A Global Optimization Method for the Molecular Replacement Problem in X-ray Crystallography*

Diane C. Jamrog† George N. Phillips, Jr.‡ Richard A. Tapia§ Yin Zhang¶

May, 2002

Abstract

The primary technique for determining the three-dimensional structure of a protein molecule is X-ray crystallography, from which the molecular replacement (MR) problem often arises as a critical step. The MR problem is a global optimization problem to locate an optimal position of a model protein, whose structure is similar to the unknown protein structure that is to be determined, so that at this position the model protein will produce calculated intensities closest to those observed from an X-ray crystallography experiment. Improving the applicability and robustness of MR methods is an important research topic because commonly used traditional MR methods, though often successful, have their limitations in solving difficult problems.

We introduce a new global optimization strategy that combines a coarse-grid search, using a surrogate function, with extensive multi-start local optimization. A new MR code, called SOMoRe, based on this strategy is developed and tested on four realistic problems, including two difficult problems that traditional MR codes failed to solve directly. SOMoRe was able to solve each test problem without any complication, and SOMoRe solved a MR problem using a less complete model than the models required by three other programs. These results indicate that the new method is promising and should enhance the applicability and robustness of the MR methodology.

Key Words. Molecular replacement problem, X-ray crystallography, global optimization, surrogate function, global search, multi-start local optimization.

1 Introduction

Knowledge of protein structures is critically useful for scientific understanding of a wide range of biological and medical processes at the molecular level, for example, for understanding the molecular basis of diseases and for designing pharmaceutical drugs.

*The first author was supported in part by the W. M. Keck Center for Computational Biology: NSF RTG BIR-94-13229, NLM RTG 2T15LM07093 and NLM 5 T15 LM07093. The second author was supported by National Institutes of Health GM-64598. The third author was supported in part DOE/LANL Contract 03891-99-23. The fourth author was supported in part by DOE Grant DE-FG03-97ER25331, DOE/LANL Contract 03891-99-23, and NSF Grant DMS-9973339.

†Department of Computational and Applied Mathematics, Rice University, Houston, Texas, USA.

‡Department of Biochemistry and Department of Computer Sciences, University of Wisconsin-Madison, Madison, Wisconsin, USA.

§Department of Computational and Applied Mathematics, Rice University, Houston, Texas, USA.

¶Department of Computational and Applied Mathematics, Rice University, Houston, Texas, USA.

X-ray crystallography has been, and still is, the primary technique for determining the detailed three-dimensional structure of a protein molecule. A difficult global optimization problem, called the phase problem, arises in X-ray crystallography. A major step towards attacking the phase problem is to solve the molecular replacement (MR) problem, which is also a global optimization problem but has a much smaller dimension than the phase problem.

In essence, the MR problem is a data-fitting problem in which one rotates and translates a so-called model protein to find an optimal position that will generate calculated intensities closest to those observed from an X-ray crystallography experiment performed on a target protein — one whose structure is to be determined. The model protein has a known structure that is more or less similar to that of the target protein. The observed intensities are experimentally measured when X-rays are scattered by the crystallized target protein. The calculated intensities are computed from the atomic coordinates of the model protein with varying rotation and translation.

Once the MR problem is solved, then a preliminary electron density map (i.e., an image) of the crystallized target protein can be generated, leading to a set of approximate atomic coordinates of the target protein suitable for further refinement and full elucidation of the target protein structure.

As more and more structures are deposited into the database of solved structures, it will be more likely that a model protein will be available for a given target protein. Hence, the use of MR methods is expected to continue to increase. However, research is still needed to improve MR methods because, although successful in solving many problems, traditional MR methods are known to have difficulty solving certain classes of MR problems, including those for which an accurate model protein is not available, and those involving a protein crystal that has a high degree of symmetry.

In this paper, we construct and validate a new global optimization method for solving MR problems, particularly those that are difficult for the traditional methods to solve. In Section 2, we introduce the background on X-ray crystallography and the MR problem. In Section 3, we review the current approaches for solving the MR problem. We then introduce the global optimization method in Section 4 and describe the current implementation of our new method in Section 5. Results from the method are given in Section 6.

2 X-ray Crystallography and Molecular Replacement

Before MR can take place, the X-ray crystallography experiment must be performed, and a model protein that is structurally similar to the crystallized target protein must be found. In this section, we discuss the basics of X-ray crystallography, including the protein crystal, the observed and calculated intensities, the phase problem, and the molecular replacement problem.

2.1 What is X-ray crystallography?

In general, X-ray diffraction techniques provide the only way to directly produce the image of molecular structures with high enough resolution to distinguish atoms. To produce an image of an object, electro-magnetic radiation, such as visible light, must be scattered by the object and recombined by a lens, as for example by a microscope lens or the lens of our eyes. However, in order to produce a detailed image, the radiation must have a wavelength equal to or smaller than the size of the object. Visible light cannot be used to produce the image of atoms because its wavelength is too long. Electron beams found in electron microscopes are generally too damaging to the fragile proteins to produce atomic images. X-rays have short enough wavelengths to detect the atomic arrangement of a protein in a crystalline state and their damage can be controlled.

However, using X-rays complicates the imaging process because they cannot be refocused to produce an image. Only the amplitudes of the scattered X-rays can be measured. The phases of the diffracted X-rays cannot be measured because of physical limits. Because an X-ray lens does not exist, the X-rays are mathematically "refocused", using a Fourier transform. However, first, good estimates for the unmeasurable phases must be found. With these estimates, the diffracted waves can be approximated, and then a Fourier transform can mathematically "refocus" the approximate waves and produce an image of the crystallized protein.

In short, X-ray crystallography comprises two main components: (i) an X-ray diffraction experiment performed on a protein crystal to measure the intensities of diffracted X-rays; and (ii) a mathematical and computational process to obtain sufficiently accurate phases. For more information on X-ray crystallography, we refer interested readers to [11, 14].

2.2 The crystal

The very first step in X-ray crystallography is to grow a crystal of the target protein whose structure is to be determined (which can be a difficult process by itself). The *protein crystal* is a three-dimensional periodic arrangement of proteins in a certain solvent. A crystal can be specified by a *unit cell*, an imaginary parallelepiped that contains the basic repeating unit of the crystal. Figure 1 is a schematic of the imaginary parallelepipeds or unit cells that are stacked three-dimensionally in the crystal. The unit cells are defined by a set of *basis vectors*, $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^3$, that also define the parallelepipeds and a lattice known as the *real lattice*. The lengths of the basis vectors are typically measured in Angstroms (Å) where 1Å$= 10^{-8}$cm.

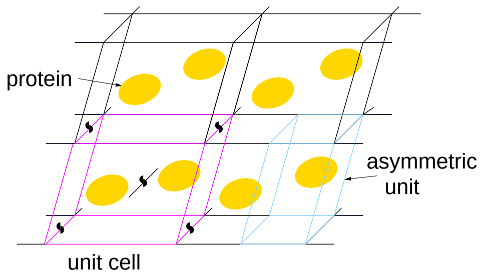

Figure 1: A schematic of a protein crystal with four unit cells delineated.

In X-ray crystallography, a convenient coordinate system is the so-called *fractional coordinate* system that is defined relative to the basis vectors $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$. The fractional coordinates $\mathbf{x} = (x, y, z)^T$ correspond to the point $\mathbf{v} = (u, v, w)^T \equiv x\mathbf{a} + y\mathbf{b} + z\mathbf{c}$ in Cartesian coordinates. Let $A = [\mathbf{a}\ \mathbf{b}\ \mathbf{c}] \in \mathbb{R}^{3\times3}$ be the transformation matrix from the fractional coordinate system to the Cartesian one, then $\mathbf{v} = A\mathbf{x}$ and $\mathbf{x} = A^{-1}\mathbf{v}$. One possible choice for A is

$$A = \begin{bmatrix} a & b\cos\gamma & c\cos\beta \\ 0 & b\sin\gamma & c(\cos\alpha - \cos\beta\cos\gamma)/\sin\gamma \\ 0 & 0 & c\sin\beta \end{bmatrix}, \tag{1}$$

where a, b, and c are the lengths of the unit cell basis vectors $\mathbf{a}, \mathbf{b}$, and $\mathbf{c}$, respectively, and α, β, and γ are the angles between $\mathbf{b}$ and $\mathbf{c}$, $\mathbf{a}$ and $\mathbf{c}$, and $\mathbf{a}$ and $\mathbf{b}$, respectively.

An important concept in crystallography is *crystallographic symmetry*. Often multiple copies of the molecule occur in the unit cell, related to each other through a set of so-called symmetry relationships. Such a group of identical molecules are said to belong to a space group. In a given space group, if the coordinates of one molecule are known, then the coordinates of the rest can be calculated through the symmetry operators of the given space group. For example, if a rotation of each molecule in the unit cell by 60 degrees about an axis superimposes a copy of the molecule onto another, then a six-fold rotational symmetry exists. The multiple copies of symmetry-related molecules are called *symmetry mates*.

Crystallographic symmetry is mathematically defined by symmetry operators, which are linear operators that are represented by pairs of matrices and vectors for rotations and translations, respectively. If $S_g \in \mathbb{R}^{3\times3}$ and $s_g \in \mathbb{R}^3$ define the gth symmetry operator and $\mathbf{x} \in \mathbb{R}^3$ are the fractional coordinates of one of the protein's atoms, then the fractional coordinates of the same atom in the gth symmetry related protein are $\mathbf{x}' = S_g\mathbf{x} + s_g$.

A portion of the unit cell that contains the largest number of molecules, unrelated by symmetry, is called the *asymmetric unit*; see Figure 1 for an illustration. An asymmetric unit usually contains only one protein, occasionally two but rarely three or more.

