Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

SOMoRe: a multi-dimensional search and optimization approach to molecular replacement

Diane C. Jamrog,[a]* Yin Zhang[a] and George N. Phillips Jr[b]

[a]Department of Computational and Applied Mathematics, Rice University, Houston, Texas, USA, and [b]Departments of Biochemistry and Computer Sciences, University of Wisconsin, Madison, Wisconsin, USA

Correspondence e-mail: djamrog@alumni.rice.edu

Received 29 July 2002
Accepted 26 November 2002

Commonly used traditional molecular-replacement (MR) methods, though often successful, have difficulty solving certain classes of MR problems. In addition, MR problems are generally very difficult global optimization problems because of the enormous number of local minima in traditionally computed target functions. As a result, a new MR program called *SOMoRe* is introduced that implements a new global optimization strategy that has two major components: (i) a six-dimensional global search of a target function computed from low-resolution data and (ii) multi-start local optimization. Because the target function computed from low-resolution data is relatively smooth, the global search can coarsely sample the MR variable space to identify good starting points for extensive multi-start local optimization. Consequently, *SOMoRe* was able to straightforwardly solve four realistic test problems, including two that could not be directly solved by traditional MR programs, and *SOMoRe* solved a problem using a less complete model than those required by two traditional programs and a stochastic six-dimensional program. Based on these results, this new strategy promises to extend the applicability and robustness of MR.

1. Introduction

When X-ray crystallography is used to determine the structure of a protein molecule, the phase problem must be overcome. One method for attacking the phase problem is to solve the more tractable molecular-replacement (MR) problem, which is a nonlinear optimization problem with dimensions much smaller than the phase problem. The goal of an MR method is to find the rotations and translations of a given protein model that produce calculated intensities closest to those observed from a crystal with unknown atomic structure. In general, the use of MR methods is expected to increase as more structures are solved and deposited in the Protein Data Bank (PDB; Berman *et al.*, 2000), because it will be more likely that an accurate model protein will be available for a given MR problem. However, research is still needed to develop more reliable and robust MR methods.

2. Current approaches

Currently, there are two basic sets of approaches for solving the MR problem: (i) traditional approaches, which separately optimize the rotational and translational degrees of freedom of the protein model in the new unit cell, and (ii) higher dimensional approaches, either six-dimensional (6D) or $6n$-dimensional ($6n$D), which simultaneously optimize these degrees of freedom for one or n copies of the protein model.

In 1962, Rossmann and Blow proposed that the MR problem be solved first by a search for optimal rotations of the model and then by a separate search for optimal translations of the oriented model (Rossmann & Blow, 1962). However, there are two main drawbacks to this approach. Firstly, for more difficult MR problems the lowest valued local minima of traditional rotation functions often do not represent the true rotation component of a MR solution (Jogl *et al.*, 2001; Kissinger *et al.*, 1999; Tong, 1996). If the optimal rotation is not found, then the optimal translation cannot be found and the traditional method fails. Secondly, the approximations inherent in traditional methods typically require models that are structurally very similar to the target protein whose structure is to be determined (Brünger, 1990, 1993, 1997).

Traditional methods are known to encounter difficulty either when the molecules are tightly packed in the crystal (Chang & Lewis, 1997; Glykos & Kokkinidis, 2000; Sheriff *et al.*, 1999; Tong, 1996), when the crystallized molecule has an elongated shape (Baker *et al.*, 1995; Chang & Lewis, 1997; Glykos & Kokkinidis, 2001) or when there are many molecules in the crystal unit cell (Baker *et al.*, 1995; Glykos & Kokkinidis, 2000; Tong, 1996). These problems result primarily because the rotation and translation variables are not optimized simultaneously in the traditional formulation. When either the molecules are tightly packed or the molecule has an elongated shape, the self-vectors and cross-vectors that characterize intensity-based target functions cannot be conveniently separated and interpretation of both the rotation and translation functions can be difficult (Chang & Lewis, 1997). In the third case, when there are many molecules in the crystal unit cell, traditional methods encounter difficulty because the ratio of the number of cross-vectors to self-vectors increases. An exception is when the molecules exhibit known high non-crystallographic symmetry which can be exploited (Tong & Rossmann, 1990).

To avoid the drawbacks associated with separately optimizing the orientation and position of the model, parallelized 6D searches (Sheriff *et al.*, 1999) and both 6D and 6*n*D stochastic optimization approaches have been proposed (Chang & Lewis, 1997; Kissinger *et al.*, 1999; Glykos & Kokkinidis, 2000). In contrast to traditional methods, 6D methods simulate scattering from the symmetry mates in the unit cell because the extra translational degrees of freedom fix the unit-cell origin, thereby allowing the symmetry mates to be positioned relative to each other. Several researchers have stated that the failures of traditional approaches are a consequence of their inability to model all Patterson vectors (Chang & Lewis, 1997; Kissinger *et al.*, 1999; Sheriff *et al.*, 1999).

3. New global optimization strategy

We implement a 6D deterministic strategy because we believe that a deterministic approach is generally more reliable than a stochastic one. However, we want a 6D approach that does not finely sample the MR variable space, because unless such a search is massively parallelized, the run times are still often prohibitive. Current deterministic 6D approaches must finely sample the variable space because of the many local minima in the MR target functions that are used to measure the disagreement between the observed and calculated intensities. The enormous number of local minima result because usually primarily medium- to high-resolution intensities are used to compute the target function.

This effect on the target function can be seen by noticing that the higher resolution intensities determine the highest frequency of the complex exponential function of a structure factor and therefore the rough landscape of the target function. The definition of the structure factor occurring at **h** is

$$F^c_{\mathbf{h}}(\Theta,\mathbf{t}) = \sum_{g=1}^{G}\sum_{j=1}^{N} f_j[d^*(\mathbf{h})]\exp\left(2\pi i\mathbf{h}\cdot\{S_g[A^{-1}\Omega(\Theta)A\mathbf{x}_j+\mathbf{t}]+s_g\}\right), \tag{1}$$

where $\Theta = (\theta_1, \theta_2, \theta_3)$ is a vector of angles, **t** is a translation in fractional coordinates, $\Omega(\Theta)$ is an orthonormal rotation matrix, G is the number of symmetry operators, N is the number of atoms in the model, $f_j[d^*(\mathbf{h})]$ is the atomic scattering factor for the jth atom for the given lattice point **h**, S_g and s_g represent the gth crystallographic symmetry operator, $\mathbf{x}_j = (x_j, y_j, z_j)^T$ are fractional coordinates of the jth atom, A^{-1} is a matrix that converts orthogonal real-space coordinates to fractional ones and A converts fractional real-space coordinates to orthogonal ones. The corresponding calculated intensity is $I^c_{\mathbf{h}} \equiv I^c_{\mathbf{h}}(\Theta,\mathbf{t}) = |F^c_{\mathbf{h}}(\Theta,\mathbf{t})|^2$. Thus, the 'angle',

$$\omega^{gj}_{\mathbf{h}} = 2\pi\mathbf{h}\cdot\{S_g[A^{-1}\Omega(\Theta)A\mathbf{x}_j+\mathbf{t}]+s_g\},$$

shows that the frequency of $I^c_{\mathbf{h}}$ is determined by **h**. The farther **h** is from the origin, the higher the spatial frequency of $F_{\mathbf{h}}$. This frequency determines the overall landscape of the target function and therefore the amount of local convexity about the global minima and the 'radius of convergence' of a local optimization method.

