Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 30 March 2000
Accepted 22 June 2000

Resolution of the phase-ambiguity problem in the centrosymmetric $P\bar{1}$ space group by Monte Carlo methods

Marc Delarue

Unité de Biochimie Structurale, Institut Pasteur, 25 rue du Dr Roux, 75015 Paris, France.
Correspondence e-mail: delarue@pasteur.fr

Simulated-annealing methods have been used to resolve the phase-ambiguity problem in the centrosymmetric $P\bar{1}$ space group. First, an energy function based on the Sayre equation is introduced and a formal comparison with classical spin systems is drawn. The energy landscape is studied in detail and the validity of several energy criteria thoroughly tested. Classical Monte Carlo methods proved to be successful using a multistart optimization of the Sayre score, along with the additional monitoring of other energetic criteria. These involved the Terwilliger map quality index in reciprocal space in the absence of envelope information, and an envelope score if the shape of the molecule is known. The inherent phase-ambiguity problem of the $P\bar{1}$ space group was therefore technically solved by Monte Carlo methods. The method should also work to resolve phase ambiguity in the SIR method of protein crystallography.

1. Introduction

In protein crystallography and X-ray structure determination, most of the phase information is still obtained experimentally, using trial-and-error heavy-atom soaking of the crystals followed by localization of the binding sites through Patterson methods.

To resolve the phase ambiguity of the Harker construct, a minimum of two independent heavy-atom-substituted crystals is needed (Blundell & Johnson, 1976). If only one heavy-atom derivative is available, structure determination is made difficult by the inherent equivalence of the two phases given by the Harker construct. The latter case is usually referred to as the SIR method (single isomorphous replacement).

The classical probabilistic treatment of this problem, as originally proposed by Blow & Crick (1959), is to use figure-of-merit (FOM) weighted structure factors and centroid phases to calculate maps and try to distinguish an envelope as well as protein features such as helices and β-sheets in them. The centroid phase might seem at first glance an odd choice, since it can be as far as 90° away from the right phase, but the figure of merit will weight down precisely those reflections for which there is a large difference between the centroid phase and any of the two possible ones.

However, it would be interesting to have a method to pick up for each reflection the right phase among the two possible ones, using some optimization method to impose physical characteristics that are known to be satisfied for real maps of proteins. Exploring all the possibilities is out of question since this would involve 2^N trials and map evaluations, where N is the number of reflections.

Since there is an analogy between the phase-ambiguity problem in the SIR method and spin systems in condensed-matter physics, and since this last problem is one of the most widely studied problems in numerical simulations of systems of interacting particles, it might seem a good idea to try methods that proved successful in this last case and apply them to the crystallographic problem. Continuing progress in the field in the past few years due to both increase in computing power and refinement of theoretical methodologies makes this possibility even more attractive (Newman & Barkema, 1999).

The formal analogy of some problems in phase determination and the Ising spin system or even the spin-glass problem (Venkatesan, 1991) is not a new idea. Monte Carlo methods have been used in the past by several authors to improve phases (Bhat, 1990; Sheldrick, 1990). The advantage of working in reciprocal space was also underlined in the recent work of Beran & Szöke (1995). The connection with statistical thermodynamics was made some years ago and used with success for small-molecule crystallography (Khachaturyan *et al.*, 1981; Semenovskaya *et al.*, 1985). In a separate paper, the analogy with the spin glass problem is drawn further by using the methods of statistical thermodynamics, especially mean-field theory (Delarue & Orland, 2000). Here, we present a successful practical implementation of Monte Carlo methods to solve the phase problem in a specific case of protein crystallography, with focus on the characterization of

the energy landscape. Special attention is devoted to testing recently developed variants of the simulated-annealing protocol to resolve phase ambiguity in the $P\bar{1}$ centrosymmetric space group.

The main difficulty encountered in this work is the definition of an effective energy function to be minimized. Following others (Weinzierl *et al.*, 1969; Coulter, 1971; Hendrickson, 1971; Main, 1990; Giacovazzo *et al.*, 1994; Mukherjee & Woolfson, 1995), the work described here originally focused on the application of the Sayre equation (Sayre, 1952). This is very similar in spirit to the recent work of Chen & Su (2000), who employed the Sayre equation and simulated-annealing methods to solve the phase problem for small-molecule crystals (up to 126 non-H atoms). Here, it is performed on a small protein, rubredoxin; it soon appeared necessary to go beyond the Sayre equation and investigate other energy terms, expressed in reciprocal space, that reflect different physical criteria characterizing real electron-density maps.

2. Theory and methods

2.1. Choice of the space group

Instead of working with a real SIR case in a specific space group, the centrosymmetric space group $P\bar{1}$ was chosen, where all structure factors are real, the phases being restricted to take one of the two values 0 and 180°. The analogy with spin systems is then even more striking. Furthermore, this system is far less unnatural than it may seem at first glance; indeed, Berg and colleagues were able a few years ago to crystallize an equal mixture of a protein and its enantiomer in space group $P\bar{1}$ (Berg & Goffeney, 1997); direct methods failed to find the solution of this crystal structure and molecular replacement was used to locate the natural enantiomer in the cell (Zawadzke & Berg, 1993). The protein used by these authors was rubredoxin, a small protein whose unnatural enantiomer was chemically synthesized. The same protein was used here but with calculated structure factors.

