Acta Crystallographica Section A
Foundations of Crystallography
ISSN 0108-7673

Received 16 October 2003
Accepted 4 February 2004

On the equivalence between a commonly used correlation coefficient and a least-squares function

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

Many objective functions have been proposed in X-ray crystallography to solve the molecular replacement (MR) problem and other optimization problems. This paper establishes the equivalence of optimizing two of these target functions, a commonly used correlation coefficient and a least-squares function. This equivalence may exist only in the neighborhoods about the global optima or the entire MR variable space depending on whether the mean values of the observed and calculated data are subtracted from the data. In addition, an argument is presented that the correlation coefficient between structure-factor magnitudes is likely to perform better than the correlation coefficient between intensities, especially when low-resolution data are used. This prediction was tested during coarse grid searches at low resolution using the MR program *SOMoRe*.

1. Introduction

A major goal in X-ray crystallography is to compare quantitatively the observed and calculated diffraction patterns for a molecular structure being solved. This comparison is central to molecular replacement (MR), the evaluation of trial models, refinements and error estimation. The measure of closeness between observed and calculated intensities (or structure-factor magnitudes) is determined by a target function. The choice of target function has been debated and much effort has been put into developing new target functions for different applications.

When X-ray crystallography is used, solving the MR problem is often a critical step in determining a molecular structure or subdomain. The MR problem is an optimization problem to determine the orientation and position of a model protein that produces calculated intensities closest to those observed from a crystal with unknown but similar atomic structure. Various target functions for MR have been discussed; see Blundell & Johnson (1976), Harada *et al.* (1981), Fujinaga & Read (1987), Brunger & DeLano (1995), Borge *et al.* (2000), Navaza (2001), Read (2001), Tong (2001), for example.

Understanding the properties of the variously used functions, including the nature of the 'landscapes' around true optima, is an essential part of developing a phenomenological approach in practical applications. In this article, we establish that maximizing a commonly used correlation coefficient is equivalent to minimizing a least-squares function when the calculated (or observed) intensities are properly scaled. In other words, these two optimization problems have the same set of global optimizers or solutions to a given crystallography problem.

2. Objective functions

We introduce the correlation coefficient and a least-squares function, and then we prove that the set of global minimizers of the least-squares function and the set of global maximizers of the correlation coefficient are identical under some mild assumptions.

2.1. Correlation coefficient

In this paper, we are interested in the standard linear correlation coefficient that has been shown by Hauptman (1982) to be a measure of the phase error between the model and the target protein and, as a result, has been widely used in MR software packages, for example Fujinaga & Read (1987), Kissinger *et al.* (1999), Grosse-Kunstleve & Adams (2001), Glykos & Kokkinidis (2001), Navaza (2001). The correlation coefficient between observed and calculated intensities is typically written as

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

where $I^o_{\mathbf{h}}$ and $I^c_{\mathbf{h}}(u)$ are the observed and calculated intensities occurring at the lattice point $\mathbf{h}$, $u \in \mathcal{R}^n$ specifies the orientation and translation of the model protein being positioned, $\sum_{\mathbf{h}}$ 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. Of course, structure-factor magnitudes, $|F_{\mathbf{h}}| = (I_{\mathbf{h}})^{1/2}$, can be used in place of intensities.

As we will see in the following proof, it is also useful to express the correlation coefficient as the following cosine function,

$$C(w(u), w^o) = \frac{w(u)^T w^o}{\|w^o\|\|w(u)\|} = \cos\langle w(u), w^o\rangle, \qquad (2)$$

where T denotes the transpose and $\cos\langle w(u), w^o\rangle$ is the cosine of the angle between the two vectors $w(u) \in \mathcal{R}^m$ and $w^o \in \mathcal{R}^m$, which can be either $|F^c(u)|^k - \langle|F^c(u)|^k\rangle$ and $|F^o|^k - \langle|F^o|^k\rangle$ or $|F^c(u)|^k$ and $|F^o|^k$ for $k = 1$ or 2. (If $k = 1$, then structure-factor magnitudes are used and, if $k = 2$, then intensities are used.)

By expressing the correlation coefficient as a cosine function, we can see clearly that $C(w(u), w^o) \in [-1, 1]$ and that it is invariant to scaling either the calculated or observed data because scaling either vector does not change the cosine of the angle between the two vectors. However, if the average values are not subtracted from the observed and calculated intensities, then $C(w(u), w^o) \in [0, 1]$ because both $w(u)$ and w^o will be non-negative. Thus, the angle between them will be between 0 and 90°.

