Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 12 February 2002
Accepted 20 May 2002

A hybrid Monte Carlo method for crystal structure determination from powder diffraction data

John C. Johnston, William I. F. David, Anders J. Markvardsen* and Kenneth Shankland

ISIS Facility, Rutherford Appleton Laboratory, Chilton, Oxon OX11 0QX, England. Correspondence e-mail: a.j.markvardsen@rl.ac.uk

A hybrid Monte Carlo algorithm for crystal structure determination from powder diffraction data is presented. The algorithm combines the key components of molecular dynamics and Monte Carlo simulations to achieve efficient sampling of phase space, allowing the crystal structure of capsaicin to be determined from powder diffraction data more effectively than by a simulated-annealing approach. The implementation of the algorithm, the choice of the simulation parameters and the performance of the algorithm are discussed.

1. Introduction

Global optimization methods of crystal structure determination from powder diffraction data have found particular utility with molecular organic crystal structures, where the known chemical connectivity of the molecule under study can be easily converted into trial three-dimensional crystal structures. The molecule under study is first parameterized as a series of rigid units connected by variable torsion angles, a task conveniently achieved by the use of internal coordinates. Thereafter, the position, orientation and conformation of the molecule within the unit cell of the crystal structure are optimized against some observed data. It is assumed that the correct crystal structure corresponds to the global minimum of some function relating the trial crystal structure to the observed data.

In the majority of cases, the level of prior chemical knowledge is such that the input molecular model is accurate and the correctness of a trial crystal structure produced in the global optimization search can be assessed in a meaningful way. This assessment is normally performed by comparing observed and calculated diffraction data, using a least-squares figure of merit. That is, either as the weighted sum of squared deviations between the observed ($y_i^{\rm obs}$) and calculated (y_i) diffraction patterns using $\chi^2_{\rm profile} = \sum_i w(y_i^{\rm obs} - y_i)^2$ (Young, 1993) or as the weighted sum of squared deviations between observed and calculated integrated intensities of the diffraction pattern:

$$\chi^2 = \sum_h \sum_k [(I_h - c|F_h|^2)(V^{-1})_{hk}(I_k - c|F_k|^2)], \qquad (1)$$

where I_h and I_k are Lorentz–polarization-corrected extracted integrated intensities from a Pawley refinement (Pawley, 1981) of the diffraction pattern, V_{hk} is the covariance matrix from the Pawley refinement, c is a scale factor, and $|F_h|$ and $|F_k|$ are the structure-factor magnitudes calculated from the trial structure. Alternatively, the correlated integrated intensities may be obtained using the iterative Le Bail method (Le Bail *et al.*, 1988; Pagola *et al.*, 2000; David *et al.*, 2002). For cases where the input molecular model is a poor approximation to the contents of the crystal structure, a maximum-likelihood figure of merit has been found to be an effective alternative to the χ^2 figure of merit mentioned above (Markvardsen *et al.*, 2002).

A number of different global optimization strategies have been applied successfully to the problem of locating the global minimum in χ^2 space (for a summary, see David *et al.*, 2002). However, many other algorithms from different research areas (see, for example, Floudas *et al.*, 1999) remain to be evaluated in respect of this particular crystallographic problem. In this paper, we investigate a hybrid Monte Carlo (HMC) method that combines the best features of Monte Carlo (MC) simulations and molecular dynamics (MD) in a single algorithm. The HMC method was introduced for numerical simulation in lattice field theory (Duane *et al.*, 1987) and has become widely used for lattice quantum-chromodynamic computations with dynamical fermions (see, for instance, Joo *et al.*, 2000). HMC has been applied to a variety of different problems including the simulation of polymer chains (Irbäck, 1994) and the conformational analysis of RNA (Fischer *et al.*, 1999).

2. The hybrid Monte Carlo method

The starting point in picturing how the HMC method can be utilized as a global optimization tool is to consider a single postulated crystal structure as a hypothetical particle in a hyperspace defined by a set of structural parameters. In the case of a molecular crystal structure solution, these parameters are the six external degrees of freedom (position and orientation of the molecule) and the internal molecular conformational degrees of freedom associated with torsion angles whose values cannot be specified in advance. The χ^2 goodness-of-fit function, given by (1), which relates observed

and calculated diffraction data, is taken to be the potential energy that the hypothetical particle possesses. The particle is assigned a kinetic energy by randomly selecting the components of its momentum from a Gaussian thermal distribution at temperature T (defined in χ^2 units) and assigning these components to each of the N structural parameters. The particle moves over the χ^2 hypersurface, following a trajectory determined by its initial momentum and the gradient of the χ^2 hypersurface. As the total energy of the system must be conserved throughout a particular particle trajectory, promising structures (*i.e.* with low values of χ^2) therefore have low potential energy and high kinetic energy. It is this high kinetic energy component that allows the particle to move uphill and thus escape from local minima.