2.2. Protein and crystallographic data

The crystallographic coordinates of rubredoxin were taken from the PDB (code 6RXN). The molecule was placed in a cubic $P\bar{1}$ cell ($a = b = c = 45$ Å) and care was taken that the packing was correct, *i.e.* that no crystallographically equivalent molecule would bump into any other molecule in the cell.

The structure factors were calculated from the atomic coordinates using the CCP4 suite of programs (Collaborative Computational Project, Number 4, 1994). The Zn^{2+} metal ion was omitted to avoid any strong bias in the Patterson map and also to meet the condition of equal atoms in the unit cell, which is at the basis of the Sayre equation. For all applications described in this article, the resolution was limited to 2.5 Å. The reason why this resolution was chosen, even though the atomicity condition is not expected to hold at this resolution, is that it is close to the resolution usually available for crystals of biological macromolecules; also, Zhang & Main (1990) have shown that the Sayre equation, when used in conjunction with other density-modification techniques such as histogram matching and solvent flattening, can actually help in phase refinement and extension even at low resolution (3 Å).

In some cases, a random error up to 30% was intentionally added to the calculated structure factors to simulate experimental errors. In this case, the structure-factor amplitudes $F(\mathbf{h})$ were replaced by $F(\mathbf{h})[1 + \text{eps rand(iseed)}]$, where rand(iseed) is a random number uniformally distributed between −1 and 1 and eps is the amplitude of the noise (10, 20, 30%, ...).

2.3. Definition of the energy to minimize

The Sayre equation is the reciprocal-space equivalent of a simple relationship in real space, $\rho(r) \propto \rho^2(r)$, which is valid for sharply peaked electron-density maps (the so-called atomicity condition). The Sayre equation reads (Sayre, 1952)

$$\mathbf{F}(\mathbf{h}) = g(\mathbf{h}) \sum_{\mathbf{k}} \mathbf{F}(\mathbf{k})\mathbf{F}(\mathbf{h} - \mathbf{k}), \tag{1}$$

where $g(\mathbf{h})$ is a resolution-dependent form factor.

Let us call $\mathbf{F}_S(\mathbf{h})$ the right-hand side of this equation, ignoring the $g(\mathbf{h})$ form factor.

$$\mathbf{F}_S(\mathbf{h}) = \sum_{\mathbf{k}} \mathbf{F}(\mathbf{k})\mathbf{F}(\mathbf{h} - \mathbf{k}), \quad i.e.\ \mathbf{F}(\mathbf{h}) = g(\mathbf{h})\mathbf{F}_S(\mathbf{h}). \tag{2}$$

The Sayre equation states that these Sayre structure factors $\mathbf{F}_S(\mathbf{h})$ should scale well with the original $\mathbf{F}(\mathbf{h})$ structure factors. To alleviate the problem of calculating explicitly the (resolution-dependent) $g(\mathbf{h})$ scaling factor, which is impossible to estimate in an analytical way at low resolution, a correlation coefficient between $\mathbf{F}(\mathbf{h})$ and $\mathbf{F}_S(\mathbf{h})$ is used. It is true that a (linear) correlation coefficient will not tackle well rapidly varying $g(\mathbf{h})$ functions but this procedure was found good enough for the purpose of this work. In particular, it was found that the correlation coefficient increases in a monotonous way for decreasing phase errors. The simplest way to derive an energy from a correlation coefficient is to define W_{Sayre} such that

$$W_{\text{Sayre}} = 1 - \text{Corr}(\mathbf{F}(\mathbf{h}), \mathbf{F}_S(\mathbf{h})), \tag{3}$$

where $\text{Corr}(A, B)$ stands for the correlation coefficient between the two quantities in parentheses (A and B) and has its usual meaning:

$$\text{Corr}(A, B) = (\langle AB\rangle - \langle A\rangle\langle B\rangle) \times (\langle A^2\rangle - \langle A\rangle^2)^{-1/2}(\langle B^2\rangle - \langle B\rangle^2)^{-1/2}. \tag{4}$$

This proved to be superior to a correlation coefficient calculated on the structure-factor modulus only:

$$W_{\text{Sayre}} = 1 - \text{Corr}(F(\mathbf{h}), F_S(\mathbf{h})). \tag{5}$$

It turns out that it is not necessary to calculate this correlation coefficient over the entire set of reflections but that a mere subset of the 1000 or so most intense ones are sufficient to obtain a very accurate Sayre score. To speed up the calculation of the energy W, the list of reflections $\mathbf{k}$ ($\mathbf{h}$) contributing to the

summation in (1) is stored once and for all for each reflection **h**.

In $P\bar{1}$, all the structure factors are real, therefore they can be written

$$\mathbf{F}(\mathbf{h}) = F(\mathbf{h})s(\mathbf{h}),$$

where $s(\mathbf{h})$ is the sign of reflection **h**.

Substituting in (1), one gets

$$s(\mathbf{h})F(\mathbf{h}) = g(\mathbf{h})\sum_{\mathbf{k}} F(\mathbf{k})F(\mathbf{h}-\mathbf{k})s(\mathbf{k})s(\mathbf{h}-\mathbf{k}) \quad (6)$$

or

$$s(\mathbf{h}) \propto \sum_{\mathbf{k}} J(\mathbf{k}, \mathbf{h}-\mathbf{k})s(\mathbf{k})s(\mathbf{h}-\mathbf{k}), \quad (7)$$

where $J(\mathbf{k}, \mathbf{h}-\mathbf{k}) = F(\mathbf{k})F(\mathbf{h}-\mathbf{k})$ can be seen as a coupling factor between reflections **k** and $\mathbf{h}-\mathbf{k}$, *i.e.* the strength of their interaction. In this form, substituting $s(\mathbf{h})$ by the magnetization of a spin located at site **h** of a lattice, it can be seen that the sign-ambiguity problem in crystallography bears strong resemblance to problems of interacting spins in condensed-matter physics, even though the energy defined in (5) is more intricate than the one encountered in the classical Ising spin problem for example.