2.1.1. Comparison to the real-space rotation function. Although we are focusing on the equivalence between the correlation coefficient and a least-squares function, it is useful to keep in mind other widely used objective functions like the real- and reciprocal-space rotation functions. Parallels are often drawn between the correlation coefficient and these functions. See, for example, Fujinaga & Read (1987), Brunger, (1993), Grosse-Kunstleve & Adams (2001) and Navaza (2001). The real-space rotation function is the integral of the product of two Patterson functions that can be computed directly from observed and calculated intensities:

$$R(\Omega) = \int_U P^o(\mathbf{u}) P^c(\Omega\mathbf{u})\, \mathrm{d}V, \qquad (3)$$

where U is a volume of integration usually spherical, P^o is the Patterson of the unknown target structure and P^c is the Patterson of the model. The reciprocal-space formulations of the Patterson functions are

$$P^o(\mathbf{u}) = \frac{1}{V}\sum_{\mathbf{h}} I^o_{\mathbf{h}} \cos[2\pi\mathbf{h}\cdot\mathbf{u}], \qquad (4)$$

$$P^c(\Omega\mathbf{u}) = \frac{1}{V}\sum_{\mathbf{p}} I^c_{\mathbf{p}} \cos[2\pi\mathbf{p}\cdot(\Omega\mathbf{u})], \qquad (5)$$

where $\mathbf{u} = (u, v, w) \in \mathcal{R}^3$ are fractional coordinates in the Patterson unit cell, V is the volume of the unit cell and $\mathbf{h}$ and $\mathbf{p}$ are lattice points. In this formulation, when the model is rotated by $\Omega \in \mathcal{R}^{3\times3}$, the same rotation is applied to Patterson space; see Drenth (1999, p. 221), for example.

To gain physical insight into the real-space rotation function, consider the corresponding real-space analog to (4):

$$P(\mathbf{u}) = \int_{\text{unit cell}} \rho(\mathbf{r})\rho(\mathbf{r}+\mathbf{u})\, \mathrm{d}\mathbf{r}, \qquad (6)$$

where $\rho(\mathbf{r})$ is the electron density at the point $\mathbf{r}$ in the crystal's unit cell. For a proof of the equivalence between (4) and (6), see Drenth (1999). Thus, if there is a peak at $\mathbf{u}$, then $\mathbf{u}$ is a vector between two atoms. If the vector is between atoms within the protein they are called *self-vectors*, while vectors between atoms in different molecules are called *cross-vectors*.

Now, we can see that the real-space rotation function measures overlap between self-vectors of the model and self-vector sets of the target protein (one set for each copy of the target protein in the unit cell). There will, of course, be a great deal of peak overlap in each individual Patterson map because there are $N^2 - N$ peaks for a molecule with N atoms. To compensate, 'sharpened' Patterson functions can be computed using normalized structure factors (Blundell & Johnson, 1976; Stout & Jensen, 1989) or other weighting schemes applied to the intensities (Dunitz & Seiler, 1973). In addition, there will be a very large peak at the origin. However, Patterson determined that subtracting the mean value of the intensities effectively removes the large origin peak; see Lattman (1985), for example. Lastly, the volume of integration U is defined in an attempt to include only self-vectors and exclude cross-vectors since self-vectors will rotate as the model rotates and cross-vectors are insensitive to rotation of the model. Using Parseval's theorem, an expression similar to the correlation coefficient can be derived from the following quotient,

$$\frac{\int_U P^o(\mathbf{u})P^c(\Omega\mathbf{u})\, \mathrm{d}V}{[\int_U P^o(\mathbf{u})^2\, \mathrm{d}V]^{1/2}[\int_U P^c(\Omega\mathbf{u})^2\, \mathrm{d}V]^{1/2}}, \qquad (7)$$

where the denominator acts to normalize the real-space rotation function in the numerator; see Navaza (2001), for example. If we take the Fourier transform of the real-space rotation function and apply Parseval's theorem, then we get the following reciprocal-space formulation of the rotation function:

$$R'(\Omega) = \frac{1}{V^2}\sum_{\mathbf{h}}\sum_{\mathbf{p}} I^o_{\mathbf{h}} I^c_{\mathbf{p}}\, G_{\mathbf{hp}}, \qquad (8)$$

where $G_{\mathbf{hp}} = \int_U \exp[2\pi i(\mathbf{h} + \Omega^T\mathbf{p})\cdot\mathbf{x}]\, \mathrm{d}\mathbf{x}$ (see Rossmann & Blow, 1962; Rossmann, 2001). In a similar comparison, Brunger (1997) notes that

