

Enhanced Macromolecular Refinement by Simulated Annealing

Axel T. Brunger^{1,2,*}, Paul D. Adams², and Luke M. Rice²

¹ The Howard Hughes Medical Institute and ² Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06511, phone: 203-432-6143, FAX: 203-432-6946, email: brunger@laplace.csb.yale.edu

* corresponding author

Introduction

The analysis of X-diffraction data generally requires sophisticated computational procedures that culminate in refinement and structure validation. The refinement procedure can be formulated as the chemically constrained or restrained non-linear optimization of a target function, which usually measures the agreement between observed diffraction data and data computed from an atomic model. The ultimate goal of refinement is to simultaneously optimize the agreement of an atomic model with observed diffraction data and with *a priori* chemical information.

The target function used for this optimization normally depends on several atomic parameters, and most importantly on atomic coordinates. The large number of adjustable parameters (typically at least three times the number of atoms in the model) give rise to a very complicated target function. This in turn produces what is known as the multiple minima problem: the target function contains many local minima in addition to the global minimum, and this tends to defeat gradient-descent optimization techniques such as conjugate gradient or least-squares methods (Press *et al.*, 1986). These methods are unable to sample molecular conformations thoroughly enough to find the most optimal model if the starting one is far from the correct structure.

The challenges of crystallographic refinement arise not only from the high dimensionality of the parameter space, but also from the phase problem. For new crystal structures, initial electron density maps must be computed from a combination of observed diffraction amplitudes and experimental phases where the latter are typically of poorer quality and/or lower resolution than the former. A different problem arises when structures are solved by molecular replacement (Hoppe, 1957; Rossman & Blow, 1962) which uses a similar structure as a search

model to calculate initial phases. In this case the resulting electron density maps can be severely “model-biased”, that is, they sometimes seem to confirm the existence of the search model without providing clear evidence of actual differences between it and the true crystal structure. In both cases, initial atomic models usually contain significant errors and require extensive refinement.

Simulated annealing (Kirkpatrick *et al.*, 1983) is an optimization technique particularly well suited to overcoming the multiple minima problem. Unlike gradient-descent methods, simulated annealing can cross barriers between minima and thus can explore a greater volume of the parameter space to find better models (deeper minima). Following its introduction to crystallographic refinement (Brunger *et al.*, 1987), there have been major improvements of the original method in four principal areas: the measure of model quality, the search of the parameter space, the target function, and the modelling of conformational variability.

For crystallographic refinement, the introduction of cross-validation and the free R value (Brunger, 1992) has significantly reduced the danger of overfitting the diffraction data during refinement. Cross-validation also produces more realistic coordinate error estimates based on the Luzzati or σ_A methods (Kleywegt & Brunger, 1996). The complexity of the conformational space has been reduced by the introduction of torsion-angle refinement methods (Diamond, 1971; Rice & Brunger, 1994), which decrease the number of adjustable parameters that describe a model approximately tenfold. The target function has been improved by using a maximum-likelihood approach which takes into account model error, model incompleteness, and errors in the experimental data (Bricogne, 1991; Pannu & Read, 1996). Cross-validation of parameters for the maximum likelihood target function was essential in order to obtain better results than with conventional target functions (Pannu & Read, 1996; Adams *et al.*, 1997; Read,

1997). Finally, the sampling power of simulated annealing has been used for exploring the molecule’s conformational space in cases where the molecule undergoes dynamic motion or exhibits static disorder (Kuryian *et al.*, 1991; Burling & Brunger, 1994; Burling *et al.*, 1996).

Cross-Validation

Cross-validation (Brunger, 1992) plays a fundamental role in the maximum likelihood target functions described below. A few remarks about this method are therefore warranted here (for reviews, see Kleywegt & Brunger, 1996; Brunger, 1997). For cross-validation, the diffraction data are divided into two sets: a large *working* set (usually comprising 90% of the data), and a complementary *test* set (comprising the remaining 10%). The diffraction data in the working set are used in the normal crystallographic refinement process, whereas the test data are not. The cross-validated (or “free”) R value computed with the test set data is a more faithful indicator of model quality. It provides a more objective guide during the model building and refinement process than the conventional R value. It also ensures that introduction of additional parameters (e.g., water molecules, relaxation of non-crystallographic symmetry restraints, or multi-conformer models) improves the quality of the model rather than increasing overfitting.

Since the conventional R value shows little correlation with the accuracy of a model, coordinate-error estimates derived from the Luzzati (1952) or σ_A (Read, 1986) methods are unrealistically low. Kleywegt and Brunger (1996) showed that more reliable coordinate errors can be obtained by cross-validation of the Luzzati or σ_A coordinate error estimates. An example is shown in Fig. 1 using the crystal structure and diffraction data of penicillopepsin (Hsu *et al.*, 1977). At 1.8 Å resolution, the model has an estimated coordinate error of \approx

0.2 Å as assessed by multiple independent refinements. As the resolution of the diffraction data is artificially truncated and the model re-refined, the coordinate error (assessed by the atomic root-mean-square difference to the refined model at 1.8 Å resolution) increases monotonically. The conventional R -value improves as the resolution decreases and the quality of the model worsens. Consequently, coordinate error estimates do not display the correct behavior either: the error estimates are approximately constant regardless of the resolution and actual coordinate error of the models. However, when cross-validation is used (i.e., the test reflections are used to compute the estimated coordinate errors), the results are much better: the cross-validated errors are close to the actual coordinate error, and they show the correct trend as a function of resolution (Fig. 1).

The Target Function

Crystallographic refinement is a search for the global minimum of the target

$$E = E_{\text{chem}} + w_{\text{xray}}E_{\text{xray}} \quad (1)$$

as a function of the parameters of an atomic model, in particular atomic coordinates. E_{chem} comprises empirical information about chemical interactions; it is a function of all atomic positions, describing covalent (bond lengths, bond angles, torsion angles, chiral centers, and planarity of aromatic rings) and non-bonded (intra-molecular as well as inter-molecular and symmetry-related) interactions (Hendrickson, 1985). E_{xray} is related to the difference between observed and calculated data, and w_{xray} is a weight appropriately chosen to balance the gradients (with respect to atomic parameters) arising from the two terms.

X-ray Diffraction Data *vs.* Model

The traditional form of E_{xray} consists of the crystallographic residual E^{LSQ} , defined as the sum over the squared differences between the observed ($|\mathbf{F}_o|$) and calculated ($|\mathbf{F}_c|$) structure factor amplitudes for a particular atomic model:

$$E_{\text{xray}} = E^{\text{LSQ}} = \sum_{hkl \in \text{working set}} (|\mathbf{F}_o| - k|\mathbf{F}_c|)^2 \quad (2)$$

where hkl are the indices of the reciprocal lattice points of the crystal, and k is a relative scale factor.

Minimization of E^{LSQ} can produce improvement in the atomic model, but it can also accumulate systematic errors in the model by fitting noise in the diffraction data (Silva & Rossman, 1985). The least-squares residual is a limiting case of the more general maximum likelihood theory, and is only justified if the model is nearly complete and error-free. These assumptions may be violated during the initial stages of refinement. Improved targets for macromolecular refinement have been obtained using the more general maximum-likelihood formulation (Bricogne, 1991; Pannu & Read, 1996; Adams *et al.*, 1997; Murshudov *et al.*, 1997). The goal of the maximum likelihood method is to determine the likelihood of the model given estimates of the model’s errors and those of the measured intensities.

A starting point for the maximum likelihood formulation of crystallographic refinement is the Sim (1959) distribution, i.e., the Gaussian conditional probability distribution of the “true” structure factors \mathbf{F} , given a partial model with structure factors \mathbf{F}_c and the model’s error (Fig. 2) (Srinivasan, 1966; Read, 1986; Read, 1990) (for simplicity we will only discuss the case of acentric reflections),

$$P_a(\mathbf{F}; \mathbf{F}_c) = \frac{1}{\pi \epsilon \sigma_\Delta^2} \exp[-(\mathbf{F} - D\mathbf{F}_c)^2 / \epsilon \sigma_\Delta^2] \quad (3)$$

where σ_Δ is a parameter that incorporates the effect of the fraction of the asymmetric unit that is missing from the model and errors in the partial structure. Assuming a Wilson distribution of intensities it can be shown that (Read, 1990)

$$\sigma_\Delta^2 = \langle |\mathbf{F}_o|^2 \rangle - D^2 \langle |\mathbf{F}_c|^2 \rangle \quad (4)$$

where D is a factor that takes into account model error: it is unity in the limiting case of an error-free model and it is zero if no model is available (Luzzati, 1952; Read, 1986). For a complete and error-free model σ_Δ therefore becomes zero and the probability distribution $P_a(\mathbf{F}; \mathbf{F}_c)$ is infinitely sharp.

