

On the Equivalence Between a Commonly Used Correlation Coefficient and a Least Squares Function

Diane C. Jamrog* George N. Phillips, Jr.† Yin Zhang*

January 2003
(revised October 2003)

Abstract

Many objective functions have been proposed in X-ray crystallography to solve the molecular replacement (MR) problem and other optimization problems. In this paper, we establish the equivalence between optimizing two target functions: a commonly used correlation coefficient and a least squares function. This equivalence may be in neighborhoods about the global optima or the entire MR variable space depending on whether the average values of the observed and calculated data are subtracted from observed and calculated data. In addition, we also present an argument that the correlation coefficient between structure factor magnitudes is likely to perform better than the correlation coefficient between intensities. This was confirmed by the MR program SOMoRe, especially when low-resolution data were used.

1 Introduction

A major goal in X-ray crystallography is to quantitatively compare the observed and calculated diffraction patterns for a molecular structure being solved. This comparison

*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.

is useful in the evaluation of trial protein models, refinements of structures, 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.

In particular, solving the molecular replacement (MR) problem is often a critical step in determining a detailed molecular structure. The MR problem is an optimization problem to determine the orientation and position of a model protein that produce calculated intensities closest to those observed from a crystal with unknown atomic structure. Various target functions for the MR problem have been discussed. See [1, 2, 4, 6, 8, 11, 12, 14], for example.

In this article, we establish that maximizing a correlation coefficient, which is a commonly used objective function for the MR problem, is equivalent to minimizing a least squares function when the calculated intensities are properly scaled and some mild assumptions are met. In other words, the two respective optimization problems have the same set of global optimizers.

2 Objective functions

We introduce a correlation coefficient and a least squares function, and then we prove that the set of global optimizers of the respective optimization problems are identical under some mild assumptions.

2.1 Correlation coefficient

The Pearson correlation coefficient is often used to solve the MR problem both because it can be interpreted in terms of Patterson functions [3, 6] and because it is scale invariant. In the X-ray crystallography literature, this correlation coefficient is typically written as:

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

where $I_{\mathbf{h}}^o$ and $I_{\mathbf{h}}^c(u)$ 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 being positioned

by MR, $\sum_{\mathbf{h}}$ is the summation over all \mathbf{h} (or intensities) 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 ($|F_{\mathbf{h}}^c|$ and $|F_{\mathbf{h}}^o|$) can be used in place of intensities; $|F_{\mathbf{h}}^c|^2 = I_{\mathbf{h}}^c$ and $|F_{\mathbf{h}}^o|^2 = I_{\mathbf{h}}^o$.

The correlation coefficient can be written in more general terms as

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

where $\cos \langle w(u), w^o \rangle$ is the cosine of the angle between the two vectors $w(u) \in \mathcal{R}^m$ and $w^o \in \mathcal{R}^m$. The vectors $w(u)$ and w^o are typically defined to be $|F^c(u)|^k - \langle |F^c(u)|^k \rangle$ and $|F^o|^k - \langle |F^o|^k \rangle$, respectively. If $k = 1$, then structure factor magnitudes are used, and if $k = 2$, then intensities are used.

Thus, the correlation coefficient is scale invariant, because the correlation coefficient is equal to the cosine of an angle, and scaling either vector does not change the cosine of the angle between the two vectors. Obviously, $C(w(u), w^o) \in [-1, 1]$. 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. Finally, the MR problem can be posed as

$$\min_u (1 - C(w(u), w^o)). \quad (3)$$

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:

$$L(w(u), \alpha) = \|\alpha w(u) - w^o\|^2, \quad (4)$$

where $\alpha \in \mathcal{R}$ is a scale factor, $w(u) \in \mathcal{R}^m$ is the vector of calculated intensities, and $w^o \in \mathcal{R}^m$ is the vector of observed intensities. In general, $w(u)$ and w^o can be either $|F^c(u)|^k$ and $|F^o|^k$, respectively, or $|F^c(u)|^k - \langle |F^c(u)|^k \rangle$ and $|F^o|^k - \langle |F^o|^k \rangle$, respectively, for $k = 1$ or 2 . If the least squares function is used, then either the calculated or the observed intensities should be scaled because the observed intensities are measured on a relative scale during the X-ray crystallography experiment. We choose to scale the calculated intensities, but the the same effect can be achieved by scaling the observed

intensities by $1/\alpha$. As a result, the MR problem can be posed as the minimization of the disagreement between observed and calculated intensities over all possible linear scale factors and all possible orientations and translations of the model protein:

$$\min_{u, \alpha} L(w(u), \alpha). \quad (5)$$

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, a computational process that is often used to refine a MR solution. Most likely, the MR problem is not posed as (5) because at face value (5) may appear to be a more difficult optimization problem than (3) due to the higher dimension of the variable space. However, the first two lemmas of the next section suggest the appropriate scale factor α . If this scale factor is used, then optimizing the above least squares function does not involve any more variables than optimizing the correlation coefficient.