2.3 The experiment, observed and calculated intensities

Observed intensities are measured from an X-ray crystallography experiment. After growing a protein crystal, the crystal is rotated in an X-ray beam, and a detector measures the X-ray diffraction pattern of the crystal that is commonly referred to as a set of observed intensities. Because the crystal is periodic, there are directions in which the scattered X-rays constructively interact so that the summation of the scattered X-rays produces a total wave with a significant amplitude that can be measured as an intensity.

The observed intensities occur at points in a three-dimensional lattice known as the *reciprocal lattice* because the molecules are arranged with respect to a real space crystallographic lattice. The reciprocal lattice is specified by a set basis vectors: $\mathbf{a}^*$, $\mathbf{b}^*$ and $\mathbf{c}^*$. In general, the coordinates of a lattice point in *reciprocal space* is given by $B\mathbf{h}$, where $B = [\mathbf{a}^*\ \mathbf{b}^*\ \mathbf{c}^*]$, and $\mathbf{h} = (h, k, l) \in \mathbb{Z}^3$ are the indices that serve as a label for this lattice point. The intensity "observed" at reciprocal lattice point $\mathbf{h}$ is denoted as $I^o_{\mathbf{h}}$. The relationship between A and B, namely between the two sets of basis vectors for the real and the reciprocal spaces, is $B = A^{-T}$ (see [14], for example). Figure 2 is a precession photograph of some observed intensities from an X-ray experiment that occur on one layer of the reciprocal lattice.

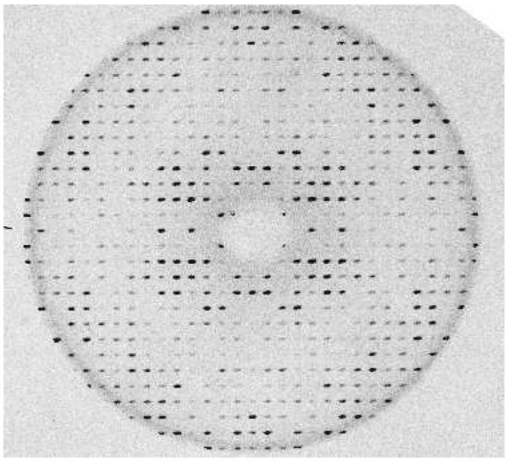

Figure 2: A precession photograph of some intensities from an X-ray experiment. The intensities are the dark spots.

We denote the calculated intensity at reciprocal lattice point $\mathbf{h}$ as $I^c_{\mathbf{h}}$, which is the magnitude squared of the so-called structure factor $F_{\mathbf{h}}$ at lattice point $\mathbf{h}$. That is, $I^c_{\mathbf{h}} = |F_{\mathbf{h}}|^2$. The structure factor is a complex number defined by the formula

$$F_{\mathbf{h}} = \sum_{g=1}^{G}\sum_{j=1}^{N} f_j(\mathbf{h}) \exp\left[2\pi i\, \mathbf{h} \cdot (\, S_g\, \mathbf{x}_j + s_g\,)\right], \tag{2}$$

where i is the imaginary unit, G is the number of symmetry mates in the unit cell, (S_g, s_g) represents the gth symmetry operator of the crystal, N is the number of atoms in the protein, $f_j(\mathbf{h})$ is the so-called scattering factor of atom j at lattice point $\mathbf{h}$, $\mathbf{x}_j$ is the fractional coordinate vector of the jth atom, and "$\cdot$" denotes the inner product in $\mathbb{R}^3$. In fact, the structure factor is nothing but a Fourier coefficient of the electron density function of the protein that is specified by the set of fractional coordinates $\{\mathbf{x}_j\}_{j=1}^N$. In an ideal situation free of modeling and experimental errors, $I^c_{\mathbf{h}}$ would be equal to $I^o_{\mathbf{h}}$ at all reciprocal lattice points given the correct set of coordinates.

2.4 The phase problem and warm start

As mentioned earlier, determining the phases of the scattered X-rays (i.e., solving the phase problem) allows the protein's image to be computed mathematically. Since only observed intensities can be measured, the best one can hope for (in a mathematical sense) is to find a set of atomic coordinates for the target protein from which the calculated intensities best match those observed. Moreover, if a set of globally optimal atomic coordinates produce the correct or nearly correct intensities, it is reasonable to expect that they also produce correct or nearly correct phases.

From a mathematical perspective, the phase problem is a global optimization problem in which the variables are the atomic coordinates of the target protein. Normally, a protein contains around a thousand or more atoms. As a result, the phase problem normally has around 3,000 or more variables. Moreover, objective functions that measure the discrepancy between the calculated and observed intensities are invariably nonlinear, non-convex, and highly oscillatory. Solving such a global optimization problem by today's technology is extremely difficult, if not impossible, without some additional information.

One common form of additional information is a good starting point, or *warm start*, that is sufficiently close to a global optimum. With such a warm start, the chances of solving the phase problem can be dramatically increased. Does such a warm start exist? Fortunately, in many cases, if not most, the answer is "yes". The answer lies in the *Protein Data Bank*, or PDB [28], in which around 18,000 protein structures have been deposited as of May of 2002, thanks to decades of research in this field.

Since a protein often has some structural similarity to related proteins, for a given target protein it is more likely than not that the structure of a related protein, a model protein, has already been resolved. One way of attacking the phase problem is to approximate the unmeasurable phases of the target protein by phases calculated from a model protein. Then, the experimentally observed intensities and the phases computed from the model can be used to compute a first approximation to the image (or more precisely, the electron density) of the crystallized target protein. However, the model must have nearly the same orientation and position as the crystallized target protein in order for the model's phases to be sufficiently accurate to produce a useful approximate electron density.

In the molecular replacement (MR) approach, such a model protein is rotated and translated, as a rigid body, to map it onto the target protein. The quality of the mapping can be judged by the agreement between the calculated intensities from the model protein (for a given rotation and translation) and the observed intensities produced by the target protein in the X-ray experiment. Obviously, the success of the MR approach depends on the degree of similarity between the model and the target protein. The greater the similarity between the two, the better the chance that the MR approach will succeed.

It should be mentioned that the linear amino acid sequence of the target protein is known *a priori* for a MR problem. Given this sequence, one can search the PDB for a protein that has a similar amino acid sequence. Similarity of the two sequences indicates that the two may have evolved from the same genetic ancestor and are likely to be structurally similar. For more information on this topic, see [8].

2.5 The MR Problem

The MR problem is a global optimization problem in which an objective function measures disagreement between the observed and calculated intensities. The MR problem is a $6n$-dimensional ($6n$D) problem, where n is the number of molecules in the the asymmetric unit of the crystal, as introduced in Section 2.2. The dimension in $6n$ because six variables are required to position each of the n molecules in the asymmetric unit: three rotation variables and three translation variables. However, MR problems are typically either 6D or 12D because n is usually equal to one or two.

The 6D MR problem can be written as

$$\min_{(\Theta,\mathbf{t})\in L} f(I^o, I^c(\Theta,\mathbf{t})), \tag{3}$$

where $L \subset \mathbb{R}^3 \times \mathbb{R}^3$ is the search region, $I^o, I^c \in \mathbb{R}^m$ are vectors of observed and calculated intensities, Θ is a vector of three angles that specify a rotation matrix that rotates the model protein, $\mathbf{t}$ is a translation vector applied to the model protein, and $f(\cdot,\cdot)$ is a function that measures the disagreement between the observed and calculated intensities. The calculated intensities are functions of the rotated and translated coordinates of the model protein, thus functions of $(\Theta, \mathbf{t})$, as indicated.

3 Current Approaches

Currently, there are basically two types of approaches for solving the MR problem: (i) traditional approaches, which separately optimize the rotational and translational degrees of freedom of the model protein, and (ii) higher-dimensional approaches, either 6D or $6n$D, which simultaneously optimize the rotation and translation of one or n copies of the model protein in an asymmetric unit.

3.1 Traditional Approach

In 1962, Rossman and Blow [31] proposed that the MR problem be solved by two separate searches: first a search to identify optimal rotations of the model, and then a search to identify optimal translations of the oriented model. At that time, 6D searches were beyond the reach of available computing capabilities. Since then, traditional MR codes have determined many molecular structures, but the method does have some drawbacks.

The optimization formulation of the traditional approach is to sequentially solve two unconstrained 3D optimization problems, which can be formalized as follows. (For more details, see [11, 31], for example.) First, try to find several approximate solutions, Θ_i, to the following problem

$$\min_{\Theta \in \mathbb{R}^3} R(I^o, I^c(\Theta, \mathbf{0})), \tag{4}$$

where the translation variable $\mathbf{t}$ is fixed at the origin, which is placed at the model protein's center of mass. Then, for each fixed Θ_i, solve

$$\min_{\mathbf{t} \in \mathbb{R}^3} T(I^o, I^c(\Theta_i, \mathbf{t})). \tag{5}$$

In the above problems, R and T are objective functions that measure disagreement between I^o and I^c and are commonly referred to as rotation and translation functions, respectively. We emphasize that in the rotation search only one copy of the model protein can be used to compute calculated intensities.

There are two main drawbacks of traditional approaches. First, for more difficult MR problems (such as those with high degrees of symmetry), the lowest valued local minima of traditional rotation functions often do not come close to the optimal rotation component of a MR solution [22, 23, 35]. If a nearly optimal rotation is not found, then a nearly optimal translation cannot be found, and the traditional method will fail. Second, traditional methods typically require good model proteins whose structural similarities with the target proteins must be relatively high. If such high-quality model proteins are not available, then traditional methods become less reliable; for example, see [1, 3].

3.2 Higher-dimensional Approaches

With today's computing capacities, more accurate formulations of the MR problem can be used. To avoid the drawbacks associated with separately optimizing the rotation and translation of the model, 6D MR methods have been designed to simultaneously optimize the two sets of MR variables. Recently, parallelized 6D grid searches [33] have been proposed that rely on massive computing power; and both 6D and $6n$D stochastic optimization approaches [6, 23, 15] have been proposed, based on genetic or simulated annealing algorithms.