Taking into account measurement errors requires multiplication of Eq. 3 with an appropriate probability distribution (usually a conditional Gaussian distribution with standard deviation σ_o) of the observed structure factor amplitudes ($|\mathbf{F}_o|$) around the “true” structure factor amplitudes ($|\mathbf{F}|$)

$$P_{\text{meas}}(|\mathbf{F}_o|; |\mathbf{F}|). \quad (5)$$

Prior knowledge of the phases of the structure factors can be incorporated by multiplication of Eq. 3 with a phase probability distribution

$$P_{\text{phase}}(\phi). \quad (6)$$

and re-writing Eq. 3 in terms of the structure factor moduli and amplitudes of $\mathbf{F} = |\mathbf{F}| \exp(i\phi)$.

The unknown variables $|\mathbf{F}|$ and ϕ in Eqs. 3, 4, and 5 have to be eliminated by integration in order to obtain the conditional probability distribution of the observed structure factor amplitudes, given a partial model with errors, the amplitude measurement errors, and prior phase information:

$$P_a(|\mathbf{F}_o|; \mathbf{F}_c) = \frac{1}{\pi \epsilon \sigma_\Delta^2} \int d\phi d|\mathbf{F}||\mathbf{F}| P_{\text{meas}}(|\mathbf{F}_o|; |\mathbf{F}|) P_{\text{phase}}(\phi) \exp[-(|\mathbf{F}| \exp(i\phi) - D\mathbf{F}_c)^2 / \epsilon \sigma_\Delta^2]. \quad (7)$$

The likelihood L of the model is defined as the joint probability distribution of the structure factors of all reflections in the working set. Assuming independent and uncorrelated structure factors, L is simply the product of the distributions in Eq. 7 for all reflections. Instead of maximizing the likelihood it is more common to minimize the negative logarithm of the likelihood,

$$E_{\text{xray}} = \mathcal{L} = - \sum_{hkl \in \text{working set}} \log[P_a(|\mathbf{F}_o|; \mathbf{F}_c)]. \quad (8)$$

Empirical estimates of σ_Δ (and D through Eq. 4) can be obtained by minimizing \mathcal{L} for a particular atomic model. It is generally assumed that σ_Δ and D show relatively little variation among neighboring reflections. Accepting this assumption, σ_Δ and D can be estimated by considering narrow resolution shells of reflections and assuming that the two parameters are constant in these shells. Minimization of \mathcal{L} can then be performed as a function of these constant shell parameters while keeping the atomic model fixed (Read, 1986; Read, 1997). Alternatively, one can assume a two-term Gaussian model for σ_Δ (Murshudov *et al.*, 1997) and minimize \mathcal{L} as a function of the Gaussian parameters. Note, that individual atomic B-factors are taken into account by the calculated model structure factors (\mathbf{F}_c).

This empirical approach to estimate σ_Δ and D requires occasional re-computation of these values as the model improves. Refinement methods that improve the model structure factors \mathbf{F}_c will therefore have a beneficial effect on σ_Δ and D . Better estimates of these values will then enhance the next refinement cycle. Thus, powerful optimization methods and maximum likelihood targets are

expected to interact in a synergistic fashion (c.f. Fig. 4). Structure factor averaging of multi-start refinement models can provide another layer of improvement by producing a better description of \mathbf{F}_c if the model shows significant variability due to errors or intrinsic flexibility (see below).

In order to achieve an improvement over the least-squares residual (Eq. 2), cross-validation was found to be essential (Pannu & Read, 1996; Adams *et al.*, 1997) for the estimation of model incompleteness and errors (σ_Δ and D). Since the test set typically contains only 10% of the diffraction data, these cross-validated quantities can show significant statistical fluctuations as a function of resolution. In order to reduce these fluctuations, Read (1997) devised a smoothing method by applying restraints to σ_A values between neighboring resolution shells where

$$\sigma_A = \sqrt{1 - \frac{\sigma_\Delta}{\langle |\mathbf{F}_o|^2 \rangle}}. \quad (9)$$

Read and Pannu (1996) have developed an efficient Gaussian approximation of Eq. 7 for the case of no prior phase information, termed the “MLF” target function. In the limit of a perfect model (i.e., $\sigma_\Delta = 0$ and $D = 1$), MLF reduces to the traditional least-squares residual (Eq. 2) with $1/\sigma_o^2$ weighting. In the case of prior phase information, the integration over the phase angles has been carried out numerically in Eq. 7, termed the “MLHL” target (Pannu *et al.*, 1998). A maximum likelihood function which expresses Eq. 7 in terms of observed intensities has also been developed, termed “MLI” (Pannu & Read, 1996).

A Priori Chemical Information

The parameters for the covalent terms in E_{chem} (Eq. 1) can be derived from average geometry and (r.m.s.) deviations observed in a small-molecule data base. Extensive statistical analyses were undertaken for the chemical moieties of

proteins (Engh & Huber, 1991) and of polynucleotides (Parkinson *et al.*, 1996) using the Cambridge crystallographic database (Allen *et al.*, 1983). Analysis of the ever increasing number of atomic resolution macromolecular crystal structures will no doubt cause some modifications of these parameters in the future.

It is common to use a purely repulsive quartic function ($E_{\text{repulsive}}$) for the nonbonded interactions which are included in E_{chem} (Hendrickson, 1985),

$$E_{\text{repulsive}} = \sum_{ij} ((cR_{ij}^{\text{min}})^n - R_{ij}^n)^m \quad (10)$$

where R_{ij} is the distance between two atoms i, j , R_{ij}^{min} is the van der Waals radius for a particular atom pair ij , $c \leq 1$ is a constant that is sometimes used to reduce the radii, and $n = 2, m = 2$ or $n = 1, m = 4$. Van der Waals attraction and electrostatic interactions are usually not included in crystallographic refinement. These simplifications are valid since the diffraction data contains information that is able to produce atomic conformations consistent with actual nonbonded interactions. In fact, atomic resolution crystal structures can be used to derive parameters for electrostatic charge distributions (Pearlman & Kim, 1990).

Searching Conformational Space

Annealing denotes a physical process wherein a solid is heated until all particles randomly arrange themselves in a liquid phase, and then is cooled slowly so that all particles arrange themselves in the lowest energy state. By formally defining the target E (Eq. 1) to be the equivalent of the potential energy of the system, one can simulate such an annealing process (Kirkpatrick *et al.*, 1983). There is no guarantee that simulated annealing will find the global minimum (Laarhoven & Aarts, 1987). However, compared to conjugate-gradient minimization where search directions must follow the gradient, simulated annealing achieves more

optimal solutions by allowing motion against the gradient (Kirkpatrick *et al.*, 1983). The likelihood of *uphill* motion is determined by a control parameter referred to as *temperature*. The higher the temperature, the more likely it is that simulated annealing will overcome barriers (Fig. 3). It should be noted that the simulated annealing temperature normally has no physical meaning and merely determines the likelihood of overcoming barriers of the target function Eq. 1.

The simulated annealing algorithm requires a mechanism to create a Boltzmann distribution at a given temperature T and an annealing schedule, that is, a sequence of temperatures $T_1 \geq T_2 \geq \dots \geq T_l$ at which the Boltzmann distribution is computed. Implementations differ in the way they generate a transition, or move, from one set of parameters to another which is consistent with the Boltzmann distribution at given temperature. The two most widely used methods are Metropolis Monte Carlo (Metropolis *et al.*, 1953) and molecular dynamics (Verlet, 1967) simulations. For X-ray crystallographic refinement, molecular dynamics has proven extremely successful (Brunger *et al.*, 1987) because it limits the search to physically reasonable “moves”.