The HMC algorithm is expressed mathematically as follows. Choose a point, **r**, in the structure parameter space. The coordinates, r_i $(i = 1, \ldots, N)$, of the point correspond to the positional, orientational and variable torsional parameter values within the crystal structure and the momentum components are denoted $p_i = m_i v_i$ $(i = 1, \ldots, N)$. According to classical mechanics, the total energy, the Hamiltonian H (*i.e.* the sum of the kinetic energy, K, and the potential energy, U) is conserved. The Hamiltonian, H, at time t, is written as

$$H(t) = \tfrac{1}{2}\sum_{i=1}^{N} m_i v_i^2(t) + U(\mathbf{r}(t)). \quad (2)$$

In principle, the mass associated with each parameter may be different but in the present implementation each mass is set to unity such that the momentum is numerically identical to the velocity, $\mathbf{p} = \mathbf{v}$. The potential energy is given by χ^2 and so the energy of the hypothetical particle that travels over the χ^2 hypersurface can be rewritten as

$$H(t) = \tfrac{1}{2}\sum_{i=1}^{N} p_i^2(t) + \chi^2(\mathbf{r}(t)), \quad (3)$$

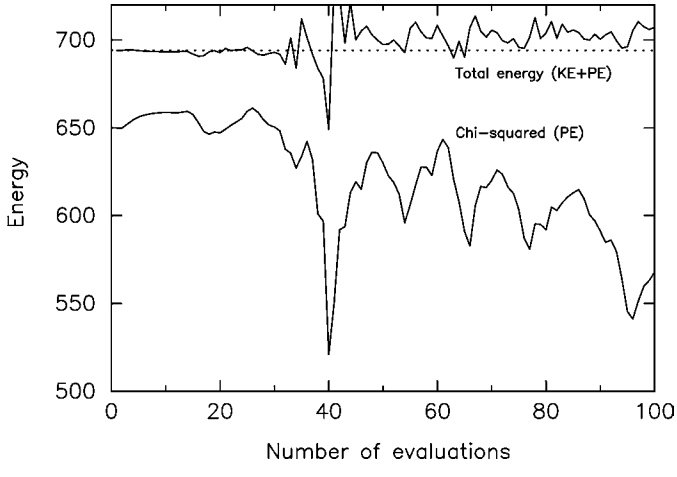

Figure 1
The potential energy (correlated integrated intensities χ^2) and total energy (kinetic energy plus potential energy) evaluated over a single MD trajectory during the crystal structure solution of capsaicin. The initial total energy is shown as a dotted line in order to highlight the total energy fluctuations arising from the finite MD step size.

where the Hamiltonian is expressed in terms of the position and momentum of the particle. Hamilton's equations of motion are given by

$$\partial r_i/\partial t = \partial H/\partial p_i = p_i, \quad i = 1, \ldots, N, \quad (4)$$

and

$$\partial p_i/\partial t = -\partial H/\partial r_i = -\partial \chi^2/\partial r_i, \quad i = 1, \ldots, N. \quad (5)$$

The initial momentum components, $p_i(t = 0)$, are sampled randomly from a Gaussian thermal distribution at temperature T, where T is expressed in χ^2 units, *i.e.* for each component,

$$\text{prob}(p_i) = \exp[-K(p_i)/T]/(2\pi)^{1/2}. \quad (6)$$

The temperature is very important for the effectiveness of the algorithm and its choice is discussed further in §6. The trajectory of the particle is calculated using a leap-frog algorithm (Hockney, 1970; Leach, 1996) where the position at one time step is used to calculate the momentum at the next and so on. The precise leap-frog algorithm used in this paper is

$$r_i(t + \Delta t) = r_i(t) + \Delta t\, p_i(t + \tfrac{1}{2}\Delta t) \quad (7)$$

$$p_i(t + \tfrac{3}{2}\Delta t) = p_i(t + \tfrac{1}{2}\Delta t) - \Delta t \left.\frac{\partial \chi^2}{\partial r_i}\right|_{(t+\Delta t)}. \quad (8)$$

The momentum components, $p_i(t = \frac{1}{2}\Delta t)$, are calculated from the initial Gaussian sample, $p_i(t = 0)$, using the equation

$$p_i(t = \tfrac{1}{2}\Delta t) = p_i(t = 0) - \tfrac{1}{2}\Delta t \left.\frac{\partial \chi^2}{\partial r_i}\right|_{(t=0)}. \quad (9)$$