In contrast to traditional methods, 6D methods simulate scattering from all the symmetry mates of the model protein in the unit cell. This is possible because the translation variables allow the symmetry mates to be positioned relative to each other. As such, the calculated intensities of a 6D method can better match the observed intensities at a solution to the MR problem. Evaluating a 6D objective function is more expensive than evaluating traditional rotation and translation functions; however, the theoretically sound approach of optimizing a 6D function should lead to a more reliable and robust solution process.

4 A New Global Optimization Strategy

As mentioned, MR objective functions are generally highly oscillatory and have a huge number of local minima. To obtain a global minimum, a brute-force, 6D fine-grid search is exceedingly time-consuming and requires massive computing power. On the other hand, the performance of stochastic optimization methods, though satisfactory on some problems, are generally unpredictable. Therefore, it is highly desirable and useful to construct a 6D MR method that is relatively fast, affordable, reliable, and deterministic.

In a deterministic procedure for global optimization of a highly oscillatory function, it is perhaps inevitable that a global search scheme must be used to gain information about the function's global landscape. Since a 6D fine-grid search is out of question because of its prohibitively high cost, we will consider a coarse-grid search based on a so-called surrogate function that is closely related to the "true" objective function but much smoother. In order for this approach to be successful, a surrogate function must capture the global behavior of the true objection function while not suffering from the curse of too many local minima. In the context of the MR problem, a natural surrogate function is one defined by a set of low-frequency intensities. Therefore, before we define the surrogate function and our algorithm, we introduce the notion of resolution and frequency of the observed and calculated intensities.

4.1 Resolutions and data sets

The resolution of an intensity $I_{\mathbf{h}}$ is defined as $1/d^*(\mathbf{h})$, where $d^*(\mathbf{h})$ is the quantity

$$d^*(\mathbf{h}) = \| B\mathbf{h} \|, \tag{6}$$

where $\|\cdot\|$ is the Euclidean norm, and $B = A^{-T} = [\mathbf{a}^* \ \mathbf{b}^* \ \mathbf{c}^*]$ as defined in Section 2.3. A resolution range is specified by a pair of low-resolution and high-resolution cut-off values that define a region in the reciprocal space between an inner sphere and an outer sphere.

In practice, a given objective function is always correlated to a set of intensities occurring at reciprocal lattice points within a given resolution range; namely, it is evaluated only using intensities, observed and calculated, within that range. As a result, the sets of observed and calculated intensities used to evaluate an objective function can be defined as $\{I^o_{\mathbf{h}} : \mathbf{h} \in \mathcal{I}\}$ and $\{I^c_{\mathbf{h}} : \mathbf{h} \in \mathcal{I}\}$ for

$$\mathcal{I} = \left\{ \mathbf{h} : \; R_{high} \leq \frac{1}{d^*(\mathbf{h})} \leq \; R_{low} \right\} \subset \mathbb{Z}^3,$$

where $R_{high}, R_{low} \in \mathbb{R}$ are the high-resolution and low-resolution cut-offs in Å, respectively. For example, in traditional methods, R_{high} is normally set to 4Å and R_{low} to 15 Å.

4.2 Spatial frequency of intensities

The observed and calculated intensities can be characterized in terms of spatial frequency. For example, if an intensity occurs at a reciprocal lattice point close to the origin, then it is a low-frequency intensity; otherwise, it is a higher frequency intensity. The frequency of an intensity can be seen by expressing $|F_{\mathbf{h}}|$ as the sum of cosine and sine functions.

Recall that the definition of the structure factor occurring at $\mathbf{h}$ is given by (2). The fractional coordinates, $\mathbf{x}_j$, of the model protein result from applying rotations and translations to the model protein as a rigid body; thus

$$\mathbf{x}_j = \left(A^{-1}\Omega(\Theta)A \right) \hat{\mathbf{x}}_j + \mathbf{t}, \tag{7}$$

where $\hat{\mathbf{x}}_j$ is the initial, fractional coordinate vector of the jth atom in the model protein, $\mathbf{t} \in \mathbb{R}^3$ is a translation in fractional coordinates, $\Omega(\Theta)$ is the rotation matrix in Cartesian system corresponding to the angles of Θ, and $A \in \mathbb{R}^{3\times 3}$ is the transformation matrix from fractional coordinates to the Cartesian ones. The transformations between orthogonal and fractional coordinates (A and A^{-1}) are necessary in order to correctly apply rigid-body rotations.

It follows from (2) and (7) that

$$F^c_{\mathbf{h}}(\Theta, \mathbf{t}) = \sum_{g=1}^{G} \sum_{j=1}^{N} f_j(\mathbf{h}) \, \exp\left[2\pi i \, \mathbf{h} \cdot \left(S_g \left(A^{-1}\Omega(\Theta)A\hat{\mathbf{x}}_j + \mathbf{t} \right) + s_g \right)\right]. \tag{8}$$

Let $F^c_{\mathbf{h}}(\Theta, \mathbf{t}) = B_{\mathbf{h}}(\Theta, \mathbf{t}) + iC_{\mathbf{h}}(\Theta, \mathbf{t})$. Then by definition

$$I^c_{\mathbf{h}}(\Theta, \mathbf{t}) = |F^c_{\mathbf{h}}(\Theta, \mathbf{t})|^2 = {B_{\mathbf{h}}}^2(\Theta, \mathbf{t}) + {C_{\mathbf{h}}}^2(\Theta, \mathbf{t}),$$

where

$$B_{\mathbf{h}}(\Theta, \mathbf{t}) = \sum_{g=1}^{G} \sum_{j=1}^{N} f_j(\mathbf{h}) \, \cos\left[2\pi \, {\omega_{\mathbf{h}}}^{gj}\right], \qquad C_{\mathbf{h}}(\Theta, \mathbf{t}) = \sum_{g=1}^{G} \sum_{j=1}^{N} f_j(\mathbf{h}) \, \sin\left[2\pi \, {\omega_{\mathbf{h}}}^{gj}\right],$$

and

$${\omega_{\mathbf{h}}}^{gj} = \mathbf{h} \cdot \left(S_g \left(A^{-1}\Omega(\Theta)A\hat{\mathbf{x}}_j + \mathbf{t} \right) + s_g \right).$$

The definition of the "angle" ${\omega_{\mathbf{h}}}^{gj}$ indicates that the frequency of $I^c_{\mathbf{h}}$ is dictated by $\mathbf{h}$. The farther $\mathbf{h}$ is from the origin (hence the larger the integer components of $\mathbf{h}$), the higher the frequency of the cosine and sine functions; that is, the higher the "frequency" of the intensity $I^c_{\mathbf{h}}$. This frequency is observed when the model's coordinates are rotated and translated by $\Omega(\Theta)$ and $\mathbf{t}$, thereby changing ${\omega_{\mathbf{h}}}^{gj}$. As a result, objective functions computed from primarily high-frequency (or high-resolution) intensities will have more local minima that those computed from primarily low-resolution intensities.

4.3 Surrogate and true objective functions

Our algorithm will evaluate a surrogate function and a more accurate, true objective function. To define these functions, we must define two sets of reciprocal lattice points, one that defines a set of primarily low-frequency intensities and another that defines a larger set of higher-frequency intensities. (Here we use the adjectives "low" and "high" rather loosely.) The low-frequency intensities will be used to compute the surrogate function, while the high-frequency ones will be used to compute the "true" objective function.

More precisely, we define two index sets

$$\mathcal{I}_k = \left\{ \mathbf{h} : \; R^k_{high} \; \le \; \frac{1}{d^*(\mathbf{h})} \le \; R^k_{low} \; \right\}, \; k = 1, 2, \tag{9}$$

where R^k_{high} and R^k_{low}, $k = 1, 2$, define the resolution ranges such that

$$0 < R^2_{high} < R^1_{high} < R^2_{low} \le R^1_{low}.$$

The index set $\mathcal{I}_1$ will define a set of low-frequency intensities and $\mathcal{I}_2$ a set of high-frequency ones, where $\mathcal{I}_1 \subset \mathcal{I}_2$ in general. Normally, we choose R^2_{high} to be close to the highest resolution of the observed data, and $R^2_{low} = R^1_{low}$ to be the lowest resolution available. The choices of these "algorithmic parameters", especially R^1_{high}, are important to the performance of our algorithm, and thus require careful consideration.

For notational convenience, let $u = (\Theta, \mathbf{t}) \in \mathbb{R}^6$, and let a function $f(w(u), w^o)$ be given where $w(u)$ and w^o are two vectors of the same length whose elements are indexed by a set of $\mathbf{h} \in \mathbb{Z}^3$ in an identical order. We use $w(u)$ and w^o in place of $I^c(u)$ and I^o, respectively, to allow the flexibility of using some other quantities. Then we define two new functions associated with f and $\mathcal{I}_k$:

$$f_k(w(u), w^o) \equiv f(\{w_{\mathbf{h}}(u) : \mathbf{h} \in \mathcal{I}_k\}, \{w^o_{\mathbf{h}} : \mathbf{h} \in \mathcal{I}_k\}), \quad k = 1, 2. \tag{10}$$

The low-frequency function f_1 will be our surrogate function, while the the high-frequency function f_2 will be the "true" objective function. The smoothness of the surrogate function depends on how many higher-frequency intensities it uses. The fewer the higher-frequency intensities it has, the smoother the surrogate function is, but also the less alike it is to the true objective function.

4.4 Algorithm for the new strategy

Now we are ready to describe our algorithm, named *SOMoRe* which stands for *Search and Optimization for Molecular Replacement.*

Algorithm SOMoRe: Given an objective function $f(\cdot,\cdot)$, select index sets $\mathcal{I}_k$, $k = 1, 2$, a set $G_{\mathcal{L}}$ of 6D grid points, and positive integers M_1 and M_2. Let f_k, $k = 1, 2$ be defined as in (10).

