Acta Cryst. (1997). D**53**, 279–289

Molecular Replacement Using Genetic Algorithms

GEOFFREY CHANG AND MITCHELL LEWIS

The Johnson Research Foundation, Department of Biophysics and Biochemistry, School of Medicine, University of Pennsylvania, Philadelphia, PA 19104–6059, USA. E-mail: lewis@crystal.med.upenn.edu

(Received 9 September 1996; accepted 2 December 1996)

Abstract

A new molecular replacement (MR) strategy is introduced which features a continuous transform and a genetic algorithm (GA) for search optimization. This strategy uses a GA to simultaneously search the rotational and translational parameters of a test model while maximizing the correlation coefficient between the observed and calculated diffraction data. This has distinct advantages over conventional MR strategies which require a cross-rotation signal. An important feature of this method is its capability to simultaneously search the overall rotation/translation of the test model in the unit cell while refining the relative orientation/position of internal subdomains. This identifies molecular replacement solutions which would otherwise be completely missed using just a static model, and greatly improve the signal-to-noise contrast.

1. Introduction

The method of molecular replacement (MR) has become a routine tool for determining crystal structures of macromolecules using models that are closely related in structure. The theory of MR and its application to protein crystallography has been described in numerous reviews (see Rossmann, 1990). As an alternative to traditional isomorphous replacement, the molecular replacement method is an elegant computational approach for establishing a preliminary set of phases. The methodology rests on the understanding that structurally similar molecules have closely related molecular transforms and that differences in transforms primarily reflect changes in the orientation and positions of the molecules. Over the years, algorithms for determining the transformations that relate the unknown and known atomic models have evolved both in terms of speed and accuracy. However, as observed by Brünger (1994), approximations made to increase computational efficiency often come at a cost of diminished accuracy.

Conventional molecular replacement methods are, for the most part, based on the properties of the Patterson function. The information that defines the orientation and position of a known structure in an unknown cell is embedded in the intensities of the diffraction data and can be extracted by superposition of the Patterson functions. The Patterson function, or $|F|^2$ synthesis, when calculated using the crystal structure amplitudes, produces a three-dimensional distribution of vectors. There are two types of vectors: self and cross. Self vectors are produced between atoms within a molecule, and cross vectors occur between the atoms of symmetry-related molecules. A Patterson function, calculated from a known atomic model of a related molecule, will produce a constellation of self vectors that are similar to those of the unknown structure in the crystal, but rotated by some as yet undefined set of angles. To ascertain these angles, a rotation function is used to explore the correlation between the observed and model Patterson functions as a function of the rotation space (Rossmann & Blow, 1962; Huber, 1985; Nordman, 1971; Crowther, 1972; Navaza, 1994).

Determining the orientation of a search model with respect to the observed diffraction data can be accomplished in real space (Huber, 1985), reciprocal space (Rossmann & Blow, 1962), or direct space (Brünger, 1994). These three methods are all useful tools and are formally equivalent, but differ in their computational accuracy and efficiency. The real- and reciprocal-space formulations rely entirely on the Patterson function. Both methods compare the observed Patterson, P_c, with that calculated from a model, P_m. The superposition of the two Patterson functions is computed for different orientations of P_m,

$$\mathrm{Rot}(\Omega, r) = \int_{U(r)} P_c(v) P_m(\Omega, v)\, \mathrm{d}v, \qquad (1)$$

and the similarity of the two Patterson functions is assessed as a function of a rotation matrix, Ω, which is specified by a set of angles that samples the angular space. The variables r and $U(r)$ are the radius and the volume of integration, respectively. To obtain the best signal-to-noise ratio, a value of r is chosen that maximizes the number of self vectors while minimizing the number of cross vectors in the volume of integration.

A direct rotation function, introduced by Brünger (1990), differs from the conventional real- and reciprocal-space methods. Instead of rotating the Patterson of the search model P_m, the orientation of the model itself is changed according to $x_i' = \Omega x_i$. Structure amplitudes

are calculated and correlated with the observed diffraction data. This technique, referred to as 'Patterson correlation', is a measure of the phase accuracy for a partial model. Brünger (1990) showed this Patterson correlation to be a better discriminator for determining the rotational parameters that specify Ω than either the real or reciprocal rotation functions. However, as currently implemented, the Patterson correlation is orders of magnitude slower to compute than the other rotation functions.

In a similar fashion, the position of a correctly oriented model is achieved using a translation function (Crowther & Blow, 1967). A translational search is conceptually analogous to the rotation function. The position of the correctly oriented search model is identified by maximizing the agreement between the observed and calculated cross Patterson vectors. Crowther & Blow (1967) showed that the translation component can be determined by correlating an observed Patterson function and a Patterson function calculated from a correctly oriented model.

$$T(t) = \int_U P_c(v)P_m(v,t)\,\mathrm{d}v, \qquad (2)$$

where t and v are the translational shift and variable of integration. In most instances, the solution of a crystal structure by molecular replacement methods is not straightforward. Failure to unambiguously determine the relationship between the known and unknown structures results from differences between the search object and the unknown molecule. In such instances, the distribution of self vectors within the Patterson function of the search object is inconsistent with that from the crystal. Difficulties also arise when the unknown molecule is irregularly shaped or when molecules are densely packed in the crystal. In these cases, the self and cross vectors cannot be conveniently segregated and the interpretation of both rotation and translation functions is difficult. As a consequence, the strategy adopted by many molecular replacement packages, such as *AMoRe* (Navaza, 1994) and *X-PLOR* (Brünger, 1992), is to analyze several potential rotational and translational maxima in an attempt to identify the correct solution.

Although rotation and translation functions are useful tools for determining the orientation and position of the unknown molecule, the correct solution is usually confirmed by calculating a correlation coefficient or a residual between the observed and calculated diffraction data. In this paper, a method is introduced for solving homologous structures without partitioning the rotation and translation search. Searching a multi-dimensional space is no longer beyond the available computational resources. Fujinaga & Read (1987) demonstrated that it was possible, in principle, to perform a six-dimensional search to find the orientation and position of the unknown molecule. However, even a full rigid-body search may fail to provide a molecular replacement solution when there are differences between the search molecule and that in the unknown structure.