3.1. A surrogate low-resolution target function

Our new strategy will initially use a target function that has been computed from primarily low-resolution intensities. In general, such a low-frequency target function will have a smoother landscape and the 6D grid search can be coarser than the search required for a target function computed from high-resolution data. The correct balance between the accuracy of this function and the coarseness of the global search was a primary topic of investigation that was determined experimentally by varying the high-resolution cutoff of this function's data set. Throughout the text, we will refer to this low-frequency function that is used only during the coarse global search as the surrogate low-resolution function.

3.2. The new strategy

The new strategy has two major components: a coarse 6D global search and multi-start local optimization. Firstly, a global search coarsely samples the surrogate low-resolution target function to identify good starting points for multi-start local optimization. Next, starting from the points with the lowest function values, multi-start local optimization is initiated using a more accurate target function computed from a more complete higher resolution set of intensities. As a result, local optimization efforts will be focused on regions of the MR variable space where MR solutions are more likely to exist, in contrast to traditional methods or 6D searches that exhaustively search a uniformly fine grid, and in contrast to 6D stochastic methods that randomly sample the variable space. Finally, all the 6D local minima are ranked according to their function values and some post-processing is performed.

Step 0. (*a*) Choose a low-resolution data set and a higher resolution one. (*b*) Define a coarse grid determined by some increments of the MR variables.

Step 1. Evaluate the surrogate low-resolution target function at every point in the coarse grid.

Step 2. Identify the 6D points that produce, for example, the 1000 lowest function values of the target function.

Step 3. Use these points as starting points for multi-start local optimization of the target function computed from the high-resolution data.

Step 4. Perform post-processing: examination of free values and crystallographic packing checks.

In general, the 6D formulation of the MR problem is used rather than the traditional one not only because the calculated intensities will be more accurate, but also because traditional rotation functions appear to perform poorly when low-resolution data are used. Brünger and coworkers report that if predominantly low-resolution data are used, then the global minima of a commonly used traditional rotation function are unlikely to correspond to MR solutions (Brünger, 1997; DeLano & Brünger, 1995). In addition, Jamrog (2002) demonstrates that a commonly used 6D target function can be more accurate than its traditional counterpart when low-resolution data are used.

3.3. Novel aspects of the approach

The new approach is very different from mainstream MR approaches because it initially uses a low-frequency target function. Currently, coarse low-resolution global searches are atypical. Low-resolution data are not typically used because they can be more difficult to measure (Evans *et al.*, 2000; Miller *et al.*, 1999) and extra modeling is required to calculate low-resolution data because their magnitudes are affected by the crystal's bulk solvent; see, for example, Urzhumtsev & Podjarny (1995). In addition, low-resolution intensities have not provided computational success when used with the traditional approaches. A few researchers implementing higher dimensional searches advocate using low-resolution data so that coarser grid searches can be used (Rabinovich *et al.*, 1998; Rabinovich & Shakked, 1984). However, such higher dimensional coarse searches have not been adopted.

One possible reason such searches are not common is that a coarse grid search alone is unlikely to identify an MR solution. In such a search, it is most likely that there will not be a grid point close enough to a global minimum to produce a function value that stands out in comparison to the other function values computed. Clearly, a coarse sampling does not necessarily produce grid points close to a minimum. Therefore, extensive local optimization of, for example, between 500 and 1000 grid points is a necessary and integral component of the new strategy. As a result, the strategy shifts the emphasis from the grid search onto local optimization and therefore is also very different from current MR strategies.

4. Implementation

Firstly, we define our target function. Next, we define the variables used to define the coarse sampling of the MR variable space and finally we discuss the local optimization method and post-processing.

4.1. Target function: correlation coefficient

We use the following standard linear correlation coefficient as our target function,

$$C(I^c, I^o) = \frac{\sum_{\mathbf{h}}(I^c_{\mathbf{h}} - \langle I^c\rangle)(I^o_{\mathbf{h}} - \langle I^o\rangle)}{\left[\sum_{\mathbf{h}}(I^c_{\mathbf{h}} - \langle I^c\rangle)^2\right]^{1/2}\left[\sum_{\mathbf{h}}(I^o_{\mathbf{h}} - \langle I^o\rangle)^2\right]^{1/2}}, \quad (2)$$

where $I^o_{\mathbf{h}}$ and $I^c_{\mathbf{h}}$ are the observed and calculated intensities occurring at the reciprocal-lattice points **h** and $\langle I^o\rangle$ and $\langle I^c\rangle$ are the average values of the observed and calculated intensities, respectively. Of course, $|F^o_{\mathbf{h}}|$ and $|F^c_{\mathbf{h}}|$ can also be used. This function is commonly used; see Grosse-Kunstleve & Adams (2001) and Navaza (2001) for examples.

4.2. Global search

The grid points in the global search are $p_i = (\Theta_i, \mathbf{t}_i)$, where Θ_i are Eulerian angles that are used to rotate the reciprocal lattice to compute the structure factors of the rotated model and $\mathbf{t}_i$ are translations of the model. The sampling of orientation space is in terms of Lattman angles (Lattman, 1972); Lattman angles are then converted to Eulerian angles because symmetry relationships among possible solutions are easily identified when the Eulerian angles are used. Lattman angles and the optimal Lattman sampling are used because they sample angular space more uniformly than Eulerian angles and because the number of Θ_i decreases by a factor of $2/\pi$ in comparison to a constant Eulerian sampling (Lattman, 1972). The definition of $\Delta\theta_2$ used by *SOMoRe* to determine the optimal sampling is

$$\Delta\theta_2 = 2\arcsin\{r_{\rm high}/[2(a+b+c)/3]\}, \quad (3)$$

where $r_{\rm high}$ is the high-resolution cutoff of the surrogate function's data set and a, b and c are the lengths of the unit-cell basis vectors. This definition is the same as that used by *CNS* v. 1.0 (Brünger *et al.*, 1998).

Similarly, the step sizes for translational variables are functions of the high-resolution cutoff,

$$\Delta t_x = r_{\rm high}/(3a), \quad \Delta t_y = r_{\rm high}/(3b), \quad \Delta t_z = r_{\rm high}/(3c). \quad (4)$$

Larger r_{high} values allow larger step sizes (lower frequencies in the surrogate function). These are the same step sizes implemented and justified by Brünger in *X-PLOR* version 3.1 (Brünger, 1992).

4.3. Local optimization

The local optimization method implemented in *SOMoRe* is the Broyden–Fletcher–Goldfarb–Shanno (BFGS) secant method (see, for example, Nocedal & Wright, 1999). We choose BFGS since an analytical expression is not readily available for the Hessian of $C(I^c, I^o)$ because an interpolation is utilized to determine a new set of structure factors for each orientation and translation in the 6D sampling, as described by Chang & Lewis (1997). For the same reason, we use a finite-difference approximation to the gradient. *Appendix A* contains the definition of the finite-difference gradient used by *SOMoRe*. We choose BFGS over the conjugate-gradient method because its convergence rate is superlinear rather than linear (see, for example, Nash & Sofer, 1996). See Jamrog (2002) for more information about the implementation of BFGS in *SOMoRe*.