Step 1. Evaluate the surrogate function $f_1(w(u), w^o)$ at every point in $G_{\mathcal{L}}$ and save the M_1 points u_i, $i = 1, \ldots, M_1$, corresponding to the M_1 lowest function values of $f_1(w(u), w^o)$.

Step 2. Use $\{u_i, i = 1, \ldots, M_1\}$ as the starting points for local minimization of the objective function $f_2(w(u), w^o)$, and save the M_2 local minima corresponding to the M_2 lowest function values of $f_2(w(u), w^o)$.

Step 3. Perform post-processing on the M_2 best local minima, including examination of free-function values and crystallographic packing checks, which will be defined in sections 5.4 and 5.5, respectively.

The precise definitions of the entities in the algorithm, such as $\mathcal{I}_k$, M_k, and $G_{\mathcal{L}}$, as well as the local optimization and post-processing components, will be given in the next two sections.

We emphasize here that the grid $G_{\mathcal{L}}$ will be a function of R^1_{high}, the high-resolution cut-off of the index set $\mathcal{I}_1$ that defines the low-frequency surrogate function f_1. For an appropriately chosen resolution range, $f_1(w(u), w^o)$ will have a smoothly varying landscape since it is computed from primarily low-frequency intensities, and at the same time will still capture the global behavior of the objective function $f_2(w(u), w^o)$. As a result, a coarse-grid search can be used to sample of the variable space in a relatively short amount of time to provide good starting points for the local optimization of the true objective function.

Due to the way the starting points are chosen in the algorithm, our local optimization efforts will be focused on regions of the MR variable space where MR solutions are more likely to exist. In comparison, either traditional methods or straightforward 6D searches exhaustively sample a uniformly fine grid. Also in comparison, 6D stochastic methods rely on random sampling of the variable space.

4.5 Distinct features of the new approach

We consider the following three features of our algorithm to be notable: (i) use of low resolution data, (ii) use of a low-frequency surrogate function, and (iii) use of extensive local optimization. The first two features are of course closely related.

In MR practice, normally only medium to high-resolution data are used. The use of low-resolution data and a coarse-grid search is quite atypical. Low-resolution data are not commonly used because (i) their measurement can be slightly more involved (but certainly measurable) [12], and (ii) low-resolution intensities are considered to be less accurate because they are more sensitive to the effects of the crystal's solvent exterior to the protein, (see [12], for example). Perhaps more importantly, low-resolution intensities have not provided computational success when used with the traditional approaches [3].

A grid search of a 6D, high-frequency objective function requires a very fine sampling of the variable space, hence and a massive amount of computation. This is the driving motivation for us to use a surrogate, low-frequency objective function in a more affordable, coarse-grid search. In this respect, the proposed new approach is very different from mainstream MR approaches.

There indeed already exist a few early works in the 1980's in which low-resolution data and coarser-grid searches were utilized to reduce the run time of higher-dimensional searches [29, 30]. However, such coarse searches, in a space of dimension greater than three, do not appear to have been embraced by the X-ray crystallography community at large. The aforementioned works are the only ones that we are aware of in which higher-dimensional, coarse-grid searches of low-frequency objective functions were reported to have succeeded in solving some MR problems.

One possible reason for such a lack of success by coarse-grid searches using low-resolution data is that a coarse-grid search alone is unlikely to identify a MR solution, as we will show later. When the grid is coarse, then it is likely that no grid point is close enough to a global minimum to produce a function value that is low enough to stand out in comparison to all the other function values evaluated. Moreover, the grid point corresponding to the lowest objective value is most likely not near an MR solution unless the model is nearly perfect. Thus, local optimization of many of the lowest-valued grid points found during the coarse-grid search, for example, between 500 to 1000 points, is essential for success. Such an extensive local optimization is an integral component our new strategy which shifts much of the emphasis of most MR methods from a grid search or searches to local optimization. In this respect, the new strategy is also very different from the strategies of current MR methods.

We note that the use of a 6D formulation and the use of low-resolution data go hand to hand. Not only will the calculated intensities be more accurate in a 6D formulation, but also evidence indicates that traditional objective functions perform poorly when low resolution data are used. Brünger and his co-workers reported that if predominantly low-resolution data were used, then the global minima of a commonly used traditional objective function were unlikely to correspond to MR solutions [3]. Furthermore, in [21] the author demonstrates that a commonly used 6D objective function is more accurate than its traditional counterpart when low-resolution data are used.

5 Implementation

In this section, we present a general discussion on several important issues about the implementation of Algorithm SOMoRe (for more details, see [21]), including the choice of an objective function and the specifics of the coarse-grid search, local optimization, and post-processing.

SOMoRe was implemented by modifying the freely distributed MR program Queen of Spades (Qs), which was developed by Glykos and Kokkinidis [15], because the code had many of the required front-end and calculation components, including an efficient structure factor calculation. The simulated annealing component of Qs code was replaced by the implementation of our algorithm.

5.1 Objective function: Correlation Coefficient

Many objective functions have been devised to measure disagreement between the observed and calculated intensities [6, 32]. A practical issue in choosing an objective function is scaling. Since the observed intensities are measured on a relative scale, the calculated intensities must be multiplied by a unknown scalar in order to match, in the best case, the observed ones. For some objective functions, such as the least squares function, this scaling factor must be determined as a variable.

The following correlation coefficient [19] is often used because it is scale invariant, meaning that it does not depend on the scaling factor. It is well known that "scaling insensitivity is very important when high-resolution data are not available and an accurate scale factor cannot be obtained" [13].

The correlation coefficient function can be written as follows (which may differ slightly from other definitions):

$$C(I^c(u), I^o) = \frac{\sum_{\mathbf{h}} (\, I^c_{\mathbf{h}}(u) - \langle\, I^c(u) \,\rangle \,)(\, I^o_{\mathbf{h}} - \langle\, I^o \,\rangle \,)}{\left[\sum_{\mathbf{h}} (\, I^c_{\mathbf{h}}(u) - \langle\, I^c(u) \,\rangle \,)^2\right]^{1/2} \left[\sum_{\mathbf{h}} (\, I^o_{\mathbf{h}} - \langle\, I^o \,\rangle \,)^2\right]^{1/2}}, \tag{11}$$

where $I^o_{\mathbf{h}}, I^c_{\mathbf{h}}(u)$ are the observed and calculated intensities occurring at the lattice point $\mathbf{h}$, $\sum_{\mathbf{h}}$ is the summation over all $\mathbf{h}$ in the resolution range, and $\langle\, I^o \,\rangle$ and $\langle\, I^c \,\rangle$ are the average values of the observed and calculated intensities, respectively; $\langle\, I^o \,\rangle = \sum_{\mathbf{h}} I^o_{\mathbf{h}}/m$ and $\langle\, I^c \,\rangle = \sum_{\mathbf{h}} I^c_{\mathbf{h}}/m$, where m is the number of observed and calculated intensities. The correlation coefficient can also be written as

$$C(w(u), w^o) \;=\; \frac{w(u)^T w^o}{\|\, w^o \,\|\; \|\, w(u) \,\|} = 1 - \cos \langle\, w(u), w^o \,\rangle, \tag{12}$$

where $w(u)$ and w^o can take one of two forms depending on whether $k = 1$ or 2:

$$w(u) = |F^c(u)|^k - \langle\, |F^c(u)|^k \,\rangle \quad \text{and} \quad w^o = |F^o|^k - \langle\, |F^o|^k \,\rangle. \tag{13}$$

The power of $k = 2$ (not superscript $k = 2$) means that the objective function is evaluated using intensities. An alternative is to use the magnitude of the structure factors, which corresponds to using $k = 1$. Since the latter choice often gives more satisfactory results, most of the results presented in the next section were obtained using $k = 1$. Further discussions on the performance of the two functions defined by the two data sets can be found in [21]. Finally, because we wish to pose the MR problem as a minimization problem, the objective function is $f(w(u), w^o) = 1 - C(w(u), w^o)$.

5.2 Global search

The first step of the new strategy is a coarse global search, which is defined by a set grid points that sample the variable space. The grid points, $p_j = (\Theta_j, \mathbf{t}_j)$, $j = 1, 2, \cdots$, are determined by the step sizes in the MR variables.

The sampling of rotation space is in terms of Lattman angles using the so-called *optimal Lattman sampling* because it samples the rotation space more uniformly than Eulerian angles [25]. When a constant Eulerian sampling is used, the unit sphere is sampled finely at the poles and coarsely at the equator [25]. If the optimal Lattman sampling is used instead of a constant Eulerian sampling, then the number of rotational grid points evaluated decreases by a factor of $2/\pi$ [25]. As mentioned, the step lengths of the Lattman angles that define the coarseness of the grid search are functions of the high-resolution cut-off of the data set used. For the definitions of the Lattman angles and their step lengths, see Appendix A.

The sampling of translation space is in terms of fractional coordinates. The step sizes are also functions of the high-resolution cut-off, R^1_{high}:

$$\Delta t_x = R^1_{high}/(3a), \qquad \Delta t_y = R^1_{high}/(3b), \qquad \text{and} \qquad \Delta t_z = R^1_{high}/(3c), \tag{14}$$

where division by a, b, and c converts units of Angstroms to fractional units. The larger R^1_{high} is, the larger the step size and the lower the frequency of the surrogate function. These are the step sizes used in X-PLOR Version 3.1 [2]. Finally, the 6D sampling can be achieved by six nested loops.

5.3 Local Optimization

The local optimization method implemented in SOMoRe is the BFGS quasi-Newton method (see [9], for example). We choose the BFGS method because an analytic expression for the Hessian of $C(I^c, I^o)$ is not readily available (since function evaluations involve fast Fourier transform and interpolations; see [6] for more information). For the same reason, we use a finite difference gradient calculation.