The molecular replacement method presented here describes a strategy for performing higher dimensional searches with current computational resources. This molecular replacement approach was developed to simultaneously search the rotational and translational degrees of freedom and, in doing so, is a departure from the classical protocol. Rather than dividing the problem into a rotation and a translation search, the orientation and position of the object are determined concurrently. Accomplishing this task requires an efficient method for computing structure amplitudes and an 'intelligent' algorithm for sampling a multi-dimensional space. The continuous transform is used to efficiently compute the structure amplitudes and a genetic algorithm is used to reduce the search space. While neither technique is new, combining the two (GA_MR) provides a very powerful tool which has distinct advantages over conventional molecular replacement strategies and also requires very little user intervention.

2. Methods

2.1. *The continuous transform*

To calculate efficiently structure amplitudes from an atomic model in a large number of different conformations requires a continuous Fourier transform (Lattmann & Love, 1970). The 'continuous transform' is functionally defined as a Fourier transform of a known molecule which has been sampled on a relatively fine grid in reciprocal space. Once the continuous transform is calculated in some reference frame, structure amplitudes can be quickly computed for any orientation of the object by a transformation of the reciprocal lattice indices. This is in contrast to the trigonometric or FFT (Fast Fourier Transform) techniques that must recompute the molecular transform for each orientation of the search molecule.

The structure-factor equation describes the relationship between atomic positions and the amplitude of a diffraction vector. Structure factors can be calculated for a molecule in any orientation given a set of atomic coordinates, x_i, and the atomic scattering f_i,

$$F^h_{\mathrm{calc}}(x,f,R,t) = \sum_{i=1}^{s} f_i \exp[-2\pi ih \cdot (\mathbf{O}^{-1}_{Xtal}\mathbf{R}\mathbf{O}_{Xtal}x_i + \mathbf{t})], \qquad (3)$$

where a rotation matrix, $\mathbf{R}$, is applied to the model coordinates (fractional), x_i and s is the number of atomic scatterers. The matrix, $\mathbf{O}_{Xtal}$, is required to orthogonalize the model coordinates and the vector, $\mathbf{t}$, is the translation of the model in the unit cell. The use of (3), however, requires that the summation must be repeated for every conformation of the molecule. The calculated

structure amplitudes can also be described in terms of its rotational and translational components.

$$F^h_{\rm calc}(x,f,R,t) = F^h_{\rm rot}(x,f,\mathbf{R})\exp(-2\pi ih\cdot t),$$

where

$$F^h_{\rm rot}(x,f,R) = \sum_{i=1}^{s} f_i \exp[-2\pi ih\cdot(\mathbf{O}^{-1}_{Xtal}\mathbf{R}\mathbf{O}_{Xtal}\,x_i)], \quad (4)$$

where the exponential term accounts for the translation of the molecule with respect to the cell axes.

Once an initial set of structure amplitudes have been calculated from a model in some arbitrary orientation, a new set of amplitudes can be computed for any rotation by applying the rotation to the reciprocal lattice indices rather than the atomic coordinates.

$$F^h_{\rm rot}(R,x,f) = F^{h'}_{\rm org}(x,f) \quad (5)$$

where $h' = R^T h$. The amplitude of $F^h_{\rm rot}$ is obtained by evaluating the original transform at the index h where R^T is the transpose of the rotation matrix. Applying a rotation to a set of integer indices, however, usually results in fractional values that do not correspond to grid points where the transform has been sampled. Therefore, to accurately evaluate the molecular scattering at these non-integral points the 'continuous' transform must be used. In practice the continuous transform, F_{mt}, is calculated by positioning the model at the origin and computing the structure amplitudes in an artificially large unit cell. From the continuous transform, the value of $F^h_{\rm rot}$ can be quickly evaluated by rotating the reciprocal lattice indices and using a linear interpolation in reciprocal space.

$$F^h_{\rm rot}(x,f,R) = F^{h'}_{\rm mt}(f,x),$$

where

$$h' = (\mathbf{O}_{\rm mt})^T R^T (\mathbf{O}^{-1}_{Xtal})^T (\mathbf{R}_{\rm Sym})^T h_{Xtal}, \quad (6)$$

where $\mathbf{O}_{\rm mt}$ and $\mathbf{O}_{Xtal}$ are the appropriate orthogonalization matrices for the continuous transform and the crystal system, respectively. The matrix $\mathbf{R}_{\rm Sym}$ is the rotational component of the crystallographic symmetry operator. The calculated structure amplitude in the crystallographic frame can be computed for a given rotation and translation by,

$$F^h_{\rm calc}(f,x,R,t,R_{\rm Sym},t_{\rm Sym}) = \sum_{i=1}^{N_{\rm sym}} F^{h'}_{\rm mt}(f,x,R,R^i_{\rm Sym})\exp[-2\pi ih\cdot(\mathbf{R}^i_{\rm Sym}t + t^i_{\rm Sym})] \quad (7)$$

where $t_{\rm Sym}$ is the translational component of the crystallographic operator and $N_{\rm Sym}$ is the number of crystallographic symmetry operators.

2.2. *The genetic algorithm*

An exhaustive sampling of rotational and translational space is computationally unrealistic given current computer resources even when using a continuous transform. A six-parameter search with modest sampling can require approximately 10^9 evaluations. Therefore, to perform a large multi-dimensional search requires an efficient search protocol. The genetic algorithms (GA's) are ideal optimization tools to search for the global minimum or maximum of functions spanning large non-linear spaces (Goldberg, 1989; Davis, 1991). Genetic algorithms belong to a family of stochastic optimization strategies which include Monte Carlo (Metropolis, Rosenbluth, Rosenbluth, Teller & Teller, 1993) and simulated annealing (Kirkpatrick, Gelatt & Verchi, 1983). GA's provide a useful alternative to these other methods because they are more robust and can explore complicated search landscapes efficiently. Genetic algorithms have also been useful tools for solving heavy-atom derivatives (Chang & Lewis, 1994) and also *ab initio* phasing of viral particles (Miller, Hogle & Filman, 1996). GA's use Darwin's principles of natural selection to find optimal solutions to complex numerical problems. In the terminology of the genetic algorithm, a chromosome evolves to maximize some defined fitness criteria.