It is also possible to express in reciprocal space the *a priori* knowledge of the form of the electron density outside the protein, which should be constant and equal to zero. One way to express it is to impose $\rho(\mathbf{r}) = \rho(\mathbf{r})\eta(\mathbf{r})$, where $\eta(\mathbf{r})$ is the characteristic function of the envelope of the molecule. If $\mathbf{G}(\mathbf{k})$ is the Fourier transform of $\eta(\mathbf{r})$, one can define another energy function, in reciprocal space:

$$W_{\rm Env} = 1 - {\rm Corr}(\mathbf{F}(\mathbf{h}), \mathbf{F}_{\rm env}(\mathbf{h})), \quad (8)$$

with

$$\mathbf{F}_{\rm env}(\mathbf{h}) = \sum_{\mathbf{k}} \mathbf{F}(\mathbf{k})\mathbf{G}(\mathbf{h}-\mathbf{k}).$$

Hence, one can try to minimize a combination of the two, controlled by the mixing parameter μ:

$$W_{\rm tot} = \mu W_{\rm Sayre} + (1-\mu)W_{\rm env}. \quad (9)$$

For the sake of completeness, it is in order to note here that the problem of using information from real-space constraints to break down the phase ambiguity in the SIR method has been re-examined recently by Gu *et al.* (1997) and Zheng *et al.* (1997).

Other types of energy were considered, mainly involving nearest-neighbour-position correlations in real space, but expressed in reciprocal space (see Bruinsma, 1988), but they were found to carry no phase constraint, all the information being contained in Patterson maps.

2.4. Terwilliger electron-density-map criterion, expressed in reciprocal space

The local roughness of the electron-density map is a good criterion to distinguish between protein and solvent regions. It can be used in real space to determine the limits of the molecule (Rees *et al.*, 1990; Jones *et al.*, 1991) and evaluate the quality of an experimental MIR map (Terwilliger & Berendzen, 1999). Its variance over the entire unit cell, σ_R^2, can be used to distinguish between good and poor maps; it can be expressed in reciprocal space *via* the formula (Terwilliger, 1999)

$$\sigma_R^2 = \sum_{\mathbf{h}\neq 0} |\mathbf{R}(\mathbf{h})|^2, \quad (10)$$

where the original notations were followed:

$$\mathbf{R}(\mathbf{h}) = \mathbf{B}(\mathbf{h})\exp[-2\pi^2\sigma^2 S(\mathbf{h})^2] - \mathbf{B}(\mathbf{h})^{\rm AVG} \quad (11)$$

with

$$\mathbf{B}(\mathbf{h}) = \sum_{\mathbf{k}} \mathbf{F}(\mathbf{k})\mathbf{F}(\mathbf{h}-\mathbf{k}) \quad (12)$$

$$\mathbf{B}(\mathbf{h})^{\rm AVG} = \sum_{\mathbf{k}} \mathbf{Q}(\mathbf{k})\mathbf{Q}(\mathbf{h}-\mathbf{k}). \quad (13)$$

In this last equation, $\mathbf{Q}(\mathbf{h})$ is defined by

$$\mathbf{Q}(\mathbf{h}) = \mathbf{F}(\mathbf{h})\exp[-2\pi^2\sigma^2 S(\mathbf{h})^2], \quad (14)$$

where

$S(\mathbf{h})$ is the inverse of the resolution and $\sigma = 2.5$ Å. (15)

2.5. Monte Carlo and simulated-annealing methods

Each time a simulation was run, the initial configuration of the system $\{s(\mathbf{h})\}$ was chosen randomly, *i.e.* a random number (either +1 or −1) was drawn for each reflection.

Monte Carlo simulations were performed in the usual way: a reflection was picked at random, then the energetic cost ΔE of flipping its sign was calculated and the move was accepted if ΔE was negative or if it satisfied the so-called Metropolis recipe (Metropolis *et al.*, 1953). Unless otherwise stated, the random generator was ran2 of *Numerical Recipes* (Press *et al.*, 1992). In some cases, other random generators were used, with similar results.

Care was taken to minimize the number of calculations involved in the evaluation of ΔE; this allowed for about 3×10E+5 sweeps of the spin system (usually 1000 reflections or so) in 12 h of CPU on a DEC Alpha PWS500 workstation.

Simulated annealing was performed (Kirkpatrick *et al.*, 1983) with an exponential cooling protocol. The energy fluctuations were monitored by plotting the specific heat, which reads:

$$C_v(T) = (\langle E^2\rangle - \langle E\rangle^2)/k_B T^2. \quad (16)$$

This quantity is useful because it helps to locate where the phase transition (if any) takes place (Kirkpatrick *et al.*, 1983). In our case, 'phase transition' means transition from a random phase set to the closest available phase set minimizing the energy; there is no guarantee that the minimum is the global one and not a local one. The reason why the $C_v(T)$ reaches a maximum is the following: intuitively, one can see that at low temperature there is no fluctuation any more, the system is frozen and the numerator ensures that C_v reaches zero. At high temperatures, there are large fluctuations but the

denominator ensures that C_v is also tending to zero. In between, it reaches a maximum at T_c, where it is recommended to decrease the temperature slowly (Newman & Barkema, 1999).