Molecular Dynamics

A suitably chosen set of atomic parameters can be viewed as generalized coordinates that are propagated in time by the classical equations of motion (Goldstein, 1980). If the generalized coordinates represent the x, y, z positions of the atoms of a molecule, the classical equations of motion reduce to the familiar Newton’s second law:

$$m_i \frac{\partial^2 \vec{r}_i}{\partial t^2} = -\nabla_i E. \quad (11)$$

The quantities m_i and \vec{r}_i are respectively the mass and coordinates of atom i , and E is given by Eq. 1. The solution of the partial differential equations

(Eq. 11) can be achieved numerically using finite-difference methods (Verlet, 1967; Abramowitz & Stegun, 1968). This approach is referred to as molecular dynamics.

Initial velocities for the integration of Eq. 11 are usually assigned randomly from a Maxwell distribution at the appropriate temperature. Assignment of different initial velocities will generally produce a somewhat different structure after simulated annealing. By performing several refinements with different initial velocities one can therefore improve the chances of success of simulated-annealing refinement. Furthermore, this improved sampling can be used to study discrete disorder and conformational variability, especially when using torsion-angle molecular dynamics (see below).

Although Cartesian (i.e., flexible bond lengths and bond angles) molecular dynamics places restraints on bond lengths and bond angles (through E_{chem} , Eq. 1), one might want to implement these restrictions as constraints, i.e., fixed bond lengths and bond angles (Diamond, 1971). This is supported by the observation that the deviations from ideal bond lengths and bond angles are usually small in macromolecular X-ray crystal structures. Indeed, fixed-length constraints have been applied to crystallographic refinement by least-squares minimization (Diamond, 1971). It is only recently, however, that efficient and robust algorithms have become available for molecular dynamics in torsion-angle space (Bae & Haug, 1987; Bae & Haug, 1988; Jain *et al.*, 1993; Rice & Brunger, 1994). We chose an approach that retains the Cartesian-coordinate formulation of the target function and its derivatives with respect to atomic coordinates so that the calculation remains relatively straightforward and can be applied to any macromolecule or their complexes (Rice & Brunger, 1994). In this formulation, the expression for the acceleration becomes a function of positions and velocities. Iterative equations of motion for constrained dynamics in this formulation can be

derived and solved by finite difference methods (Abramowitz & Stegun, 1968). This method is numerically very robust and has a significantly increased radius of convergence in crystallographic refinement compared to Cartesian molecular dynamics (Rice & Brunger, 1994).

Temperature Control

Simulated annealing requires the control of the temperature during molecular dynamics. The current temperature of the simulation (T_{curr}) is computed from the kinetic energy

$$E_{\text{kin}} = \sum_i^{n_{\text{atoms}}} \frac{1}{2} m_i \left(\frac{\partial r_i}{\partial t} \right)^2 \quad (12)$$

of the molecular dynamics simulation,

$$T_{\text{curr}} = \frac{2E_{\text{kin}}}{3nk_b}. \quad (13)$$

Here n is the number of atoms, m_i is the mass of the atom, and k_b is Boltzmann’s constant. One commonly used approach to control the temperature of the simulation consists of coupling the equations of motion to a heat bath through a “friction” term (Berendsen *et al.*, 1984). Another approach is to periodically rescale the velocities in order to match T_{curr} with the target temperature.

Annealing Schedules

The simulated annealing temperature needs to be high enough to allow conformational transitions but not so high that the model moves too far away from the correct structure. The optimal temperature for a given starting structure is a matter of trial and error. Starting temperatures were determined for a variety of simulated annealing protocols (Brunger, 1988; Adams *et al.*, 1997) which work

for the average case. However, it might be worth trying a different temperature if a particularly difficult refinement problem is encountered. In particular, significantly higher temperatures are attainable using torsion–angle molecular dynamics. Note, that each simulated annealing refinement is subject to “chance” by using a random number generator to generate the initial velocities. Thus, multiple simulated annealing runs can be carried out in order to increase the success rate of the refinement. The best structure(s) (as determined by the free R value) among a set of refinements using different initial velocities and/or temperatures can be taken for further refinement or structure factor averaging (see below).

The annealing schedule can in principle be any function of the simulation step (or “time” domain). The two most commonly used protocols are linear slow–cooling or constant–temperature followed by quenching. A slight advantage is obtained with slow–cooling (Brunger *et al.*, 1990). The duration of the annealing schedule is another parameter. Too short a protocol does not allow sufficient sampling of conformational space. Too long a protocol may waste computer time since it is more efficient to run multiple trials as opposed to one long refinement protocol (unpublished results).

An Intuitive Explanation of Simulated Annealing

The goal of any optimization problem is to find the global minimum of a target function. In the case of crystallographic refinement, one searches for the conformation or conformations of the molecule that best fit the diffraction data and that simultaneously maintain reasonable covalent and non–covalent interactions. Simulated annealing refinement has a much larger radius of convergence than conjugate–gradient minimization (see below). It must therefore be able to find a lower minimum of the target E (Eq. 1) than the local minimum found by simply moving along the negative gradient of E .

It is most easy to visualize this property of simulated annealing in the case of a one-dimensional problem where the goal is to find the global minimum of function with multiple minima (Fig. 3). An intuitive way to understand a molecular dynamics simulation is to envision a ball rolling on this one-dimensional surface. When the ball is far from the global minimum it gains a certain inertia which allows it to cross barriers of the target function (Eq. 13). Slow-cooling temperature control ensures that the ball will eventually reach the global minimum rather than just bouncing across the surface. The initial temperature must be large enough to overcome smaller barriers but low enough to ensure that the system will not escape the global minimum if it manages to arrive there.

While temperature itself is a global parameter of the system, temperature fluctuations arise principally from local conformational transitions - for example from an amino acid sidechain falling into the correct orientation. These local changes tend to lower the value of the target E , thus increasing the kinetic energy, and hence the temperature, of the system. Once the temperature control has removed this excess kinetic energy through "heat dissipation", the reverse transition is very unlikely, since it would require a localized increase in kinetic energy where the conformational change occurred in the first place (Fig. 3). Temperature control maintains a sufficient amount of kinetic energy to allow local conformational corrections, but does not supply enough to allow escape from the global minimum. This explains the observation that on average the agreement with the diffraction data will improve rather than worsen with simulated annealing.

Examples

Many examples have shown that simulated annealing refinement starting from initial models (obtained by standard crystallographic techniques) produces significantly better final models compared to those produced by least-squares or conjugate-gradient minimization (Brunger *et al.*, 1987; Brunger, 1988; Kuriyan *et al.*, 1989; Fujinaga & Gros, 1989; Rice & Brunger, 1994; Adams *et al.*, 1997). In another realistic test case (Adams & Brunger, in preparation), a series of models for the aspartic proteinase penicillopepsin was generated from homologous structures present in the Protein Data Bank. The sequence identity among these structures ranged from 100% to 25%, thus providing a set of models with increasing coordinate error compared to the refined structure of penicillopepsin. These models, after truncation of all residues to alanine, were all used as search models in molecular replacement against the native penicillopepsin diffraction data. In all cases the correct placement of the model in the penicillopepsin unit cell was found.

Both conjugate gradient minimization and simulated annealing were carried out in order to compare the performance of the E^{LSQ} least-squares residual (Eq. 2), MLF (the maximum-likelihood target using amplitudes), and MLHL (the maximum-likelihood target using amplitudes and experimental phase information). In the latter case, phases from single-isomorphous replacement were used. A very large number of conjugate gradient cycles were carried out in order to make the computational requirements equivalent for both minimization and simulated annealing. The conjugate gradient minimizations were converged, i.e., there was no change when further cycles were carried out.

For a given target function, simulated annealing always outperformed minimization (Fig. 4). For a given starting model, the maximum-likelihood tar-

gets outperformed the least-squares residual target for both minimization and simulated annealing, producing models with lower phase errors and higher map correlation coefficients when compared to the published penicillopepsin crystal structure (Fig. 4). This improvement is illustrated in σ_A -weighted electron density maps obtained from the resulting models (Fig. 5). The incorporation of experimental phase information further improved the refinement significantly despite the ambiguity in the SIR phase probability distributions. Thus, the most efficient refinement will make use of simulated annealing, and phase information in the MLHL maximum likelihood target function.