The GA functions by mimicking nature, performing genetic operations such as cross over and point mutation which introduce variation in the population of chromosomes. The variables to be optimized by the GA are encoded as a bit string that are grouped into a chromosome. These 'parent' chromosomes are evaluated for their fitness, mutagenized, mated, and their genetic information is passed on to a new generation of 'children'. After several generations, a population of chromosomes will evolve which has a higher level of fitness than its ancestors. The fitness function in the genetic algorithm plays the same role as the environment plays in natural evolution. Chromosomes that perform well on a particular task will survive and have a higher probability of passing their genetic material to the next generation. Over several generations, chromosomes will emerge with high fitness which represents a solution to a particular problem.

The genetic algorithm used for molecular replacement also has a Lamarckian component. Parents can undergo extensive modification (*i.e.*, they are shaped by their environment) and pass these traits directly to the next generation. This is accomplished by randomly choosing a small number of parent chromosomes across the entire population and optimizing their fitness using gradient minimization. As implemented, less than 2% of the individuals in any generation are influenced by the environment and this optimization is performed only during the latter half of the GA evolution. Optimization is accomplished by varying each search parameter

(rotation and translation) by a steepest descent. The refined parameters are then encoded into the chromosome and reintroduced into the population. As a result of 'doping' the gene pool, convergence to the global minimum can often be obtained more quickly.

3. Methods and results

To illustrate how a genetic algorithm, GA_MR, can be used to solve an unknown crystal structure, four different molecular replacement problems are presented. These examples are representative of the types of problems frequently encountered in molecular replacement and are presented in order of increasing level of difficulty. The first example demonstrates how the genetic algorithm can be used to solve a single-body molecular replacement problem. The second example illustrates how a GA_MR can be used to search the translation of a model while performing rigid-body refinement on selected subdomains. The third example introduces additional degrees of freedom by incorporating non-crystallographic symmetry in the search. And the fourth example demonstrates how the genetic algorithm can be used to search all of rotation and translational space of two molecules in the asymmetric unit where the structure of the unknown molecule deviates from the search model. In these examples, the variables to be varied are encoded into a chromosome whose viability or fitness is assessed by correlating the observed and calculated structure intensities,

$$\mathrm{Corr}(F_{\mathrm{calc}}, F_{\mathrm{obs}}, N) = \frac{N\sum_{i=1}^{N} F^2_{\mathrm{calc}}(i)F^2_{\mathrm{obs}}(i) - \sum_{i=1}^{N} F^2_{\mathrm{obs}}(i)\sum_{i=1}^{N} F^2_{\mathrm{calc}}(i)}{\left\{N\sum_{i=1}^{N} F^4_{\mathrm{obs}}(i) - \left[\sum_{i=1}^{N} F^2_{\mathrm{obs}}(i)\right]^2\right\}^{1/2}\left\{N\sum_{i=1}^{N} F^4_{\mathrm{calc}}(i) - \left[\sum_{i=1}^{N} F^2_{\mathrm{calc}}(i)\right]^2\right\}^{1/2}}. \quad (8)$$

box with cell dimensions $a = b = c = 150$ Å which is approximately four times the largest dimension of the molecule. The electron density, with 1 Å sampling, was back transformed to produce a complete set of $P1$ structure factors. The six-dimensional search space that describes the rotation and three translation parameters was encoded into a 35-bit chromosome. A total of 20 bits were used to explore the entire rotational space that was sampled in 3° intervals; seven bits were used to define two Euler angles, θ_1 and θ_3, and six bits were allocated θ_2. The remaining 15 bits of the chromosome were used to sample the translational space in intervals of 1/32 of the asymmetric unit's cell edge. The population size used in the GA is related to the number of bits in the chromosome as described by Davis (1987) and was set to 500 in this example. The evaluation function serves multiple functions; it extracts the rotation and translation parameters encoded in the chromosome, calculates structure amplitudes by interpolation of the continuous transform (6) and assesses a fitness by correlating the observed and calculated structure factors for data with d spacings from 10 to 5 Å. To explore systematically the entire search space requires over 10^9 evaluations. By using the genetic algorithm the correct solution was found in only 10^4 trials. The best chromosome had a fitness score (*i.e.* a correlation coefficient) of 68%. The performance of the GA is plotted as a function of generation in Fig. 1.

3.1. *Example 1. Molecular replacement of a single rigid body*

A genetic algorithm was used to determine the six parameters necessary to orient and position a rigid model of a homologous molecule in the unit cell of an unknown crystal. Crystals of 3α-hydroxysteroid dehydrogenase were grown in space group $C222_1$ with cell dimensions of $a = 51.3$, $b = 89.5$ and $c = 143.3$ Å. The structure was solved by molecular replacement with the program *AMoRe* (Navaza, 1994) using aldose reductase (PDB file 1DLA) as a search probe (Hoog, Pawloshi, Alzari, Penning & Lewis, 1994). To illustrate this new methodology, this structure determination was repeated using the GA search protocol. As a first step, a continuous transform was computed from a polyalanine model of aldose reductase. The coordinates were transformed so that the center of mass was positioned at the origin of a $P1$

3.2. *Example 2. Molecular replacement of a molecule with flexible domains*

A frequently encountered problem for molecular replacement, illustrated in Fig. 2, occurs when a protein contains well conserved but loosely connected structural domains. In many instances, the individual domains are identical to those in the search molecule but the relative orientation of the domains change. When the relative differences are modest, the cross rotation function may provide a signal that describes an 'averaged' or a composite orientation. However, a translational search is more sensitive to small orientational differences and the correct solution is often lost in the noise. In some cases, a Patterson correlation (PC) protocol can correct these structural deviations (Brünger, 1990). When the conformational change is large, GA_MR provides an alternate approach. In this second example, a multi-

dimensional search was performed in which the relative orientation and position of the domains were treated as variables, as well as the translational vector that defines the position of the molecules with respect to the crystallographic axes.

