# The Effect of the Separation of Variables on the Molecular Replacement Method

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

[a] *Department of Computational and Applied Mathematics, Rice University, Houston, Texas, U.S.A., and* [b] *Department of Biochemistry and Cell Biology, Rice University, Houston, Texas, U.S.A. E-mail: georgep@rice.edu*

## Abstract

Traditional approaches for solving the molecular replacement problem separate a six-dimensional optimization problem into two three-dimensional ones in order to reduce the computational cost. There are, however, serious drawbacks in such a separation of the rotational and translational degrees of freedom. In this paper, we present computational experiments indicating that even under ideal conditions the separation can fail to preserve the correspondence between the global minima of a target function and the correct rotations when low resolution data are used. This phenomenon is a reason why only high resolution data are used in traditional approaches for solving the molecular replacement problem. In this paper, we provide a theoretical explanation for this phenomenon.

In order to solve difficult molecular replacement problems, we believe that low resolution terms should be utilized because they generate smooth, shape-defining components in a target function, making it more amenable to global optimization. This study indicates that in order to utilize low resolution data in the molecular replacement method, we need to consider all degrees of freedom simultaneously. The full-dimensional optimization formulation, once a prohibitive procedure due to its high

computational cost, should now be feasible given the current state of computational resources and algorithms.

## 1. Introduction

The use of the molecular replacement method for solving new macromolecular structures is certain to increase as the set of known structures becomes even larger. The molecular replacement method seeks one or more optimal orientations and translations of a known search model to best fit the observed diffraction data obtained from an unknown structure in a crystallographic asymmetric unit. This problem can be stated as a $6n$-dimensional optimization problem where $n$ is the number of molecules in the asymmetric unit, with three rotational and three translational variables for each molecule. More often than not, $n$ is one and the number of degrees of freedom is six.

Traditional approaches implementing the molecular replacement method separate a six-dimensional optimization problem into two three-dimensional ones (Rossman & Blow, 1962): first searching for a number of rotations of the model structure as promising candidates; then searching for the best translational vector corresponding to each rotation candidate. Multiple candidates are needed since in practice the correct rotation may not be a global optimizer of the rotation function, hence other promising local optimizers must be included in the translational search.

Unfortunately, the traditional approaches are too often unable to find correct solutions. So far, the understanding of the reasons for the ineffectiveness of the traditional approaches remains limited (Brunger, 1997). Various techniques have been proposed in order to improve the effectiveness of these approaches. For example, recently researchers have proposed to search the rotation function space more carefully (Tong, 1996), to perform Patterson correlation refinement before the translational search (Brunger, 1990; Brunger, 1997), to use simulated annealing or maximum-likelihood refinement (Adams *et al.*, 1999) to improve the solutions, and to check for steric clashes (Tong, 1996) to weed out impossible solutions. Further improving the effectiveness of the molecular replacement method remains an important research goal.

It has been realized that if the search for rotation and translation is separated, there is a decrease in signal to noise for problems involving a large number of short intermolecular vectors, i.e. problems with a large number of molecules per unit cell and crystals with low solvent content (Glykos & Kokkinidis, 2000). Other researchers have noticed that the signal to noise ratio in the direct rotation function decreases as the upper limit of the resolution range is decreased (Brunger, 1997; Brunger & DeLano, 1995).

In this paper, we compare rotation functions derived with and without the separation of the rotational and translational variables. Our comparisons are done for a small molecule under carefully controlled conditions where we take the "unknown" structure to be exactly the same as the search model and we use simulated diffraction data without any model inaccuracy or noise. Different sets of the simulated diffraction data are used in our comparisons. We then provide a theoretical explanation of the results of our comparison.

## 2. Materials and Methods

We describe our experiments in this section.