Given that finite step sizes are used when calculating the trajectory across the parameter space, systematic errors inevitably occur and can accumulate within the system as the simulation progresses. Fig. 1 shows both the total energy and the potential energy throughout a single trajectory of 100 MD steps for the capsaicin example discussed in §4. As expected, the trajectory over the hypersurface involves numerous downhill and uphill moves. The total energy fluctuates considerably around the relatively deep local minimum found after 40 moves, highlighting the problems of performing MD with finite step sizes. With smaller steps, these fluctuations in the total energy would be reduced, but so would the overall distance travelled by the particle and thus the extent of the parameter space sampled would also be reduced, decreasing the efficiency of the search. The leap-frog algorithm is at its least exact when the potential energy (*i.e.* the χ^2) is changing most rapidly. If no corrections were applied, then the total energy would become increasingly incorrect and the sampling of the parameter space would not follow a true MD path. In the HMC approach, this problem is dealt with by comparing the initial and final total energy after a number of MD steps and insisting upon detailed balance with respect to sampling of a canonical ensemble. Thus, after a specified number of MD steps, the trajectory is accepted if the final total energy, E_M, is lower than the initial total energy, E_0. If E_M is higher than E_0 (as is the case for the example shown in Fig. 1), then the trajectory is accepted with a Boltzmann probability,

$$\text{prob(accept)} = \exp[-(E_M - E_0)/T]. \quad (10)$$

If a trajectory is accepted, then new momentum components are chosen randomly from a Gaussian thermal distribution and the next trajectory begins at the endpoint of the accepted trajectory. If the trajectory is rejected, then the new trajectory begins at the same starting point as the rejected trajectory, again with new momentum components chosen randomly from the Gaussian thermal distribution. Determination of the 'correct' step size in the hypothetical time frame is clearly important and this is performed at the beginning of the HMC process. The initial time step is large but is dynamically reduced in size until ~95% of the trajectories are accepted. This scheme represents a reasonable compromise between accuracy and efficiency.

The HMC approach thus combines, in one algorithm, the best features of MD and MC simulations. The MD component uses Hamilton's equations of motion to move quickly through parameter space whilst, in the MC component, a Metropolis criterion ensures detailed balance and the momentum components take on the role of the random variables. Denoting the number of MD steps within a given trajectory by $N_{\rm MD}$ and the number of MC steps (Metropolis evaluations) by $N_{\rm MC}$, then a HMC run consists of $N_{\rm MC}$ Monte Carlo steps each containing $N_{\rm MD}$ molecular dynamics steps. The simulation runs until the χ^2 goodness of fit is lower than a predetermined