Proteins within the *LacI* family have sequence homology and are structurally similar to the periplasmic binding proteins that have two flexible domains. The purine repressor (PurR) is a member of this family of proteins. This repressor crystallized in space group $C222_1$ with cell dimensions $a = 175.85$, $b = 94.79$ and $c = 81.84$ Å and has been solved to atomic resolution (Schumacher, Choi, Zalkin & Brennan, 1994). PurR is structurally quite similar to the periplasmic ribose-binding protein, however, as depicted in Fig. 2, the relative orientation of these domains are not completely conserved. Nonetheless, a cross rotation function calculated using a polyalanine model of the periplasmic binding protein (1RBP) produced a peak (4.9σ) that correctly oriented the model with respect to the crystallographic frame of the purine repressor. However, a Crowther–Blow translation function calculated to identify the correct position of the search object with respect to the crystallographic axes failed. In fact, the

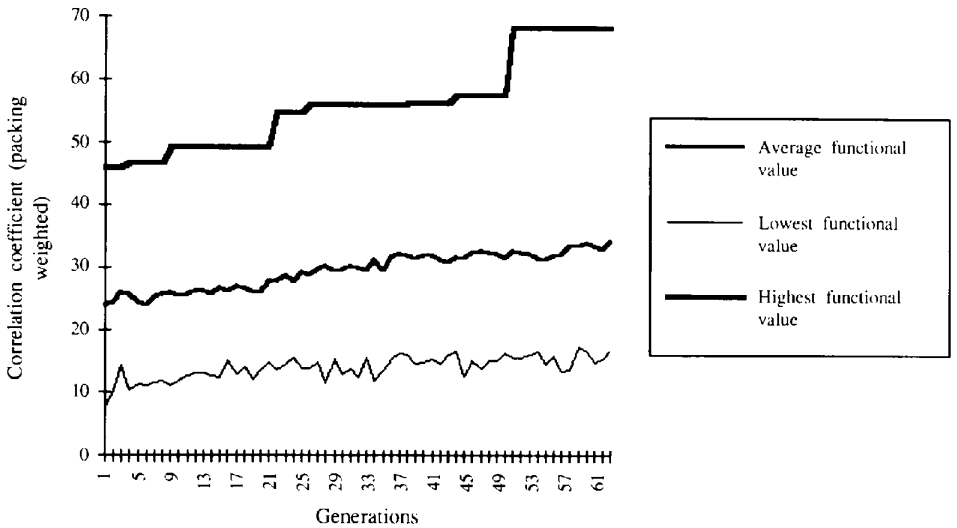

Fig. 1. Performance of genetic algorithm for solving 3α-hydroxysteroid dehydrogenase using an aldose reductase model. The details are described in the text. Shown are the average, lowest and highest functional values (correlation coefficient between F_{obs} and F_{calc}) as a function of population generation of binary chromosomes. The correct solution in this particular case is 68% found at generation 51.

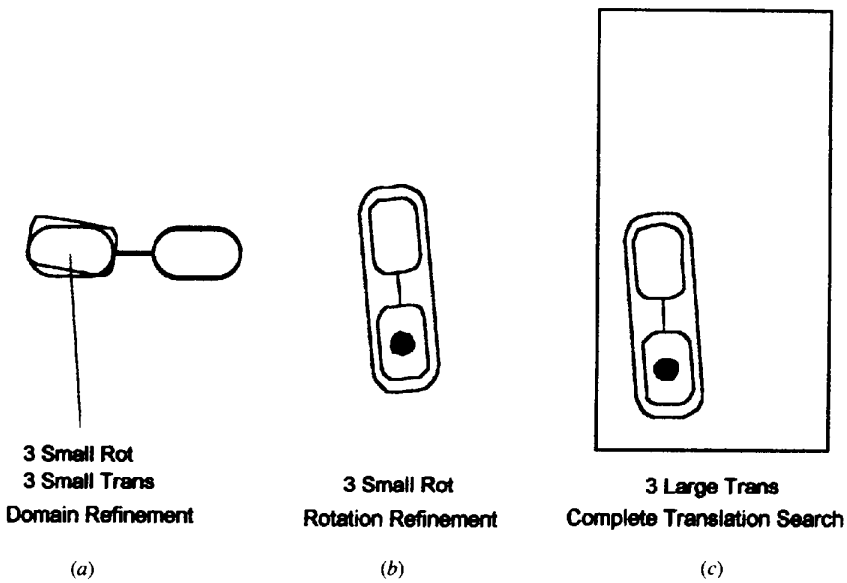

Fig. 2. Schematic illustration of MR case using ribose-binding protein to solve PurR. (*a*) shows that there is a small subdomain movement which requires three small rotational and three small translational parameters to properly model. (*b*) shows that the orientation of the complete molecule (highlighted in green) also needs to be refined because there is a crude cross-rotation solution in this case. The point of rotation is indicated by the blue dot. (*c*) shows that the position of the molecules needs to be searched which requires three large parameters.

correct solution was not within the top 50 peaks and less than 1σ. A search was repeated using the molecular replacement package in *X-PLOR* with the PC protocol and again the correct solution was not obvious.

Given that conventional MR methods failed to yield a solution, a genetic algorithm was constructed to determine the relative orientations of the two domains within purine repressor and to position this altered molecule with respect to the crystallographic unit cell. To describe the conformation of the two subdomains better, the overall rotation R^o, determined from the cross rotation function, was altered using two perturbation matrices $\mathbf{R}_1^{\Delta}$ and $\mathbf{R}_2^{\Delta}$. In addition to these rotational operators, GA_MR also determine the displacement vector $\mathbf{t}^{\Delta}$, which was needed to correctly relate the domains. The structure amplitudes were calculated by summing the two transforms and their crystallographically related mates such that,

$$F^h_{\text{calc}}(\mathbf{R}^{\mathbf{o}}, R_1^{\Delta}, R_2^{\Delta}, x, t, f) = \sum_{j=1}^{N_{\text{sym}=4}} \sum_{i=1}^{2} F^{h_i'}_{mt}(x, f) \times \exp(-2\pi i h \cdot t_{i,j}),$$

where

$$h_i' = (O_{mt})^T (R^o R_i^{\Delta})^T (O_{Xtal}^{-1})^T (R^j_{\text{Sym}})^T h_{Xtal},$$

and

$$t_{i,j} = (R^j_{\text{Sym}} O_{Xtal}^{-1} R^o R_i^{\Delta} O_{Xtal}) \cdot (t_i^o + t_i^{\Delta} + t) + t^j_{\text{Sym}},$$

and

$$\mathbf{t}_1^{\Delta} = 0. \tag{9}$$

The displacement vector, $\mathbf{t}_1^{\Delta}$, is equal to zero because it is incorporated into the monomer's translation, t. The pertubation matrices explored a limited angular space defined by the rotational parameters ($\Delta\alpha$, $\Delta\beta$, $\Delta\gamma$) and the angles were allowed to deviated $\pm 20^{\circ}$ about the observed conformation in increments of 2.5°. Three translational parameters (Δx, Δy, Δz) were used to adjust the relative positions, t_i^{Δ}, of the subdomains sampling a space that was ± 5 Å about the known displacements, t_i^o, in intervals of 0.625 Å. The sampling of these nine perturbation variables (six small rotations and three small translations for 36 bits), along with three overall translation parameters (for 17 bits) required a chromosome of 53 bits. A systematic search of this multi-dimensional pace would have required 2^{53} evaluations. The genetic algorithm, however, converged in approximately 7×10^4 trials producing a chromosome with a fitness or correlation of 51%. The optimization was terminated when the number of generations exceeded a user specified value. The performance of the genetic algorithm is shown as a function of generation in Fig. 3.