3 Proof of Equivalence

In this section, we present four lemmas and a theorem establishing the equivalence between minimizing $L(w(u), \alpha)$ and $1 - C(w(u), w^o)$. In other words, (u^*, α^*) is a global minimizer of $L(w(u), \alpha)$ if and only if u^* is also a global minimizer of $1 - C(w(u), w^o)$. Two optimization problems will be referred to as *equivalent* if the two sets of *global* optimizers are identical. This equivalence will be symbolically denoted as \Leftrightarrow .

The sequence of theoretical results begins with a lemma that uses the result that if $w(u) \neq 0$, then the optimal scale factor for the least squares function is

$$\beta(u) = w(u)^T w^o / (w(u)^T w(u)).$$

The second lemma establishes that minimizing $L(w(u), \alpha)$ is equivalent to minimizing $L(w(u), \beta(u))$ when the above optimal scale factor $\beta(u)$ is used. The third lemma shows that $1 - C^2(w(u), w^o) = L(w(u), \beta(u)) / \|w^o\|^2$ provided $\|w^o\| \neq 0$ and $\|w(u)\| \neq 0$. Finally, the theorem ties all these results together to show the equivalence between minimizing $L(w(u), \alpha)$ and $1 - C(w(u), w^o)$ under mild assumptions. The role the assumptions play with respect to the regions of equivalence are discussed following the theorem.

3.1 Theoretical results

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

$$\min_{\alpha \in \mathcal{R}} \|\alpha u - v\|^2 = \left\| \frac{u^T v}{u^T u} u - v \right\|^2 = \|v\|^2 (1 - \cos^2 \langle u, v \rangle). \quad (6)$$

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. Now, using this optimal scale factor,

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

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

$$v^T v \left(1 - \frac{(u^T v)^2}{v^T v u^T u} \right) = \|v\|^2 (1 - \cos^2 \langle u, v \rangle). \quad (7)$$

◇

Lemma 2 Let $L(w(u), \alpha)$ be the least squares function as defined in (4), 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

$$w(u)^T w^o > 0. \quad (8)$$

Then

$$\min_{u, \alpha} L(w(u), \alpha) \Leftrightarrow \min_u L(w(u), \beta(u)), \quad (9)$$

where

$$\beta(u) = \frac{w(u)^T w^o}{\|w(u)\|^2}. \quad (10)$$

Proof: Let

$$f(u, \alpha) = \|\alpha w(u) - w^o\|^2 \quad \text{and} \quad g(v) = \|\beta(v)w(v) - w^o\|^2. \quad (11)$$

To prove the lemma, we show

$$(u^*, \alpha^*) \in U^* = \{ (\tilde{u}, \tilde{\alpha}) \text{ such that } f(\tilde{u}, \tilde{\alpha}) \leq f(u, \alpha) \forall (u, \alpha) \in \mathcal{R}^n \times \mathcal{R} \} \quad (12)$$

if and only if

$$u^* \in V^* = \{ \tilde{v} \text{ such that } g(\tilde{v}) \leq g(v) \forall v \in \mathcal{R}^n \} \quad (13)$$

and

$$\alpha^* = \beta(u^*). \quad (14)$$

Let $(u^*, \alpha^*) \in U^*$. Assumption (8) implies $\|w(u^*)\| \neq 0$. Hence, as shown in Lemma 1, the unique solution to

$$\min_{\gamma} \|\gamma w(u^*) - w^o\|^2 \quad (15)$$

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

$$g(u^*) = \|\beta(u^*) w(u^*) - w^o\|^2 \leq \|\alpha^* w(u^*) - w^o\|^2 = f(u^*, \alpha^*) \leq \|\beta(v) w(v) - w^o\|^2, \quad (16)$$

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

$$f(u^*, \alpha^*) = \|\alpha^* w(u^*) - w^o\|^2 \leq \|\beta(u^*) w(u^*) - w^o\|^2 = g(u^*), \quad (17)$$

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 (15). 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, $(v^*, \beta(v^*)) \in U^*$. \diamond