In terms of CPU, 250 temperature steps could be performed in 4 h CPU, with 400 sweeps of the 1000 reflections system at each step, a grand total of 10E+8 attempts to change phases. A typical 2% decrease of the temperature at each step was applied, spanning roughly two orders of magnitude, from 0.0003 to 0.000003. This range was determined by looking at the $C_v(T)$ curve in a quick preliminary simulation using a larger rate of temperature decrease, just to spot the transition temperature T_c.

To get an idea of the number of cycles to do at each temperature, a longer simulation was performed to calculate the auto-correlation function $\chi(t)$ and estimate its characteristic decay time τ (Newman & Barkema, 1999) at different temperatures.

$$\chi(t)/\beta = \int \mathrm{d}t' \, [m(t') - \langle m(t')\rangle][m(t+t') - \langle m(t+t')\rangle], \quad (17)$$

where $\langle m(t)\rangle$ is the mean sign of the reflection, per reflection, at time t of the simulation.

The correlation time τ was estimated by fitting the $\chi(t)$ function to the theoretical curve

$$\chi(t) = \chi(0)\exp(-t/\tau). \quad (18)$$

In some test cases designed to study the energy landscape and check out the algorithm, a certain percentage of phases was imposed to their 'native' values. Those reflections whose phase (sign) were imposed and set to their native value for the rest of the simulation were chosen randomly. The percentage of imposed reflections refers to the fraction of the working set (usually 1000 reflections) being unrefined and correct.

Unless otherwise stated, three phase reflections were always imposed to remove origin definition problems.

2.6. Extension to *P*1 and other space groups

It is also possible to perform Monte Carlo simulations in any space group, sampling the phases at 45, 135, 225 and 315°. In this case, the most efficient method is the so-called rejectionless Monte Carlo method, whereby at each step the four energies $E(\varphi = 45°)$, $E(\varphi = 135°)$, $E(\varphi = 225°)$ and $E(\varphi = 315°)$ are evaluated for the reflection under consideration and their Boltzmann factors calculated. Their sum is normalized to one and the resulting normalized weights are used to choose just one of them according to the outcome of a random generator. This method is also known as the Gibbs sampling method (Newman & Barkema, 1999).

3. Results

3.1. Validity of the energy function

First, the validity and consistency of the energy function have to be established. In order to do so, the energy defined in (5), $W_{\rm Sayre}$, was calculated for different percentages of imposed phases. As expected, it was found that the energy is decreasing with an increasing percentage of correct reflections (Fig. 1). Evidently, this is a necessary but not sufficient condition if one wants to use powerful minimization techniques. Next, the amount of computing time needed to evaluate the energy was reduced by working with the minimum number of reflections, while still giving a sufficiently accurate result. In practice, it is enough to work with the top 1000 most intense reflections (Fig. 2*a*). Alternative definitions of the energy function were also considered. Working with a correlation coefficient calculated on the modulus of structure factors instead of the full (signed) structure factors turned out to be less discriminative (Fig. 1). The influence of noise on the data was also investigated (Fig. 2*b*) and found to be small; below 20% for the noise amplitude.

As a complement, the same dependency on the percentage of imposed phases of two other energy criteria was investigated, namely the envelope score (Fig. 1), whose definition (8) in reciprocal space follows the same kind of formalism as the Sayre score, and the local roughness of the map, the Terwilliger score, σ_R^2, as defined recently (Terwilliger, 1999, data not shown). Both also satisfy the necessary conditions of being a monotonously decreasing function for more and more correct phase sets. They are also based on different physical properties of the electron-density maps and could therefore be used in combination with each other, in the hope of reducing the number of false (local) minima in the energy landscape.

However, the calculation of the envelope score necessitates the knowledge of the envelope characteristic function, while the Terwilliger criterion does not; the former is therefore a somewhat less artificial and more general score than the latter.

3.2. Simulated-annealing and related protocols

Simulated-annealing protocols were tried at different percentages of imposed phases as a way to understand the energy landscape of the system. It was found that relatively simple protocols were able to find the right solution [*i.e.* the right $\{s(\mathbf{h})\}$ phase ensemble] for percentages of imposed

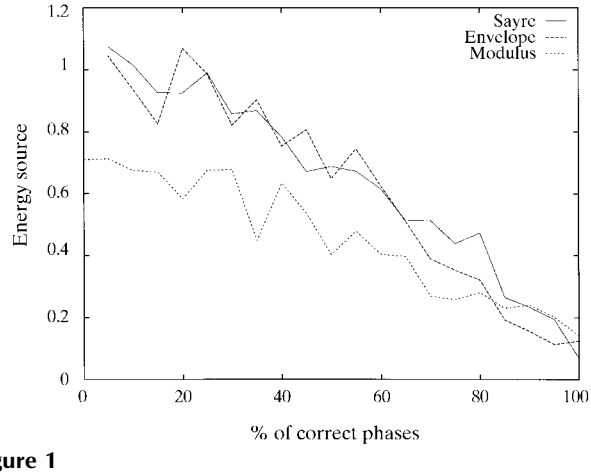

Figure 1
Influence of the percentage of imposed phases on the Sayre score, as calculated from equation (3) (modulus and phase) and as calculated from equation (5) (modulus only); the envelope score defined in equation (8) is also shown for comparison.