4.4. Post-processing

There are two post-processing components: (i) examination of free values and (ii) crystallographic packing checks. A researcher primarily relies on a list of the lowest valued local minima. The greater the contrast between the function's values, the more confidence the researcher will have that the lowest valued points are solutions. However, we also use *free values* or function values that are computed from a randomly chosen 10% of the data set. These values are not used during the global search or local optimization. During the global search, the lowest M function values are retained regardless of the associated free values.

Points output by an MR code can be immediately dismissed if the packing of symmetry mates of the positioned model is bad; that is, they interpenetrate to a significant degree. To determine the degree of interpenetration, every intra-atomic distance between the symmetry mates is computed and then compared with a threshold to see if any *distance violations* or inter-atomic distances smaller than the threshold occur (Tong, 1996; Jamrog, 2002).

5. Criterion for evaluating the results

Because each test problem uses experimental data for which the crystal structure has been determined and deposited in the PDB, one measure to judge the output of *SOMoRe* is the root-mean-square deviation (RMSD) between the coordinates of the reoriented model and the coordinates of the known structure. A point identified by an MR method may map the model onto any one of the target's symmetry mates. Thus, we define the RMSD to be the minimum of all RMSDs computed between the target structure and the reoriented model's symmetry mates:

$$\mathrm{RMSD} = \min\left[\left(\left\{\sum_{j=1}^{N} ||A\mathbf{x}_j - A[S_g\mathbf{x}'_j(\Theta, \mathbf{t}) + s_g]||^2\right\}/N\right)^{1/2}, \quad g = 1, \ldots, G\right], \tag{5}$$

where $\mathbf{x}_j$ and $\mathbf{x}'_j(\theta, \mathbf{t})$ are the jth fractional coordinates of the target protein and the repositioned model, respectively.

Furthermore, the symmetry mates directly produced by the symmetry operators may be in unit cells other than the particular unit cell of the target structure. Thus, before calculating the RMSD, each symmetry mate should be moved an integer number of unit-cell basis-vector translations so that it is the closest symmetry mate of its kind to the target structure to ensure that the RMSD will be as small as possible. A pseudo code for determining the closest symmetry mate is given in *Appendix B*.

We note that RMSDs cited in the MR literature are computed by trying to determine the best fit between two given structures and should be close to the smallest RMSD possible. Thus, we call such an RMSD 'optimal' and judge the quality of $(\theta, \mathbf{t})$ by comparing the RMSD defined by (5) with the optimal one.

6. Test problems and results

We analyze the numerical results produced by *SOMoRe* on four test problems: a problem with a very good model, two problems that have either defeated or severely challenged traditional MR software and finally a problem with protein models that range from being complete to only 37% complete.

6.1. General description of experiments

All test problems were taken from articles that introduce new MR algorithms or software. For each test problem, two global searches were performed: one using all available data between ∞ and 8 Å resolution and another using all available data between ∞ and 10 Å. These two resolution ranges were chosen because the run times were reasonable and the results demonstrate that 8 Å appears to be a safe high-resolution cutoff. It is important to approximately determine the largest possible high-resolution cutoff that will allow the surrogate function to identify good starting points because the larger the high-resolution cutoff, the faster the run time. After each global search, local optimization is performed using data between ∞ and 4 Å.

We summarize the information for each test problem in Table 1. Because of crystallographic symmetry, the translation $\mathbf{t}$ is restricted to the subset of the unit cell listed in Table 1, known as the Cheshire-group unit cell (Hirshfeld, 1968), except for 6rhn. The range for 6rhn is not the Cheshire-group unit cell, but it is consistent with the space group. In addition, each problem was 6D because there was only one molecule in the asymmetric unit.

6.1.1. Parameters. *SOMoRe* is based on the fast target-function evaluation algorithm of *Qs* (Glykos & Kokkinidis, 2000), described by Chang & Lewis (1997). Like *Qs*, *SOMoRe*

Table 1
Relevant information for each test problem.

The size of the protein refers to its number of amino acids. The last column gives the lowest RMSDs computed for each test problem, showing that the new strategy successfully solved each problem.

Problem name		Size	Space group	No. symmetry operators	Resolution range (Å)	Translation range	'Optimal' RMSD	Computed RMSD
1aki	Lysozyme	129	$P2_12_12_1$	4	15–1.5	$\frac{1}{2}\mathbf{a} \times \frac{1}{2}\mathbf{b} \times \frac{1}{2}\mathbf{c}$	1.2	1.01
1cgn	Cytochrome c'	128	$P6_522$	12	20–2.2	$\mathbf{a} \times \mathbf{b} \times \frac{1}{2}\mathbf{c}$	1.3	1.38
1b6q	Helical bundle	56	$C222_1$	4	40.8–1.8	$\frac{1}{2}\mathbf{a} \times \frac{1}{2}\mathbf{b} \times \frac{1}{2}\mathbf{c}$	0.2	0.39
6rhn	Histidine	126	$P4_32_12$	8	35.1–2.2	$\frac{1}{2}\mathbf{a} \times \frac{1}{2}\mathbf{b} \times \frac{1}{4}\mathbf{c}$	0.3	0.32

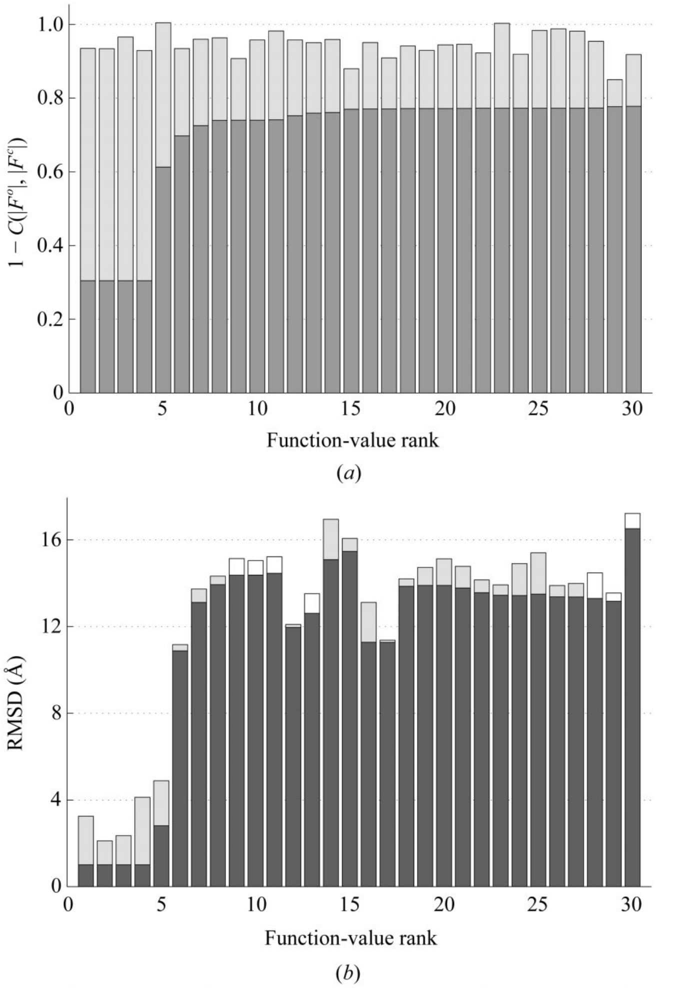

Figure 1
Local optimization results from an 8 Å global search for 1aki. Bar charts indicating function values (*a*) and RMSDs (*b*) before (light gray) and after optimization at 4 Å (dark gray). The contrast in function value between the first four points and the fifth is an accurate indicator that the remaining local minima are not solutions. A white bar in (*b*) indicates that the RMSD increased as a result of optimization by the height of the white bar; that is, the value before optimization would be the height of the dark gray bar minus the height of the white bar. Most importantly, an increase in RMSD has not been observed when a starting point is close to a solution.