Cross-validation is essential in the calculation of the maximum likelihood target (Kleywegt & Brunger, 1996; Pannu & Read, 1996; Adams *et al.*, 1997). Maximum-likelihood refinement without cross-validation gives much poorer results, as indicated by higher free R -values, higher $R_{\text{free}} - R$ differences, and phase errors (Adams *et al.*, 1997). It should be noted that the final normal R value is in general increased compared to refinements with the least-squares target when using the cross-validated maximum likelihood formulation. This is a consequence of the reduction of overfitting by this method.

Multi-start Refinement and Structure Factor Averaging

Multiple simulated annealing refinements starting from the same model, termed “multi-start” refinement, will generally produce somewhat different structures. Even well-refined structures will show some variation consistent with the estimated coordinate error of the model (c.f., results for 1.8 Å resolution in Fig. 1). More importantly, the poorer the model, the more variation is observed (Brunger, 1988). Some of the models resulting from multi-start refinement may be better

than others, for example as judged by the free R value. Thus, if computer time is available, multi-start refinement has several advantages. A more optimal single model than that produced by a single simulated annealing calculation can usually be obtained. Furthermore, each separate model coming from a multi-start refinement fits the data slightly differently. This could be the result of intrinsic flexibility in the molecule (see below), or the result of model building error. Regions in the starting model that contain significant errors often show increased variability after multi-start refinement, and a visual inspection of the ensemble of models produced can be helpful in identifying these incorrectly modelled regions.

To better identify the correct conformation, structure factors from each of the models can be averaged (Rice, Shamoo, & Brunger, in preparation). This averaging tends to reduce the effect of local errors (noise) that are presumably different in each member of the family. The average structure factor can produce phases that contain less model bias than phases computed from a single model. It should also produce better estimates of σ_{Δ} and D for maximum likelihood targets and σ_A weighted electron density maps because \mathbf{F}_c is used in the computation of these parameters (Eq. 7). Because it is inherently a noise-reducing technique, multi-start refinement followed by structure factor averaging should be most useful in situations where there is significant noise, namely when the data to parameter ratio is low (e.g., if only moderate resolution diffraction data are available).

Ensemble Models

In cases of conformational variability or discrete disorder, there is not a single correct solution to the global minimization of Eq. 1. Rather, the X-ray diffraction data represent a spatial and temporal average over all conformations that

are assumed by the molecule. Ensembles of structures, which are simultaneously refined against the observed data, may thus be a more appropriate description of the diffraction data. This has been used for some time when alternate conformations are modelled locally. Alternate conformations can be generalized to global conformations (Gros *et al.*, 1990; Kuriyan *et al.*, 1991; Burling & Brunger, 1994), i.e., the model is duplicated n -fold, the calculated structure factors corresponding to each copy of the model are summed and this composite structure factor is refined against the observed X-ray diffraction data. Each member of the family is chemically “invisible” to all other members. The optimal number n can be determined by cross-validation (Burling & Brunger, 1994; Burling *et al.*, 1996).

An advantage of a multi-conformer model is that it directly incorporates many possible types of disorder and motion (global disorder, local sidechain disorder, local wagging and rocking motions). Furthermore, it can be used to automatically detect the most variable regions of the molecule by inspecting the atomic r.m.s. difference around the mean as a function of residue number. Thermal factors of single conformer models may sometimes be misleading because they underestimate the degree of motion or disorder (Kuriyan *et al.*, 1986) and, thus, the multiple-conformer model can be a more faithful representation of the diffraction data. A disadvantage of the multi-conformer model is that it introduces many more parameters in the refinement.

Although there are some similarities between averaging structure factors of individually refined structures and performing multi-conformer refinement, there are also fundamental differences. For example, multi-start averaging seeks to improve the calculated electron density map by averaging out the noise present in the individual models, each of which is still a good representation of the diffraction data. This method is most useful at the early stages of refinement when the model still contains errors. In contrast, multi-conformer refinement seeks to

create an ensemble of structures at the final stages of refinement which, taken together, best represent the data. It should be noted that each individual conformer of the ensemble does not necessarily remain a good description of the diffraction data since the whole ensemble is refined against the data. Clearly, multi-conformer refinement requires a high observable-to-parameter ratio.

Conclusions

Simulated annealing has dramatically improved the efficiency of crystallographic refinement. A case in point is the combination of torsion angle molecular dynamics with cross-validated maximum-likelihood targets. These two independent developments interact synergistically to produce less model bias than any other method to date. The combined method dramatically increases the radius of convergence, allowing the productive refinement of poor initial models, e.g., those obtained by weak molecular replacement solutions (Rice & Brunger, 1994; Adams *et al.*, 1997; Adams & Brunger, in preparation).

Simulated annealing can also be used to provide new physical insights into molecular function which may depend on conformational variability. The sampling characteristics of simulated annealing allow the generation of multi-conformer models which can represent molecular motion and discrete disorder, especially when combined with the acquisition of high-quality data (Burling *et al.*, 1996). Simulated annealing is thus also a stepping stone towards development of improved models of macromolecules in solution and in the crystalline state.

Acknowledgments

The computational developments discussed in this review are implemented in the software suite “Crystallography and NMR System” (Brunger *et al.*, 1998). A pre-release of the software suite is available upon request. L.M.R. is an HHMI predoctoral fellow. This work was funded in part by grants from the National Science Foundation to A.T.B. (BIR 9514819 and ASC 93-181159).

Bibliography

Abramowitz, M. and Stegun, I. (1968). Handbook of mathematical functions, Applied Mathematics Series, vol. 55, Dover Publications, New York, pp. 896.

Adams, P.D., Pannu, N.S., Read, R.J. and Brunger, A.T. (1997). Cross-validated maximum likelihood enhances crystallographic simulated annealing refinement. *Proc. Natl. Acad. Sci. USA* **94**, 5018–5023.

Allen, F.H., Kennard, O. and Taylor, R. (1983). Systematic analysis of structural data as a research technique in organic chemistry. *Acc. Chem. Res.* **16**, 146-153.

Bae, D.-S. and Haug, E.J. (1987). A recursive formulation for constrained mechanical system dynamics: Part I. Open loop systems. *Mech. Struct. and Mach.* **15**, 359-382.

Bae, D.-S. and Haug, E.J. (1988). A recursive formulation for constrained mechanical system dynamics: Part II. Closed loop systems. *Mech. Struct. and Mach.* **15**, 481-506.

Berendsen, H.J.C., Postma, J.P.M., van Gunsteren, W.F., DiNola, A. and Haak, J.R. (1984). Molecular dynamics with coupling to an external bath. *J. Chem. Phys.* **81**, 3684-3690.

Bricogne, G. (1991). A multisolution method of phase determination by combined maximization of entropy and likelihood. III. Extension to powder diffraction data. *Acta Cryst. A* **47**, 803-829.

Brunger, A.T. (1988). Crystallographic refinement by simulated annealing: application to a 2.8 Å resolution structure of aspartate aminotransferase. *J. Mol. Biol.* **203**, 803-816.

Brunger, A.T. (1992). The free R value: a novel statistical quantity for assessing

- the accuracy of crystal structures. *Nature* **355**, 472-474.
- Brunger, A.T. (1997). Free *R* value: cross-validation in crystallography, *Methods in Enzym.* **277**, 366-396.
- Brunger, A.T., Kuriyan, J. and Karplus, M. (1987). Crystallographic *R* factor refinement by molecular dynamics. *Science* **235**, 458-460.
- Brunger, A.T., Krukowski, A. and Erickson, J. (1990). Slow-cooling protocols for crystallographic refinement by simulated annealing. *Acta Cryst. A* **46**, 585-593.
- Brunger, A.T., Adams, P.D., Clore, G.M., Gros, P., Grosse-Kunstleve, R.W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N.S., Read, R.J., Rice, L.M., Simonson, T., Warren, G.L. (1998). Crystallography and NMR system (CNS): A new software system for macromolecular structure determination, *Acta Cryst. D*, in press.
- Burling, F.T. and Brunger, A.T. (1994). Thermal motion and conformational disorder in protein crystal structures: comparison of multi-conformer and time-averaging models. *Israel Journal of Chemistry* **34**, 165-175.
- Burling, F.T., Weis, W.I., Flaherty, K.M. and Brunger, A.T. (1996). Direct observation of protein solvation and discrete disorder with experimental crystallographic phases, *Science* **271**, 72-77.
- Diamond, R. (1971). A real-space refinement procedure for proteins. *Acta Cryst. A* **27**, 436-452.
- Engh, R.A. and Huber, R. (1991). Accurate bond and angle parameters for X-ray structure refinement. *Acta Cryst. A* **47**, 392-400.
- Fujinaga, M., Gros, P. & van Gunsteren, W.F. (1989). Testing the method of crystallographic refinement using molecular dynamics. *J. Appl. Cryst.* **22**, 1-8.