3.3. *Example 3. MR with non-crystallographic symmetry and flexible domains*

Many proteins crystallize with more than one molecule in the asymmetric unit. Often the transformation between these molecules can be determined from a self rotation function. This information can be readily incorporated in a GA search by modifying the equation that relates the crystallographic indices to the indices of the continuous transform,

$$h' = (\mathbf{O}_{mt})^T (NR)^T (\mathbf{O}_{Xtal}^{-1})^T h_{Xtal}, \tag{10}$$

where N is a non-crystallographic symmetry operator. This third example describes how a genetic algorithm

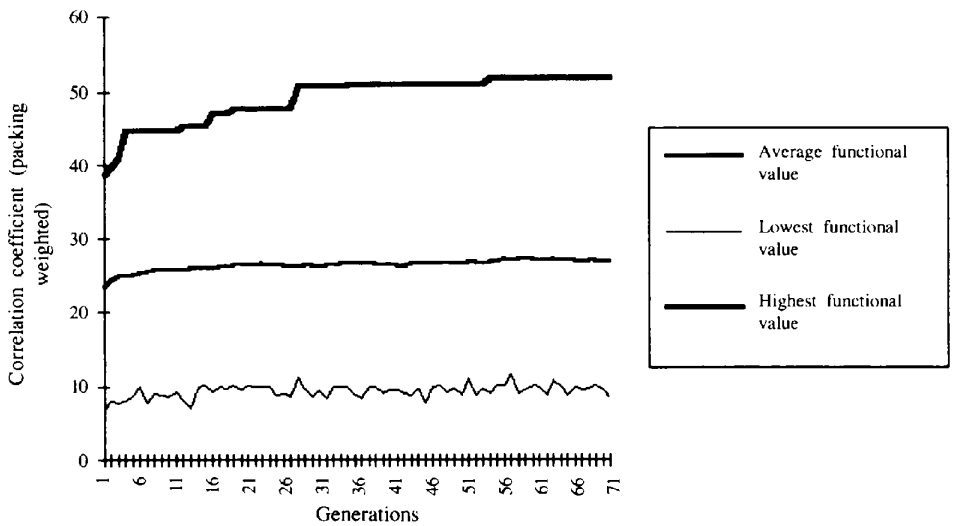

Fig. 3. Performance of genetic algorithm for solving the purine repressor complexed to DNA with ribose-binding protein model. The details are described in the text. Shown are the average, lowest and highest functional values (correlation coefficient between F_{obs} and F_{calc}) as a function of population generation of binary chromosomes. The correct solution in this particular case is 52% found at generation 55.

was developed to solve the structure of the tetrameric *lac* repressor, another member of the *LacI* family, using PurR as a search model.

The uncomplexed form of the lactose repressor (LacR) is a homotetramer that crystallized in space group $C2$ with unit-cell dimensions $a = 160.3$, $b = 73.3$, $c = 147.8$ Å, and $\beta = 120.3°$. A self-rotation function calculated with data between 10 and 4 Å produced two prominent peaks, approximately half the value of the origin. These peaks were observed on the $\kappa = 180°$ section with polar angles $\varphi = 60$, $\psi = 90°$ and $\varphi = 150$, $\psi = 90°$ suggesting there is a non-crystallographic twofold axis in the xz plane. In addition, satellite peaks at $\varphi = 130$, $\psi = 100°$ and $\varphi = 170$, $\psi = 80°$ were observed flanking the $\varphi = 150$, $\psi = 90°$ operator on this section. A cross rotation function, calculated using the dimeric purine repressor as a model, produced two large peaks; the primary peak was 7σ and second highest peak 5σ. The rotation function correctly oriented the PurR dimer such that the twofold axes was coincident with the satellite peaks observed in the self rotation search. It was, therefore, completely plausible that the large peak observed on the $\kappa = 180°$ section related the dimers of the tetramer.

Several attempts were made to translationally position the dimers in the Lac repressor cell without success. As in example 2, it was likely that the orientation of the domains within a monomer was different. A multi-dimensional GA was constructed to (i) explore small rotational and translational pertubations of the domains, (ii) to refine the non-crystallographic symmetry transformations and (iii) to position both dimers in the asymmetric unit (Fig 4). In this example, eight objects must be correctly placed in the asymmetric unit. Therefore, in order calculate a set of structure amplitudes, the scattering for eight transforms and their crystallographic mates need to be summed.

$$F^h_{\text{calc}}(R,x,t,f) = \sum_{l=1}^{N_{\text{sym}}=4} \sum_{k=1}^{2} \sum_{j=1}^{2} \sum_{i=1}^{2} F_{mt}^{h'_{i,j,k,l}}(x,f) \times \exp(-2\pi i h \cdot t_{i,j,k,l}),$$

where

$$h'_{i,j,k,l} = (O_{\text{mt}})^T (N_k M_j R^o R_i^\Delta)^T (O_{Xtal}^{-1})^T (R^I_{\text{Sym}})^T h_{Xtal},$$

and

$$t_{i,j,k,l} = (R^I_{\text{Sym}} O_{Xtal}^{-1} N_k M_j R^o R_i^\Delta O_{Xtal}) \cdot (t_i^o + t_i^\Delta + t_k) + t^I_{\text{Sym}}. \quad (11)$$

In order to obtain a more accurate representation of the LacR molecule, each monomer had to be divided into two domains and allowed to move in an independent fashion like in example 2. This is shown as the right most summation ($\sum_i$) in (11). A crude transformation that describes the orientation of the dimer model was obtained from cross rotation function, R^o, and the initial displacement of these domains is t_i^o. The matrices, $\mathbf{R}_1^\Delta$ and $\mathbf{R}_2^\Delta$, and the vectors, $\mathbf{t}_1^\Delta$ and $\mathbf{t}_2^\Delta$, described the relative changes between these two domains.