has several parameters that affect the accuracy and efficiency of the structure-factor calculations. For all experiments, the interpolation scheme was linear and the size of the model's artificial unit cell was increased by a factor of 4.0 to accurately sample the model's Fourier transform. In addition, all structure factors were scaled by an exponential term to simulate the bulk-solvent environment of the crystal: $1 - k_{\rm SOL} \exp[B_{\rm SOL}(d^*\mathbf{h})^2/4)]$, where $k_{\rm SOL} = 0.785$ and $B_{\rm SOL} = 205.0$ (Glykos & Kokkinidis, 2001). Furthermore, all calculated intensities were scaled by the linear scale factor $\alpha = \sum_{\mathbf{h}} |F^o_{\mathbf{h}}|^k / \sum_{\mathbf{h}} |F^c_{\mathbf{h}}|^k$. This scale factor appears to work well with the current computation of the finite-difference gradient.

6.1.2. Terminology. When we refer to an '8 Å search' or a '10 Å search', we are referring to the global search using all available data between ∞ and 8 Å and ∞ and 10 Å, respectively. Secondly, we call a global search 'successful' if after optimization a minimum is found that has an associated RMSD close to the 'optimal' one. In addition, it is 'successful' only if this RMSD is associated with a target-function minimum that is either the lowest valued minimum or the lowest valued minimum after other minima are ruled out. A minimizer that produces a repositioned model that has an RMSD that is within 1 Å of the optimal RMSD is considered to be a solution.

6.2. Results for test problem 1aki

We present the results for test problem 1aki because experience was gained in determining the appropriate high-resolution cutoff of the surrogate function's data set. This problem appears in Glykos & Kokkinidis (2000). The data are the observed intensities deposited with the coordinates of a chicken egg-white lysozyme (PDB code 1aki). The model is quail lysozyme (PDB code 2ihl), reported to have an optimal RMSD of 1.2 Å from 1aki (Glykos & Kokkinidis, 2000).

The 8 Å global search using $1 - C(I^o, I^c)$ was a success. However, the 10 Å global search was not. During the 8 Å search, the points that produced the 1000 lowest function values were identified for use as starting points for local optimization. Of these points, the closest grid point to a global minimum was the 108th lowest valued point because this point had the lowest associated RMSD of 2.11 Å. This point is 108th primarily because the grid search is coarse, but also because the surrogate low-resolution function is not as accurate as a target function computed using more higher resolution data. However, the grid point is close enough for the new strategy to succeed. During local optimization, the starting points with the four lowest RMSDs converged to local minima with associated RMSDs of 1.0 Å; when the 1000 local minima are ranked according to their function values, the minima with associated RMSDs of 1.0 Å are at the top of the list.

The leftmost bar chart in Fig. 1(*a*) shows the function values of the starting points (light gray bars) that converge to the 30 lowest valued minima and the function values of the 30 minima (dark gray bars). Several starting points converged to the same global minima. The rightmost bar chart in Fig. 1(*b*)

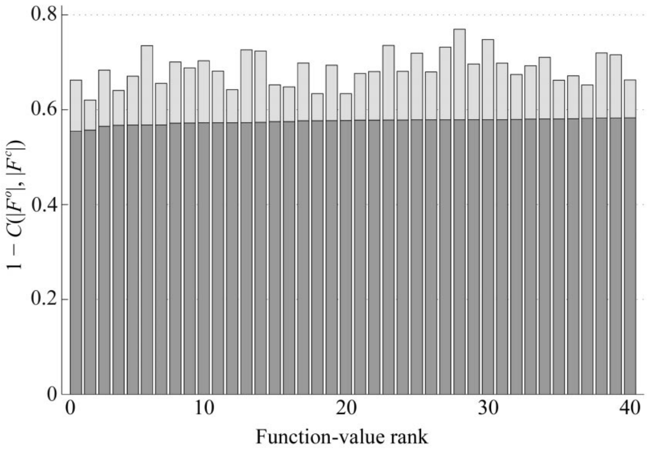

Figure 2
Function values before (light gray) and after (dark gray) optimization at 4 Å resolution of 8 Å resolution global search results for 1cgn.

shows the corresponding RMSDs, demonstrating that the lowest valued local minima are solutions.

6.3. Results for two difficult problems

These test problems are problems that either could not be solved using traditional MR software or for which the solution to the problem was not immediately obvious using such software.

6.3.1. Test problem 1cgn. Problem 1cgn appears in Kissinger *et al.* (1999). The data are the observed intensities deposited with the coordinates of cytochrome c' from the bacteria *Alcaligenes denitrificans* (PDB code 1cgn). The model is the polyalanine part of cytochrome c' from the bacteria *Rhodospirillum molischianum*, specifically amino acids 3–125 plus the heme group (PDB code 2ccy). This model was one of the models used in the original structure determination.

We estimate the optimal RMSD between 2ccy and 1cgn to be roughly 1.3 Å, using the RMSD computed between 2ccy and 1cgo, which is 1.27 Å (Baker *et al.*, 1995), and the RMSD between the backbones of 1cgo and 1cgn, which is 0.17 Å (Dobbs *et al.*, 1996). The atoms we use in our RMSD calculation are similar to those used by Baker *et al.* (1995). To compute the RMSDs between 1cgn and 2ccy, amino acids 4–30, 41–57, 80–95 and 104–120 plus the heme from 1cgn were paired with amino acids 4–30, 40–56, 83–98 and 106–122 plus the heme from 2ccy, thereby excluding the three loop regions. *SEQUOIA* (Bruns *et al.*, 1999) finds the RMSD between 119 of the amino acids (that it finds to be equivalent) in 2ccy and 1cgn to be 1.9 Å, demonstrating the difference in the loops of the two structures.

The original structure determination was difficult. *X-PLOR* (Brünger, 1992) and the rotation function of *ALMN* (Collaborative Computational Project, Number 4, 1994) failed (Baker *et al.*, 1995). Rotation searches using *AMoRe* (Navaza, 1994), four different models and two different resolution ranges also failed to indicate a solution that Baker *et al.* (1995) hoped

> would appear in most, if not all, of the experiments. . . even if it was not necessarily the top solution in each case.

In the end, the problem was solved using *AMoRe*, but to do so required a lot of supplementary information, including the anomalous scattering of the Fe atom, density-modification methods and an isomorphous data set. In contrast, the solution is easily identifiable using *SOMoRe* without using any of the supplementary information.

6.3.2. 8 Å search for 1cgn. For this problem if $C(I^o, I^c)$ is used, then the 8 Å search is not successful. Based on some numerical results, we suspect that the target function $C(|F^o|, |F^c|)$ is likely to be more accurate than $C(I^o, I^c)$ when low-resolution data is used (Jamrog, 2002). The lowest RMSD calculated from the grid points with the lowest 10 000 function values was approximately 3.27 Å, too far from a solution to converge to it using local optimization. However, the problem can be solved using an 8 Å search and $C(|F^o|, |F^c|)$.