Lemma 3 *Let $C(w(u), w^o)$ be the correlation function as defined in (2), where $w^o \in \mathcal{R}^m$ and $w : \mathcal{R}^n \rightarrow \mathcal{R}^m$. Let $L(w(u), \alpha)$ be the least squares function as defined in (4), and $\beta(u)$ be the scale factor as defined in (10). Assume $w(u) \neq 0$ and $w^o \neq 0$. Then*

$$1 - C^2(w(u), w^o) = \frac{L(w(u), \beta(u))}{\|w^o\|^2}. \quad (18)$$

Proof: Since $w(u) \neq 0$, by Lemma 1,

$$L(w(u), \beta(u)) = \|w^o\|^2 \left(1 - \cos^2 \langle w(u), w^o \rangle\right) = \|w^o\|^2 \left(1 - C^2(w(u), w^o)\right). \quad (19)$$

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

$$1 - C^2(w(u), w^o) = \frac{L(w(u), \beta(u))}{\|w^o\|^2}. \quad (20)$$

◇

Lemma 4 *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 continuous function on a compact set $D \subset \mathcal{R}^n$. Assume*

$$\gamma_1 = \min_u \cos\langle w(u), w^o \rangle, \quad \gamma_2 = \max_u \cos\langle w(u), w^o \rangle, \quad \text{and } |\gamma_1| < \gamma_2, \quad (21)$$

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

$$\min_u 1 - C^2(w(u), w^o) \Leftrightarrow \min_u 1 - C(w(u), w^o). \quad (22)$$

Proof: Clearly,

$$\min_u 1 - C^2(w(u), w^o) \Leftrightarrow \max_u C^2(w(u), w^o) \Leftrightarrow \max_u \cos^2\langle w(u), w^o \rangle. \quad (23)$$

Similarly,

$$\min_u 1 - C(w(u), w^o) \Leftrightarrow \max_u \cos\langle w(u), w^o \rangle. \quad (24)$$

Now, given assumption (21), u^* is a global maximizer of $\cos\langle w(u), w^o \rangle$ if and only if $\cos\langle w(u^*), w^o \rangle = \gamma_2$. Similarly, 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 $\gamma_2^2 > \gamma_1^2$. ◇

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 (4), and $\beta(u)$ be the scale factor as defined in (10). Assume there exists $u \in \mathcal{R}^n$ such that*

$$w(u)^T w^o > 0, \quad (25)$$

and

$$\gamma_1 = \min_u \cos\langle w(u), w^o \rangle, \quad \gamma_2 = \max_u \cos\langle w(u), w^o \rangle, \quad \text{and } |\gamma_1| < \gamma_2, \quad (26)$$

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

$$\min_{u, \alpha} L(w(u), \alpha) \Leftrightarrow \min_u 1 - C(w(u), w^o). \quad (27)$$

Proof: Given (25), $w(u^*) \neq 0$, and by Lemma 2,

$$\min_{u,\alpha} L(w(u), \alpha) \Leftrightarrow \min_u L(w(u), \beta(u)). \quad (28)$$

Given (25), by Lemma 3,

$$\min_u L(w(u), \beta(u)) \Leftrightarrow \min_u 1 - C^2(w(u), w^o). \quad (29)$$

Finally, given (26), by Lemma 4,

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

◇

3.2 Regions of equivalence

The assumptions of the lemmas and theorem are satisfied for the observed and calculated intensities, either in neighborhoods about a global minimizer u^* or for all u in the MR variable space.

First for the MR problem, assumption (25) should always be satisfied because $w(u) = I^c(u) \neq 0$ and $w^o = I^o \neq 0$. The calculated and observed intensities should not all be equal to zero. 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 grows [13, p. 165]. For the same reason, $w^o \neq 0$.

Second, whether assumption (26) 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 (26) implies that

$$\gamma_1 \leq \cos \langle w(u), w^o \rangle \leq \gamma_2, \quad (31)$$

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 (26) may be satisfied only in a neighborhood of a global minimizer u^* rather than for all u .

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 (26). 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 (26).

We now give a concrete example that shows that if the means are subtracted, then there may be regions for which the function $1 - C(w(u), w^o)$ 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 the cosine will always be non-negative, and assumption (26) and Lemma 4 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.)

Finally, we note that for the least squares function there does not seem to be a justification for subtracting $\langle I^o \rangle$ and $\langle I^c \rangle$ from the observed and calculated intensities. On the other hand, when the correlation coefficient is used as a traditional rotation function, Brunger notes that the numerator of the correlation coefficient is equal to the real space rotation function given some assumptions [3], and for the real space rotation function, the very large spurious origin peak is damped by subtracting the average values of the intensities; see [10, 13], for example.