phases as low as 5% (see Fig. 3). The transition was easy to spot using the $C_v(T)$ curve (Kirkpatrick *et al.*, 1983). However, no slowing down of the correlation time τ as defined in (18) was observed around T_c. Convergence at 10% of imposed phases in the working set necessitated at least 400 sweeps at each temperature step and a decrease in temperature of 0.98 at each of the simulated-annealing steps. Using this procedure at 5% of imposed phases in the working set, convergence was sometimes (but not always) achieved, depending on the initial configuration. At 2% of imposed phases, longer and more sophisticated cooling protocols had to be used, with only limited success: for instance, 500 temperature cycles with a decreasing rate of 0.99 and 1000 sweeps each time, with a grand total of 5.0 × 10E+8 attempts to switch phases, lead to convergence only once. Because this procedure is more CPU time consuming, only a limited number of starting configurations could be tried.

The influence of noise on the structure factors was also investigated, at 10% of imposed phases, with the amplitude of noise varying from 0 to 20%, in steps of 5%, all simulations converged to the right solution (data not shown).

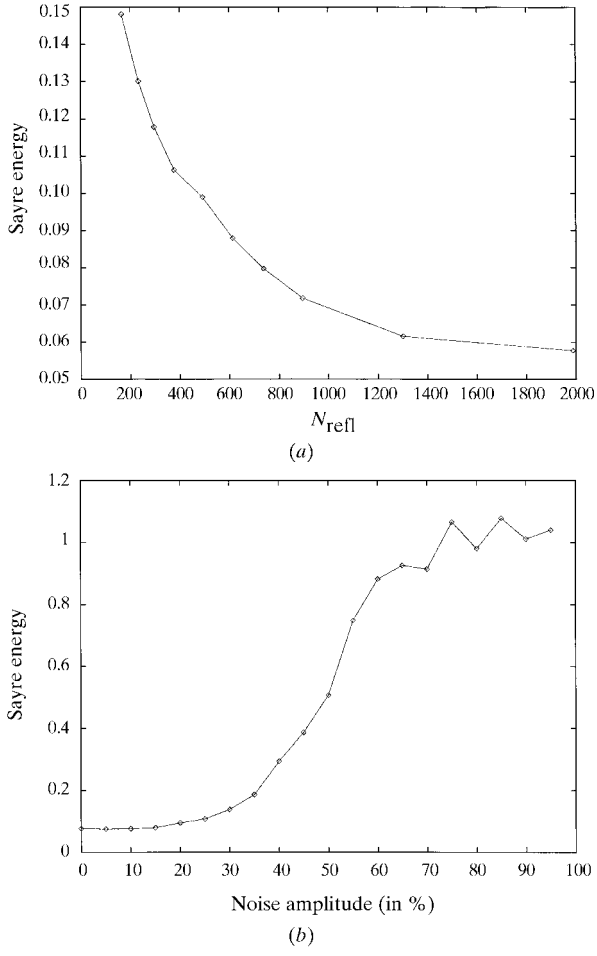

Figure 2
(*a*) Influence of the number of the most intense reflections $N_{\rm refl}$ included in the summation involved in equations (1) and (3) on the Sayre score. For most applications in this paper, $N_{\rm refl}$ was set to 1000. (*b*) Influence of the noise amplitude (in %, *x* axis) on the Sayre score (*y* axis).

A number of different simulated-annealing-related protocols (see Berne & Straub, 1997) were then implemented and tested, at lower percentages of imposed phases.

Among them, the so-called 'threshold accepting' algorithm was at least as successful as simulated annealing. In this method, a move is accepted depending on the sign of the difference between ΔE and an 'accepting threshold', which is gradually decreased in the simulation (Berg, 1993, and references therein). The initial value of the 'accepting threshold' has to be adjusted by trial and error.

A variant of the Metropolis criterion, based on the Tsallis entropy, was also tried. It is supposed to allow for longer steps to be taken in the energy landscape, as in Levy flights (Andricioaei & Straub, 1996).

Another variant is to do a normal simulated-annealing simulation, to locate the transition temperature using the $C_v(T)$ curve and to reiterate the cooling procedure at a lower rate (say half the original rate) several times, starting just below T_c, for instance at $0.8T_c$; this is called 'simulated bouncing' (Schneider *et al.*, 1998). To locate the transition temperature in noisy $C_v(T)$ curves, a non-linear fit (Press *et al.*, 1992) with a sum of two exponentials was performed and gave good results.

Finally, a procedure called simulated tempering was also tried, where the system is free to adjust its temperature by choosing among a discret set of possible values (usually around 20) so that each of them is sampled more or less uniformly in the simulation. This procedure is more lengthy since it is necessary to run the simulation at least twice. The first run is used just to determine the weights for each temperature of the ensemble, using histograms of visits; these weights are determined recursively (Marinari & Parisi, 1992; Hansmann & Okamoto, 1998).