The solution can be found by first examining the free function values and then the distance violations from the packing checks. Fig. 2 shows the function values of the 40 lowest valued local minima. As shown in Fig. 3(*a*), 16 minima have low free values. The local minima with relatively high free values should be ruled out. Then, if the distance violations of these 16 minima are taken into consideration, every minima except two can be ruled out, as shown in Fig. 3(*b*). The other minima produce interpenetration of the symmetry mates that was detected by packing checks using a threshold of 2 Å. The last two remaining minima are solutions with RMSDs of 1.4 Å, as shown in Fig. 3(*c*).

6.3.3. 10 Å search for 1cgn. In contrast, a 10 Å search using $1 - C(I^o, I^c)$ was successful. When the same protocol is used to eliminate local minima that are not solutions, as before, only the solutions remain. We do not show the bar charts because they are very similar (Jamrog, 2002).

6.3.4. Test problem 1b6q. Test problem 1b6q appears in Glykos & Kokkinidis (2001). The data and accurate model were both kindly supplied to us by N. Glykos (Glykos & Kokkinidis, 1999). The target structure is a two-α-helical bundle (PDB code 1b6q). The search model is an 'essentially perfect' polyalanine model of the two helices that was refined by a simulated-annealing procedure (Glykos & Kokkinidis, 1999). In fact, the optimal RMSD between this model and 1b6q is reported to be less than 0.2 Å (Glykos & Kokkinidis, 2001).

However, Glykos & Kokkinidis (2001) report that even though the

> search model is exceptionally accurate and the data of high quality, conventional methods (program *MOLREP*) could not identify the correct solution during the default run.

This MR problem is more difficult for traditional approaches than for a 6D approach because the assumption that the cross-vectors and self-vectors are 'topologically separate' is false (Glykos & Kokkinidis, 2001). Because the protein consists of α-helices there will be long self-vectors and because the crystal contains only 30% solvent there will be short cross-vectors. Therefore, the standard trick of traditional approaches to choose an appropriate volume of integration to try to include only self-vectors and exclude cross-vectors will not work.

In contrast, *SOMoRe* efficiently finds a solution to this MR problem using either an 8 or 10 Å search. The function values of the lowest valued minima are shown in Fig. 4(*a*), while the corresponding RMSDs are shown in Fig. 4(*b*). Again, there is a jump in the correlation-coefficient values that distinguishes solutions from non-solutions. The bar charts for the 10 Å search results are very similar (Jamrog, 2002); therefore, we do not show them.

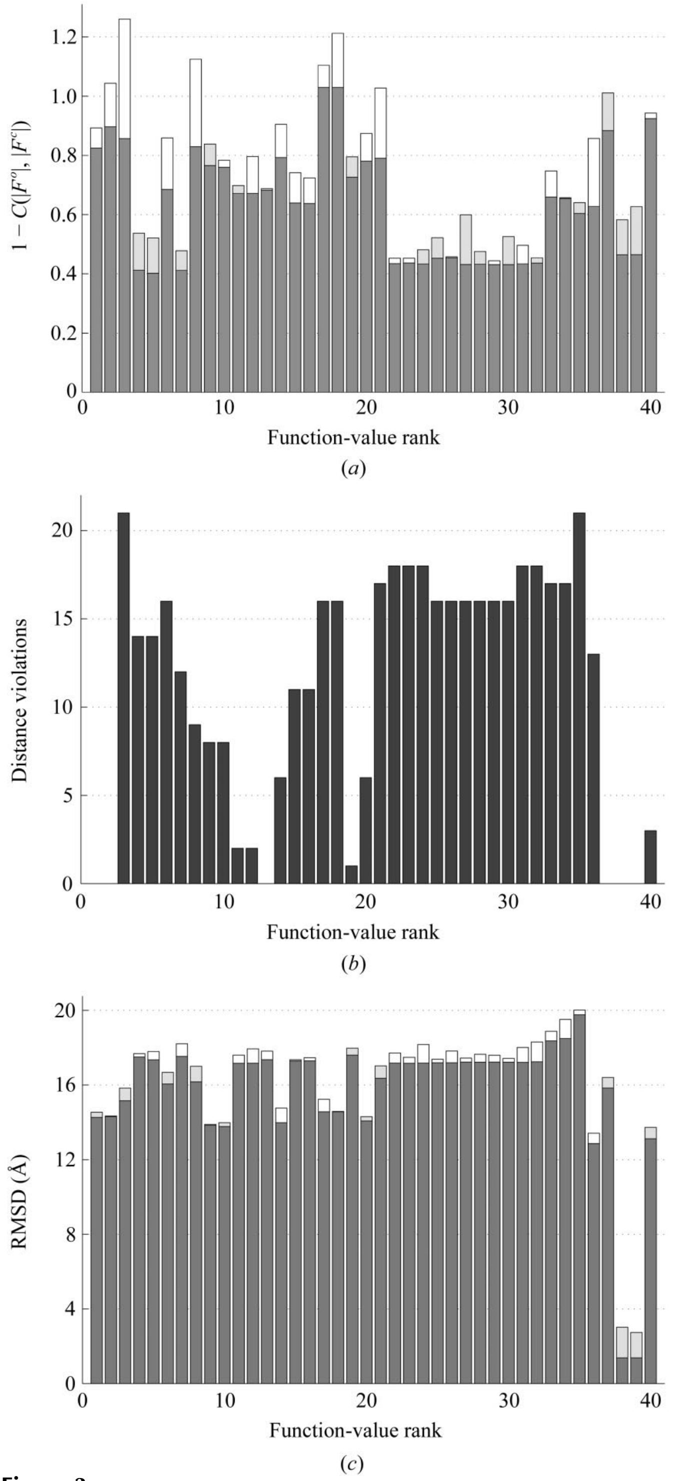

Figure 3
Results from 4 Å optimization of the initial 8 Å search results for 1cgn. (*a*) shows free values before optimization (light gray) and after optimization (dark gray). (*b*) shows the distance violations associated with each of the 40 lowest valued local minimum and (*c*) shows RMSDs before (light gray) and after optimization (dark gray). Again, a white bar indicates a value that has increased owing to optimization, such that the value before optimization is the height of the dark gray bar minus the height of the white bar.

6.4. Results for a problem using increasingly incomplete models

Test problem 6rhn is defined in Kissinger *et al.* (1999, 2001). This test problem was designed to determine how much of the model protein could be removed without preventing the MR problem from being solved. In the first article, the new 6D stochastic approach *EPMR* is compared with the traditional approaches *X-PLOR* and *AMoRe*. In the second article, the relationship between increased model truncation and decreased search efficiency of *EPMR* is discussed.