The orientation and position of the LacR monomers were interdependent and related by two levels of non-crystallographic symmetry which are expressed in (11) as summations ($\sum_j$ and $\sum_k$). To create the dimer, these structural domains must be transformed by the non-crystallographic symmetry operator, M_j, where $j = 1, 2$. An initial estimate for this symmetry operator was obtained from self and cross rotation function using the PurR dimer model. To account for deviations in the direction of the twofold axis from that observed in the rotation function, the angles the define the direction of the twofold were allowed to explore a limited angular space. Once the dimer was constructed, a tetramer was created by applying a second operation, N_k, which also

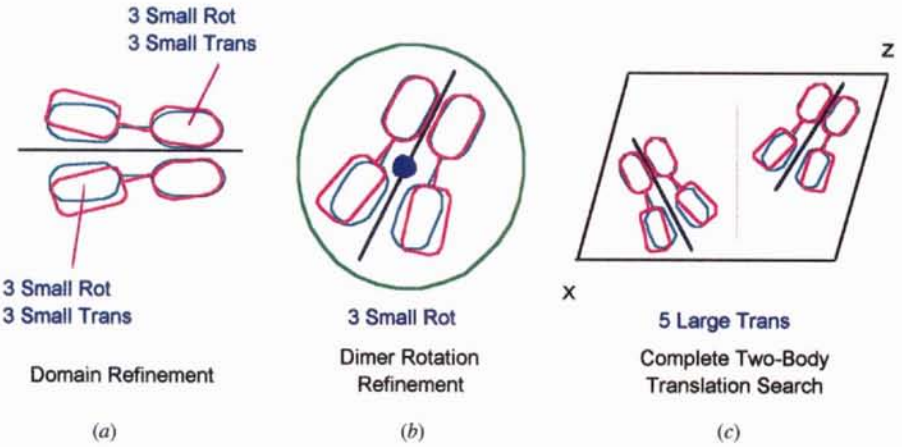

Fig. 4. Schematic illustration of MR case using purine–DNA (PurR) core to solve the lactose repressor (LacR). (*a*) shows that there is subdomain movement that is constrained by a non-crystallographic twofold axis. This requires 12 parameters to be properly modeled. (*b*) shows that the overall orientation of the dimer needs to be refined because there is a cross-rotation solution in this case. The blue dot indicates the point of rotation. (*c*) shows that the translation of both dimers needs to be searched in the asymmetric unit. The orientation of the second dimer is related by the non-crystallographic axis (shown in purple) which is also refined.

Table 1. *Parameters for the LacR molecular replacement cases*

The rotational and translational parameters and their respective sampling ranges are shown.

Subdomain	Rotation range (°)	Rotation sampling (°)	Translation range (Å)	Sampling (Å)	Number of parameters
N-terminal domain (Monomer)	[−20,20]	2	[−5,5]	0.2	6
C-terminal domain (Monomer)	[−20,20]	2	[−5,5]	0.2	6
Dimer 1	[−10,10]	2	[0.0,0.5] (Fractional)	0.25	5
Dimer 2	[−2.5,2.5] (Non-crystallographic)	0.25	[0.0,1.0]	0.25	5

had three degrees of freedom to describe deviations in the twofold axis that relate the dimers.

Thus, in total, the search space for this particular problem had 24 degrees of freedom. The conformation of each domain was allowed three rotations and three translations. In addition, three degrees of freedom were allocated to M_j to account for deviations in the direction of the twofold axis from that observed in the rotation function. The second NCS operator, N_k, also had three degrees of freedom to describe the direction of the twofold axis that relates the dimers. In addition, five degrees of freedom are needed to position the two dimers in the crystallographic $C2$ cell. These variables were encoded in a 71 bit chromosome. Since the conformation of the domains was allowed to change, as in example 2, two continuous transforms were calculated using a polyalanine model of PurR. The range of parameter values are listed in Table 1 and the performance of the genetic algorithm is shown in Fig. 5. A chromosome that encoded all of these parameters evolved in approximately 2.0×10^6 trials and had a fitness or correlation of 45%. By comparison, the next highest correlation for a chromosome at the end of the experiment was 30%. Inspection of the properly positioned tetramer in the unit cell revealed good packing. Moreover, when this polyalanine model was refined with *X-PLOR*, imposing strict crystallographic constraints, the residual dropped to below 30% and phases calculated from this model were sufficient for locating heavy-atom positions in isomorphous and anomalous difference Fourier transforms (Lewis *et al.*, 1996).

3.4. *Example 4. MR of a molecule with flexible domains and no cross rotation signal.*

In the first three examples the conformational search space could be reduced by using information about the

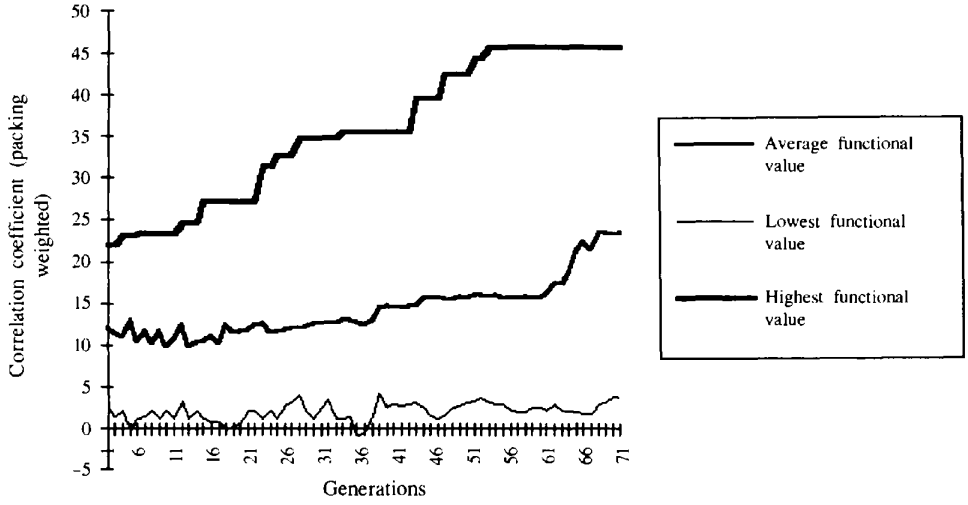

Fig. 5. Performance of genetic algorithm for solving the Lac repressor using purine repressor–DNA core model. The details are described in the text. Shown are the average, lowest and highest functional values (correlation coefficient between F_{obs} and F_{calc}) as a function of population generation of binary chromosomes. The correct solution in this particular case is 45% found at generation 52.

orientation of the molecule from a conventional rotation function. Without the benefit of initial cross rotation signal, traditional molecular replacement techniques usually will fail. These difficult molecular replacement problems can be solved, in some instances, using the genetic algorithm. As an example, a molecular replacement solution of PurR could be found using the ribose binding protein as the search model. While the number of variables needed to solve this problem is less than in example 3, the actual search space that was explored is larger.