In addition, we implement two line searches: the standard Armijo backtracking line search and a specialized line search scheme. The former is used whenever the norm of the search direction is less than or equal to 0.5, which occurs in most iterations, and the specialized line search is always used for the first iteration.

In most cases, the Armijo backtracking line search works very well. However, several instances were observed when the search direction was very large. For such large directions, if a full BFGS step is taken, then the next iterate would be far from the good local neighborhood determined by the global search. In particular, this troubling behavior was observed for starting points that were close to a MR solution. However, once the more specialized line search was implemented for such large search directions, these iterates converged to the nearby MR solution.

The specialized line search scheme begins with a very small step length and then increments this step length until an increase in function value is detected. As such, it locates one of the first, if not the first, local minima along the given search direction. A pseudo code for this special "first minimum line search" is given in [21].

Finally, because these line search schemes do not necessarily satisfy the Wolfe conditions (see [27], for example), we skip an update whenever we detect that the positive definiteness of the Hessian approximation is not guaranteed. In our computation, however, such skips are quite infrequent.

5.4 Function values and "free" values

In MR, researchers primarily rely upon the examination of a number of rotation-translation pairs that have the lowest objective function values. Ideally, a researcher would hope that the lowest function values reported by a MR code correspond to solutions of the MR problem. In addition, the greater the contrast between the lowest functions values and remaining function values, the more confidence the researcher will have on the probability of having indeed obtained solutions to the MR problem [1, 26].

However, often objective function values alone are not discriminating enough to delineate solutions from non-solutions. This is the case especially when the quality of the model protein is not sufficiently high. A common remedy for this situation is the use of so-called *free values* – function values that are computed from a small percentage of intensities set aside for the purpose of cross-validation; see [15, 17], for example. Using SOMoRe, we randomly select 10% of the data, in each resolution range, to be set aside for free value calculations.

5.5 Crystallographic packing check

In reality, symmetry mates of a protein never overlap in the unit cell of a crystal. This fact can be used to check the correctness of MR solution candidates by a procedure known as a crystallographic packing check. A solution candidate can be dismissed if the so-called "packing of the model" in the crystal has some symmetry mates that inter-penetrate each other. To determine whether two symmetry mates are inter-penetrated, we compute every intra-atomic distance and then compare them to a threshold to see if any *distance violations* occur, that is, inter-atomic distances smaller than the threshold. This *a posteriori* packing check is described in [21] and [35].

6 Test Problems and Results

In this section, we describe a set of four test problems and numerical results produced by SOMoRe on these problems.

6.1 Test problems

SOMoRe has been tested on a number of problems, but we consider four test problems to be the most meaningful and representative: one has a very good model, two have either defeated or severely challenged traditional MR software, and one simulates a range of models that are complete to only 37% complete.

All test problems were taken from articles that introduce new MR software. (References are provided in the discussion of the results.) Overall, these problems are designed to answer two questions: (1) is our new approach more effective than traditional approaches on difficult MR problems? and (2) how incomplete can the model be and a MR solution still be found?

We summarize the information for the four test problem in Table 1. For each test problem, there is one molecule in the asymmetric unit so each problem is 6D. The first-three columns of the table are self-explanatory.

In the fourth column, the number of symmetry operators is listed because the time required to calculate an intensity is determined by the number of symmetry operators. In addition, MR problems involving crystals with high symmetry, that is, with symmetry

Table 1: Test Problems. The number of atoms does not include hydrogen atoms.

problem name (PDB ID	Name of protein	No. of atoms	No. of sym. ops	translation range	optimal RMSD (Å)
1AKI	lysozyme	1,001	4	$\frac{1}{2}\mathbf{a} \times \frac{1}{2}\mathbf{b} \times \frac{1}{2}\mathbf{c}$	1.2
1CGN	cytochrome c'	953	12	$\mathbf{a} \times \mathbf{b} \times \frac{1}{2}\mathbf{c}$	1.27
1B6Q	Rop	575	4	$\frac{1}{2}\mathbf{a} \times \frac{1}{2}\mathbf{b} \times \frac{1}{2}\mathbf{c}$	0.2
6RHN	histidine	878	8	$\frac{1}{2}\mathbf{a} \times \frac{1}{2}\mathbf{b} \times \frac{1}{4}\mathbf{c}$	0.3

specified by a large number of symmetry operators, are typically more difficult for traditional approaches than those with lower symmetry [1, 15, 35].

In the fifth column, we list the translation range for each problem. Due to crystallographic symmetry, the domain of the translation variables is often a proper subset of the unit cell, known as the Cheshire-group unit cell [20]. For example, for problem 1CGN, it is only necessary to search half way in the **c**-direction (or z-direction in the fractional coordinates).

The sixth column gives the so-called "optimal" Root Mean Squared Deviation (RMSD). In general, an RMSD is a norm measuring the deviation of two structures. Because each test problem uses experimental data for which the crystal structure has already been solved and deposited into the PDB, we can compute the RMSD between the coordinates of the reoriented model protein produced by SOMoRe and the coordinates of the target protein, which is known to us, and compare it to the optimal RMSD. The definition we use to compute an RMSD is given in Appendix B.

The optimal RMSD values in Table 1 are taken from the literature and are an estimate of the smallest possible RMSD between two structures, computed by different methods. (See [18, p.601] or [37], for example.) In the case that the model and the target structures are the same, then the optimal RMSD should be zero. However, typically the model and the target structures are different so the optimal RMSD is positive.

6.2 Resolution ranges and computed solution

Since the larger the high-resolution cut-off, the larger the step sizes defining the 6D grid, and the faster the grid search, it is important to know approximately the largest possible high-resolution cut-off that will still allow the surrogate function to identify good starting points.

To test the robustness of our algorithm, for each test problem we performed two global searches: one using data between ∞ and 8Å and another using data between ∞ and 10Å, except the 10% data that has been set aside for computing free values. We will refer to these two search as an *8Å search* or a *10Å search*, respectively. All available low-resolution data are used, as specified by the low-resolution cut-off of ∞. After each global search, local optimization is performed using data between ∞ and 4Å. In terms of our algorithm, these resolution cut-off values are

$$R^1_{high} = 8, \;\; R^1_{low} = \infty, \;\; R^2_{high} = 4, \;\; R^2_{low} = \infty.$$

Now we introduce a precise definition for a *computed solution*. We will call a global search *successful* if after optimization a minimum is found that has an associated RMSD within .75Å of the optimal RMSD. In addition, it is *successful* only if this RMSD is associated with a minimum that is either the lowest valued minimum or the lowest valued minimum after other minima are ruled out because they produce relatively high free values or bad crystallographic packing. Therefore, a minimizer that produces a repositioned model that has an RMSD within .75Å of the optimal RMSD is a *computed solution*.

There are many other issues, some biological and some computational, that we choose not to discuss in this paper because of space limitations and other considerations. For more details on our numerical experiments, we refer interested readers to [21].

6.3 Results for test problem 1AKI

Problem 1AKI was taken from articles [15] and [17]. The data are the observed intensities deposited with the coordinates of the protein lysozyme from chicken-egg-white (PDB ID 1AKI). The model is lysozyme from quail (PDB ID 2IHL). The optimal RMSD is reported to be 1.2Å [15, 17].

The 8Å global search was successful. However, the 10Å global search was not. During the 8Å search, the points that produced the 1,000 lowest function values were identified and used as starting points for local optimization. Of these points, the closest grid point to a global minimum was the 108th lowest valued point. This point had the lowest RMSD of 2.11Å.

During multi-start local optimization, the starting points with the four lowest RMSDs converged to local minima with associated RMSDs of 1.01Å. Furthermore, when the resulting 1,000 local minima were ranked in ascending order according to their function values, the minima with associated RMSDs of 1.01Å were at the top of the list of local minima, where they are expected to be.

The leftmost bar chart in Figure 3 shows the "true" objective 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). In addition, the right most bar chart in Figure 3 shows the corresponding RMSDs, demonstrating that the lowest valued local minima are solutions. The white bars in the RMSD bar chart indicates that the RMSD increased as a result of optimization by the height of the white bar. Most importantly, an increase in RMSD has not been observed when an starting point is close to a solution. For every bar chart, the horizontal axis is the *function value rank* of the minima when the function values are ranked in ascending order.

6.4 Results for difficult problems

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

6.4.1 Test problem 1CGN

Problem 1CGN was taken from [23]. The data are the observed intensities deposited with the coordinates of a protein known as cytochrome c' from a bacteria (PDB ID 1CGN).

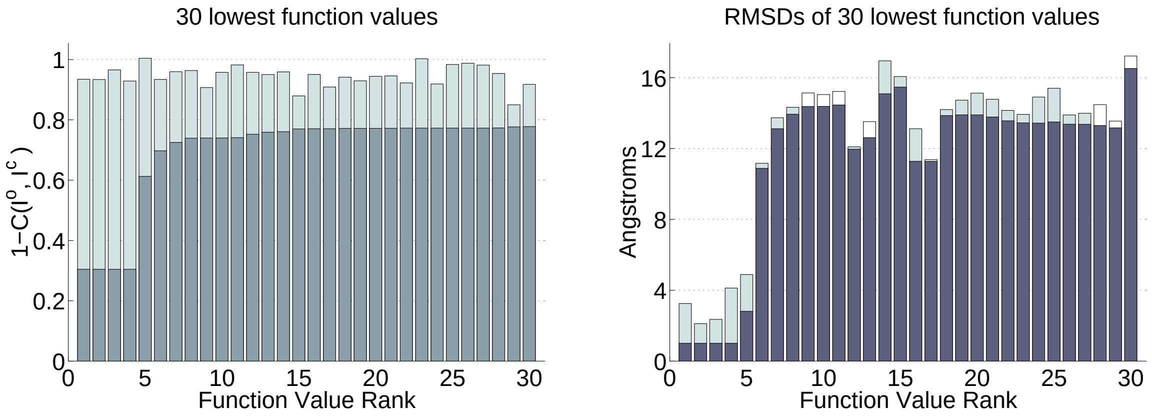

Figure 3: Local optimization results from an 8Å global search for 1AKI. Bar charts indicating function values and RMSDs before (light gray) and after optimization (dark gray). The contrast in function value between the first four points and the fifth is an accurate indicator that the remaining minima are not solutions.