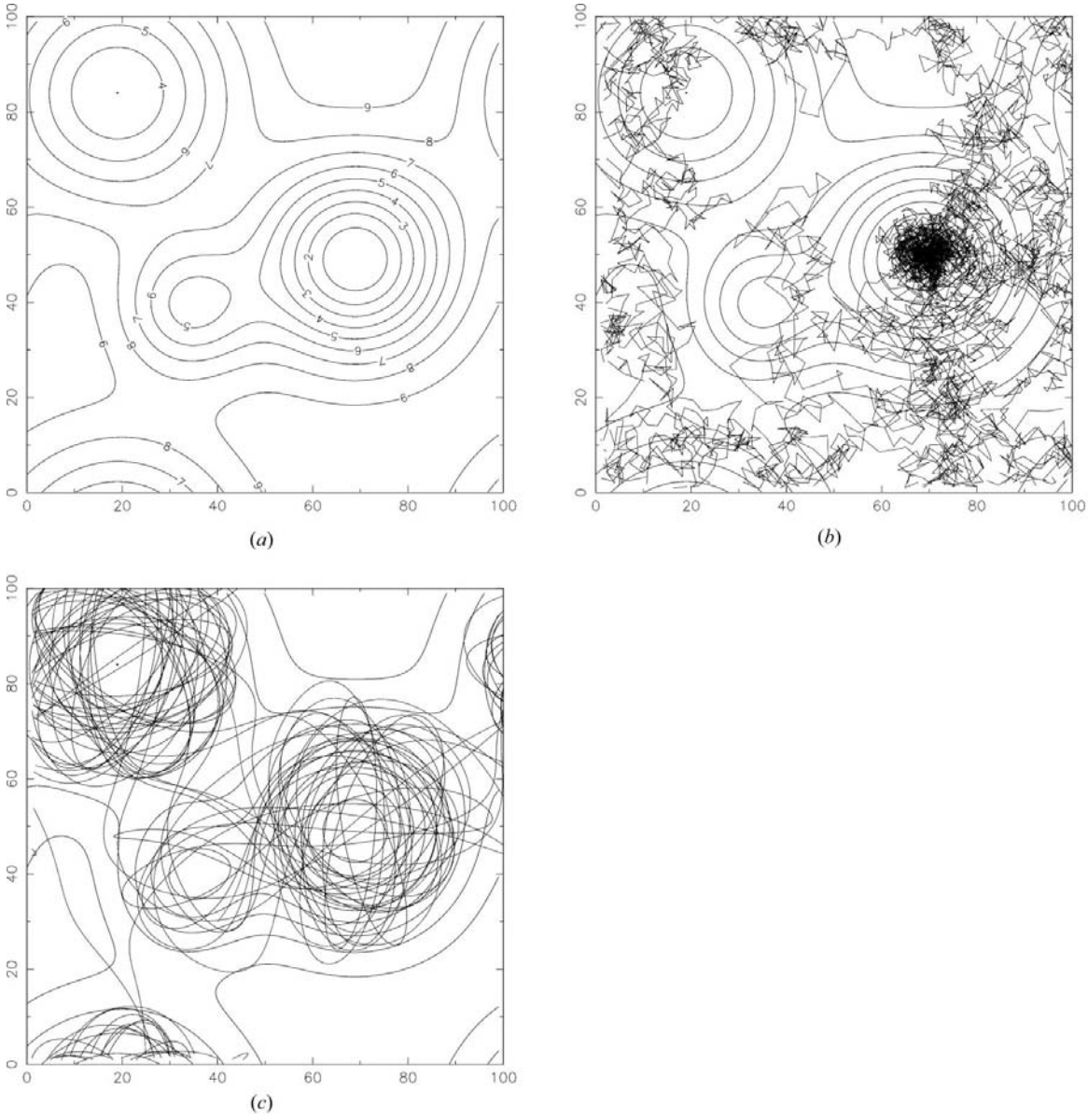

Figure 2
(*a*) A simple periodic potential energy function constructed over a square cell. The maximum value of the potential energy is 10, with uniform contour levels shown at $1, 2, 3, \ldots, 9$. The function has three minima with values 0, 3 and 4.9 at positions (70, 50), (20, 83) and (35, 40), respectively. (*b*) A simulated-annealing solution for the test function shown in (*a*). The SA run consists of 1000 moves and it rapidly locates the global minimum. The figure highlights the random-walk nature of the SA algorithm. (*c*) A molecular dynamics solution for the test function shown in (*a*). The MD run consists of 1000 steps and all three minima are visited many times. The smooth trajectory is a consequence of the deterministic nature of MD equations of motion.

Table 1
The lowest correlated integrated intensities χ^2 values obtained during each of the SA and HMC runs.

Column 3 lists the χ^2 values obtained following conjugate-gradient minimization of the HMC determined crystal structures.

Run No.	χ^2 (SA)	χ^2 (HMC)	χ^2 (CGM)
1	130.4	88.2	85.7
2	214.6	89.0	84.1
3	118.2	90.9	85.4
4	85.0	90.0	82.5
5	85.6	141.3	134.7
6	130.0	88.3	82.9
7	168.7	87.1	82.2
8	81.5	236.2	228.6
9	82.8	87.7	83.9
10	223.4	87.7	81.9
11	201.0	87.4	84.6
12	134.0	267.3	267.1
13	130.8	182.7	177.1
14	115.5	88.6	83.5
15	167.7	181.2	175.9
16	150.9	86.1	83.5
17	135.5	87.7	82.4
18	142.8	87.3	82.0
19	148.1	90.6	83.7
20	107.4	89.5	86.8

value or the maximum number of χ^2 evaluations (equal to $N_{MC} \times N_{MD}$) is exceeded. The HMC algorithm presented in this section is referred to as 'standard' HMC, as it provides an effective algorithm for sampling a canonical ensemble at constant temperature. The HMC method has also been modified to sample other distributions such as the multicanonical ensemble (Arnold *et al.*, 2000) and the mixed-canonical ensemble (Fischer *et al.*, 1999). The HMC method may be further modified by combining it with a simulated-annealing (SA) strategy or a form of tempering such as simulated or parallel tempering (see, for instance, Boyd, 1998, and references therein).

3. Molecular dynamics compared with simulated annealing

The essential difference between a MD run and a SA run is illustrated for the simple case of locating the global minimum of the function shown in Fig. 2(*a*). The 'directed' random walk of the SA algorithm is apparent in Fig. 2(*b*). The algorithm has the ability to climb and cross the χ^2 hills in its search for the low-lying χ^2 basins, eventually settling in the global minimum as the overall temperature of the system is lowered. In contrast, the smooth trajectory-like nature of a MD run is shown in Fig. 2(*c*). Given an appropriate level of kinetic energy, the algorithm traces a smooth path throughout the function space, sampling the function minima along its trajectory. The choice of the 'appropriate level of kinetic energy' is discussed in more detail in §6.

4. HMC implementation for crystal structure determination

The HMC algorithm outlined in §2 was implemented in a *C++* computer program designed to solve crystal structures from powder diffraction data using an agreement factor as specified in (1). The performance of the algorithm was benchmarked against the SA algorithm implemented in the *DASH* computer program (David *et al.*, 2001), which utilizes the same agreement factor. Unlike simulated annealing, molecular dynamics requires the calculation of derivatives and this overhead means that each χ^2 evaluation for the HMC takes approximately twice as long as its SA counterpart.

The test crystal structure selected was that of capsaicin (Fig. 3), $P2_1/c$, a = 12.2234, b = 14.7900, c = 9.4691 Å, β = 93.9754°, T = 100 K. Capsaicin was one of the early test-case structures for the SA algorithm in *DASH* (David *et al.*, 1998) and remains a challenging problem for global optimization methods, with 15 degrees of freedom and only C, H, N and O atoms present.