For this MR problem, the model is the polyalanine part of a rabbit histidine-triad nuclear-binding protein (PDB code 4rhn). The data are the experimentally observed structure-factor magnitudes deposited with the coordinates of the same protein that crystallized with different symmetry (PDB code

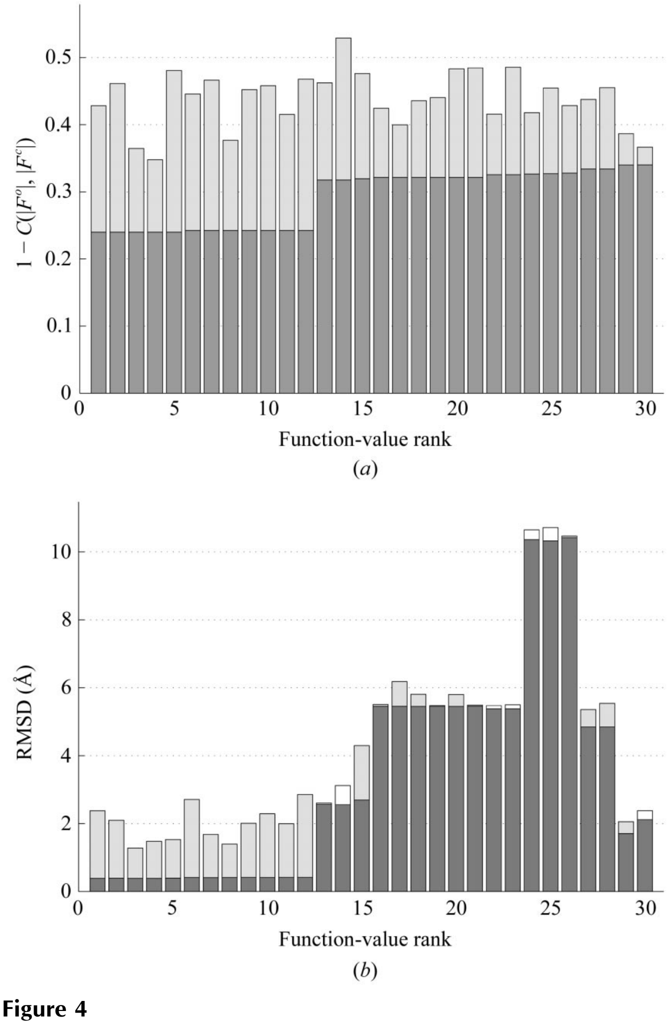

Figure 4
(*a*) Function values and (*b*) RMSDs before (light gray) and after optimization (dark gray) of the 8 Å global search results for 1b6q.

Table 2
The maximum amount of model truncation tolerated by the MR approaches where the model is the polyalanine part of 4rhn, which has 115 amino acids.

MR code	No. of amino acids in the least complete model	Truncation of the poly-Ala model (%)
SOMoRe	42	63
EPMR	44	62
X-PLOR	~62	~46
AMoRe	~67	~42

6rhn). The optimal RMSD between the polyalanine parts of 4rhn and 6rhn is cited as 0.3 Å (Kissinger *et al.*, 2001).

In the first article, an initial model, consisting of the polyalanine part of 4rhn containing 104 out of 115 residues, was truncated by five or six amino acids at a time from the C-terminal end (Kissinger *et al.*, 1999). In the second article, amino acids were removed from the polyalanine model one at a time until *EPMR* could not find a solution; that is, the highest correlation coefficient obtained after 100 searches by *EPMR* did not correspond to a solution. *SOMoRe* was similarly tested, using models that contained 104, 99, 93, 88, 82, 77, 71, 66, 60, 55, 49 and 44 residues, and then, because the 8 Å search was successful using the 44-residue model and the least complete model that *EPMR* could use to find a solution contained 44 residues, the 44-residue model was truncated one residue at a time until *SOMoRe* failed. For the 8 and 10 Å global searches, 1 − $C(|F_o|, |F_c|)$ was used. Kissinger *et al.* (1999) also used this target function. However, all available low-resolution data to 4 Å was used for local optimization, while Kissinger *et al.* (1999) used only the data in the resolution range 15–4 Å.

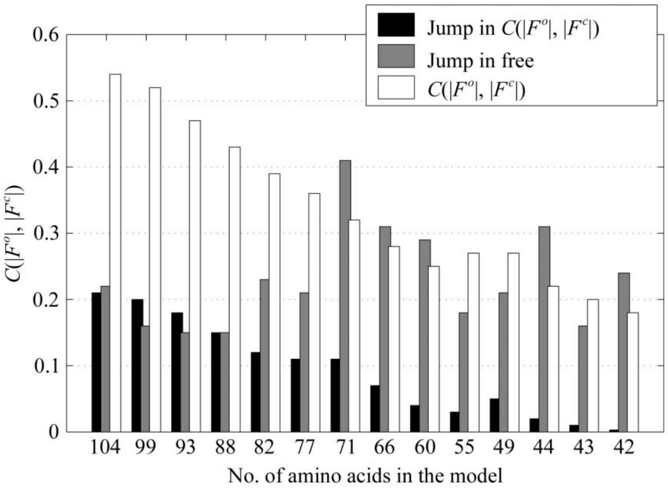

Figure 5
Bar charts showing the lowest function value found and the jumps in $C(|F_o| - |F_c|)$ and the free function values between solutions and non-solutions after optimization of the 8 Å search results.

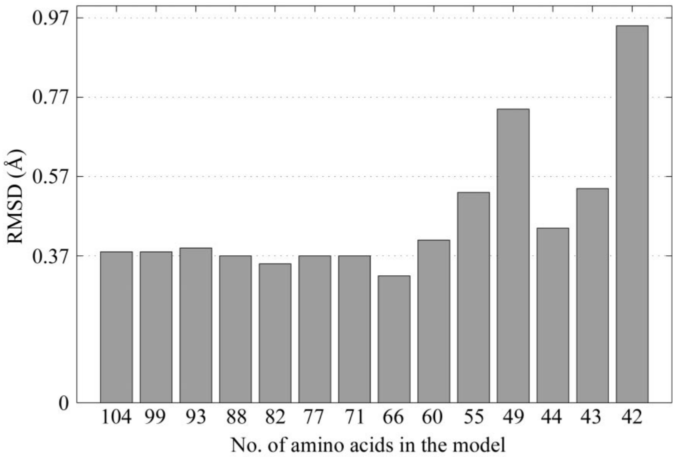

Figure 6
RMSDs associated with the best function value found after optimization of the 8 Å search results.

SOMoRe finds a solution to the MR problem using a model that contains only 42 residues. The least complete models that allowed *EPMR*, *X-PLOR* and *AMoRe* to succeed contained 44 residues (Kissinger *et al.*, 2001) and approximately 62 and 67 residues, respectively (Kissinger *et al.*, 1999). The authors state that the models for *X-PLOR* and *AMoRe* could be truncated by approximately 40 and 35%, respectively. Assuming that 100% of the model is the first 104 residues of 4rhn, then the models could be truncated by 42 residues for *X-PLOR* and 37 residues for *AMoRe* (Kissinger *et al.*, 1999). Table 2 summarizes these results.

Furthermore, according to Kissinger *et al.* (2001), if the search model has been truncated by 60% (leaving a 46-residue polyalanine model), then the search efficiency of *EMPR* is approximately 5% (only five out of 100 runs were successful). In contrast, the search efficiency of *SOMoRe* is always 100% because the approach is deterministic.