4 Intensities verses structure factor magnitudes

During our development of the MR program SOMoRe [9], we used sets of low-resolution intensities and structure factor magnitudes to compute “surrogate” functions, that is, functions that could be sampled relatively quickly to identify regions of the MR variable space where solutions might exist. Based on our numerical experimentation with SOMoRe, we feel that $C(|F^o|, |F^c|)$ is likely to be more accurate than $C(I^o, I^c)$ especially when low-resolution data is used. Similarly, Glykos and Kokkindis also advocate the general use of structure factor magnitudes over intensities [7].

For example, during some of SOMoRe’s deterministic searches of the surrogate function, good starting points for local optimization could not be found when $C(I^o, I^c)$ was used but could be found when $C(|F^o|, |F^c|)$ was used. (By good starting points, we mean starting

points that were sufficiently close to a global minimizer such that the local optimization method BFGS could converge to the solutions of the MR problems.)

Besides the numerical evidence presented in [9], we believe that $C(|F^o|, |F^c|)$ is more accurate because this function essentially incorporates a weighting of the intensities according to the error in their measurements. The observation of the diffraction intensities during an X-ray crystallography experiment is a stochastic process with underlying Poisson statistics. Thus, the error in the measurements is proportional to the square root of the intensities. A proper point-wise weighting scheme of $w_{\mathbf{h}}I_{\mathbf{h}}$ would have a multiplier of $w_{\mathbf{h}} = 1/\sqrt{I_{\mathbf{h}}}$, and this weighting effectively produces $C(|F^o|, |F^c|)$ from $C(I^o, I^c)$.

Acknowledgments

Diane Jamrog 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. George N. Phillips, Jr. was supported by the National Institute of Health GM-64598. Yin Zhang was supported in part by DOE Grant DE-FG03-97ER25331, DOE/LANL Contract 03891-99-23 and NSF Grant DMS-9973339.

References

- [1] T. Blundell and L. Johnson. *Protein Crystallography*. Academic Press, 1976.
- [2] J. Borge, C. Alvarez-Rua, and S. Garcia-Granda. A new vector-search rotation function: image seeking functions revisited in macromolecular crystallography. *Journal of Molecular Biology*, D56:735–746, 2000.
- [3] A. T. Brunger. Patterson correlation searches and refinement. *Methods in Enzymology*, 276:558–580, 1997.
- [4] A. T. Brunger and W. DeLano. The direct rotation function: Rotational patterson correlation search applied to molecular replacement. *Acta Crystallographica*, D51:740–748, 1995.
- [5] M. Crisma, G. Valle, V. Monaco, F Formaggio, and C. Toniolo. n^α -benzyloxycarbonyl- α -aminoisobutryl-glycyl-l-isoleucyl-l-luceince methyl ester monohydrate. *Acta Crystallographica*, C50:563–565, 1994.
- [6] M. Fujinaga and R. J. Read. Experiences with a new translation-function program. *Journal of Applied Crystallography*, 20:517–521, 1987.
- [7] N. M. Glykos and M. Kokkinidis. Multidimensional molecular replacement. *Acta Crystallographica*, D57:1462–1473, 2001.
- [8] Y. Harada, A. Lifchitz, J. Berthou, and P. Jolles. A translation function combining packing and diffraction information: An application to lysozyme (high-temperature form). *Acta Crystallographica*, A37:398–406, 1981.
- [9] D. C. Jamrog. *A New Global Optimization Strategy for the Molecular Replacement Problem*. PhD thesis, Rice University, 6100 Main Street, Houston, Texas, April 2002. Technical Report TR-0208 at <http://www.caam.rice.edu/>.
- [10] E. E. Lattman. Use of rotation and translation functions. *Methods in Enzymology*, 115:55–77, 1985.

- [11] J. Navaza. Implementation of molecular replacement in amore. *Acta Crystallographica*, D57:1367–1372, 2001.
- [12] R. J. Read. Pushing the boundaries of molecular replacement with maximum likelihood. *Acta Crystallographica*, D57:1373–1382, 2001.
- [13] G. Stout and L. Jensen. *X-ray Structure Determination, A Practical Guide*. John Wiley & Sons, Inc., second edition, 1989.
- [14] L. Tong. How to take advantage of non-crystallographic symmetry in molecular replacement: 'locked' rotation and translation functions. *Acta Crystallographica*, D57:1383–1389, 2001.