An internal-coordinate description of the capsaicin molecule was generated using standard bond lengths, bond angles and bond torsions. The diffraction data (λ = 0.6528 Å) used had previously been collected at 100 K from a 0.7 mm capillary filled with capsaicin powder and mounted on the diffractometer at BM16 of the ESRF in Grenoble (David *et al.*, 1998). The data were Pawley-fitted over the range 2.7–22.5° 2θ ($\sim$1.7 Å), in order to extract 379 correlated integrated intensities, resulting in a $\chi^2_{profile}$ value of 13.1.

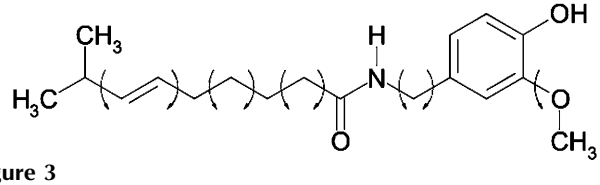

Figure 3
The molecular structure of capsaicin. The arrows denote bonds for which torsion angles cannot be assigned correctly in advance of a structure determination.

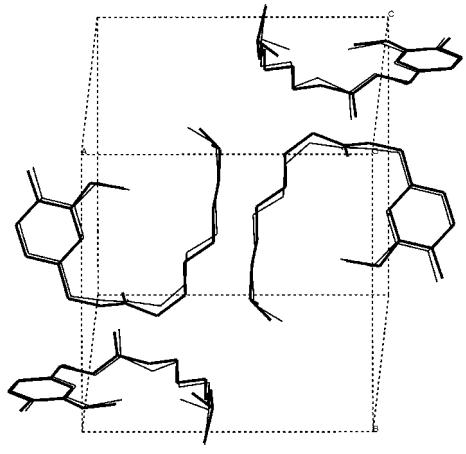

Figure 4
The crystal structure corresponding to HMC solution No. 10 (thin lines), superimposed upon the single-crystal structure of capsaicin (bold lines).

Twenty *DASH* structure-solution runs were performed using default values for the initial system temperature and the cooling rate. All runs were set to terminate after a maximum of 3×10^6 SA moves. Twenty HMC runs were then performed using values of T = 10, N_{MD} = 100, N_{MC} = 10000 and an acceptance rate of ~95%. The choice of T is discussed in detail in §6. At the end of each of the 20 HMC runs, a conjugate-gradient minimization (CGM) of the structure was invoked in the same parameter and data space. Both programs were run on an 800 MHz Pentium-III-based PC running Windows NT Version 4.

5. Results

5.1. Success rates for each method

The results of the SA and HMC runs (Table 1) show that both methods were able to solve the crystal structure of capsaicin repeatedly. The correct solutions are easily identified by their low (< 90) correlated integrated intensities χ^2 values and Fig. 4 shows the excellent agreement between one such HMC solution and the known single-crystal structure. In each case where $\chi^2 < 90$, the structure was compared with the single-crystal structure and confirmed as solved to a good degree of accuracy. Using this χ^2 value as a cut-off point, the success rates in obtaining structure solutions for capsaicin were 20% for the SA runs and 75% for the HMC runs. Including SA solutions 3, 14 and 20, which also show convincing agreement with the single-crystal structure, increases the overall SA success rate to 35%.

5.2. Efficiency of the methods

Fig. 5(*a*) shows a typical 'ridge and cliff' plot of χ^2 *versus* the number of χ^2 evaluations for the 20 HMC runs, whilst Fig. 5(*b*) shows an expanded plot for the first 100000 evaluations.