### *2.1. Search model and diffraction data*

In this study the rotation functions are evaluated using generated diffraction data from the coordinates of all non-hydrogen atoms of a peptidic analogue of the antibiotic molecule, trichogin A IV, crystallized in space group $P2_1$, in a monoclinic cell with

$a = 14.56$, $b = 11.769$, $c = 9.473$Å and $\beta = 104.1°$ (Crisma *et al.*, 1994). We denote the generated or "observed" intensities as $I^o$ with similar notation for other "observed" quantities. In order to study the effect of the separation of variables exclusively, we take the model and unknown to be the same molecule, i.e. one solution corresponds to no rotation and no translation. To plot sections of the rotation functions involved in the traditional formulation, we use the direct rotation function (Brunger, 1997; Brunger & DeLano, 1995), where the structure factors are computed from rotating the search model (without symmetry mates) in a P1 unit cell with the same dimensions and angles as the crystallographic cell. Let the rotation operator be denoted as $\Omega(\theta_1, \theta_2, \theta_3) \in R^{3\times3}$ an orthogonal matrix where $\theta_1, \theta_2$ and $\theta_3$, in this work, are Eulerian angles and let $I^c$ denote the calculated intensities, then $I^c \equiv I^c(\Omega X)$ where $X = \{\mathbf{x}_j, j = 1 \ldots N\}$, $\mathbf{x}_j$ are the atomic coordinates of the $j$th atom and $N$ is the number of atoms in the molecule. To plot the rotation function involved in the full dimensional formulation, we place the search model and symmetry mate, in P$2_1$ unit cell isomorphic to the target crystal. In general for the full dimensional formulation $I^c \equiv I^c(\Omega_1 X + t_1, \ldots, \Omega_n X + t_n)$ where $\Omega_i$ and $t_i$ are the rotations and translations for the $i$th molecule in the asymmetric unit. In our case, for this crystal of trichogin, $n = 1$ and $I^c \equiv I^c(\Omega X + t)$.

### 2.2. *Objective functions*

We use two different objective functions to measure the fit of the observed and calculated diffraction data: a least squares function and a linear correlation coefficient (Brunger & DeLano, 1995; Hauptman, 1982). Let $E^o$ and $E^c$ denote the observed and calculated normalized structure factors. They can be thought of as vectors with components:

$$E_{hkl} \equiv \frac{F_{hkl}}{\sqrt{\langle |F|^2 \rangle}},$$

where the average intensity $\langle |F|^2 \rangle$ is computed, in this work, over one shell:

$$\langle |F|^2 \rangle = \frac{\sum_{hkl} |F_{hkl}|^2}{\sum_{hkl} 1}.$$

Then the Patterson correlation coefficient is defined by:

$$PC(E^o, E^c) = \frac{\langle |E^o|^2 |E^c|^2 - \langle |E^o|^2 \rangle \langle |E^c|^2 \rangle \rangle}{\sqrt{(\langle |E^o|^4 - \langle |E^o|^2 \rangle^2 \rangle \langle |E^c|^4 - \langle |E^c|^2 \rangle^2 \rangle)}} \tag{1}$$

where the average normalized intensity, in this work, is computed over one shell:

$$\langle |E|^2 \rangle = \frac{\sum_{hkl} |E_{hkl}|^2}{\sum_{hkl} 1}.$$

and multiplication of $E^o$ and $E^c$ is component-wise.

The least squares function is defined by:

$$LS(I^o, I^c) = \|sI^o - I^c\|_2,$$

where $I^o$ and $I^c$ can also be thought of as vectors with components $I^o_{hkl}$ and $I^c_{hkl}$ and $s$ is the least squares solution to

$$\min_{s \in R} \quad \| \, sI^o - I^c \, \|_2^2,$$

which is

$$s = \frac{I^{oT} I^c}{I^{cT} I^c}.$$

Thus $s$ is calculated for each $\Omega(\theta_1, \theta_2, \theta_3)$ sampled.

## 3. Results and Discussion

In this section, we present results from our computational experiments, offer a theoretical explanation to the results, and propose a new optimization strategy that utilizes low resolution diffraction data.

### *3.1. Results*

All graphics presented are two-dimensional sections of $LS$ and $1-PC$ in $\theta_1\theta_2$-space, where $\theta_1$ and $\theta_2$ are Eulerian angles. For all calculations of $I^c$, $\theta_3$ is fixed to zero. For the calculation of the six-dimensional rotation functions, i.e. rotation functions calculated using intensities which are dependent on both rotation and translation of the molecule in the asymmetric unit, the translation vector $t$ is fixed to zero.

In Figure 1 the functions were computed using diffraction data with a high resolution cutoff of 9Å and with no low resolution cutoff. In order to more clearly see the landscape of the lower function values, the color is scaled according to the square root of the actual values of the functions. Low values are indicated by red, with higher values indicated by blue. The rotation functions for this data set are relatively smooth with a small number of local minima residing in relatively wide, locally convex basins. The correct rotation solutions are indicated by the lowest valued minimizers at (0,0) and $(\pi, \pi)$.