$$PC = \frac{\langle|E^o|^2|E^c(\Omega)|^2 - \langle|E^o|^2\rangle\langle|E^c(\Omega)|^2\rangle\rangle}{\langle|E^o|^4 - \langle|E^o|^2\rangle^2\rangle^{1/2}\langle|E^c(\Omega)|^4 - \langle|E^c(\Omega)|^2\rangle^2\rangle^{1/2}} \qquad (9)$$

is equal to the Fourier transform of the origin-subtracted real-space rotation function using normalized structure factors, $|E^o|$ and $|E^c|$, when the average is taken over a single shell.

So, in general, the correlation coefficient and rotation function, real or reciprocal, should behave similarly. However, in practice, when computing different formulations of the rotation function, different approximations are used, thereby introducing discrepancies of varying magnitude between these functions; see Lattman (1985), Brunger (1993), Grosse-Kunstleve & Adams (2001), for example.

2.2. Least-squares function

A natural target function to measure the disagreement between the observed and calculated intensities is the following least-squares function:

$$L(w(u), \alpha) = \|\alpha w(u) - w^o\|^2, \tag{10}$$

where $\alpha \in \mathcal{R}$ is a scale factor, $w(u) \in \mathcal{R}^m$ is the vector of calculated data and $w^o \in \mathcal{R}^m$ is the vector of observed data. Again, $w(u)$ and w^o can be either $|F^c(u)|^k - \langle |F^c(u)|^k \rangle$ and $|F^o|^k - \langle |F^o|^k \rangle$ or $|F^c(u)|^k$ and $|F^o|^k$ for $k = 1$ or 2. If the least-squares function is used, then either the calculated or the observed data should be scaled because the observed data are measured on a relative scale during the X-ray crystallography experiment. We choose to scale the calculated data, but the same effect can be achieved by scaling the observed data by $1/\alpha$. As a result, a crystallography optimization problem can be posed as the minimization of the disagreement between observed and calculated data over all possible linear scale factors and all possible parameters that position the model protein:

$$\min_{u,\alpha} L(w(u), \alpha). \tag{11}$$

The least-squares function is generally not used as an objective function for the MR problem but has been used as an objective function for rigid-body refinement. In fact, the optimal value of α (in a least-squares sense), which we will use to prove the equivalence between the two objective functions, is commonly used in rigid-body refinement programs such as *CNS* and *X-PLOR* (Brunger, 1992). As will be shown, this optimal value for α is $\sum_{\mathbf{h}} I^c_{\mathbf{h}} I^o_{\mathbf{h}} / (\sum_{\mathbf{h}} I^c_{\mathbf{h}} I^c_{\mathbf{h}}) = I^{cT} I^o / I^{cT} I^c$.

3. Proof of equivalence

In this section, we present a theorem establishing the equivalence between minimizing $L(w(u), \alpha)$ and maximizing $C(w(u), w^o)$. Since maximizing the correlation coefficient is equivalent to minimizing $1 - C(w(u), w^o)$, we will use these optimization problems interchangeably. Thus, we will show that (u^*, α^*) is a global minimizer of $L(w(u), \alpha)$ if and only if u^* is also a global minimizer of $1 - C(w(u), w^o)$. The two optimization problems will be referred to as *equivalent* if the two sets of *global* minimizers are identical. This equivalence will be symbolically denoted as $\Leftrightarrow$. The role the assumptions play with respect to local *versus* global equivalence are discussed following Theorem 1.

3.1. Theoretical results

Lemma 1. For $u, v \in \mathcal{R}^m$ and $u \neq 0$,

$$\min_{\alpha \in \mathcal{R}} \|\alpha u - v\|^2 = \|v\|^2 (1 - \cos^2 \langle u, v \rangle). \tag{12}$$

Proof. For fixed u and v, the optimal scale factor is $\alpha^* = u^T v / (u^T u)$ or the solution to the normal equations for the minimization problem above. [It is easy to see that the solution to the more general problem, $\min_{x \in \mathcal{R}^n} \|Ax - b\|^2$ for $A \in R^{m \times n}$ with full rank and $b \in R^m$, must satisfy the normal equations $A^T A x = A^T b$ because $f(x) = \|Ax - b\|^2$ is convex and differentiable; see Demmel (1997), for example.] Now, using the optimal scale factor α^*,

$$\begin{aligned}
&\left(\frac{u^T v}{u^T u} u - v\right)^T \left(\frac{u^T v}{u^T u} u - v\right) \\
&= \left(\frac{u^T v}{u^T u}\right)^2 u^T u - 2 \frac{u^T v}{u^T u} u^T v + v^T v \\
&= v^T v - \frac{(u^T v)^2}{u^T u} \\
&= v^T v \left(1 - \frac{(u^T v)^2}{v^T v u^T u}\right).
\end{aligned}$$