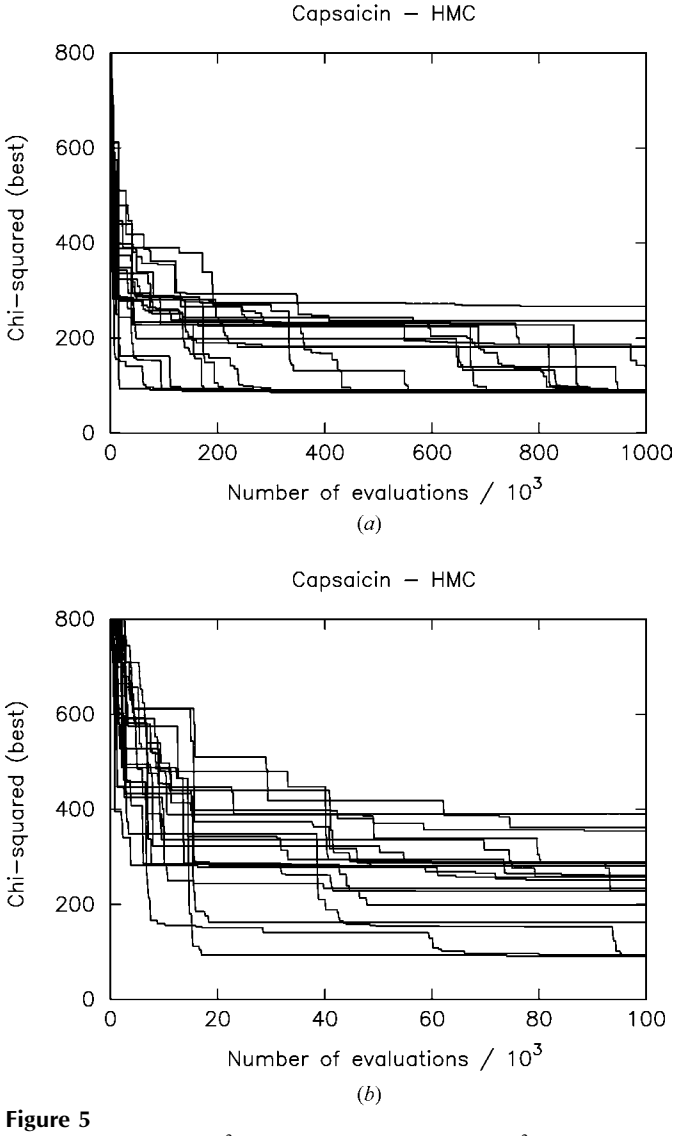

Figure 5
(*a*) A plot of the best χ^2 values *versus* the number of χ^2 evaluations for the 20 HMC runs. (*b*) A plot of the first 100000 evaluations for the runs shown in (*a*).

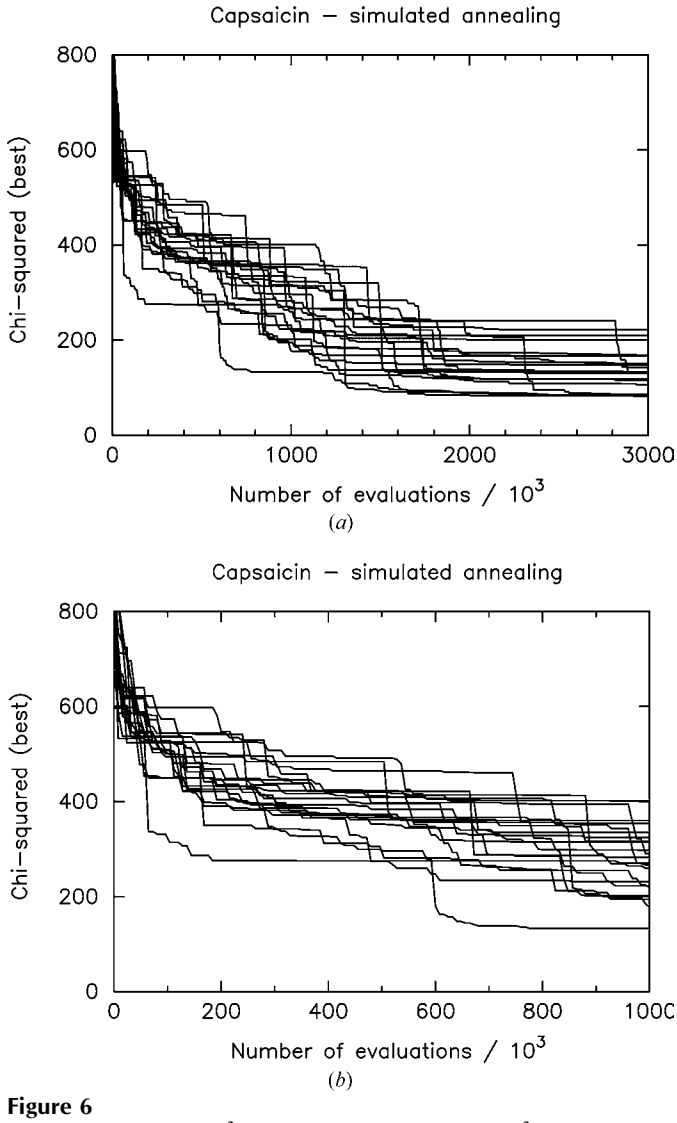

Figure 6
(*a*) A plot of the best χ^2 values *versus* the number of χ^2 evaluations for the 20 SA runs. (*b*) A plot of the first 1000000 evaluations for the runs shown in (*a*).

Correct solutions were obtained in ~18000 to 1000000 χ^2 evaluations. This is a significant gain over the ~1.6 to 2.6 million χ^2 evaluations for the successful SA runs (Figs. 6*a*, *b*) and the fact that eight HMC solutions were obtained in less than 500000 evaluations serves to underline the efficiency of the HMC algorithm.