Figure 2 shows the same six-dimensional rotation functions computed using diffraction data between 5 and 2Å, i.e. excluding some of the low resolution data. The landscapes in Figure 1 and Figure 2 are significantly different. Many local minimizers have been introduced in Figure 2 with a narrowing of the entrance into the correctly positioned global minimizers. Global optimization with respect to $\theta_1$ and $\theta_2$, even with a multi-start method is a formidable job.

Figure 3 shows the direct rotation functions computed using diffraction data with a high resolution cutoff of 9Å and with no low resolution cutoff. Again when using the low resolution data only, the rotation functions are relatively smooth. However, the positions of the global minimizers do not correspond to the correct rotations $(\theta_1, \theta_2)$ = (0,0) or $(\pi, \pi)$. The lowest function value occurs at (1.47, 1.57) and (4.61, 1.57) in $\theta_1\theta_2$-space for $LS$ and (.34, 2.90) and (3.49, .25) in $\theta_1\theta_2$-space for $1-PC$. Again, other researchers have also noted the incorrect position of optimizers for low resolution rotation functions (Brunger & DeLano, 1995; Brunger, 1997).

Using higher resolution data, the section of the direct rotation function in Figure 4 is similar to the section of the six-dimensional rotation function in Figure 2. The surface has many local minimizers and the openings into the global minimizers have narrowed relative to the low resolution result in Figure 3. However, the positions of the global minimizers now indicate the correct solution of the rotation subproblem. Thus, in order to use the rotation subproblem to indicate correct solutions one must use high resolution data. For this small molecule, in order to use the benefit that low resolution data provides, i.e. fewer local minima and larger areas of local convexity, the full six-dimensional rotation functions must be used.

### *3.2. Theoretical explanation*

The reason for the loss of signal-to-noise ratios in the classical rotation function arises primarily from the failure of the method to account for intermolecular atomic vectors (Rossman & Blow, 1962). We motivate this reasoning by considering the real space rotation function function and noticing that the numerator of the Patterson correlation coefficient (1) is the Fourier transform of the origin-subtracted real space rotation function using normalized structure factors (Brunger, 1997). The real space rotation function can be written as

$$R(\Omega) = \int_U P^o(\mathbf{u})\ P^c(\Omega\mathbf{u})\ dV \tag{2}$$

where $P^o(\mathbf{u})$ is the Patterson function of the unknown protein determined by $|F^o_{hkl}|^2$ and $P^c(\mathbf{u})$ is the Patterson of the model protein determined by $|F^c_{hkl}|^2$. Because each Patterson function $P(\mathbf{u})$ can be described as the set of inter-atomic vectors, it follows that many will be intramolecular vectors, but depending on the number of different orientations of neighbor molecules, the set of intermolecular vectors can be quite large.

Separation of the inter- from intra-atomic vectors is not easy and is the primary cause of the loss of signal to noise in traditional Patterson based searches when compared to full dimensional searches.

Early attempts to deal with this problem used the equivalent of limiting the integration volume $U$ in (2) to radii in Patterson space less than the maximum expected inter-atomic radius and greater than some value to avoid the origin peak, sometimes using origin removal approximations to do so. This practice of limiting the integration volume, sometimes excludes valuable intra-atomic vectors and fails to account for short inter-atomic vectors remaining in $U$. In using the direct rotation function, Brunger & DeLano (1995) remedy the first problem, but not the second. The power of a full dimensional search over the direct search method is illustrated by the following calculations.

Consider the following definition of the Patterson function

$$P(\mathbf{u}) = \mathcal{FT}^{-1}(|F_{\mathbf{h}}|^2) \equiv \frac{1}{V}\sum_{\mathbf{h}} |F_{\mathbf{h}}|^2 \exp(-2\pi i\,(\mathbf{h}\cdot\mathbf{u})), \qquad (3)$$

where $\mathbf{u}$ is a vector in Patterson space, $V$ is the volume of the unit cell, $\mathcal{FT}^{-1}$ is the inverse Fourier transform operator and "$\cdot$" is the inner product operator. To compute the Patterson function using only low resolution diffraction data, the limits of the summation could be changed to include only $\mathbf{h} = (hkl)$ such that the length of the reciprocal lattice vector, $d^*(\mathbf{h})$, is below some upper limit, or the limits of summation can remain the same and the intensities can be modified, so that after a certain distance from the origin they are essentially zero. Consider multiplying the intensities by a Gaussian to obtain the following modified intensities