Finally, using the definition $\cos \langle u, v \rangle = u^T v / (\|u\| \|v\|)$,

$$v^T v \left(1 - \frac{(u^T v)^2}{v^T v u^T u}\right) = \|v\|^2 (1 - \cos^2 \langle u, v \rangle). \tag{13}$$

□

Lemma 2. Let $L(w(u), \alpha)$ be the least-squares function as defined in (10), where $w^o \in \mathcal{R}^m$ and $w : \mathcal{R}^n \rightarrow \mathcal{R}^m$ and $\alpha \in \mathcal{R}$. Assume there exists $u \in \mathcal{R}^n$ such that

$$w(u)^T w^o > 0. \tag{14}$$

Then,

$$\min_{u,\alpha} L(w(u), \alpha) \Leftrightarrow \min_{u} L(w(u), \beta(u)), \tag{15}$$

where the scale factor for the second optimization problem is

$$\beta(u) = \frac{w(u)^T w^o}{\|w(u)\|^2}. \tag{16}$$

Proof. Let

$$f(u, \alpha) = \|\alpha w(u) - w^o\|^2 \quad \text{and} \quad g(v) = \|\beta(v) w(v) - w^o\|^2. \tag{17}$$

To prove the lemma, we show

$$(u^*, \alpha^*) \in U^* = \{(\tilde{u}, \tilde{\alpha}) \text{ such that } f(\tilde{u}, \tilde{\alpha}) \leq f(u, \alpha) \ \forall \ (u, \alpha) \in \mathcal{R}^n \times \mathcal{R}\} \tag{18}$$

if and only if

$$u^* \in V^* = \{\tilde{v} \text{ such that } g(\tilde{v}) \leq g(v) \ \forall \ v \in \mathcal{R}^n\} \tag{19}$$

and

$$\alpha^* = \beta(u^*). \tag{20}$$

Let $(u^*, \alpha^*) \in U^*$. Assumption (14) implies $\|w(u^*)\| \neq 0$. Hence, as shown in Lemma 1, the unique solution to

$$\min_{\gamma} \|\gamma w(u^*) - w^o\|^2 \tag{21}$$

is well defined as $\gamma^* = w(u^*)^T w^o / \|w(u^*)\|^2 = \beta(u^*)$. Therefore,

$$\begin{aligned}
g(u^*) &= \|\beta(u^*) w(u^*) - w^o\|^2 \\
&\leq \|\alpha^* w(u^*) - w^o\|^2 = f(u^*, \alpha^*) \\
&\leq \|\beta(v) w(v) - w^o\|^2,
\end{aligned} \tag{22}$$

that is, $g(u^*) \leq g(v)$ for arbitrary v. Thus, $u^* \in V^*$. Moreover,

$$f(u^*, \alpha^*) = \|\alpha^* w(u^*) - w^o\|^2 \leq \|\beta(u^*) w(u^*) - w^o\|^2 = g(u^*) \quad (23)$$

because (u^*, α^*) is a global minimizer of $f(u, \alpha)$. Thus, $\alpha^* = \beta(u^*)$, since $\|\alpha^* w(u^*) - w^o\|^2 = \|\beta(u^*) w(u^*) - w^o\|^2$ and $\beta(u^*)$ is the unique minimizer of (21). In addition, $g(u^*) = f(u^*, \alpha^*)$.

Now, let $v^* \in V^*$ and suppose $f(v^*, \beta(v^*)) > f(u^*, \alpha^*)$. This inequality implies $g(v^*) > g(u^*)$, a contradiction. Therefore, $(v^*, \beta(v^*)) \in U^*$. □