The protein crystallized with twelve symmetry mates in the unit cell, that is, with high-symmetry. The model is a part of another cytochrome c' from a related bacteria (PDB ID 2CCY), which was one of the models used in the original structure determination.

The optimal RMSD between 2CCY and 1CGN is somewhat high in comparison to the other optimal RMSDs. In this respect, the model was poor. We estimate that the best possible RMSD between 2CCY and 1CGN to be 1.27Å, using the optimal RMSDs cited between 2CCY and 1CGO [1] and between 1CGO and 1CGN [10]. Our RMSDs calculations are similar to those described in [1]; for more information, see [21].

The original determination of this protein structure required a great deal of extra effort and supplemental information. The first attempts at solving the structure using traditional software were unsuccessful. X-PLOR [2] failed, and the rotation function of the MR program ALMN [7] produced "no convincing" solution [1]. Subsequently, searches using AMoRe [26] were performed using four models and two different resolution ranges, 10 to 4Å and 15 to 3.5 Å, with the "expectation [being] that the correct solution would appear in most, if not all, of the experiments, . . . even if it was not necessarily the top solution in each case" [1]. However, this too proved to be unsuccessful. In the end, Baker et al. did solve the problem using AMoRe but not without using considerably more information and additional techniques [1].

SOMoRe was successful using both an 8Å search and a 10Å search. We present only results for the former (even though the results for the latter are more impressive). Figure 4 shows the function values of the 40 lowest valued local minima. The function values themselves are not discriminating, so we need to consider the free values that are shown in the top bar chart of Figure 5. In this chart, sixteen minima have low free values. The local minima with relatively high free values should be ruled out from consideration.

If the distance violations of these sixteen minima, calculated from packing checks, are taken into consideration, every minima except two can be ruled out, as the middle bar chart

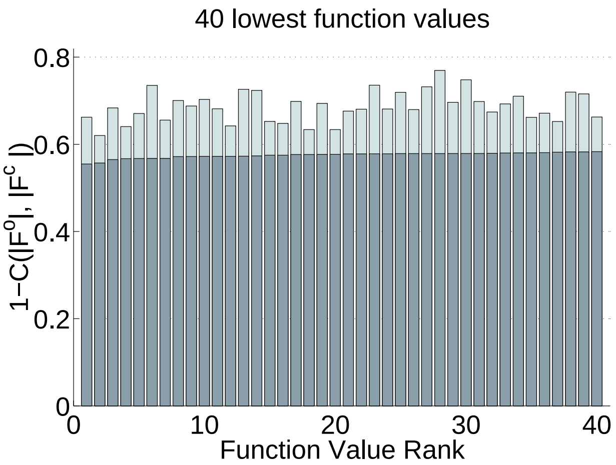

Figure 4: Function values before (light gray) and after optimization (dark gray) of 8Å global search results for 1CGN.

in Figure 4 shows. The other minima produce inter-penetration of the symmetry mates that was detected by the packing check, which used a distance threshold of 2Å. In fact, the last two remaining minima are solutions with RMSDs of 1.38Å, as shown in the bottom bar chart.

Two inter-penetrated symmetry mates, which have fourteen distance violations, are shown in Figure 6. These symmetry mates correspond to the second distance violation bar in Figure 5, which is associated with the fourth lowest valued minima.

6.4.2 Test problem 1B6Q

Test problem 1B6Q appears in [17]. The data and accurate model were both supplied to us by Nicholas Glykos [16]. The target structure is a small protein composed of two helices (PDB ID 1B6Q). The search model is an "essentially perfect" polyalanine model [16]. The optimal RMSD between this model and 1B6Q is reported to be less than 0.2Å [17].

However, Glykos and Kokkinidis 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" [17]. In general, this MR problem is more difficult for traditional approaches to solve than 6D approaches because the molecule has an elongated shape and the crystal contains relatively little solvent, only 30%. Traditional approaches are known to have difficulty on MR problems that involve such molecules [1, 6, 17] and crystals [6, 15, 33, 35]. These difficulties essentially arise because traditional approaches split the MR optimization problem into two three-dimensional problems [15, 21, 23].

In contrast, SOMoRe efficiently finds a solution to this MR problem, using either an 8Å or a 10Å search. The leftmost bar chart of Figure 7 shows the function values of the lowest valued minima found by optimizing 500 starting points that were identified by the 8Å global search. The function values of the starting points that converged to these minima

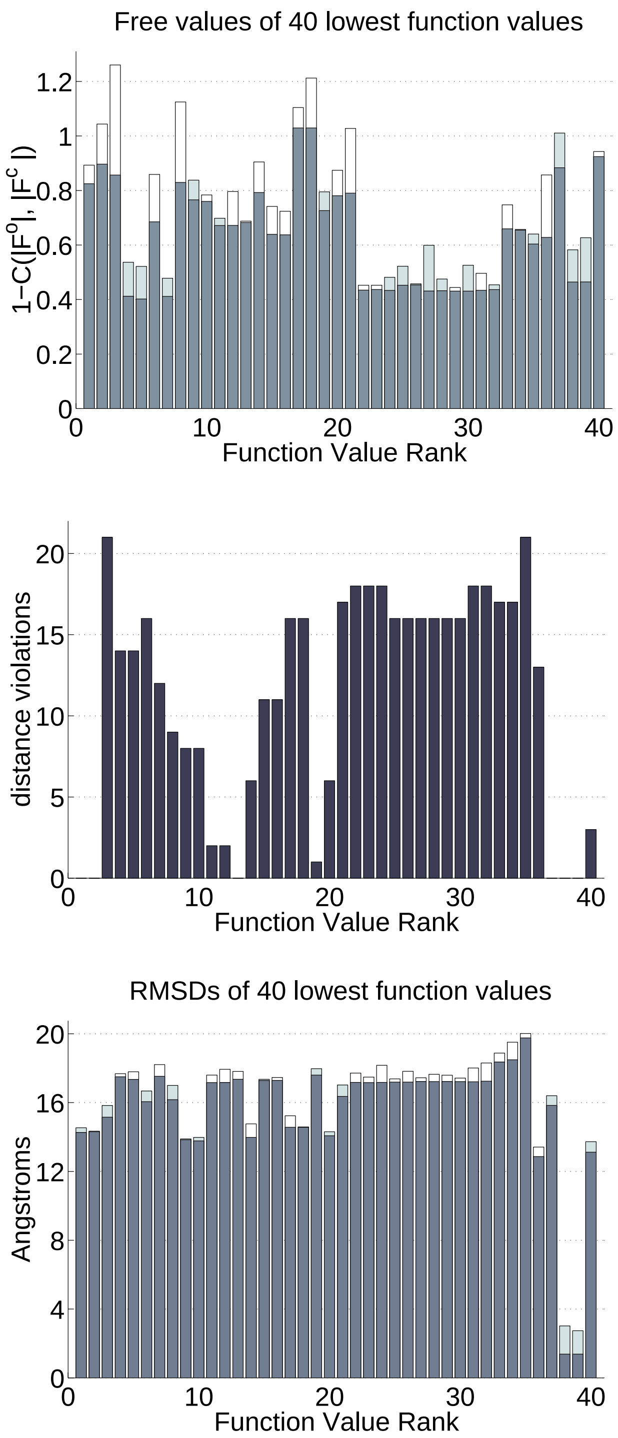

Figure 5: Free function values, distance violations, and RMSDs from optimization of 8Å search results for 1CGN, showing that the MR solution can be found if only minima with low free values and no distance violations are considered.

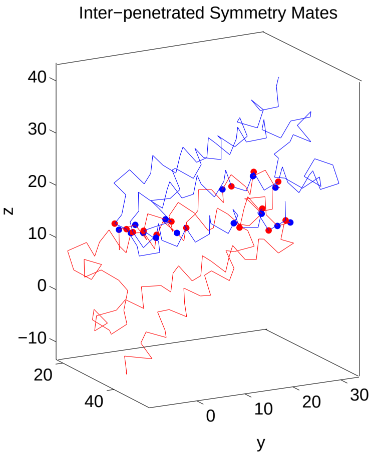

Figure 6: Two inter-penetrated symmetry mates of the repositioned model 2CCY. The lines trace the backbone of the protein, and the atom pairs that have inter-atomic distances smaller than the threshold are indicated by the large dots.

are also depicted. The rightmost bar chart in Figure 7 show the corresponding RMSDs of these minima and starting points. One can clearly see that there is a jump, or contrast, in the function values that accurately distinguishes solutions from non-solutions.

6.5 A test problem using incomplete models

Test problem 6RHN is defined in [23] and [24]. This test problem is designed to determine how much of the model protein can be removed without preventing the MR problem from being solved. In the first article, the 6D stochastic approach EPMR is compared to 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 histidine protein from a rabbit (PDB ID 4RHN). The data are the experimentally observed structure factor magnitudes deposited with the coordinates of the same protein (PDB ID 6RHN), except these coordinates were determined from a crystal with different symmetry. The optimal RMSD between the polyalanine parts of 4RHN and 6RHN is cited as 0.30Å [24].