Goldstein, H. (1980). *Classical Mechanics* 2nd ed., Addison-Wesley Pub. Co., Reading, Massachusetts.

Gros, P., van Gunsteren, W.F. and Hol, W.G.J. (1990). Inclusion of thermal motion in crystallographic structures by restrained molecular dynamics. *Science* **249**, 1149-1152

Jain, A., Vaidehi, N. and Rodriguez, G. (1993). A fast recursive algorithm for molecular dynamics simulation. *J. Comp. Phys.* **106**, 258-68.

Hendrickson, W.A. (1985). Stereochemically restrained refinement of macromolecular structures. *Meth. Enzymol.* **115**, 252-270.

Hoppe, W. (1957). Die Faltmolekülmethode - eine neue Methode zur Bestimmung der Kristallstruktur bei Ganz oder Teilweise bekannter Molekülstruktur. *Acta Cryst.* **10**, 750-751.

Hsu, I.N., Delbaere, L.T.J., James, M.N.G. & Hoffman, T. (1977). Penicillopepsin from *Penicillium janthinellum* crystal structure at 2.8 Å and sequence homology with porcine pepsin. *Nature* **266**, 140-145.

Kirkpatrick, S., Gelatt, C.D. and Vecchi, Jr., M.P., (1983). Optimization by simulated annealing. *Science* **220**, 671-680.

Kleywegt, G.J. and Brunger, A.T. (1996). Cross-validation in crystallography: practice and applications. *Structure* **4**, 897-904.

Kuriyan, J., Petsko, G.A., Levy, R.M. and Karplus, M. (1986). Effect of anisotropy and anharmonicity on protein crystallographic refinement. *J. Mol. Biol.* **190**, 227-254.

Kuriyan, J., Brunger, A.T., Karplus, M., & Hendrickson, W.A. (1989). X-ray refinement of protein structures by simulated annealing: test of the method on

myohemerythrin, *Acta Cryst.* **A45**, 396–409.

Kuriyan, J., Ösapay, K., Burley, S.K., Brunger, A.T., Hendrickson, W.A. and Karplus, M. (1991). Exploration of disorder in protein structures by X-ray restrained molecular dynamics. *Proteins* **10**, 340-358.

Laarhoven, P.J.M. and Aarts, E.H.L.(eds.) (1987). Simulated annealing: Theory and applications. Dordrecht: D. Reidel Publishing Company, pp. 187.

Luzzati, V. (1952). Traitement statistique des erreurs dans la détermination des structures cristallines. *Acta Crystallogr.* **5**, 802–810.

Metropolis, N., Rosenbluth, M., Rosenbluth, A. Teller, A. and Teller, E. (1953). Equation of state calculations by fast computing machines. *J. Chem. Phys.* **21**, 1087-1092.

Murshudov, G.N., Vagin, A.A., & Dodson, E.J. (1997). Refinement of macromolecular structures by the maximum-likelihood method. *Acta Cryst. D* **53**, 240–255.

Pannu, N.S. and Read, R.J. (1996). Improved structure refinement through maximum likelihood. *Acta Cryst. A* **52**, 659-668.

Pannu, N.S., Murshudov, G.N., Dodson, E.J., & Read, R.J. (1998). Incorporation of prior phase information strengthens maximum likelihood structural refinement. *Acta Cryst. D*, in press.

Parkinson, G., Vojtechovsky, J., Clowney, L., Brunger, A.T. and Berman, H.M. (1996). New parameters for the refinement of nucleic acid containing structures. *Acta Cryst. D* **52**, 57-64.

Pearlman, D.A. and Kim, S.-H. (1990). Atomic charges for DNA constituents derived from single-crystal X-ray diffraction data. *J. Mol. Biol.* **211**, 171-187.

- Press, W.H., Flannery, B.P., Teukolosky, S. A. and Vetterling, W.T. (eds.) (1986). Numerical recipes, Cambridge, Cambridge University Press, 498-546.
- Read, R.J. (1986). Improved Fourier coefficients for maps using phases from partial structures with errors. *Acta Crystallogr. A* **42**, 140-149.
- Read, R.J. (1990). Structure-factor probabilities for related structures. *Acta Cryst. A* **46**, 900-912.
- Read, R.J. (1997). Model phases: probabilities and bias. *Meth. Enzymol.* **278**, 110-128.
- Rice, L.M. and Brunger, A.T. (1994). Torsion angle dynamics: reduced variable conformational sampling enhances crystallographic structure refinement. *Proteins: Structure, Function, and Genetics* **19**, 277-290.
- Rossmann, M.G. and Blow, D.M. (1962). The detection of sub-units within the crystallographic asymmetric unit. *Acta Cryst. A* **15**, 24-51.
- Silva, A.M. and Rossmann, M.G. (1985). The refinement of southern bean mosaic virus in reciprocal space. *Acta Cryst. B* **41**, 147-157.
- Sim, G.A. (1959). The distribution of phase angles for structures containing heavy atoms. II. A modification of the normal heavy-atom method for non-centrosymmetric structures. *Acta Cryst.* **12**, 813-815.
- Srinivasan, R. (1966). Weighting functions for the use in the early stages of structure analysis when a part of the structure is known. *Acta Cryst.* **19**, 143-144.
- Suguna, K, Bott, R.R., Padlan, E.A., Subramanian, E., Sheriff, S., Cohen, G.H., and Davies, D.R. (1987). Structure and refinement at 1.8 Å resolution of the aspartic proteinase from *Rhizopus chinensis*. *J. Mol. Biol.* **196**, 877-900.

Verlet, L. (1967). Computer Experiments on classical fluids. I. Thermodynamical properties of Lennard-Jones molecules. *Phys. Rev.* **159**, 98-105.

Figure Captions

Figure 1: Effect of resolution on coordinate-error estimates: accuracy as a function of resolution. Refinements were begun with the crystal structure of penicillopepsin (Hsu *et al.*, 1977) with water molecules omitted and with uniform temperature factors. The low-resolution limit was set to 6 Å. Inclusion of all low resolution diffraction data does not change the conclusions (Adams *et al.*, 1997). The penicillopepsin diffraction data were artificially truncated to the specified high-resolution limit. Each refinement consisted of simulated annealing using a Cartesian-space slow-cooling protocol starting at 2000 K, overall B-factor refinement, and individual restrained B-factor refinement. All refinements were carried out with 10% of the diffraction data randomly omitted for cross-validation. (a) Coordinate-error estimates of the refined structures using the methods of Luzzati (1952) and Read (1986). All observed diffraction data were used, i.e., no cross-validation was performed. The actual coordinate errors (rms. differences to the original crystal structure) are shown for comparison. (b) Cross-validated coordinate-error estimates. The test set was used to compute the coordinate-error estimates (Kleywegt & Brunger, 1996).

Figure 2: The Gaussian probability distribution forms the basis of maximum likelihood targets in crystallographic refinement. The conditional probability of the true structure factor \mathbf{F} given model structure factors is a Gaussian in the complex plane (Eq. 3). The expected value of the probability distribution is $D\mathbf{F}_c$ with variance σ_Δ , where D and σ_Δ account for missing or incorrectly placed atoms in the model.

Figure 3: Illustration of simulated annealing for minimization of a one-dimensional function. The kinetic energy of the system (a “ball” rolling on the one-dimensional surface) allows local conformational transitions with barriers smaller than the kinetic energy. If a larger drop in energy is encountered the excess kinetic energy is dissipated. It is thus unlikely that the system can climb out of the global minimum once it has reached it.