$$|\tilde{F}_{\mathbf{h}}|^2 = |F_{\mathbf{h}}|^2 \exp(-\alpha\, d^*(\mathbf{h})^2)$$

where $\alpha > 0$. The greater the value of $\alpha$, the smaller the contribution will be from high resolution diffraction data. The Patterson function computed from $|\tilde{F}_{\mathbf{h}}|^2$ is

$$\tilde{P}(\mathbf{u}) = \mathcal{FT}^{-1}\left(|F_{\mathbf{h}}|^2 \exp(-\alpha\, d^*(\mathbf{h})^2)\right). \tag{4}$$

The convolution theorem states that $\mathcal{FT}^{-1}\left(\mathcal{FT}(f)\mathcal{FT}(g)\right) = f * g$, where "$*$" is the convolution operator. So (4) becomes

$$\tilde{P}(\mathbf{u}) = \mathcal{FT}^{-1}(|F_{\mathbf{h}}|^2) * \mathcal{FT}^{-1}\left(\exp(-\alpha\, d^*(\mathbf{h})^2)\right) = P(\mathbf{u}) * g(d(\mathbf{u})),$$

where (3) is used,

$$g(d(\mathbf{u})) = \mathcal{FT}^{-1}\left(\exp(-\alpha\, d^*(\mathbf{h})^2)\right) = C\exp\left(-\frac{d(\mathbf{u})^2}{4\alpha}\right),$$

and $C$ is the resulting constant. This calculation shows that to compute the Patterson function using low resolution diffraction data is essentially to convolute it with the Gaussian kernel $g(d(\mathbf{u}))$. As a result of the convolution, the Patterson function is smoothed out and the inter and intramolecular Patterson peaks are smeared together in a degree determined by $\alpha$. The greater the value of $\alpha$, the more pronounced the smoothing and smearing. In other words, the fewer low resolution data used, the more severe the loss of signal to noise.

Since the smearing of these two sets causes deterioration of the signal to noise ratio, the search is not helped by restricting the resolution of the analysis. The more common practice of using fewer high resolution terms is equivalent to multiplying Fourier space by a three-dimensional "top hat" function, which is equivalent to convolution of the peaks corresponding to the vector sets with a damped but oscillating function, which only further confuses the vector sets. On the other hand, a $6n$-dimensional search can operate at lower resolution effectively, due to the large areas of local convexity in the target function, making it amenable to fast, accurate, optimization algorithms.

Other researchers are also implementing searches in higher dimensional space in order to improve the accuracy of the method. Sheriff *et al.* (1999) implement a parallel six-dimensional search, stepping through rotational space and inputing it into the

AMoRe fast translation function, in order to search for clear solutions. Kissinger *et al.* (1999) propose a six-dimensional search using an evolutionary optimization algorithm which can be used to sequentially search for multiple molecules in the asymmetric unit. Glykos & Kokkinidis (2000) implement a simultaneous $6n$-dimensional optimization for the positioning of $n$ molecules in the unit cell, using reverse Monte-Carlo simulation. While overcoming the problems associated with the separation of variables, these methods are either computationally inefficient or suffer from stochastic approaches that do not guarantee complete coverage of the solution space.

### *3.3. A new global optimization strategy*

There are multiple factors that characterize the difficulty of a particular global optimization problem, the dimensionality being one and the functional geometry being another. With similar geometry, a function of more variables is of course more difficult than one of less (the curse of dimension). On the other hand, a function of three variables with ten thousand local minima, for example, can be more difficult to minimize than a function of ten thousand variables but with only three local minima.

In the case of the traditional molecular replacement methods, only using high resolution data produces target functions of three variables that are extremely difficult to optimize. By introducing low resolution data, we can obtain better target functions, but we will need to optimize with respect to all six variables simultaneously.

We now propose the following global optimization strategy for molecular replacement. We start with a coarse grid at low resolution which is just enough to cover the "basins" of all local minima. By local optimization, we locate locally optimal orientations (whose number should be relatively small) that fit the low resolution diffraction terms. We then add higher resolution terms and do another local grid search and optimization around the best solutions identified in the previous stage with an appropriate grid increment. The process is repeated until a suitable set of diffraction terms are included. Unlike random sampling methods, this procedure should be able to capture the correct solutions with a high likelihood, while much less computationally intensive than a six or higher dimensional, uniformly fine grid search.