In both articles, the polyalanine part of 4RHN is truncated. In the first article, amino acids are truncated either by five or six amino acids at a time from the initial model that contained 104 out of 115 amino acids [23]. As a result, an approximate upper bound was determined for the maximum amount of model truncations that EPMR, X-PLOR and AMoRe could tolerate and still find a MR solution. In the second article, amino acids were removed from the 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 an 8Å search, on models that contained: 104, 99, 93, 88, 82, 77, 71, 66, 60, 55, 49, and 44 amino acids or *residues*. Because the 8Å search

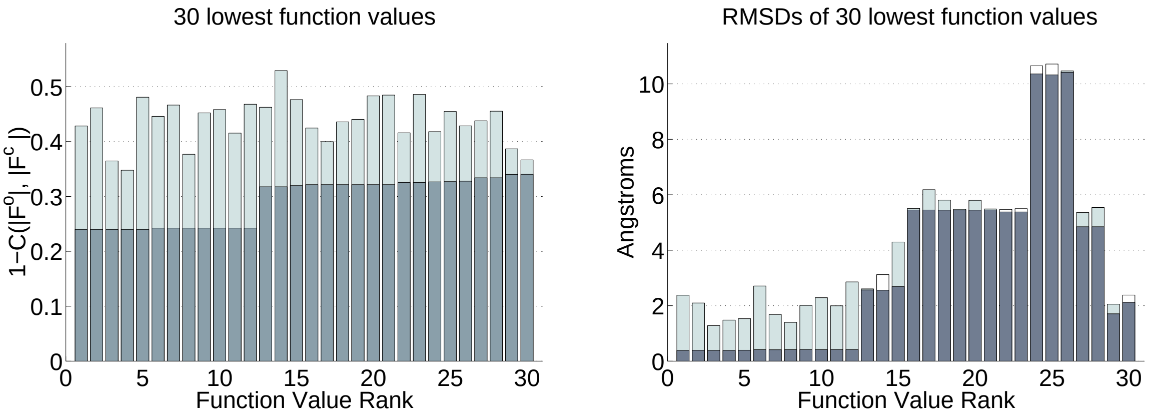

Figure 7: Function values and RMSDs before (light gray) and after optimization (dark gray) of the 8Å global search results for 1B6Q.

Table 2: Maximum amount of model truncation tolerated by MR methods

MR code	No. of amino acids in the least complete model	Truncation of the model (115 amino acids)
SOMoRe	42	63%
EPMR	44	62%
X-PLOR	≈ 62	$\approx 46\%$
AMoRe	≈ 67	$\approx 42\%$

was successful using the polyalanine model containing 44 residues and because the least complete model that EPMR could use to solve the MR problem contained 44 residues, the model containing 44 residues was truncated one residue at a time until SOMoRe failed.

The least complete model that allowed EPMR to succeed contained 44 residues [24], while the least complete model that allowed X-PLOR and AMoRe to solve the problem contained approximately 62 and 67 residues respectively [23]. In comparison, SOMoRe finds a solution to the MR problem using a model containing only 42 residues. The number of residues for X-PLOR and AMoRe are approximate because the polyalanine models for X-PLOR and AMoRe could be truncated by approximately 40% and 35%, respectively, where 100% of the model is the first 104 residues of 4RHN [23]. Table 6.5 summarizes these results.

Furthermore, according to the second article [24], if the search model has been truncated by 60% (leaving a 46 residue polyalanine model), then the search efficiency for EPMR is approximately 5% (i.e., 5 out of 100 runs were successful) [24]. In contrast, SOMoRe's search efficiency using the 44 amino acid model is 100% because it is deterministic rather than stochastic.

6.6 Summary of results

SOMoRe was successful on every test problem. Table 3 lists the RMSDs of solutions found by SOMoRe and the optimal RMSD from the literature.

Table 3: The Best RMSDs computed for each test problem, showing that the new strategy successfully solved each problem. For test problem 6RHN, the best RMSD computed when the least complete model was used was 0.95 Å.

problem name/ PDB ID	best RMSD computed (Å)	optimal RMSD (Å)
1AKI	1.01	1.2
1CGN	1.38	1.27
1B6Q	0.39	0.2
6RHN	0.32 (0.95)	0.3

In our tests, the grid points associated with the lowest RMSDs are generally not at the top of the list produced by the first step of the SOMoRe algorithm. This is so not only because the low-frequency surrogate function is not as accurate as its high-frequency counterpart, but also because the grid search is a coarse sampling of the variable space which cannot guarantee that grid points will lie very close to a global minimum. The test results demonstrate the necessity of extensive local optimization to increase contrast in function values so that MR solutions can stand out amongst non-solutions. In practice, an accurate ranking of solution candidates is essential; otherwise a researcher would be forced to carefully investigate a large number of solution candidates in order not to miss a solution.

In general, run time is a function of the number of intensities in the resolution range, and the number of symmetry operators. The more symmetry mates in the unit cell, 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 length 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.

The run times for SOMoRe are quite reasonable given that 6D searches are performed, as shown in Table 6.6. All experiments were run at Rice University on a 300MHZ R12000 processor of an SGI Origin2000 machine. Table 6.6 lists run times for both 8Å and 10Å searches. In Table 6.6, the asterisk following 1AKI indicates that the 10Å search was unsuccessful. The run times for 6RHN are the average of all run times over the range of incomplete models. In Table 6.6, the third column lists the number of intensities in the prescribed resolution range; the fourth column gives one of the angular increments in the search grid; M_1 is the number of starting point in the local optimization; and the meanings of the other columns should be clear. As can be seen, the time for local optimization is almost negligible in comparison to that of global search. Another observation is that while 8Å search is safer, 10Å search can be much faster. For example, 1CGN was solved in a little over one day in 10Å search, while it took the 8Å search almost nine days.

Finally, to show the efficiency of our approach over a straightforward 6D fine-grid search,

Table 4: Global search and optimization run times for each test problem.

Prob. name	No. of sym. ops.	No. $I^o_{\mathbf{h}}$ ∞–8Å	$\Delta\theta_2$	No. function evaluations	Search time (hrs.)	M_1	Opt. time (min.)
1AKI	4	143	8.7°	21,919,248	3.35	1,000	49
1CGN	12	99	4.8°	747,367,992	213.36	1,000	102
1B6Q	4	72	8.9°	16,720,896	1.25	500	9
6RHN	8	165	6.2°	58,832,256	19.14	1,000	111
		∞–10Å					
1AKI*	4	70	10.9°	5,982,075	0.45	1,000	49
1CGN	12	45	5.9°	210,458,470	27.12	1,000	98
1B6Q	4	41	11.2°	5,104,190	0.22	500	10
6RHN	8	86	7.7°	18,286,653	3.08	1,000	114

Table 5: Estimated run times for fine 6D global searches of objective functions computed using all data between ∞ and 4Å.

Prob. name	No. of $I^o_{\mathbf{h}}$ ∞–4Å	$\Delta\theta_2$	No. of function evaluations	Factor	Estimated search time (days)
1AKI	1133	4.4°	1,317,513,600	476.2	67
1CGN	727	2.4°	43,360,941,130	426.1	3,788
1B6Q	515	4.5°	1,009,536,576	431.9	22
6RHN	1168	3.1°	3,584,438,784	431.3	344

we estimate the run time for a 6D fine-grid search of an objective function that is computed using data between ∞ and 4Å. Because SOMoRe calculates the structure factors according to the method described by Chang and Lewis [6], the run time is linear in the number of reflections [17] and in the number of grid points. (The structure factor calculations are identical to those implemented in the MR program Qs [15, 17].) Thus, to compute the estimated run time, we determine g_2, the number of grid points that would be in such a fine search, and d_2, the number of intensities in the resolution range ∞ to 4Å (both of which are computed by SOMoRe). Then, for each problem, we compute

$$\text{factor} = \frac{g_2}{g_1} \cdot \frac{d_2}{d_1}, \tag{15}$$

where g_1 is the number of grid points in the 8Å global search grid and d_1 is the number of intensities in the resolution range ∞ and 8Å. Finally, we multiply the run time for the 8Å search by (15) and list the results in Table 5. Obviously, 6D fine-grid searches using high-frequency objective functions are still out of reach for most problems, unless a massively parallelized search is performed.

7 Conclusions

Our strategy was able to successfully and straightforwardly solve all test problems, including two that could not be directly solved by traditional codes and one with a less complete model than required by three other codes. These results suggest that the new global optimization method can extend the applicability and improve the robustness of the MR methodology.

The strengths of our method lie in the effective use of low-resolution data and a low-frequency surrogate function, and in the novel integration of a coarse-grid global search and multi-start local optimization. Unlike traditional methods, our method spends more computational effort in promising areas of the variable space where solutions are more likely to occur. Also, unlike stochastic 6D methods, our method is deterministic in nature. We predict that as computing resources improve, more accurate and robust approaches such as ours will become increasingly more attractive to the X-ray crystallographic community not only for solving more difficult problems, but for general use as well.

More recently, we have just used SOMoRe to solve a new protein structure that has not been previously determined by any other method. The lowest valued local minimizer found by SOMoRe is indeed a solution verified by inspection of electron density maps. The protein structure is currently being refined at the atomic level by researchers at the University of Wisconsin-Madison.

Acknowledgment

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.

References

[1] E. N. Baker, B. F. Anderson, and A. J. Dobbs. Use of iron anomalous scattering with multiple models and data sets to identify and refine a weak molecular replacement solution: Structure analysis of cytochrome c' from two bacterial species. *Acta Crystallographica*, D51:282–289, 1995.

[2] A. T. Brunger. *X-PLOR. A system for X-ray crystallography and NMR.* Yale University Press, New Haven, CT, 1992.

[3] A. T. Brunger. Patterson correlation searches and refinement. *Methods in Enzymology*, 276:558–580, 1997.