6. Discussion

One of the attractive features of the HMC approach is the simplicity of the underlying algorithm. In common with SA, the dominant algorithmic parameter is the system temperature and temperatures that lie far from an optimal value normally result in failure of HMC to obtain a correct solution within a reasonable time scale. For example, if the temperature is set too low, the hypothetical particle has too little kinetic energy to escape local minima. In practice, the mean kinetic energy ($\langle K\rangle = \frac{1}{2}NT$, where N is the number of degrees of freedom) is a more intuitive property to work with than the temperature because of its close interrelationship with the potential energy (χ^2) through the equations of motion. The choice of the optimal mean kinetic energy has been found empirically to correspond to the kinetic energy at which fluctuations in χ^2, observed over a series of trajectories, are highest. This is illustrated in Fig. 7, in which fluctuations in the correlated integrated intensities χ^2 values obtained for the capsaicin structure are plotted as a function of the average kinetic energy of the system. It is a straightforward matter to calculate this distribution *via* a short series of MD runs and the optimal value of T is then fixed for the duration of the HMC run. In Fig. 7, the largest fluctuations occur at $\langle K\rangle \approx 90$, indicating that $T \approx 10$ is a good choice for this system. In marked contrast to SA, it is not a prerequisite to decrease T further in order to increase the probability of sampling low χ^2 values, as the HMC trajectory will necessarily visit these minima.

Fig. 8(*a*) shows that the evolution of χ^2 with the total number of MD steps possesses a very definite long-time-scale structure in addition to the obvious rapid fluctuations. Deep minima are very uncommon in this structure solution hyperspace and the initial 10000 MD steps traverse regions of space with relatively high χ^2 values. Shortly afterwards, over the space of a very small number of evaluations, there is a precipitous decrease in χ^2 followed by several smaller decreases, two of which are shown in greater detail in Fig. 8(*b*). The increased frequency of χ^2 fluctuations after the second drop indicates that the HMC trajectory is exploring the parameter space around the best solution. The ability of the HMC method to 'fine tune' both the internal and external degrees of freedom of capsaicin results in a structure that is sufficiently close to the global minimum that subsequent CGM brings about only a very small further reduction in the best χ^2 value (see also Table 1).

It is therefore not surprising to find that the overall reproducibility in the HMC solutions is very good. Fig. 9 shows an overlay of the first eight HMC solutions with $\chi^2 \leq 90$. The small differences between the structures highlight the extent

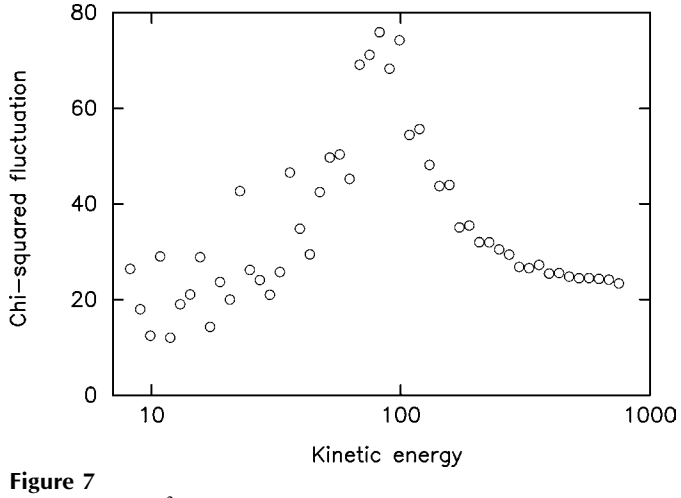

Figure 7
Fluctuations in χ^2 plotted as a function of the average kinetic energy for the capsaicin problem. In simulated annealing, the dependence of χ^2 fluctuations upon temperature takes a similar form.