## 4. Conclusion

With the current state of computational resources, this study lends support to the view that the molecular replacement method should, in most cases, be solved without the separation of the rotational and translational variables. Since a six-dimensional, uniformly fine grid search can still be computationally prohibitive, an adaptive grid search combined with local optimization techniques seems to be a promising strategy. In this strategy, the use of low resolution data is important. Not only do low resolution data lead to relatively smooth functions with large regions of local convexity, but more importantly these functions have global minima approximately (or exactly in the ideal case) corresponding to the correct orientations. On the other hand, traditional direct rotation functions derived from the separation of variables cannot preserve this correspondence for low resolution data as was shown by our experiments.

The implementation and other issues involved in the utilization of low resolution diffraction data are topics of our current research.

**Acknowledgments:** Diane Jamrog is supported by a training fellowship from the Keck Center for Computational Biology, (NSF Research Training Grant BIR-94-13229 and National Library of Medicine Training Grant 2T15LM07093) and acknowledges the AGEP program at Rice University. George N. Phillips, Jr. acknowledges the support of the National Institute of Health (AR 40252), the Robert A. Welch Foundation (C-1142), Intel Corporation and the Kegge Foundation. Yin Zhang is supported in part by DOE Grant DE-FG03-97ER25331, DOE/LANL Contract 03891-99-23 and

NSF Grant DMS-9973339.

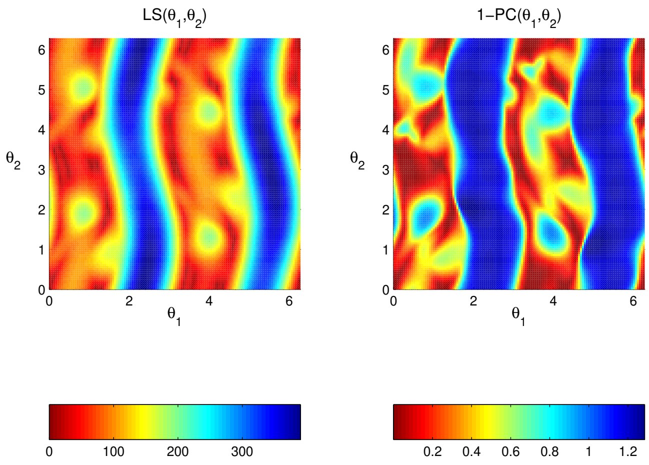


Fig. 1. Sections of the six-dimensional functions $LS$ and $1-PC$ in $\theta_1\theta_2$-space, computed using $I^c(\Omega X + t)$ where $\Omega$ and $t$ are the rotations and translations for the single molecule in the asymmetric unit. Using all low resolution data up to 9Å, these functions have the correctly positioned lowest valued minima and larger regions of local convexity. They have values less than $1 \times 10^{-11}$ at (0,0) and $(\pi, \pi)$.

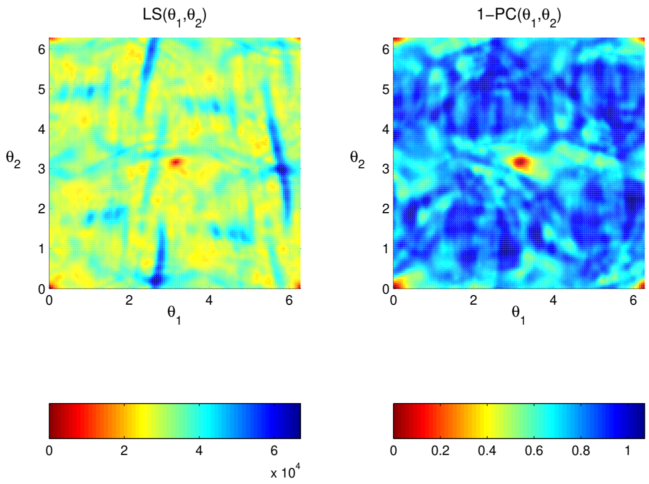


Fig. 2. Sections of the six-dimensional functions $LS$ and $1-PC$, in $\theta_1\theta_2$-space, computed using $I^c(\Omega X + t)$. Using diffraction data between 5 and 2Å, they have very deep minima but do not have large regions of local convexity. They do have minima at the correct positions (0,0) and $(\pi, \pi)$ with values less than $10^{-11}$.

