoms cannot be situated closer than 1.2 Å. After the atom is selected and its new position is determined by the guiding function, we decide whether this move is to be accepted by the usual Metropolis Monte Carlo scheme used in the traditional SA method. Thus the function ρ guides the system toward regions in the configuration space where ρ is large. Having low energy is no longer the only criterion for selecting new configurations.

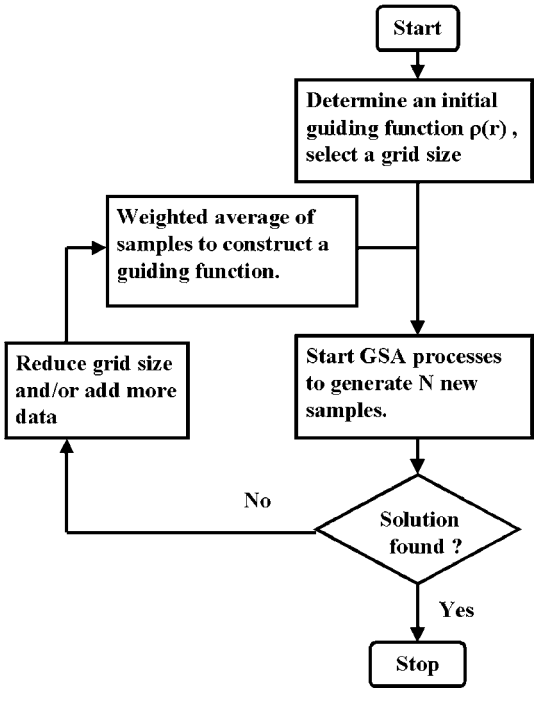

Figure 1
A flowchart of the GSA process.

After the guiding function is obtained from the low-energy E_1 configurations, in the next cycle of simulation we add E_2 to the energy function. At the same time, we increase the spatial resolution by decreasing the size of the grid or the Gaussian width σ. Thus a more refined charge-density distribution ρ or the guiding function is obtained.

Sometimes we will take several runs with smaller and smaller width while keeping the energy function the same until the system becomes trapped in a certain local minimum. In Fig. 1, the algorithm discussed above is illustrated by a flow chart.

To make the above general description of the methodology clearer, further details are discussed in connection with the following specific examples.

3. Examples

3.1. Isoleucinomycin ($C_{60}H_{102}N_6O_{18}$)

This structure was solved by Pletenev *et al.* (1992). The space group is $P2_12_12_1$. The cell constants are a = 11.516, b = 15.705, c = 39.310 Å and $\alpha = \beta = \gamma = 90^\circ$. There are four formula units per cell. The structure is shown in Fig. 2. We used 2000 reflections separated into four subgroups. Hydrogen atoms are neglected. In Fig. 3, the charge-density distributions obtained by the GSA method are shown in eight panels with