[4] A. T. Brunger, P. D. Adams, G. M. Clore, W. L. Delano, P. Gros, R. W. Grosse-Kunstleve, J.-S.Jiang, J. Kuszewski, M. Nilges, N. S. Pannu, R. J. Read, L. M. Rice, T. Simonson, and G. L. Warren. Crystallography and nmr system (cns): A new software system for macromolecular structure determination. *Acta Crystallographica*, D54:905–921, 1998.

[5] C.M. Bruns, I. Hubatsch, M. Ridderstrom, B. Mannervik, and J.A. Tainer. Human glutathione transferase a4-4 crystal structures and mutagenesis reveal the basis of high

catalytic efficiency with toxic lipid peroxidation products. *Journal of Molecular Biology*, 288:427–439, 1999.

[6] G. Chang and M. Lewis. Molecular replacement using genetic algorithms. *Acta Crystallographica*, D53:279–289, 1997.

[7] Number 4 Collaborative Computational Project. The ccp4 suite: Programs for protein crystallography. *Acta Crystallographica*, D50:760–763, 1994.

[8] T. Creighton. *Proteins, Structures and Molecular Properties.* John Wiley & Sons, Inc., 1959.

[9] J. E. Dennis and R. B. Schnabel. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations.* Society for Industrial and Applied Mathematics, 1996.

[10] A. J. Dobbs, B. F. Anderson, H. R. Faber, and E. N. Baker. Three-dimensional structure of cytochrome c' from two alcaligenes species and the implications for four-helix bundle structures. *Acta Crystallographica*, D52:356–368, 1996.

[11] J. Drenth. *Principles of Protein X-ray Crystallography.* Springer-Verlag, second edition, 1999.

[12] G. Evans, P. Roversi, and G. Bricogne. In-house low-resolution x-ray crystallography. *Acta Crystallographica*, D56:1304–1311, 2000.

[13] M. Fujinaga and R. J. Read. Experiences with a new translation-function program. *Journal of Applied Crystallography*, 20:517–521, 1987.

[14] J. Glusker and K. Trueblood. Crystal Structure Analysis, A Primer. Oxford University Press, 1985, 2nd Ed.

[15] N. Glykos and M. Kokkinidis. A stochastic approach to molecular replacement. *Acta Crystallographica*, D56:169–174, 2000.

[16] N. M. Glykos and M. Kokkinidis. Meaningful refinement of polyalanine models using rigid-body simulated annealing: application to the structure determination of the a31p rop mutant. *Acta Crystallographica*, D55:1301–1308, 1999.

[17] N. M. Glykos and M. Kokkinidis. Multidimensional molecular replacement. *Acta Crystallographica*, D57:1462–1473, 2001.

[18] G. H. Golub and C. F. Van Loan. *Matrix Computations.* Johns Hopkins University Press, third edition, 1996.

[19] H. Hauptman. On integrating the techniques of direct methods and isomorphous replacement. i. the theoretical basis. *Acta Crystallographica*, A38:289–294, 1982.

[20] F. L. Hirshfeld. Symmetry in the generation of trial structures. *Acta Crystallographic*, A24:301–311, 1968.

[21] D. C. Jamrog. *A New Global Optimization Strategy for the Molecular Replacement Problem.* Ph.D. thesis, Rice University, 6100 Main Street, Houston, Texas, April 2002. Also CAAM Technical Report TR02-08, 2002. (`http://www.caam.rice.edu/`).

[22] G. Jogl, X. Tao, Y. Xu, and L. Tong. Como: a program for combined molecular replacement. *Acta Crystallographica*, D57:1127–1134, 2001.

[23] C. Kissinger, D. Gehlhaar, and D. Fogel. Rapid automated molecular replacement by evolutionary search. *Acta Crystallographica*, D55:484–491, 1999.

[24] C. Kissinger, D. Gehlhaar, B. A. Smith, and D. Bouzida. Molecular replacement by evolutionary search. *Acta Crystallographica*, D57:1474–1479, 2001.

[25] E. E. Lattman. Optimal sampling of the rotation function. *Acta Crystallographica*, B28:1065–1068, 1972.

[26] J. Navaza. Implementation of molecular replacement in AMoRe. *Acta Crystallographica*, A57:1367-1372, 2001.

[27] J. Nocedal and S. J. Wright. *Numerical Optimization.* Springer, 1999.

[28] H. M. Berman and J. Westbrook and Z. Feng and G. Gilliland and T. N. Bhat and H. Weissig and I. N. Shindyalov and P. E. Bourne. The Protein Data Bank. *Nucleic Acids Research*, 28:235-242, 2000.

[29] D. Rabinovich, H. Rozenberg, and Z. Shakked. Molecular replacement: the revival of the molecular fourier transform method. *Acta Crystallographica*, D54:1336–1342, 1998.

[30] D. Rabinovich and Z. Shakked. A new approach to structure determination of large molecules by multi-dimensional search methods. *Acta Crystallographica*, A40:195–200, 1984.

[31] M. Rossman and D. Blow. The detection of sub-units within the crystallographic asymmetric unit. *Acta Crystallographica*, 15:45–52, 1962.

[32] M. G. Rossman, editor. *The Molecular Replacement Method, A Collection of Papers on the use of Non-Crystallographic Symmetry.* International Science Review Series. Gordon and Breach, Science Publishers, 1972.

[33] S. Sheriff, H. Klei, and M. Davis. Implementation of a six-dimensional search using the amore translation function for difficult molecular replacement problems. *Journal of Applied Crystallography*, 32:98–101, 1999.

[34] G. Stout and L. Jensen. X-ray Structure Determination, A Practical Guide. John Wiley & Sons, Inc., 1989, 2nd Edition.

[35] L. Tong. Combined molecular replacement. *Acta Crystallographica*, A52:782–784, 1996.

[36] A. Vagin and A. Teplyakov. Molrep: an automated program for molecular replacement. *Journal of Applied Crystallography*, 30:1022–1025, 1997.

[37] J. Yoon, Y. Gad, and Z. Wu. Mathematical modeling of protein structure using distance geometry. CAAM Technical Report TR00-24, Rice University, Houston, TX, 77005, 2000. (http://www.caam.rice.edu/).

A Lattman Angles

Lattman angles are defined in terms of Eulerian angles, which represent consecutive counter clockwise rotations about three axes. We choose the convention used by Rossman and Blow [31]: θ_1 is a rotation about the z-axis; θ_2 is a rotation about the x'-axis, the rotated x-axis; and θ_3 is a rotation about the z'-axis, the rotated z-axis. As a result, the Eulerian rotation matrix $\Omega(\theta_1, \theta_2, \theta_3)$ is

$$\begin{bmatrix} -\sin\theta_1\cos\theta_2\sin\theta_3 + \cos\theta_1\cos\theta_3 & \cos\theta_1\cos\theta_2\sin\theta_3 + \sin\theta_1\cos\theta_3 & \sin\theta_2\sin\theta_3 \\ -\sin\theta_1\cos\theta_2\cos\theta_3 - \cos\theta_1\sin\theta_3 & \cos\theta_1\cos\theta_2\cos\theta_3 - \sin\theta_1\sin\theta_3 & \sin\theta_2\cos\theta_3 \\ \sin\theta_1\sin\theta_2 & -\cos\theta_1\sin\theta_2 & \cos\theta_2 \end{bmatrix},$$

and $\Omega(\theta_1 + \pi, -\theta_2, \theta_3 + \pi) = \Omega(\theta_1, \theta_2, \theta_3)$.

Lattman angles, θ^+, θ_2 and θ^-, have the following simple relationship with Eulerian angles:

$$\theta^+ = \theta_1 + \theta_3, \qquad \theta_2 = \theta_2, \qquad \theta^- = \theta_1 - \theta_3. \tag{16}$$

To produce all possible rotations of the model, the ranges of the angles are

$$0 \le \theta^+ \le \pi, \qquad 0 \le \theta_2 \le \pi, \qquad \text{and} \qquad 0 \le \theta^- \le 2\pi.$$

Moreover, Lattman determined the optimal sampling of Lattman angular space to be:

$$\Delta\theta^+(\theta_2) = \Delta\theta_2/\cos(\theta_2/2), \quad \Delta\theta^-(\theta_2) = \Delta\theta_2/\sin(\theta_2/2),$$

where $\Delta\theta_2$ remains constant during the global search [25]. The definition of $\Delta\theta_2$ used by SOMoRe is

$$\Delta\theta_2 = 2\arcsin\left(R^1_{high}/(2(a+b+c)/3)\right), \tag{18}$$

where R^1_{high} is the high-resolution cut-off of the data set. This is also the definition used by CNS Version 1.0 [4].

B Root Mean Squared Deviation (RMSD)

Let $\{\mathbf{x}_j\}_1^N$ be the fractional coordinates of the model protein after it has been rotated and translated where $\mathbf{x}_j \in \mathbb{R}^{3\times N}$ contains the coordinates of the jth atom as in (7). In addition, let $\{\mathbf{y}_j\}_1^N$ be the fractional coordinates of the known target structure where $\mathbf{y}_j$ contains the coordinates of the jth atom. Then, we define the RMSD to be the minimum of all RMSDs computed between the known target structure and the symmetry mates of the repositioned model structure, namely,

$$\text{RMSD} = \min_{g=1,\cdots,G}\left(\frac{1}{N}\sum_{j=1}^{N}\|A\mathbf{y}_j - A(S_g\,\mathbf{x}_j + s_g)\|^2\right)^{1/2},$$

where N is the number of atoms being compared and G is the number of symmetry mates in the unit cell. We define the RMSD as such because a rotation-translation pair identified by a MR method may map the model structure onto any one of the symmetry mates of the known target structure.

In addition, some of the symmetry mates may end up in unit cells other than the unit cell of the target structure. Therefore, before calculating the RMSD, each symmetry mate should be moved an integer number of basis-vector translations so that it is the closest symmetry mate of its kind to the target structure to ensure the RMSD will be as small as possible. A pseudo code for determining the closest symmetry mate is presented in [21].