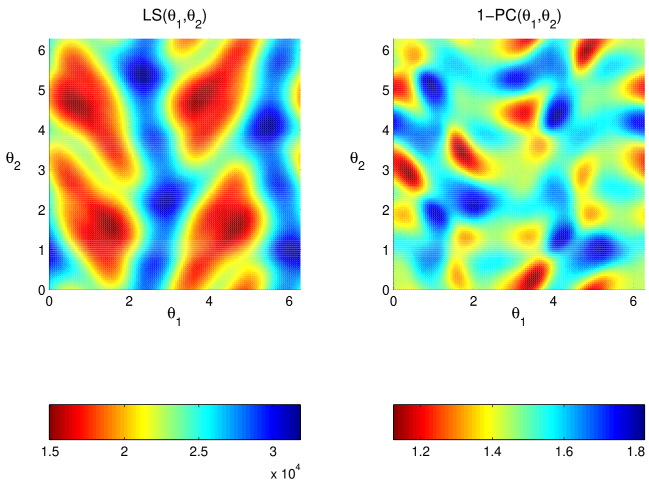


Fig. 3. Sections of the three-dimensional direct rotation functions, in $\theta_1\theta_2$-space, computed using $I^c(\Omega X)$. Using all low resolution data up to 9Å, they have large regions of convexity. However, the minima are in the wrong positions. The lowest functions values were $LS(1.47, 1.57) = LS(4.61, 1.57) \approx 1.495354\times10^4$ and $1 - PC(.34, 2.90) = 1 - PC(3.49, .25) \approx 1.12$.

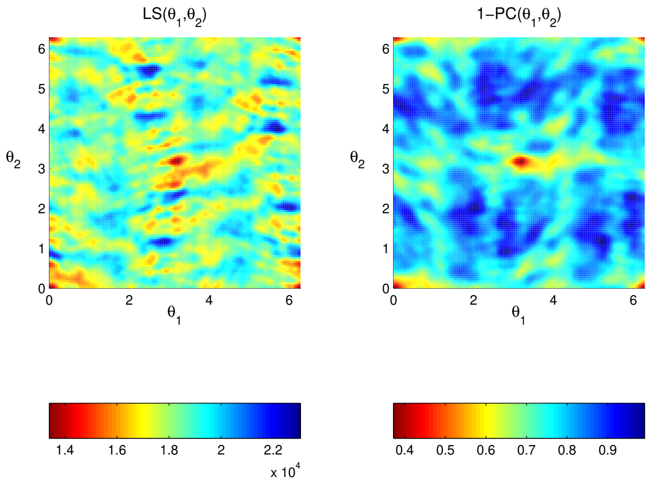


Fig. 4. Sections of the three-dimensional direct rotation functions, in $\theta_1\theta_2$-space computed using $I^c(\Omega X)$. Using diffraction data between 5 and 2Å, they have deep minima and a rough surface with many local minimizers. They do have their lowest function values in the correct positions. $LS(0, 0) = LS(\pi, \pi) \approx 1.340068\ \times10^4$, $1 - PC(0, 0) = 1 - PC(\pi, \pi) \approx .37$.

## References

Adams, P., Pannu, N., Read, R. & Brunger, A. (1999). *Acta Crystallographica*, D**55**, 181–190.

Brunger, A. (1990). *Acta Crystallographica,* A**46**, 46–57.

Brunger, A. (1997). *Methods in Enzymology,* **276**, 558–580.

Brunger, A. & DeLano, W. (1995). *Acta Crystallographica,* D**51**, 740–748.

Crisma, M., Valle, G., Monaco, V., Formaggio, F. & Toniolo, C. (1994). *Acta Crystallographica,* C**50**, 563–565.

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

Hauptman, H. (1982). *Acta Crystallographica,* A**38**, 289–294.

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

Rossman, M. & Blow, D. (1962). *Acta Crystallographica,* **15**, 45–52.

Sheriff, S., Klei, H. & Davis, M. (1999). *Journal of Applied Crystallography,* **32**, 98–101.

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

**Synopsis**

Molecular replacement methods can be improved by utilizing low resolution data and by a good management of local versus global optimizations. In order to do so, it is necessary to use a full-dimensional formulation without the separation of variables.