The core of PurR-apo crystallized in space group $P2_1$ with unit-cell dimensions $a = 38.04$, $b = 125.26$, $c = 61.29$ Å, $\beta = 100.7°$ and solved to atomic resolution (Schumacher, Choi, Lu, Zalkin & Brennan, 1995). Unlike example 2, this crystal formed has a dimer in the asymmetric unit. A self rotation function calculated using data between 10 and 4 Å clearly showed a large peak on the $\kappa = 180°$ section, indicating a twofold rotation relates the monomer in the asymmetric unit. A cross rotation function (*AMoRe*, Navaza, 1994), calculated using RBP as the search model, however was unsuccessful and the correct solution was not observed in the top 100 peaks. RBP is not a good model of PurR-Apo and the structures deviate significantly. Indeed, it was later found that the relationship of the domains had significant deviations. Analogous to example 2, a perturbation matrix and vector were required to orient and position the two subdomains. Solving this particular problem required 14 variables. Although the number of variables needed to solve this problem was less than half the number required in example three, the search space was actually larger. The entire rotational space had to be explored since there was no prior knowledge of the orientation of the molecule from a rotation function.

In this example, a GA was constructed to (i) sample the parameters that describe the relationship of the subdomains, (ii) perform a complete rotational search for each domain, (iii) orientation of the second molecule in the asymmetric unit was generated from the self rotation function, and (iv) to position of both monomers in the symmetric unit (Fig. 6). The total search space was encoded in a 74 bit chromosome which, if systematically sampled, would require 2^{74} (1.88×10^{22}) evaluations. The results of this search as a function of generation is shown in Fig. 7. In approximately 2.5×10^6 trials a chromosome evolved with a correlation coefficient of approximately 50% that was consistent with the correct solution (Schumacher *et al.*, 1995).

4. Conclusions

Traditional molecular replacement methods have been and will continue to be useful for solving crystal structures. However, when these methods fail, a more powerful approach is needed. For homologous structures, the cross rotation functions are robust and provide the relationship between the known and unknown molecules. However, when there are structural differences between the known model and the unknown crystal structure, a large number of cross rotation peaks will need to be examined to identify the correct solution. In our experience, the translation function is even less forgiving and creates the bottleneck in these difficult structure determinations.

The molecular replacement method described here sidesteps many of these problems by simultaneously searching for the rotational and translational parameters. The only disadvantage of this method is that it is computationally more demanding. Two features were

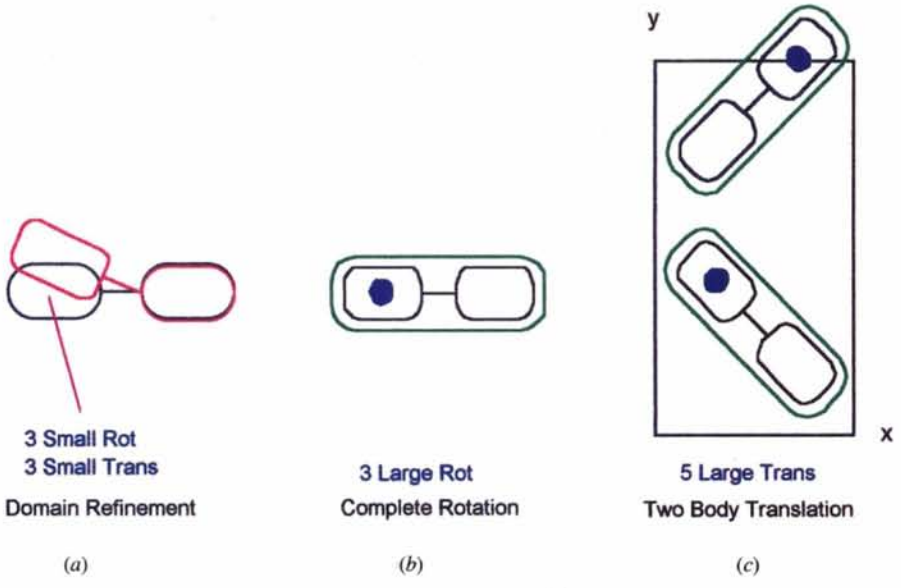

Fig. 6. Schematic illustration of MR case using the ribose-binding protein (RBP) to solve the apo form of the purine repressor (PurR-apo). (*a*) shows that there is large subdomain movement which requires six parameters to properly model. (*b*) shows that the orientation of the molecule needs to be searched. There is no cross-rotation signal in this particular case. (*c*) shows that a two-body translation search is required to properly position the molecules in the asymmetric unit. The orientation of the second molecule is related by the non-crystallographic operator from self-rotation analysis.

employed to make this structure solving method more practical. First, a genetic algorithm was used to efficiently explore this complicated search landscape. Although a single GA_MR run may not guarantee the maximum correlation coefficient, multiple runs can be executed in parallel to increase the probability of achieving the 'best' molecular replacement solution. A continuous transform was also used to greatly increase the efficiency of the model structure-factor calculations. In example 1, for instance, the correct molecular replacement solution was found with an average time per GA run of 15 min on an SGI R4000 Indigo. Although the interpolation of the rotational component of the scattering is fast, the iterative computation of potentially thousands of possible conformations for more difficult molecular replacement cases (examples 2–4) can be more expensive. Upon completion of the examples presented here, it became obvious that the search procedure would be more efficient if a subset of the data were used in the calculations. As a test, the first example was repeated using different fractions of randomly selected data between 15 and 4 Å. Using only 25% of the observed reflections decreased the computational time by almost a factor of four while producing the same results.