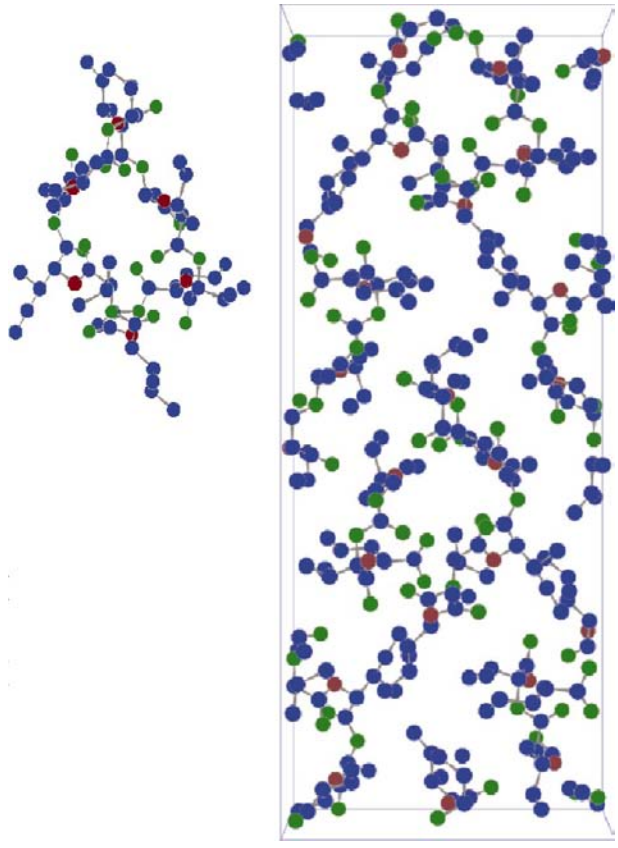

Figure 2
The molecular structure and unit-cell picture of isoleucinomycin. Blue = carbon, green = oxygen, red = nitrogen.

different resolutions or Gaussian width σ. In this figure, we project the three-dimensional charge density onto a two-dimensional plane perpendicular to the c axis. In the first panel where σ = 2.0 Å, the non-uniformity of the charge density is barely noticeable. As the resolution is increased with the grid size or σ reduced to 1.2 Å, the general shape of the charge distribution begins to emerge. After about 20 cycles or loops in the flow chart, the final result of the structure is obtained. The atom positions agree with the results published by Pletenev *et al.* The R factor is 0.22. This calculation took about one week on 20 Pentium III PC computers.

3.2. Tetraundecylpentacyclooctacosadodecaenoctol tetraethanol solvate ($C_{72}H_{112}O_8 \cdot 4C_2H_6O$)

The second structure was solved by Hibbs *et al.* (1998). There are 92 non-H atoms in this molecule and the space group is $P\bar{1}$. The cell constants are a = 12.533, b = 12.649, c = 25.319 Å and α = 84.79, β = 80.74, γ = 83.84°. There are two formula units per cell as shown in Fig. 4. In contrast to the last example, here we increase the number of Monte Carlo steps in our GSA computing process. The temperature is also reduced more slowly. Hence we spent more computing time during each cycle of annealing but better samples are obtained in each step. In Fig. 5, charge-density distributions for three different resolutions are shown. In the first panel where σ = 1.0 Å, the non-uniformity of charge density is already significant. As the resolution is increased with σ reduced to 0.6 Å, the general shape of the charge distribution is very close to the known result. The final result of the structure is obtained after only three steps. The atom positions agree with the results published by Hibbs *et al.* The R factor is 0.15. This calculation took about 3 days on 20 Pentium III PC computers. One of the reasons this structure is easier to resolve than the previous example is probably that the structure factors for this molecule are all real.