Theorem 1. Let $C(w(u), w^o)$ be the correlation function as defined in (2), where $w^o \in \mathcal{R}^m$ and $w : \mathcal{R}^n \rightarrow \mathcal{R}^m$ is a continuous function on a compact set $D \subset \mathcal{R}^n$. Let $L(w(u), \alpha)$ be the least-squares function as defined in (10) and $\beta(u)$ be the scale factor as defined in (16). Assume that $w(u) \neq 0$, $w^o \neq 0$, and there exists $u \in \mathcal{R}^n$ such that

$$w(u)^T w^o > 0, \quad (24)$$

and assume that

$$\gamma_1 = \min_u \cos\langle w(u), w^o\rangle, \quad \gamma_2 = \max_u \cos\langle w(u), w^o\rangle, \quad |\gamma_1| < \gamma_2, \quad (25)$$

where the minimum and maximum are taken over the set D. Then over the set D

$$\min_{u,\alpha} L(w(u), \alpha) \Leftrightarrow \min_u L(w(u), \beta(u)) \Leftrightarrow \min_u 1 - C(w(u), w^o). \quad (26)$$

Proof. Given that (24) holds, $w(u^*) \neq 0$ and, by Lemma 2,

$$\min_{u,\alpha} L(w(u), \alpha) \Leftrightarrow \min_u L(w(u), \beta(u)). \quad (27)$$

Now, since $w(u) \neq 0$, by Lemma 1,

$$L(w(u), \beta(u)) = \|w^o\|^2[1 - \cos^2\langle w(u), w^o\rangle] = \|w^o\|^2[1 - C^2(w(u), w^o)]. \quad (28)$$

Thus, because $w^o \neq 0$,

$$\frac{L(w(u), \beta(u))}{\|w^o\|^2} = 1 - C^2(w(u), w^o). \quad (29)$$

So

$$\min_u L(w(u), \beta(u)) \Leftrightarrow \min_u 1 - C^2(w(u), w^o). \quad (30)$$

Now, clearly,

$$\min_u 1 - C^2(w(u), w^o) \Leftrightarrow \max_u [C^2(w(u), w^o) = \cos^2\langle w(u), w^o\rangle]. \quad (31)$$

Similarly,

$$\min_u 1 - C(w(u), w^o) \Leftrightarrow \max_u \cos\langle w(u), w^o\rangle. \quad (32)$$

Now, given the assumption that $\gamma_2 = \max_u \cos\langle w(u), w^o\rangle$, u^* is a global maximizer of $\cos\langle w(u), w^o\rangle$ if and only if $\cos\langle w(u^*), w^o\rangle = \gamma_2$. In addition, u^* is a global maximizer of $\cos^2\langle w(u), w^o\rangle$ if and only if $\cos\langle w(u^*), w^o\rangle = \gamma_2$ since $|\gamma_1| < \gamma_2 \Rightarrow \gamma_2^2 > \gamma_1^2$. Thus,

$$\min_u 1 - C^2(w(u), w^o) \Leftrightarrow \min_u 1 - C(w(u), w^o). \quad (33)$$

□

As a result of Theorem 1, the least-squares function can also be compared to the rotation function just as the correlation coefficient was. One could also argue that there is justification for subtracting the means, $\langle|F^o|^k\rangle$ and $\langle|F^c(u)|^k\rangle$, from the respective data sets when computing the least-squares function because the very large spurious origin peak of the real-space rotation function is damped by subtracting these average values. In addition, it is interesting to note that, when the least-squares function values are normalized by $\|w^o\|^2$, the least-squares function is equal to $1 - C^2(w(u), w^o)$ so that the landscape should be somewhat 'sharper' than the correlation coefficient.

3.2. Regions of equivalence

The assumptions of the lemmas and theorem are satisfied for the observed and calculated intensities (or structure-factor magnitudes) either in neighborhoods about a global minimizer u^* or for all u in the variable space. First, for a crystallography optimization problem, assumption (24) should always be satisfied because $w(u) = I^c(u) \neq 0$ and $w^o = I^o \neq 0$. Similarly, if $w(u) = I^c(u) - \langle I^c(u)\rangle$ and $w^o = I^o - \langle I^o\rangle$, then $w(u) \neq 0$ because the calculated intensities become less bright at a 'fairly rapid rate' as their distance from the origin in reciprocal space grows (Stout & Jensen, 1989, p. 165). For the same reason, $w^o \neq 0$.

Second, whether assumption (25) holds for any u in the optimization variable space D depends on the definition of $w(u)$ and w^o. {For example, in MR, u may be equal to $(\theta_1, \theta_2, \theta_3, x, y, z)$ and $D = [0, 2\pi]^3 \times [0, 1]^3$.} Assumption (25) implies that

$$\gamma_1 \leq \cos\langle w(u), w^o\rangle \leq \gamma_2, \quad (34)$$

where $|\gamma_1| < \gamma_2$. If $w(u) = I^c(u) - \langle I^c\rangle$ and $w^o = I^o - \langle I^o\rangle$, then (25) may be satisfied only in a neighborhood of a global minimizer u^*. That is, there will only be local equivalence between the two functions in a neighborhood of the global minimum.