6.4.1. 8 Å searches. Overall, despite varying amounts of model truncation, the lowest valued local minimum is a solution. In general, as the model is truncated there is a linear decrease in the contrast between function values associated with solutions and non-solutions and a linear decrease in the function values of the lowest valued minimum, as shown in Fig. 5. Interestingly, however, the contrast between free values does not decrease linearly, as also shown in Fig. 5. The free values clearly differentiate solutions from non-solutions and play a larger role in cross-validating possible solutions for the least complete models. Model truncation has a similar effect on the RMSDs associated with these minima, as shown in Fig. 6. For the last six most incomplete models, there appears to be a linear increase in the associated RMSDs.

6.4.2. 10 Å search 6rhn. In contrast, the 10 Å searches were successful only using models containing 104, 99, 93 and 88 residues. The successful search results are very similar to the corresponding 8 Å results (Jamrog, 2002). Thus, if the model represents a small portion of the target structure, it may be advantageous to use an 8 Å global search.

6.5. Run times

The run times for *SOMoRe* are very reasonable given that 6D searches are performed, as shown in Table 3. All experiments were run at Rice University on a 300 MHZ R12000 processor of an SGI Origin2000 machine, which has 10 Gb of RAM. Table 3 also indicates the unsuccessful 10 Å searches

Table 3
Global search and optimization run times for each test problem.

An asterisk indicates an unsuccessful search.

	No. of symmetry operators	No. of $I^o_{\mathbf{h}}$	$\Delta\theta_2$	No. of grid points	Search time (h)	No. of starting points	Opt. time ∞–4 Å (min)
8 Å structures							
1aki	4	143	8.7	21919248	3.35	1000	49
1cgn	12	99	4.8	747367992	213.36	1000	102
1b6q	4	72	8.9	16720896	1.25	500	9
6rhn	8	165	6.2	58832256	19.14	1000	111
10 Å structures							
1aki*		70	10.9	5982075	0.45	1000	49
1cgn		45	5.9	210458470	27.12	1000	98
1b6q		41	11.2	5104190	0.22	500	10
6rhn		86	7.7	18286653	3.08	1000	114

with an asterisk. The run times for 6rhn are the average of all run times.

In general, run time is a function of the number of intensities in the resolution range and the number of symmetry operators. Given a set of proteins that are roughly the same size and crystals with the same liquid content, the more symmetry mates the larger the unit cell and the longer the run time, because the step lengths in the MR variables are inversely related to the average of the lengths of the unit-cell basis vectors. In addition, the larger the unit cell, the smaller the spacing of the diffraction pattern and the more data in a given resolution range.

We note that it will be more time-consuming to solve the structures of larger sized proteins using *SOMoRe* and that this will be a topic of future investigation. The time-consuming grid-search part of the calculation is easily parallelizable owing to the independence of the calculations at each grid point and the program is being adapted for cluster computing for this purpose.

Finally, to show the efficiency of our approach over a simple high-resolution 6D search, we estimate the run time for a 6D search of a target function computed using data between ∞ and 4 Å. Using the method for calculating structure factors described by Chang & Lewis (1997) and implemented in *Qs* (Glykos & Kokkinidis, 2000), the run time is only linearly dependent on the number of reflections and, of course for our method, linearly dependent on the number of grid points. To compute the estimated run time, we determine the number of grid points in the fine search, f_2, and the number of intensities between ∞ to 4 Å, d_2. (We run *SOMoRe* to obtain this information and then kill the run.) Then, for each problem, we multiply the run time for the 8 Å search by

$$\text{factor} = \frac{f_2}{f_1} \cdot \frac{d_2}{d_1}, \tag{6}$$

where f_1 is the number of grid points in the 8 Å global search and d_1 is the number of intensities between ∞ and 8 Å. As shown in Table 4, unless a massively parallelized search is performed, 6D grid searches are still out of reach for most MR problems.

Table 4
Estimated run times for fine 6D global searches of target functions computed using all data between ∞ and 4 Å.

	No. of $I^o_{\mathbf{h}}$ ∞–4 Å	$\Delta\theta_2$	No. of grid points	Factor	Estimated search time (d)
1aki	1133	4.4	1317513600	476.2	67
1cgn	727	2.4	43360941130	426.1	3788
1b6q	515	4.5	1009536576	431.9	22
6rhn	1168	3.1	3584438784	431.3	344

7. New protein structure

We have recently used *SOMoRe* to solve the structure of adenylate kinase from *Bacillus subtilis*. This new protein structure had not been previously determined. The lowest valued local minimizer was a solution for a majority of the protein structure, as verified by inspection of the electron-density maps. The coordinate set used for the search model has a high degree of sequence homology; however, it also has an unusual position of the so-called 'lid' domain, comprising 25 or so amino acids. To determine the correct position of the lid domain relative to the remainder of the model, we again used *SOMoRe*, except this time we performed a local 6D search about the initial position of the lid, which was specified by lowest valued local minima found for the entire model, and then local optimization of the best local grid points. As a result, we found the correct position of this domain. Attempts to determine the position for the lid domain using rigid-body refinement in the *CNS* program had failed. During the process, we used the same high-resolution cutoffs for the surrogate function and the local optimization that we used for the above-described test problems. The protein structure has been successfully refined and the structure determination will be published in detail.

8. Conclusions

Our new strategy was successful on every test problem. Table 1 lists the estimated 'optimal' RMSD and the RMSDs computed from the lowest valued minima, or in the case of 1cgn, the lowest valued minimum after other minima were ruled out.

As a result, *SOMoRe* promises to extend the applicability of MR because it straightforwardly solved test problems 1cgn and 1b6q. '*AMoRe*, which is commonly recognized as the best in the field' (Vagin & Teplyakov, 1997) could not solve test problem 1cgn in a straightforward way, while the traditional code *MOLREP* could not solve test problem 1b6q, despite the availability of a very good model. In addition, our method solved a problem using a less complete model than the models required by *EPMR*, *X-PLOR* and *AMoRe*. The results from this test problem corroborate our hypothesis that six-dimensional approaches should be able to solve MR problems using models that are less complete than those required by traditional approaches.

In general, we have demonstrated that a coarse 6D global search can identify starting points that will converge to MR solutions. Secondly, we have shown that 8 Å appears to be a safe limit for the high-resolution cutoff, given the current step lengths in the MR variables. Thirdly, we have shown that optimization is essential to the new approach because it increases the contrast between the function values associated with points close to MR solutions, improving these points and producing local minima with the very lowest function values. Finally, clearly the new approach works well even if some of the lowest resolution data are missing.

The major strength of our new strategy is the novel integration of a coarse 6D search, using a surrogate low-resolution function, and a multi-start local optimization process, using a higher resolution target function. The coarse search is relatively fast and cost-efficient compared with fine 6D grid searches. Unlike traditional methods, our new strategy spends more computational effort in promising areas of the variable space where solutions are likely to occur. Also, unlike stochastic 6D methods, it is deterministic in nature and can be completely parallelized. We believe that as computing resources improve, more accurate and robust approaches like *SOMoRe* will become increasingly more attractive not only for solving more difficult problems, but for general use as well.

The program *SOMoRe* will be available upon request from the authors. In the future, we anticipate that it will be available for download from the World Wide Web.