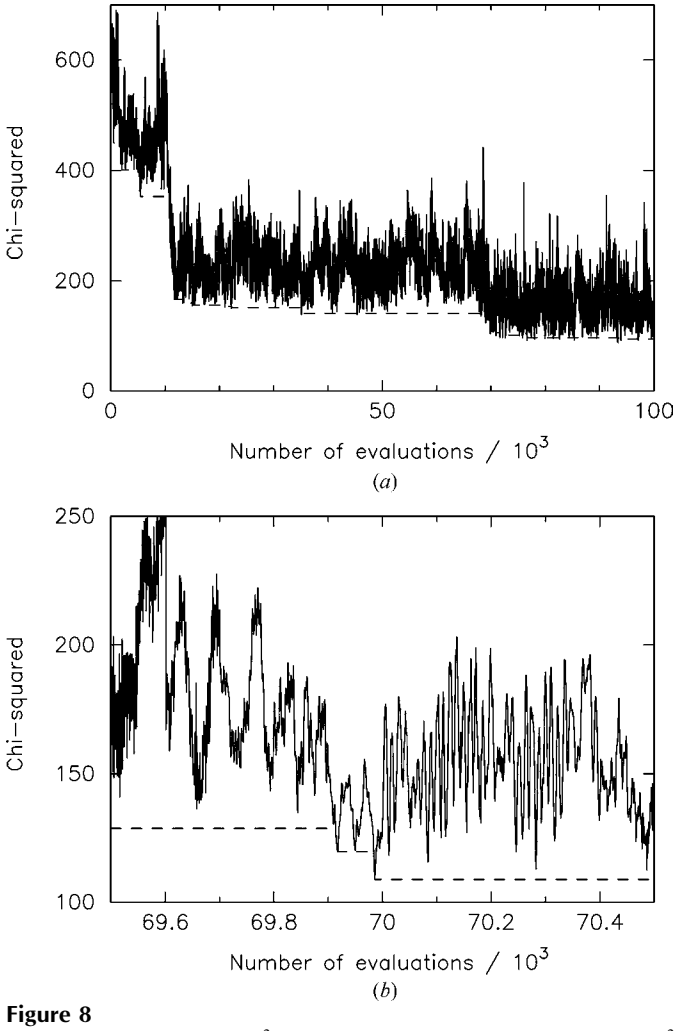

Figure 8
(*a*) The potential energy (χ^2) plotted as a function of the number of χ^2 evaluations for the HMC structure solution of capsaicin. The dotted line corresponds to the lowest χ^2 value obtained thus far. (*b*) An expanded region of the plot shown in (*a*). The best χ^2 for this run was 99.0. Subsequent CGM reduced this value to 95.2.

to which the individual structural parameters are correlated, *i.e.* each structure gives essentially the same fit to the diffraction data despite having slightly different positions, orientations and conformations. This is a consequence both of the e.s.d.'s on the structure factors and the relatively low resolution (∼1.7 Å) of the diffraction data.

7. Conclusions

The hybrid Monte Carlo method outlined in this paper is an effective and efficient global search method for the determination of crystal structures from powder diffraction data. The overall success rate in solving the crystal structure of capsaicin is high for a molecule of this conformational complexity and is significantly better than that achieved with SA optimization. In common with simulated annealing, the algorithm has a limited number of control parameters and is thus well suited to routine use.

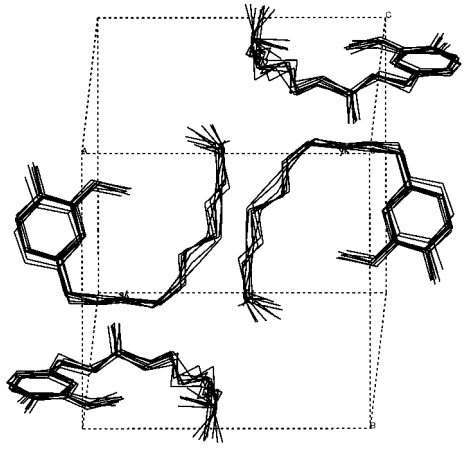

Figure 9
An overlay of the capsaicin crystal structures corresponding to HMC solutions 1, 2, 3, 4, 6, 7, 9 and 10, following conjugate-gradient minimization.

We gratefully acknowledge Dr Andy Fitch of the ESRF for his help in collecting the capsaicin diffraction data and Dr Norman Shankland of CrystallografX Ltd for his critical reading of the manuscript.

References

Arnold, G., Schilling, K. & Lippert, T. (2000). *Nucl. Phys. B Proc. Suppl.* **83–4**, 768–770.
Boyd, G. (1998). *Nucl. Phys. B Proc. Suppl.* **60**A, 341–344.
David, W. I. F., Shankland, K., Cole, J., Maginn, S., Motherwell, W. D. S. & Taylor, R. (2001). *DASH User Manual*, Cambridge Crystallographic Data Centre, Cambridge, England.
David, W. I. F., Shankland, K., McCusker, L. & Baerlocher, C. (2002). Editors. *Structure Determination from Powder Diffraction Data.* Oxford University Press.
David, W. I. F., Shankland, K. & Shankland, N. (1998). *Chem. Commun.* pp. 931–932.
Duane, S., Kennedy, A. D., Pendleton, B. J. & Roweth, D. (1987). *Phys. Lett. B*, **195**, 216–222.
Fischer, A., Cordes, F. & Schütte, C. (1999). *Comput. Phys. Commun.* **121–122**, 37–39.
Floudas, C. A., Klepeis, J. L. & Pardalos, J. L. (1999). *Global Optimization Approaches in Protein Folding and Peptide Docking*, *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, edited by F. Roberts, Vol. 47, pp. 141–171. Providence, RI: American Mathematical Society.
Hockney, R. W. (1970). *Methods Comput. Phys.* **9**, 136–211.
Irbäck, A. (1994). *J. Chem. Phys.* **101**, 1661–1667.
Joo, B., Pendleton, B., Kennedy, A. D., Irving, A. C., Sexton, J. C., Pickles, S. M. & Booth, S. P. (2000). *Phys. Rev. D*, **62**, Article 114501.
Leach, A. R. (1996). *Molecular Modelling Principles and Applications.* London: Longman.
Le Bail, A., Duroy, H. & Fourquet, J. L. (1988). *Mater. Res. Bull.* **23**, 447–452.
Markvardsen, A. J., David, W. I. F. & Shankland, K. (2002). *Acta Cryst.* A**58**, 316–326.
Pagola, S., Stephens, P. W., Bohle, D. S., Kosar, A. D. & Madsen, S. K. (2000). *Nature (London)*, **404**, 307–310.
Pawley, G. S. (1981). *J. Appl. Cryst.* **14**, 357–361.
Young, R. A. (1993). Editor. *The Rietveld Method.*. Oxford University Press.