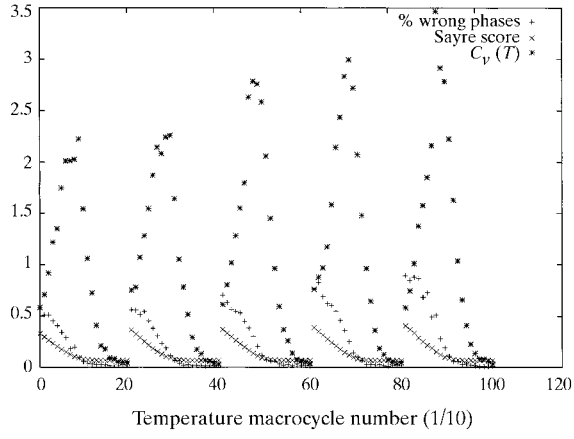

Figure 3
Results of several simulated-annealing simulations at different percentages of imposed phases. From left to right: 25, 20, 15, 10, 5% of imposed phases. The horizontal axis represents the temperature cycle number multiplied by 10, with 400 sweeps of the phases being performed at each temperature step. Each simulation includes 210 temperature cycles. The Sayre score is the quantity being minimized, while the percentage of correct phases is monitored as well as the specific heat $C_v(T) = (\langle E^2\rangle - \langle E\rangle^2)/k_BT^2$.

In some cases, threshold accepting gave better results than plain simulated annealing, especially in 2% of imposed phases, but in no case was any algorithm capable of finding the right solution at 0% of imposed phases. Therefore, it appears that, in the absence of any phase restraint, the Sayre score is not enough to distinguish the right solution from an ensemble of other solutions with scores as low as the correct solution, but completely wrong. This is not entirely unexpected on the basis of a recent detailed theoretical study aimed at characterizing the energy landscape of popular energy terms for phase improvement and/or refinement (Baker *et al.*, 1993).

To further understand the energy landscape, the scores of flipping one phase at a time were calculated, starting from the native configuration. This provides a picture of the neighbourhood of the right solution in the energy landscape. While most of the flips did not vary the energy very much, it was noticed that a couple of them have a disastrous effect on the score; this can be seen in the histogram presented in Fig. 4(*a*). Reciprocally, this means that the system can have a score quite far away from the minimum score and be actually only one flip away from it. Therefore, the correlation between the difference in scores of two configurations and their Hamming distance is far from being perfect. This is illustrated in Fig. 4(*b*).

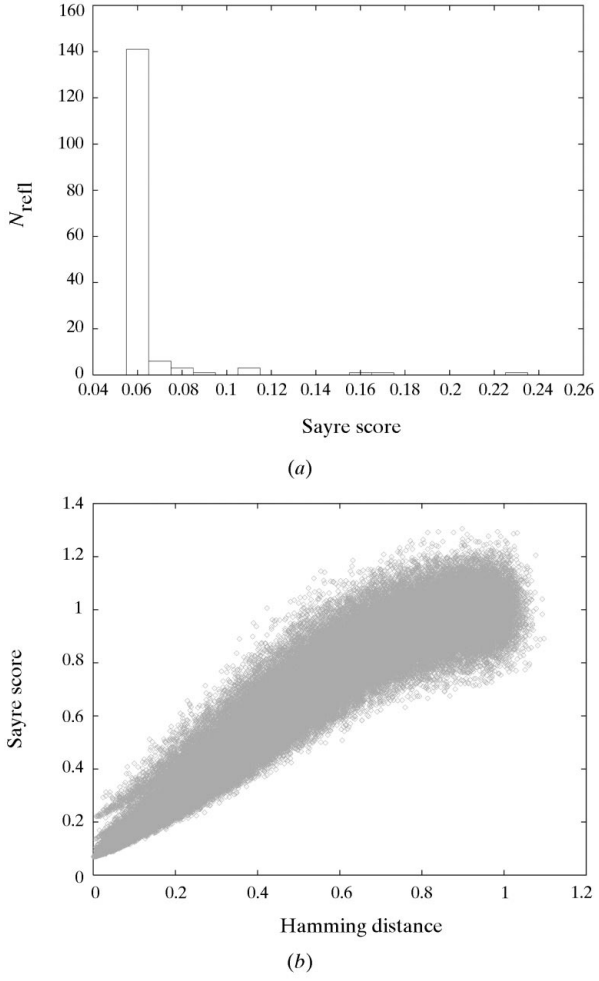

Figure 4
Characterization of the energy landscape and phase relationships. (*a*) Histogram of the number of reflections having a given Sayre score after the flip of just one phase, starting from the minimal conformation; there is one reflection whose flip brings the Sayre score from 0.0575 to 0.235. The mean value of the new score for one (and only one) wrong phase sign is 0.0590, with a r.m.s.d. of 0.0063. $N_{\rm refl}$ was set to 2500. (*b*) Correlation between the Sayre score and the percentage of wrong phases in randomly generated different phase-set configurations, expressed as the Hamming distance from the true solution.

3.3. Multistart Monte Carlo simulations

It was noticed, in the course of the numerous different simulated-annealing protocols that were tried at low percentages of imposed phases (*i.e.* below 5%), that convergence depended a lot on the initial configuration (which is chosen randomly, in the limit of imposed phases). It was therefore decided to try multistart simulations, as this is also commonplace in direct-methods strategies (*e.g.* Debaerdemaeker *et al.*, 1988; Mukherjee, 1999). In practice, a rather crude simulated-annealing protocol was chosen, in exchange for a large number of trial initial configurations. As judged both through the envelope score and the percentage of correct phases, this strategy proved successful, with a rate equal to 2/23 (Fig. 5*a*).