Figure 4: Simulated annealing produces better models than extensive conjugate gradient minimization. Map correlation coefficients were computed before and after refinement against the native penicillopepsin diffraction data (Hsu *et al.*, 1977) for the poly-alanine model derived from Rhizopuspepsin (Suguna *et al.*, 1987, PDB code 2APR). Correlation coefficients are between σ_A weighted maps calculated from each model and from the published penicillopepsin structure. The observed penicillopepsin diffraction data was in spacegroup C2 with cell dimensions $a=97.37\text{\AA}$, $b=46.64\text{\AA}$, $c=65.47\text{\AA}$, and $\beta=115.4^\circ$. All refinements were carried out using diffraction data from the lowest resolution limit of 22.0\AA up to 2.0\AA . The MLHL refinements used single isomorphous phases from a $\text{K}_3\text{UO}_2\text{F}_5$ derivative of the penicillopepsin crystal structure, which covered a resolution range of 22.0\AA to 2.8\AA . Simulated annealing refinements were repeated 5 times with different initial velocities. The numerical averages of the map correlation coefficients for the 5 refinements are shown as the hashed bars. The best map correlation coefficients from simulated annealing are shown as the white bars.

Figure 5: Maximum likelihood targets significantly decrease model bias in simulated annealing refinement. σ_A -weighted electron density maps contoured at 1.25σ for models from simulated annealing refinement with different targets. Residues 233 to 237 are shown with the published penicillopepsin crystal structure (Hsu *et al.*, 1977) in solid lines, and the model with the lowest free R value from 5 independent refinements in dashed lines.

Fig. 1

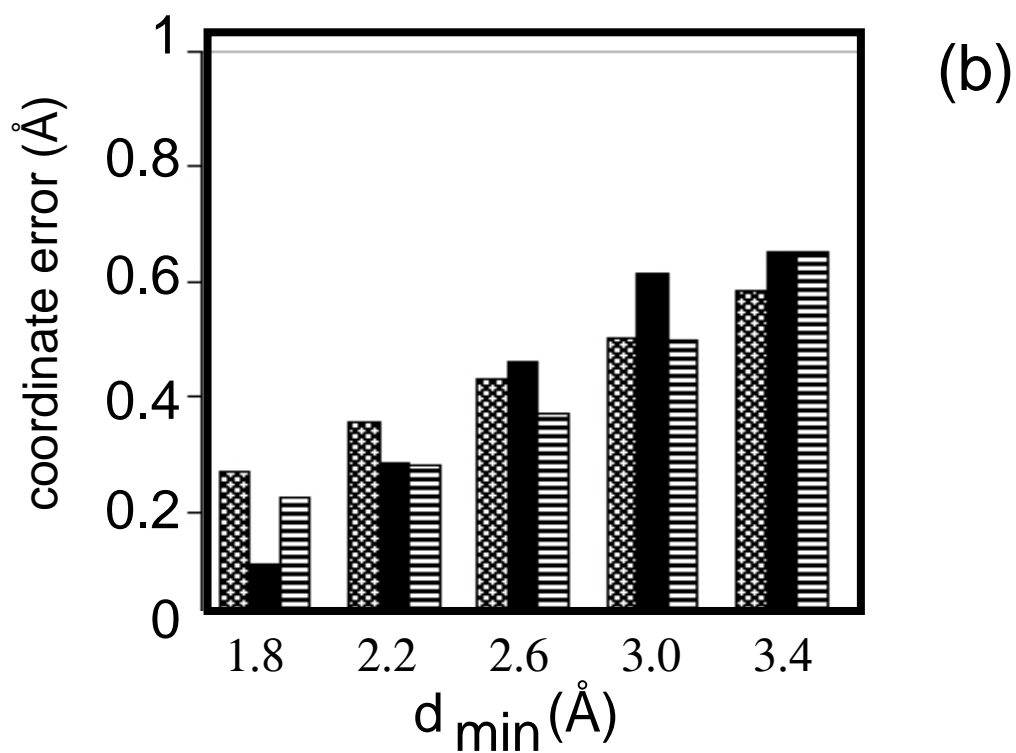
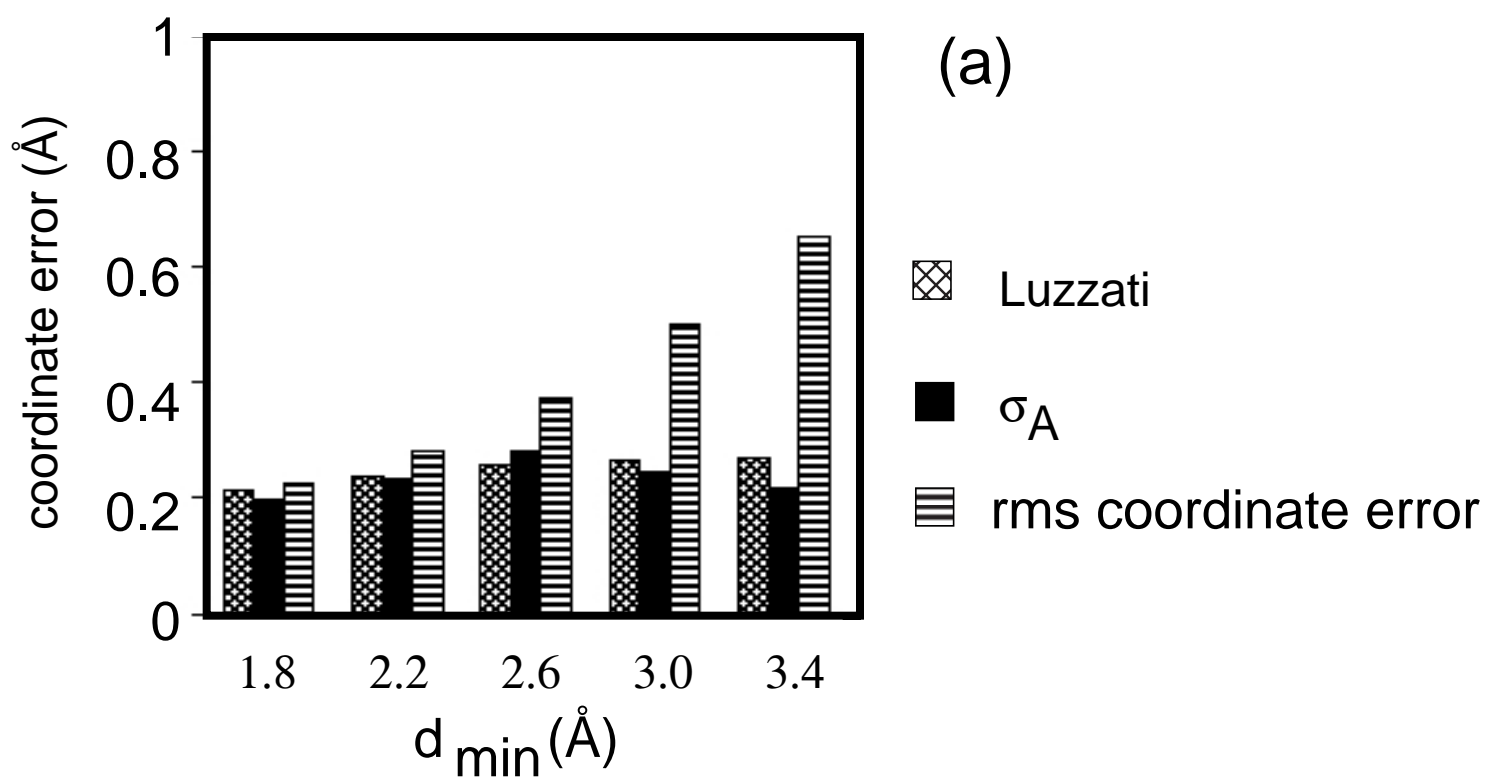