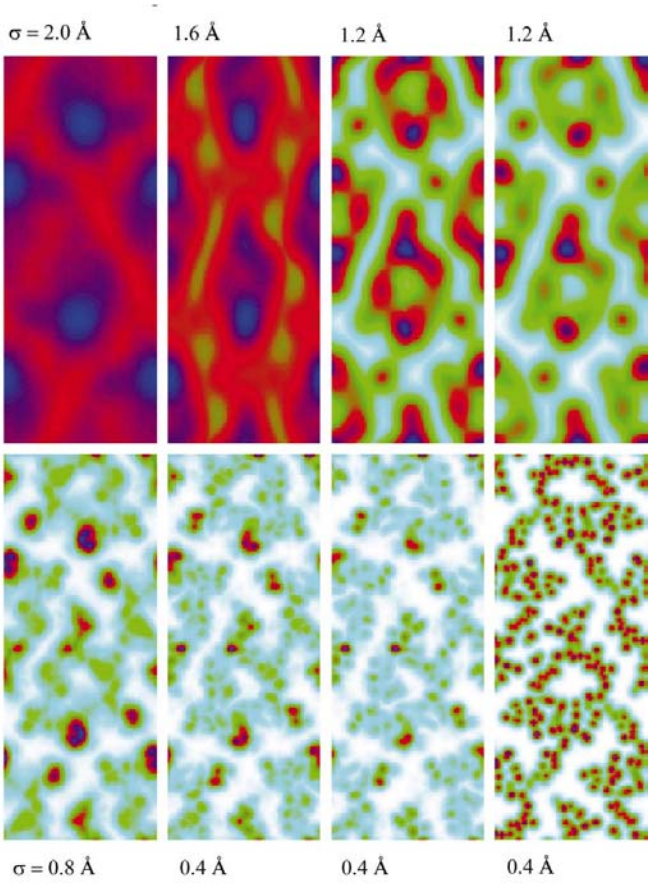

Figure 3
The guiding functions (the charge-density distributions) of isoleucinomycin in eight panels with different Gaussian width σ.

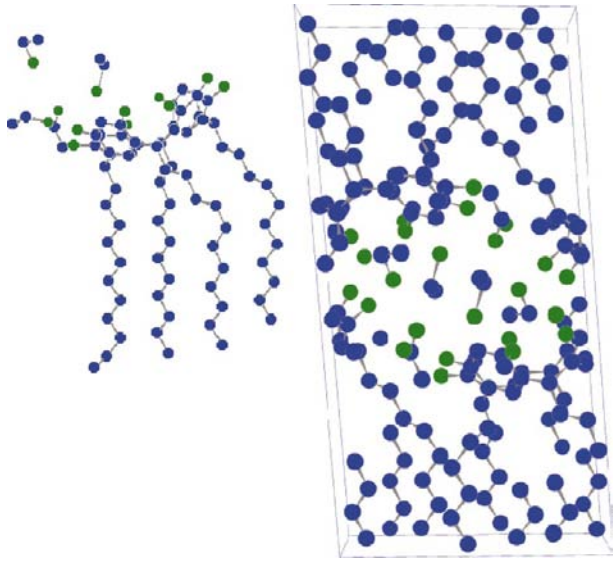

Figure 4
The molecular structure and unit-cell picture of tetraundecylpentacyclooctacosadodecaenoctol tetraethanol solvate. Blue = carbon, green = oxygen.

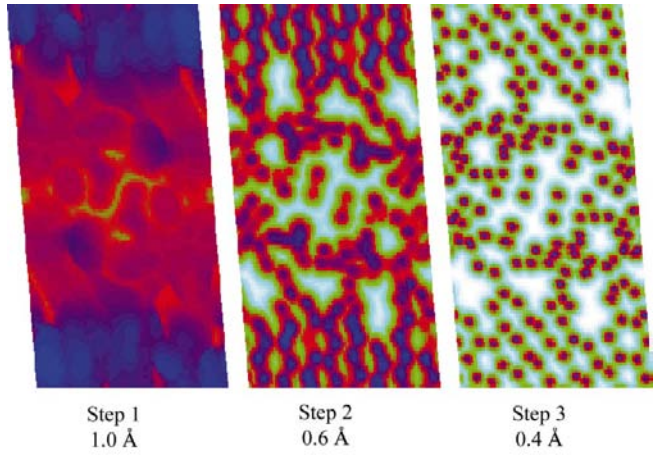

Figure 5
The guiding function (the charge-density distributions) of tetraundecylpentacyclooctacosadodecaenoctol tetraethanol solvate in three panels with different Gaussian width σ.

4. Summary and discussion

A new *ab initio* method, the GSA method, for X-ray crystallography is presented. This method tries to find the positions of atoms to fit the observed diffraction intensities directly instead of using the statistical distribution of phases to find phases of the structure factors and then the structure itself as in the direct method.

This GSA method has several new features. Firstly, the energy or cost function is modified. The structure factors are grouped into subsets according to their observed intensities. These sets contribute to the total energy function with equal weight and they are taken into consideration in succession. This ensures the effect of smaller structure factors is not completely overlooked. Secondly, we introduce a guiding function in the traditional simulated annealing method. In addition to the use of the Boltzmann factor to select the low-E configurations, here the guiding function determines new configurations of atoms to be sampled. This guiding function is constructed by using the histograms of the distribution of atoms in the unit cell during the annealing processes. Lastly, the GSA method is a multiresolution algorithm. In the first few stages, we choose the large grid size to get a coarse-grained distribution of atoms. Then the grid size is gradually reduced in the later stages so that we may find more accurate atomic positions.