Since both these criteria require some *a priori* knowledge of the right solution, there is a need to discriminate between right and wrong configurations in the absence of any other information. One possibility is to express in reciprocal space an index measuring fluctuations in the actual electron-density map corresponding to the phase set being tested. This criterion has been recently expressed in reciprocal space by T. C. Terwilliger and has been implemented here (see *Theory and methods*). It readily identifies the correct solution (Fig. 5*b*), in the sense that the correct solution is the one corresponding to the absolute minimum of this energy function, among all the different simulations with different starting configurations.

4. Discussion

We have shown here that it is possible to resolve the phase-ambiguity problem in the $P\bar{1}$ space group by applying a widely used numerical technique in condensed-matter physics, namely simulated annealing in the multistart mode.

The main difficulty encountered in this work is the definition of an energy function for which the right solution is the absolute minimum, with an energy spectrum such that the 'native' configuration is well detached from all the other (wrong) ones. Such a criterion does not exist to our knowledge, meaning that it would be meaningless to refine one trial configuration against any existing energy function (see Baker *et al.*, 1993). The only thing that can be performed therefore is to minimize one physically sound energy, while monitoring another (independent) one allowing one to distinguish whether or not the converged solution is correct. In other words, optimization of one criterion provides 'admissible' maps, among which the right one ought to be selected out by another criterion. The Terwilliger σ_R^2 index proved successful

to carry out this program and the phase ambiguity was indeed broken in our test case.

Knowledge of the shape of the molecule is a plus since it allows for the use of an envelope score, which also readily identifies the right set of phases. Even though this kind of information is not always available, there is hope that it will be more so in the near future because significant progress has been recently accomplished in the field of low-resolution three-dimensional structure reconstruction of biological macromolecules from experimental SAXS data in solution (Chacon *et al.*, 1998; Svergun, 1999; Walther *et al.*, 2000). However, even if the shape of the envelope is known, the problem remains to locate it in the unit cell; this amounts to a low-resolution molecular replacement problem, which could be solved using standard techniques.

In space group $P\bar{1}$ (as in any other space group for that matter), the set of imposed phases could come from experimentally determined phases for some reflections (Weckert & Hummer, 1997; Shen, 1998). Simulations along these lines, where the phases of a small subset of reflections are imposed, have already been reported for protein crystallography (Mo *et al.*, 1996). It is not unreasonable to foresee that these methods will be more widely used in the near future.

Another promising avenue of research would be to impose connectivity in the map, as originally suggested by Baker *et al.* (1993). Lunin *et al.* (1999) recently reported some work in this direction, albeit at low resolution only.

Finally, the method has been applied to break down phase ambiguity in the SIR method of protein crystallography. Preliminary tests in space group $P2_12_12_1$ showed that the correct phases are quickly recovered, with the centric reflections playing the role of reflections with imposed phases.

It is a pleasure to thank P. Koehl, T. Garel and H. Orland as well as S. Doniach for helpful and stimulating discussions. Thanks also are due to P. Dumas for careful reading of the manuscript.

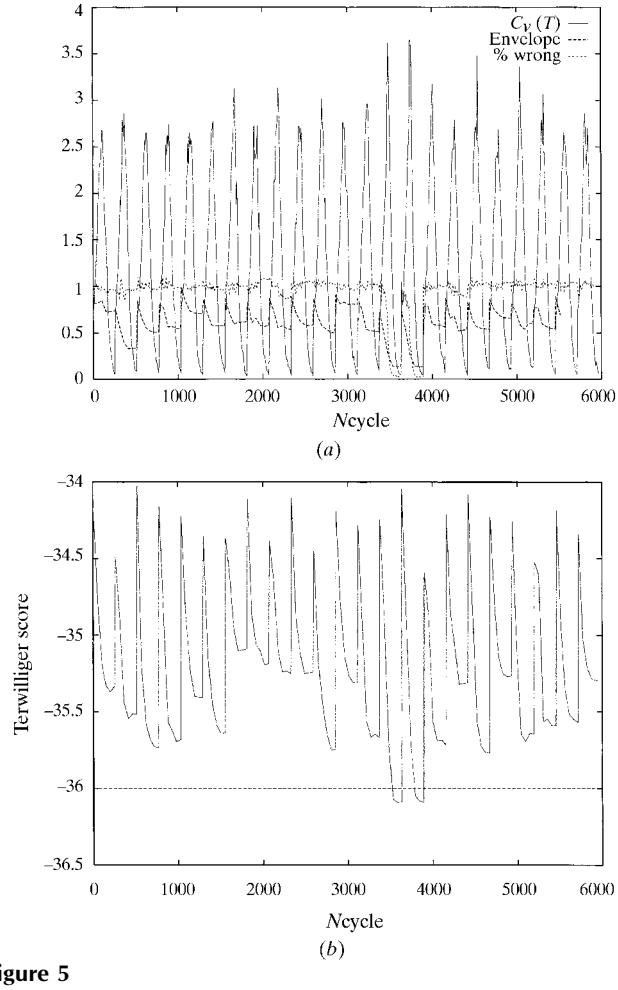

Figure 5
(*a*) Multistart simulated annealing at 0% of imposed phases. The minimized Sayre score, the specific heat curve $C_v(T)$ as well as the percentage of correct phases are indicated. (*b*) Same as (*a*) but monitoring the Terwilliger reciprocal-space map quality index σ_R^2. 23 different optimizations with different starting points are represented, with just 2 of them being successful.