Fig. 2

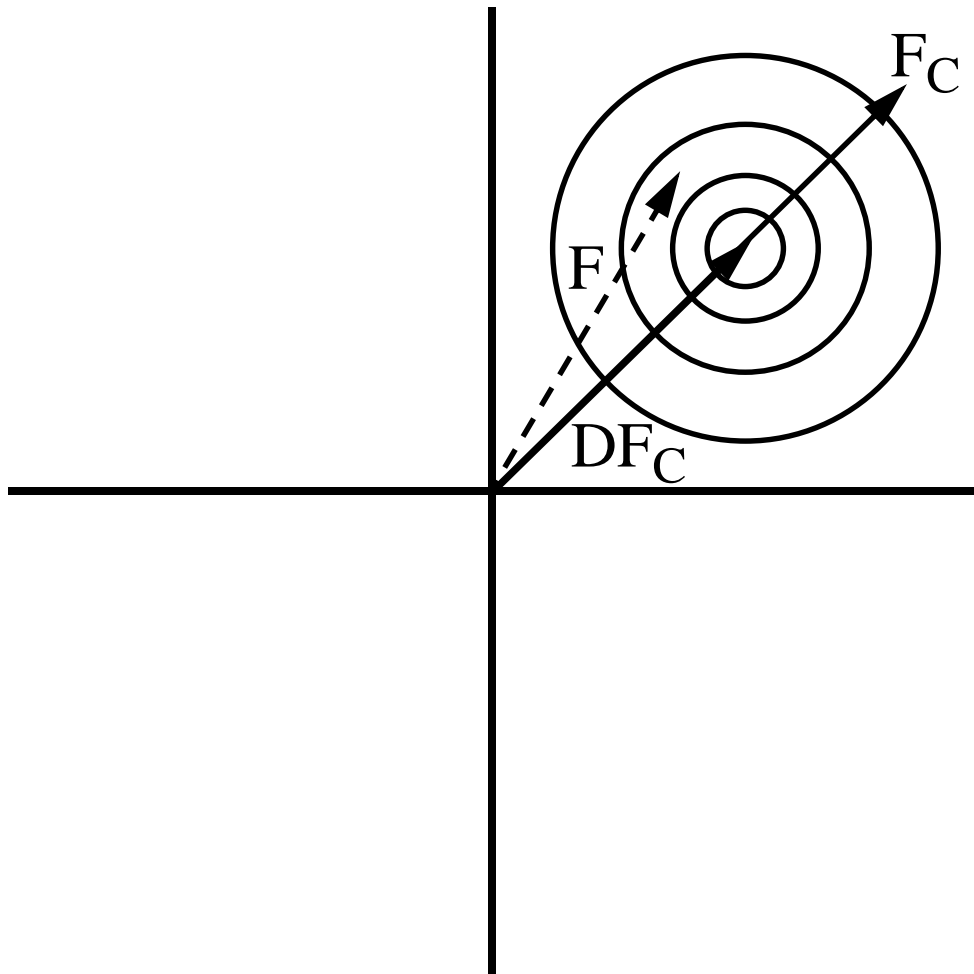


Fig. 3

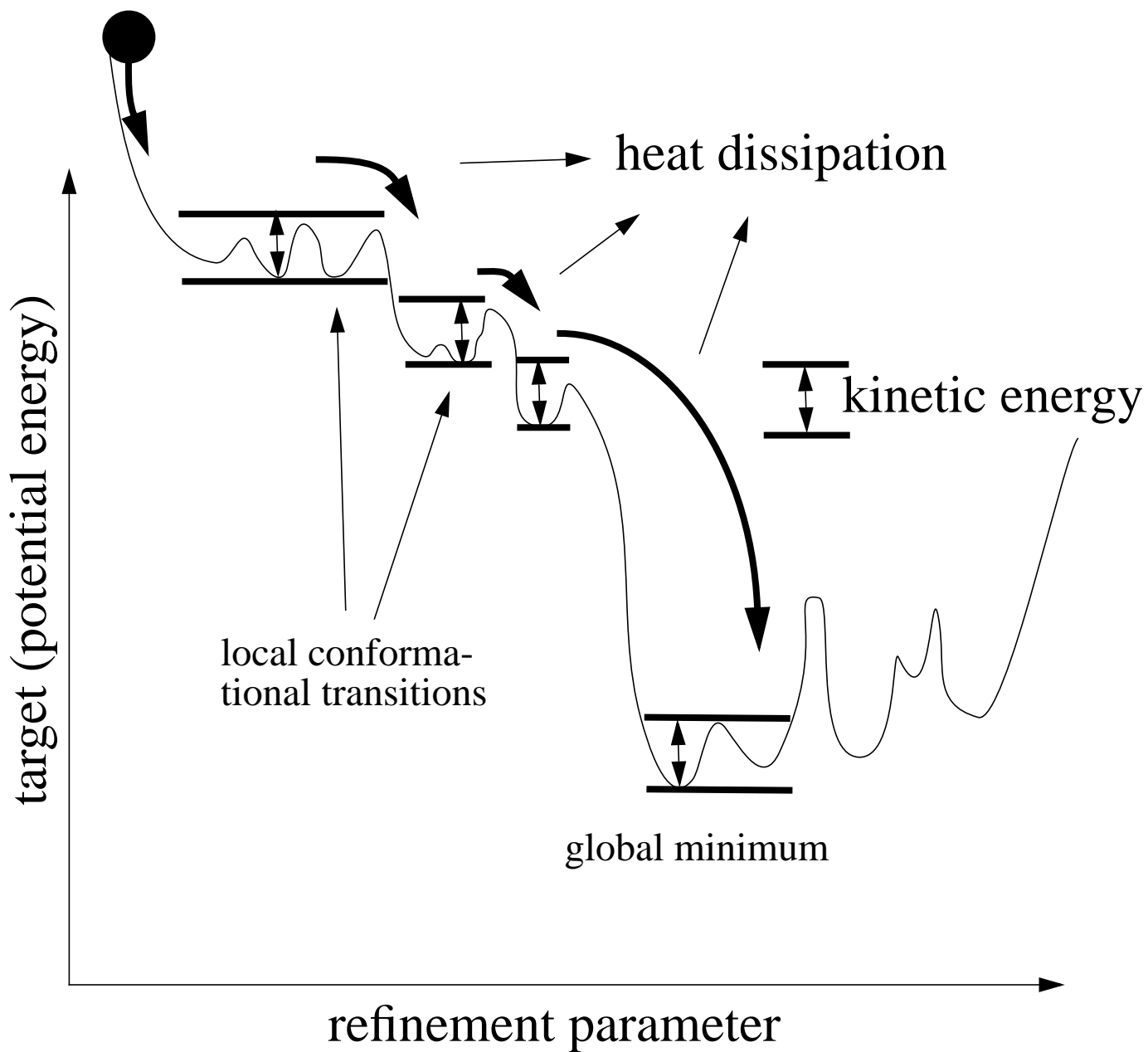


Fig. 4

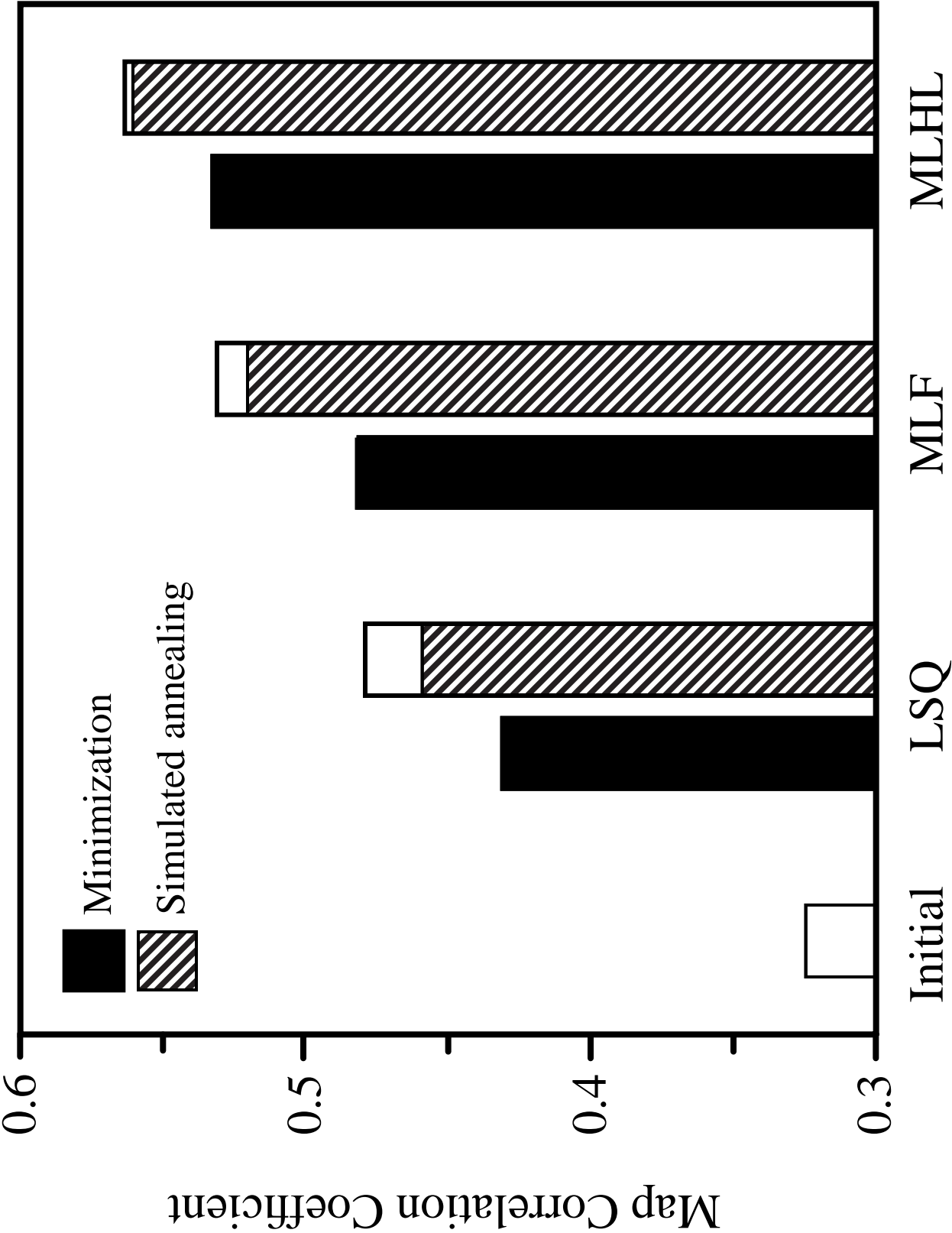