If the average values are subtracted, then the cosine of the angle between the two vectors $w(u)$ and w^o may be large and violate assumption (25). However, if the model protein is accurate enough, then, in a neighborhood of the global minimizer u^*, the initial angle between the observed and calculated data should be small enough so that subtracting the average values will not increase the angle so much as to violate (25).

We now give a concrete example that shows that if the means are subtracted then there may be regions for which $1 - C(w(u), w^o)$ and the least-squares function are not equivalent. Suppose the means are subtracted and $C(w(u), w^o)$

has a local minimum at u^* such that $C(w(u^*), w^o) < 0$. Then, $1 - C(w(u), w^o)$ will have a local maximum at u^*, but $\|w^o\|^2[1 - C^2(w(u), w^o)] = L(w(u), \beta(u))$ will have a local minimum at u^*. Thus, optimization of the two functions will not be equivalent near u^*.

In contrast, if the means are not subtracted, then $\cos\langle w(u), w^o\rangle$ will always be non-negative and assumption (25) will hold for all u; that is, equivalence between the two functions will hold for the entire optimization variable space D. (Of course, the above arguments are the same if structure-factor magnitudes are used in place of intensities.)

3.3. Some numerical results

To demonstrate graphically the equivalence between these two objective functions, we compute two-dimensional slices of $1 - C(w(u), w^o)$ and the least-squares function for a MR problem. The data were measured from a crystal of a peptidic analog of the antibiotic molecule trichogin A IV. There was only one molecule in the asymmetric unit of a relatively small unit cell ($a = 14.56$, $b = 11.759$ and $c = 9.473$ Å), and the space group is $P2_1$. Thus, the MR problem is five-dimensional. The model molecule is the structure originally determined from the experimental data, that is, the analog's 38 non-hydrogen atoms (Crisma *et al.*, 1994), so the model is exact. As a result, one MR solution corresponds to no rotation and no translation of the model. In addition, because the molecule crystallized according to space group $P2_1$, there is one symmetry-related solution at the Eulerian angles $(\theta_1, \theta_2, \theta_3) = (\pi, \pi, 0)$ and the translation $\mathbf{t} = (0, 0, 0)$.

We compute two-dimensional slices (level sets) of the five-dimensional functions in the $\theta_1\theta_2$-plane, where θ_1 and θ_2 are Eulerian angles. To produce the level sets, θ_3 is held fixed at zero and $\mathbf{t} = (0, 0, 0)$. All intensities with resolutions between ∞ and 7 Å, that is nine intensities, were used to compute the target functions so the landscapes are relatively smooth. The level sets are shown in Fig. 1. Clearly, the global minima of the two functions occur in the same positions, and the two functions are highly similar even though the least-squares function was computed using intensities from which the mean value was not subtracted, while the correlation coefficient was computed using intensities from which the mean value was subtracted [or using equation (1)].

Recent developments in molecular replacement and other crystallographic optimizations have included target functions based on maximum-likelihood estimates in a Bayesian approach; see Read (2001), Murshudov *et al.* (1997), for example. These approaches involve statistical terms that depend on the fraction of the unknown structure depicted by a model and also the ability of the model to account for accurate estimates of the observed diffraction data. These targets are hard to relate mathematically to the targets analyzed in this report. Furthermore, because of the dependence on large numbers or terms for adequate sampling, the maximum-likelihood target is not calculable for the example given here. Ongoing studies of the newer target functions seem appropriate and are under way.

3.3.1. Intensities *versus* structure-factor magnitudes. Now that we have shown that the two functions are equivalent, one must also choose to work with either intensities or structure-