References

Andricioaei, I. & Straub, J. E. (1996). *Phys. Rev. E*, **53**, R3055–R3058.
Baker, D., Krukowski, A. E. & Agard, D. A. (1993). *Acta Cryst.* D**49**, 186–192.
Beran, P. & Szöke, A. (1995). *Acta Cryst.* A**51**, 20–27.
Berg, B. A. (1993). *Nature (London)*, **361**, 708–710.
Berg, J. M. & Goffeney, N. W. (1997). *Methods Enzymol.* **276**, 619–627.
Berne, B. & Straub, J. E. (1997). *Curr. Opin. Struct. Biol.* **7**, 181–189.
Bhat, T. N. (1990). *Acta Cryst.* A**46**, 735–742.
Blow, D. M. & Crick, F. H. C. (1959). *Acta Cryst.* **12**, 794–802.
Blundell, T. & Johnson, L. (1976). *Protein Crystallography.* New York: Academic Press.
Bruinsma, R. (1988). *Phys. Rev. Lett.* **61**, 1966–1970.
Chacon, P., Moran, F., Diaz, J. F., Pantos, E. & Andreu, J. M. (1998). *Biophys. J.* **74**, 2760–2775.
Chen, Y. & Su, W. P. (2000). *Acta Cryst.* A**56**, 127–131.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Coulter, C. L. (1971). *Acta Cryst.* B**27**, 1730–1740.
Debaerdemaeker, T., Germain, G., Main, P., Refaat, L. S., Tate, C. & Woolfson, M. M. (1988). *MULTAN88. A System of Computer Programs for the Automatic Solution of Crystal Structures from X-ray Diffraction Data.* Universities of York, England, and Louvain, Belgium.
Delarue, M. & Orland, H. (2000). *Acta Cryst.* A**56**, 562–574.
Giacovazzo, C., Siliqi, D. & Ralph, A. (1994). *Acta Cryst.* A**50**, 503–510.
Gu, Y. X., Zheng, C. D., Zhao, Y. D., Ke, H. M. & Fan, H. F. (1997). *Acta Cryst.* D**53**, 792–794.
Hansmann, U. E. & Okamoto, Y. (1998). *J. Comput. Chem.* **18**, 920–933.
Hendrickson, W. A. (1971). *Acta Cryst.* B**27**, 1474–1475.
Jones, E. Y., Walker, N. P. C. & Stuart, D. I. (1991). *Acta Cryst.* A**47**, 753–770.
Khachaturyan, A., Semenovskaya, S. & Vainshtein, B. (1981). *Acta Cryst.* A**37**, 742–754.
Kirkpatrick, S., Gellatt, C. D. Jr & Vecchi, M. P. (1983). *Science*, **220**, 671–680.
Lunin, V. Y., Lunina, N. L. & Urzhumstev, A. G. (1999). *Acta Cryst.* A**55**, 916–925.
Main, P. (1990). *Acta Cryst.* A**46**, 372–377.

Marinari, E. & Parisi, G. (1992). *Europhys. Lett.* **19**, 451–458.
Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. & Teller, E. (1953). *J. Chem. Phys.* **21**, 1087–1092.
Mo, F., Mathiesen, R. H., Hauback, B. C. & Adman, E. T. (1996). *Acta Cryst.* D**52**, 893–900.
Mukherjee, M. (1999). *Acta Cryst.* D**55**, 820–855.
Mukherjee, M. & Woolfson, M. M. (1995). *Acta Cryst.* D**51**, 626–628.
Newman, M. E. J. & Barkema, G. T. (1999). *Monte Carlo Methods in Statistical Physics.* Oxford University Press.
Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. (1992). *Numerical Recipes: the Art of Scientific Computing*, 2nd ed. Cambridge University Press.
Rees, B., Bilwes, A., Samama, J.-P. & Moras, D. (1990). *J. Mol. Biol.* **214**, 281–297.
Sayre, D. (1952). *Acta Cryst.* **5**, 60–65.
Schneider, J., Morgenstern, I. & Singer, J. M. (1998). *Phys. Rev. E*, **58**, 5085–5092.
Semenovskaya, S., Khachaturyan, K. A. & Khachaturyan, A. G. (1985). *Acta Cryst.* A**41**, 268–273.
Sheldrick, G. M. (1990). *Acta Cryst.* A**46**, 467–473.
Shen, Q. (1998). *Phys. Rev. Lett.* **80**, 3268–3272.
Svergun, A. (1999). *Biophys. J.* **76**, 2879–2886.
Terwilliger, T. C. (1999). *Acta Cryst.* D**55**, 1174–1178.
Terwilliger, T. C. & Berendzen, J. (1999). *Acta Cryst.* D**55**, 501–505.
Venkatesan, R. (1991). *Acta Cryst.* A**47**, 400–404.
Walther, D., Cohen, F. E. & Doniach, S. (2000). *J. Appl. Cryst.* **33**, 350–363.
Weckert, E. & Hummer, K. (1997). *Acta Cryst.* A**53**, 108–143.
Weinzierl, J. E., Eisenberg, D. & Dickerson, R. E. (1969). *Acta Cryst.* B**25**, 380–387.
Zawadzke, L. E. & Berg, J. E. (1993). *Proteins*, **16**, 301–305.
Zhang, K. Y. J. & Main, P. (1990). *Acta Cryst.* A**46**, 377–381.
Zheng, X. F., Zheng, C. D., Gu, Y. X., Fan, H. F. & Hao, Q. (1997). *Acta Cryst.* D**53**, 49–55.