APPENDIX *A*
Finite-difference gradient

An analytic expression is not readily available for the gradient of the correlation coefficient because interpolation is used to determine the appropriate structure factors for each orientation and translation of the model protein considered during MR. However, the gradient can be approximated using finite differences. The *j*th component of the finite-difference approximation to the gradient of $C[I^c(\mathbf{u}), I^o]$ is

$$\nabla C_j = \frac{C[I^c(\mathbf{u} + he_j), I^o] - C[I^c(\mathbf{u}), I^o]}{h}, \qquad (A1)$$

where $\mathbf{u} = (\Theta, \mathbf{t})$, e_j is the *j*th standard Euclidean basis vector, $h = h_\varepsilon \operatorname{sign}(\mathbf{u}_j)$, $\operatorname{sign}(\mathbf{u}_j)$ is the sign of $\mathbf{u}_j$ (either 1 or -1) and h_ε is typically a small number. When angles are expressed in radians and translations expressed in fractional coordinates, $h_\varepsilon = 10^{-4}$ allows the local optimization method BFGS to perform well; that is, to converge in most cases after a reasonable number of iterations and produce a final gradient with a relatively small norm. Finally, $h = h_\varepsilon \operatorname{sign}(\mathbf{u}_j)$ to prevent possible roundoff error from subtracting two numbers that are close in magnitude. Two general references are Dennis & Schnabel (1996) and Nocedal & Wright (1999).

APPENDIX *B*
Determining the closest symmetry mate

A pseudo code for determining the closest symmetry mate is given below. Let the center of mass of a molecule be $c = (\sum_{j=1}^{N} x_j/N, \sum_{j=1}^{N} y_j/N, \sum_{j=1}^{N} z_j/N)$, where N is the number of atoms and (x_j, y_j, z_j) are the orthogonal coordinates of the *j*th atom.

Pseudo code for determining the closest mate

Step 1. Compute the centers of mass of symmetry mates *i* and *j*, c_i and c_j.

Step 2. Compute $d_c = c_i - c_j$.

Step 3. Convert d_c to fractional coordinates; that is, compute $d_f = A^{-1}d_c$.

Step 4. Add the translation $b_\mathbf{t}$ (which is in fractional coordinates) to the fractional coordinates of symmetry mate *j*. A pseudo code for determining the appropriate basis vector translation, $b_\mathbf{t}$, is below. Let $d_f(k)$ and $b_\mathbf{t}(k)$ refer to the *k*th component of d_f and $b_\mathbf{t}$ and let floor(), ceil() and mod() be the standard floor, ceiling and modulo functions, respectively.

Pseudo code for determining $b_\mathbf{t}$

```
for $k$ = 1 to 3
   if $d_f(k) > 0$ (mate $i$ is to the right of mate $j$ with respect to basis vector $k$),
      $b_\mathbf{t}(k)$ = floor($d_f(k)$)
      if mod($d_f(k)$,1.0) > .5,
         $b_\mathbf{t}(k) = b_\mathbf{t}(k) + 1$
   else
      $b_\mathbf{t}(k)$ = ceil($d_f(k)$)
      if mod($d_f(k)$,1.0) < $-$.5,
         $b_\mathbf{t}(k) = b_\mathbf{t}(k) - 1$
   end
end
```

We would like to thank Nicholas Glykos, the author of *Qs*, for graciously sharing his code online and for all his helpful discussions and explanations. DCJ was supported in part by a training fellowship from the Keck Center for Computational Biology (NSF GRT BIR92-56580, NSF BIR-94-13229 and NLM LM07093) and NSF DMS-9973339. GNP was supported by the National Institutes of Health GM-64598. YZ was supported in part by DOE Grant DE-FG03-97ER25331, DOE/LANL 03891-99-23 and NSF DMS-9973339.

References

Baker, E. N., Anderson, B. F. & Dobbs, A. J. (1995). *Acta Cryst.* D**51**, 282–289.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Brünger, A. T. (1990). *Acta Cryst.* A**46**, 46–57.

Brünger, A. T. (1992). *X-PLOR. A System for X-ray Crystallography and NMR.* New Haven, CT, USA: Yale University Press.

Brünger, A. T. (1993). *ImmunoMethods*, **3**, 180–190.

Brünger, A. T. (1997). *Methods Enzymol.* **276**, 558–580.

Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., 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). *Acta Cryst.* D**54**, 905–921.

Bruns, C., Hubatsch, I., Ridderstrom, M., Mannervik, B. & Tainer, J. (1999). *J. Mol. Biol.* **288**, 427–439.

Chang, G. & Lewis, M. (1997). *Acta Cryst.* D**53**, 279–289.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

DeLano, W. & Brünger, A. T. (1995). *Acta Cryst.* D**51**, 740–748.

Dennis, J. E. & Schnabel, R. B. (1996). *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Philadelphia: Society for Industrial and Applied Mathematics.

Dobbs, A. J., Anderson, B. F., Faber, H. R. & Baker, E. N. (1996). *Acta Cryst.* D**52**, 356–368.

Evans, G., Roversi, P. & Bricogne, G. (2000). *Acta Cryst.* D**56**, 1304–1311.

Glykos, N. M. & Kokkinidis, M. (1999). *Acta Cryst.* D**55**, 1301–1308.

Glykos, N. M. & Kokkinidis, M. (2000). *Acta Cryst.* D**56**, 169–174.

Glykos, N. M. & Kokkinidis, M. (2001). *Acta Cryst.* D**57**, 1462–1473.

Grosse-Kunstleve, R. W. & Adams, P. D. (2001). *Acta Cryst.* D**57**, 1390–1396.

Hirshfeld, F. L. (1968). *Acta Cryst.* A**24**, 301–311.

Jamrog, D. C. (2002). PhD thesis, Rice University, Houston, Texas, USA. Technical Report TR-0208 at http://www.caam.rice.edu/.

Jogl, G., Tao, X., Xu, Y. & Tong, L. (2001). *Acta Cryst.* D**57**, 1127–1134.

Kissinger, C., Gehlhaar, D. & Fogel, D. (1999). *Acta Cryst.* D**55**, 484–491.

Kissinger, C., Gehlhaar, D., Smith, B. A. & Bouzida, D. (2001). *Acta Cryst.* D**57**, 1474–1479.

Lattman, E. E. (1972). *Acta Cryst.* B**28**, 1065–1068.

Miller, S. T., Genova, J. D. & Hogle, J. M. (1999). *J. Appl. Cryst.* **32**, 1183–1185.

Nash, S. G. & Sofer, A. (1996). *Linear and Nonlinear Programming.* New York: McGraw–Hill.

Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.

Navaza, J. (2001). *Acta Cryst.* D**57**, 1367–1372.

Nocedal, J. & Wright, S. J. (1999). *Numerical Optimization.* New York: Springer.

Rabinovich, D., Rozenberg, H. & Shakked, Z. (1998). *Acta Cryst.* D**54**, 1336–1342.

Rabinovich, D. & Shakked, Z. (1984). *Acta Cryst.* A**40**, 195–200.

Rossmann, M. & Blow, D. (1962). *Acta Cryst.* **15**, 24–31.

Sheriff, S., Klei, H. & Davis, M. (1999). *J. Appl. Cryst.* **32**, 98–101.

Tong, L. (1996). *Acta Cryst.* A**52**, 782–784.

Tong, L. & Rossmann, M. G. (1990). *Acta Cryst.* A**46**, 783–792.

Urzhumtsev, A. & Podjarny, A. (1995). *Acta Cryst.* D**51**, 888–895.

Vagin, A. & Teplyakov, A. (1997). *J. Appl. Cryst.* **30**, 1022–1025.