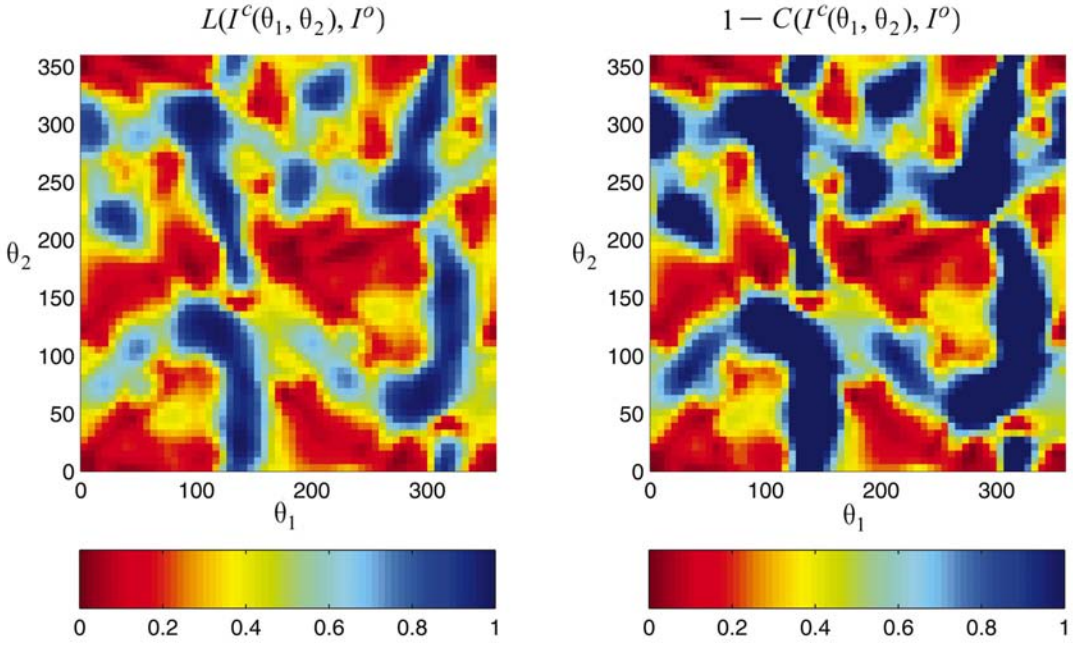

Figure 1
The level sets of the low-resolution five-dimensional correlation coefficient and least-squares function that has been normalized by dividing by $\|I^o\|^2$. Clearly, the global minima occur in the same positions and the two functions are highly similar even though the least-squares function was computed using intensities from which the mean value was not subtracted, while the correlation coefficient was computed using intensities from which the mean value was subtracted. All intensities between 500 and 7 Å, that is nine intensities, were used to compute the target functions so the landscapes are relatively smooth.

factor magnitudes and whether or not to subtract the means. On the practical issue of using intensities or structure-factor magnitudes, during the development of the MR program *SOMoRe* (search and optimization for molecular replacement), we noticed that the correlation coefficient computed using structure factors tended to perform better (Jamrog *et al.*, 2003). This is not surprising since using structure factors is, in essence, using the intensities weighted according to the errors in their measurement. Because the experimental observation of a diffraction intensity is a stochastic process with underlying Poisson statistics, the estimated error in the measurement is proportional to the square root of the intensity. As a result, a proper weighting scheme for each intensity would have a multiplier of $1/(I_{\mathbf{h}})^{1/2}$ transforming $C(I^c(u), I^o)$ into $C(|F^c(u)|, |F^o|)$. Otherwise more weight will be given to the innermost intensities (which are likely to be the most inaccurate), especially when computing a 'low-resolution' target function.

To estimate the amount of data sufficient to compute a reasonably accurate low-resolution correlation coefficient (or 'surrogate' function) that can be used to identify regions of the MR variable space where solutions are likely to exist, *SOMoRe* was used to compute both $C(I^c(u), I^o)$ and $C(|F^c(u)|, |F^o|)$ using equation (1) and, during some coarse searches of these surrogate functions, viable starting points for multi-start local optimization could only be found when $C(|F^c(u)|, |F^o|)$ was used. By viable starting points, we mean points that were sufficiently close to a global minimizer such that the local optimization method BFGS (Broyden–Fletcher–Goldfarb–Shanno) converged to a solution of the MR problem. As a result, we feel $C(|F^c(u)|, |F^o|)$ is likely to be more accurate than $C(I^c(u), I^o)$, especially when low-resolution data are used. Similarly, Glykos & Kokkinidis (2001) and Kissinger *et al.* (1999) have also advocated the use of structure-factor magnitudes over intensities.