The multiresolution process we used here greatly reduces the possibility of being trapped in a local minimum. It also has an advantage in computing time. In the initial several cycles, the energy landscape varies smoothly and the minima are shallow. Unless we impose a very low temperature, the system samples configurations quite freely. Since the resolution requirement is low, there is no need to go to very low temperatures in the annealing process. Thus little computing time is needed. At later stages, the grid size is reduced and the requirement for higher resolution demands more samples and more computing time. However, the higher resolution of the guiding function reduces the regions in the unit cell for atoms to be placed. Alternatively, the region of configuration space allowed by the guiding function gets smaller and smaller. Thus the requirement for computing time does not increase significantly.

This method has several other advantages. Its Monte Carlo nature makes it very easy to be used in a parallel-computing environment. The energy or cost function used in this work could easily be modified to include other considerations such as chemical knowledge, known phases or similar structures *etc.* Also, since this is now an optimization problem, the GSA method could be easily applied to other systems, such as the Lennard-Jones cluster problem (Jones & Ingham, 1925), the Thomson problem (Whyte, 1952) and the generalized Frenkel–Kontorova problem (Frenkel & Kontorova, 1938). These works will be reported elsewhere.

References

DeTitta, G. T. Weeks, C. M., Thuman, P., Miller, R. & Hauptman, H. A. (1994). *Acta Cryst.* A**50**, 203–210.

Frenkel, Y. I. & Kontorova, T. (1938). *Zh. Eksp. Teor. Fiz.* **8**, 1340; Engl. transl: (1938). *Sov. Phys. JETP*, **13**, 1.

Hauptman, H. (1986). *Science*, **233**, 178–183.

Hauptman, H. (1995). *Acta Cryst.* B**51**, 416–422.

Hibbs, D. E., Hursthouse, M. B., Abdul Malik, K. M., Adams, H., Stirling, C. J. M. & Davis, F. (1998). *Acta Cryst.* C**54**, 987–992.

Jones, J. E. & Ingham, A. E. (1925). *Proc. R. Soc. London Ser. A*, **107**, 636.

Karle, J. (1991). *Proc. Natl Acad. Sci. USA*, **88**, 10099.

Kirkpatrick, S., Gelatt, C. D. Jr & Vecchi, M. P. (1983). *Science*, **220**, 671–680.

Ladd, M. F. C. & Palmer, R. A. (1977). *Structure Determination by X-ray Crystallography*. New York: Plenum Press.

Liu, X. & Su, W. P. (2000). *Acta Cryst.* A**56**, 525–528.

Metropolis, N., Rosenbluth, M., Rosenbluth, A., Teller, A. & Teller, E. (1953). *J. Chem. Phys.* **21**, 1087–1092.

Pletenev, V. Z., Ivanov, V. T., Langs, D. A., Strong, P. & Duax, W. L. (1992). *Biopolymers*, **32**, 819–827.

Su, W. P. (1995*a*). *Acta Cryst.* A**51**, 845–849.

Su, W. P. (1995*b*). *Physica (Utrecht)*, A**221**, 193–201.

Weeks, C. M., DeTitta, G. T., Hauptman, H. A., Thuman, P. & Miller, R. (1994). *Acta Cryst.* A**50**, 210–220.

Weeks, C. M. & Miller, R. (1999). *Acta Cryst.* D**55**, 492–500.

Whyte, L. L. (1952). *Am. Math. Mthly*, **59**, 606–611.

Wille, L.T. (1986). *Nature (London)*, **324**, 46–48.

Xu, H., Weeks, C. M., Deacon, A. M., Miller, R. & Hauptman, H. (2000). *Acta Cryst.* A**56**, 112–118.