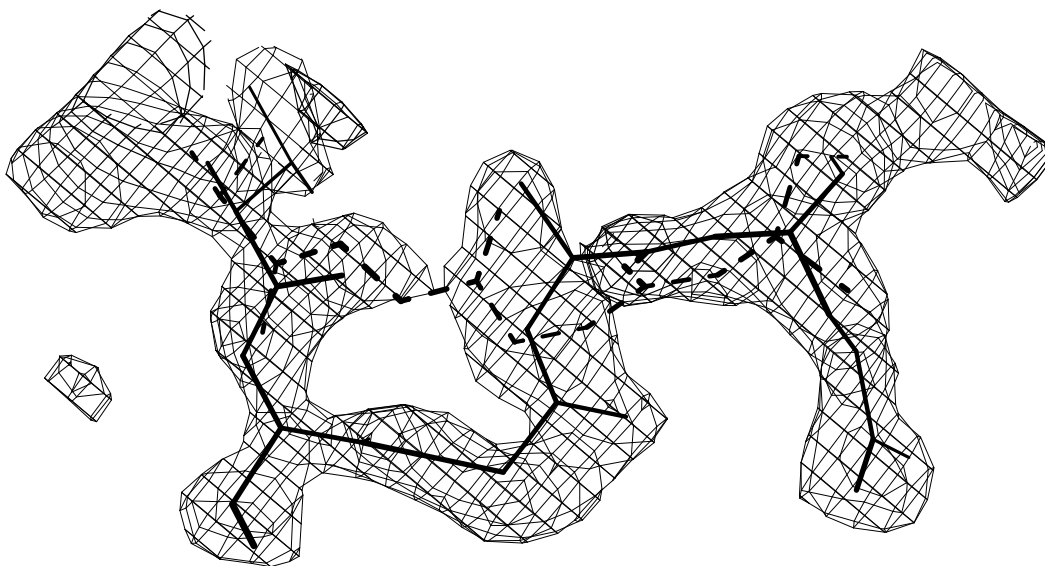
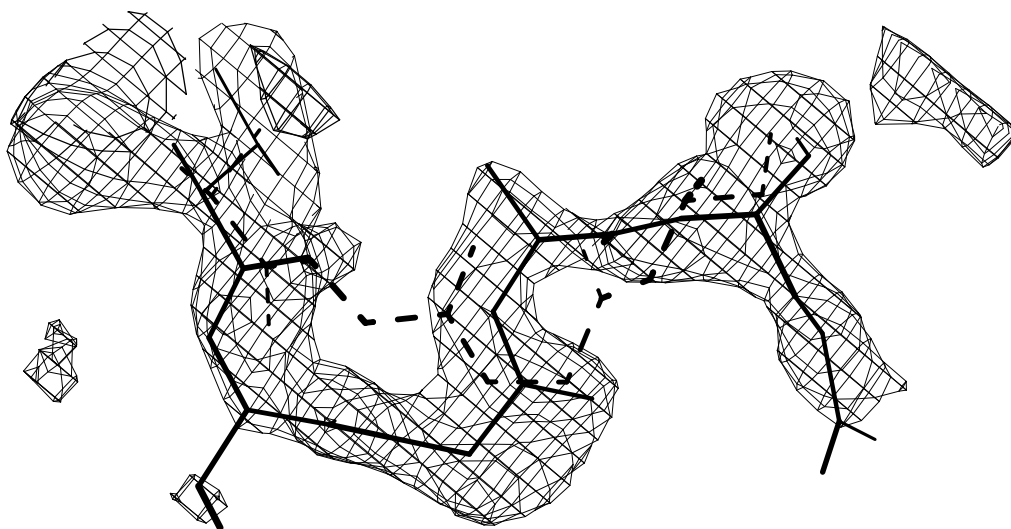


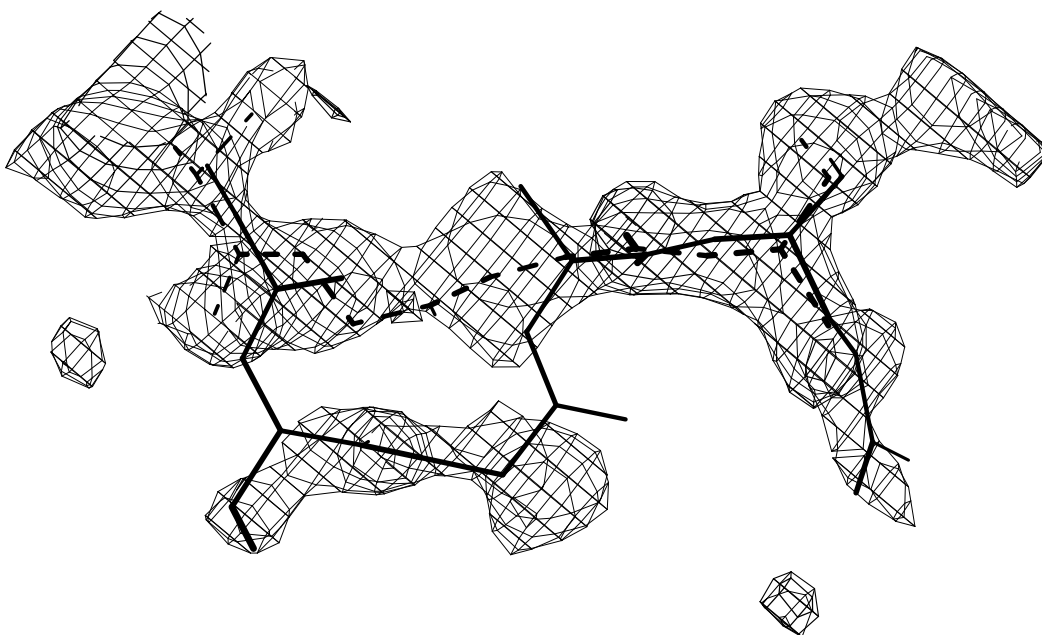
Fig. 5



MLHL



MLF



LSQ