The strength of genetic algorithms in molecular replacement is its ability to simultaneously refine the orientation and position of molecular fragments or domains while searching the entire translation space. A variety of different fitness functions can be used to accomplish this task. For example, an evaluation function was constructed to maximize the peak height in a difference Fourier (from isomorphous differences of anomalous data) using phases from potential molecular replacement solutions generated by the GA. This can be very useful if the positions of the heavy atoms are known.

Difficult molecular replacement cases require robust and flexible search tools. The methods presented here are not intended to replace other molecular replacement methodologies but rather to augment these techniques and provide a simple way to incorporate known information into the search procedure.

It has been shown that globular proteins are organized in an hierarchical fashion with well defined molecular volumes (Nichols, Rose & Ten Eyck, 1995). Proteins are created by structural domains which can be decomposed into subdomains, and so forth. It is, therefore, reasonable to represent any search object in molecular replacement as a collection of small well packed units. A continuous transform of each of these smaller units can be calculated and allowed a limited degree of flexibility. Clearly as the number of transforms used to represent the object increases so do the computational costs. The genetic algorithm is, however, a beautifully parallel process that is ideally suited for distributed processing which can be achieved using *Parallel Virtual Machine* (*PVM*) software (Sunderan, Geist, Dongano & Manchek, 1994).

We wish to thank Richard Brennan and Maria Schumacher for providing us with diffraction data collected on the purine repressor. This work was funded by a National Institutes of Health grant GM-44617 and an Army Research Office grant DAAL03-92-G-0713, and a National Institutes of Health Molecular Biophysics Training Grant 2-T32-GM082745.

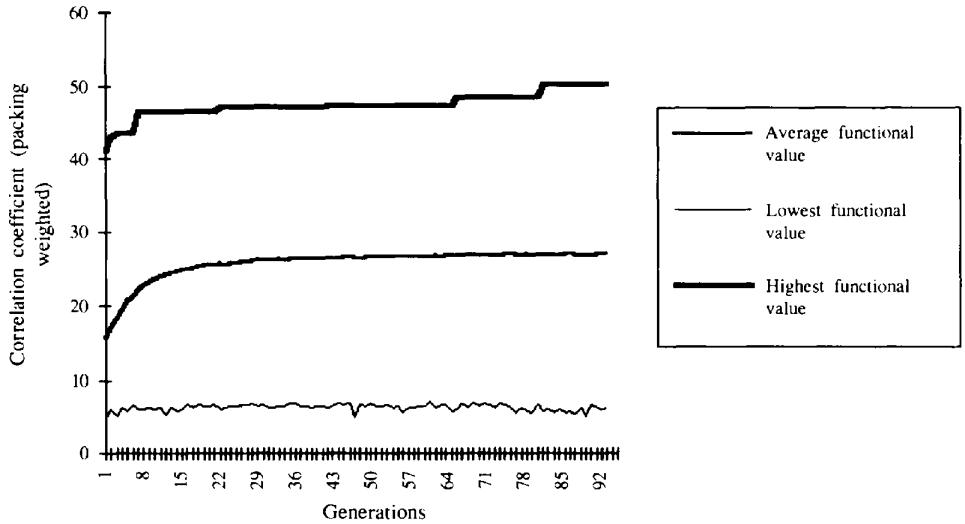

Fig. 7. Performance of genetic algorithm for solving the purine repressor (apo form) with ribose-binding protein model. The details are described in the text. Shown are the average, lowest and highest functional values (correlation coefficient between F_{obs} and F_{calc}) as a function of population generation of binary chromosomes. The correct solution in this particular case is 51% found at generation 82.

References

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

Brünger, A. T. (1992). *X-PLOR.* Version 3.1. *A System for X-ray Crystallography and NMR.* New Haven, CT: Yale University Press.

Brünger, A. T. (1994). *Acta Cryst.* D**51**, 740–748.

Chang, G. & Lewis, M. (1994). *Acta Cryst.* D**50**, 667–674.

Crowther, R. A. (1972). *The Molecular Replacement Method*, edited by M. G. Rossmann, *Int. Sci. Rev. Ser.* No. 13, p 10. New York: Gordon & Breach.

Crowther, R. A. & Blow, D. M. (1967). *Acta Cryst.* **23**, 544–548.

Davis, L. (1987). *Genetic Algorithms and Simulated Annealing*, edited by L. Davis, pp. 42–60. London: Pitman.

Davis, L. (1991). *Handbook of Genetic Algorithms.* Van Nostrand Reinhold, New York.

Fujinaga, M. & Read, R. (1987). *J. Appl. Cryst.* **20**, 517–521.

Goldberg, D. (1989). *Genetic Algorithms in Search, Optimization, and Machine Learning.* London: Addison-Wesley.

Hoog, S. S., Pawloski, J. E., Alzari, P. M., Penning. T. M. & Lewis, M. (1994). *Proc. Natl Acad. Sci. USA*, **91**(7), 2517–2521.

Huber, C. P. (1985). *Proceedings of Daresbury Study Weekend.* Warrington: Daresbury Laboratory.

Kirkpatrick, S., Gelatt, C. D. Jr & Vecchi, M. P. (1983). *Science*, **220**, 671–680.

Lattmann, E. E. & Love, W. E. (1970). *Acta Cryst.* B**26**, 1854–1857.

Lewis, M., Chang, G., Horton, N., Kercher, M., Pace, H, Schumacher, M., Brennan, R. G. & Lu, P. (1996). *Science*, **271**, 1247–1254.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. & Teller, E. (1953). *J. Chem. Phys.* **21**, 1087–1092.

Miller, S. T., Hogle, J. M. & Filman, D. J. (1996). D**52**, 235–251.

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

Navaza, J. & Vernsoslava, E. (1994). *Acta Cryst.* A**51**, 445–449.

Nordman, C. E. (1971). *Acta Cryst.* A**28**, 134–143.

Nichols, W. L., Rose, G. D. & Ten Eyck, L. F. (1995). *Protein Struct. Funct. Genet.* **23**, 38–48.

Rossmann, M. G. (1990). A**46**, 73–82.

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

Schumacher, M. A., Choi, K. Y., Lu, F., Zalkin, H. & Brennan, R. G. (1995). *Cell*, **83**, 147–155.

Schumacher, M. A., Choi, K.Y., Zalkin, H. & Brennan, R. (1994). *Science*, **266**, 763–770.

Sunderan, V. S., Geist, G. A., Dongano, J. & Manchek, R. (1994). *Parallel Comput.* **20**(4), 531–545.