3.3.2. Subtracting the means. We performed two experiments using *SOMoRe* to determine if there is any appreciable difference in accuracy between the correlation coefficient $C(|F^c(u)|, |F^o|)$ defined by (1), computed using data from which the means have been subtracted, and the following correlation coefficient:

$$C'(|F^c(u)|, |F^o|) = \frac{\sum_{\mathbf{h}} |F^c_{\mathbf{h}}(u)||F^o_{\mathbf{h}}|}{[\sum_{\mathbf{h}} |F^c_{\mathbf{h}}(u)|^2]^{1/2}[\sum_{\mathbf{h}} |F^o_{\mathbf{h}}|^2]^{1/2}}. \qquad (35)$$

For the first experiment, both functions were used to optimize the same set of starting points [found from a coarse 8 Å global search using $C(I^c(u), I^o)$]. Then, to determine the accuracy of the global minimizers found, the root-mean-square deviations (RMSDs) between the atomic coordinates of the target protein and the model positioned according to the two sets of global minimizers were computed. The difference between the two sets of RMSDs were in the third decimal place of an Å. Thus, using $C(|F^c(u)|, |F^o|)$ or $C'(|F^c(u)|, |F^o|)$ did not have an appreciable effect on the accuracy of the local optimization performed (Jamrog, 2002). However, the function values at the global minimums were 0.08 for $C'(|F^c(u)|, |F^o|)$ and 0.24 for $C(|F^c(u)|, |F^o|)$. Thus, the cosine of the angle between the vectors grew because the mean value was subtracted from both data sets (growing from about 23 to 40.5°). A similar comparison was also performed for a more difficult MR test problem, more difficult because the model protein was not as accurate. Both functions were used to perform a coarse global search of the surrogate function and local optimization and, again, from a practical point of view, we did not see an appreciable difference. Both functions were accurate enough to provide solutions to the MR problem.

We would like to thank Gary Wesenberg for his helpful comments. DJ was supported in part by a training fellowship from the Keck Center for Computational Biology (NSF GRT Grant BIR92-56580, NSF RTG BIR-94-13229 and NLM 5 T15 LM07093) and NSF Grant DMS-9973339. YZ was supported in part by DOE Grant DE-FG03-97ER25331, DOE/LANL Contract 03891-99-23 and NSF Grant DMS-9973339. GNP was supported by the National Institute of Health GM-64598, NSF Grant ACI-0082645 and a Vilas Associate Award from the University of Wisconsin–Madison.

References

Blundell, T. & Johnson, L. (1976). *Protein Crystallography*. New York: Academic Press.
Borge, J., Alvarex-Rua, C. & Garcia-Granda, S. (2000). *J. Mol.Biol.* D**56**, 735–746.
Brunger, A. T. (1992). *X-PLOR. A System for X-ray Crystallography and NMR*. New Haven, CT: Yale University Press.
Brunger, A. T. (1993). *ImmunoMethods*, **3**, 180–190.
Brunger, A. T. (1997). *Methods Enzymol.* **276**, 558–580.
Crisma, M., Valle, G., Monaco, V., Formaggio, F. & Toniolo, C. (1994). *Acta Cryst.* C**50**, 563–565.
DeLano, W. & Brunger, A. T. (1995). *Acta Cryst.* D**51**, 740–748.
Demmel, J. W. (1997). *Applied Numerical Linear Algebra*. Philadelphia, PA: Society for Industrial and Applied Mathematics.
Drenth, J. (1999). *Principles of Protein X-ray Crystallography*, 2nd ed. Berlin: Springer-Verlag.
Dunitz, J. D. & Seiler, P. (1973). *Acta Cryst.* B**29**, 589–595.
Fujinaga, M. & Read, R. J. (1987). *J. Appl. Cryst.* **20**, 517–521.
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.
Harada, Y., Lifchitz, A., Berthou, J. & Jolles, P. (1981). *Acta Cryst.* A**37**, 398–406.
Hauptman, H. (1982). *Acta Cryst.* A**38**, 289–294.
Jamrog, D. C. (2002). PhD thesis, Rice University, Houston, Texas, USA. Technical Report TR-0208 at http://www.caam.rice.edu/.
Jamrog, D. C., Zhang, Y. & Phillips, G. N. Jr (2003). *Acta Cryst.* D**59**, 304–314.
Kissinger, C., Gehlhaar, D. & Fogel, D. (1999). *Acta Cryst.* D**55**, 484–491.
Lattman, E. E. (1985). *Methods Enzymol.* **115**, 55–77.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Navaza, J. (2001). *Acta Cryst.* D**57**, 1367–1372.
Read, R. J. (2001). *Acta Cryst.* D**57**, 1373–1382.
Rossmann, M. (2001). *Acta Cryst.* D**57**, 1360–1366.
Rossmann, M. & Blow, D. (1962). *Acta Cryst.* **15**, 24–31.
Stout, G. & Jensen, L. (1989). *X-ray Structure Determination, a Practical Guide*, 2nd ed. New York: John Wiley and Sons.
Tong, L. (2001). *Acta Cryst.* D**57**, 1383–1389.