Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Rapid automated molecular replacement by evolutionary search

Charles R. Kissinger,[a]* Daniel K. Gehlhaar[a] and David B. Fogel[b]

[a]Agouron Pharmaceuticals, Inc., 3565 General Atomics Court, San Diego, CA 92121, USA, and [b]Natural Selection, Inc., 3333 North Torrey Pines Court, Suite 200, La Jolla, CA 92037, USA

Correspondence e-mail: crk@agouron.com

Received 21 July 1998
Accepted 28 September 1998

A new procedure for molecular replacement is presented in which an efficient six-dimensional search is carried out using an evolutionary optimization algorithm. In this procedure, a population of initially random molecular-replacement solutions is iteratively optimized with respect to the correlation coefficient between observed and calculated structure factors. The sensitivity and reliability of the method is enhanced by uniform sampling of the rotational-search space and the use of continuously variable rotational and translational parameters. The process is several orders of magnitude faster than a systematic six-dimensional search, and comparisons show that it can identify solutions using significantly less accurate or less complete search models than is possible with two existing molecular-replacement methods. A program incorporating the method, *EPMR*, allows the rapid and highly automated solution of molecular-replacement problems involving single or multiple molecules in the asymmetric unit. *EPMR* has been used to solve a number of difficult molecular-replacement problems.

1. Introduction

Crystallographic structure determination by molecular replacement (MR) makes use of an approximate structural model to provide initial phase estimates. If the model is sufficiently accurate and can be placed correctly in the unit cell, the final structure can be obtained through rebuilding and refinement. MR is now widely used as a method for macromolecular structure determination, and opportunities for its use will continue to increase as the number of known protein structures grows and homology-modeling methods improve.

MR requires the identification of the correct orientation and position of the structural model in the asymmetric unit. The success of this procedure depends critically on the quality of the model used. Although MR has been accomplished using models comprising of a very small fraction of the total scattering matter in the asymmetric unit (Oh, 1995; Bernstein & Hol, 1997), experience has shown that the coordinate differences from the target molecule must be small for the procedure to be successful. Even when an accurate structural model is available, a solution is not always obtained. As a result, there is considerable interest in methods that can increase the reliability and extend the range of applicability of MR.

The identification of the three rotational and three translational parameters that define the orientation and position of the molecule in the asymmetric unit has been traditionally accomplished using the Patterson superposition methods pioneered by Rossmann & Blow (1962). These methods are based on finding the maximum correlation between observed Patterson vectors and those calculated for the search model in an artificial cell. To reduce the computational cost over that of a comprehensive six-dimensional search, the procedure is divided into an initial search for the correct rotation of the search model, followed by a search for the correct translation. The rotation search can be formulated in real space (Huber, 1965) or reciprocal space (Rossmann & Blow, 1962), but is conceptually performed by examining only Patterson vectors up to a certain length so that primarily intramolecular vectors (self-vectors) are considered. The translation search (Crowther & Blow, 1967) then involves finding the maximum correlation between the remaining intermolecular Patterson vectors (cross-vectors).

Division of the MR problem into separate rotation and translation searches reduces the computational cost dramatically over a comprehensive six-dimensional search. However, this approach has disadvantages. At the optimum in the rotation search, only a subset of all Patterson vectors is superimposed, and thus the signal-to-noise ratio in this step is inherently low. In some cases, the highest correlation found in the rotation search does not correspond to the correct solution, and alternative rotation solutions must also be examined. Furthermore, the translation search can be highly sensitive to any error in the orientation obtained from the rotation search.

Numerous techniques have been introduced to circumvent these problems. For example, the program *AMoRe* (Navaza, 1994) allows the automated examination of multiple rotation-function peaks in the translation search. The program *BRUTE* (Fujinaga & Read, 1987) allows refinement of the orientation during the translation search. Brünger (1990) has introduced the technique of 'Patterson correlation' refinement to optimize the search-model orientation prior to the translation search. An alternative form of rotation search, termed the 'direct' rotation function, has been described by DeLano & Brünger (1995), in which the search model itself is rotated rather than the corresponding Patterson map, and the correlation coefficient between observed and calculated structure factors is determined at each trial orientation. This method has the advantage that all data, and thus all self-vectors, are considered.

Many of the limitations inherent in separate rotation and translation searches are avoided by a comprehensive six-dimensional search, which should ultimately be a more sensitive and reliable method of obtaining MR solutions. Limited forms of six-dimensional search have been used to solve several difficult MR problems (Rabinovich & Shakked, 1984; Tong, 1996). However, despite continuing increases in computer speeds, systematic six-dimensional searches have not generally been feasible because of the very large number of trial positions that must be evaluated. Contemporary non-linear optimization techniques provide a potential solution to this problem. Chang & Lewis (1997) have recently demonstrated the use of genetic algorithms for multidimensional MR searches. In an independently developed procedure, we have employed a similar stochastic search technique, evolutionary programming (Fogel *et al.*, 1966), to simultaneously optimize the orientation and position of a search model with respect to the correlation coefficient between the observed and calculated structure factors. We show that this method is capable of identifying MR solutions using less accurate or less complete search models than is possible with two existing MR methods. The method has been incorporated into a computer program, *EPMR* (evolutionary programming for molecular replacement), that allows rapid automated identification of MR solutions.

2. Molecular replacement by evolutionary search

Evolutionary programming has been shown to be an efficient algorithm for determining the global optimum of a variety of complex non-linear search spaces (Fogel, 1995; Gehlhaar *et al.*, 1995; Bowie & Eisenberg, 1994). By analogy with evolutionary processes in biology, this stochastic search algorithm acts through the iterative optimization of a population of trial solutions. The population 'evolves' through competition among its members for survival, followed by production of 'offspring' by the surviving members of the population. The relative 'fitness' of each population member is calculated using a mathematical objective function. Competition can be performed through simple rank ordering of the members of the population by their objective function score and then discarding some fraction of the lower scoring individuals. More commonly, a stochastic tournament is employed, whereby the fitness of each member of the population is compared to that of a small number of other randomly chosen individuals, and the population is ranked according to the number of competitions that each member wins. This non-deterministic ranking serves to maintain a greater degree of diversity in the surviving solutions. Surviving members of the population produce offspring so as to restore the population to its original size. Offspring are produced by introducing small random variations in the values of the parameters comprising a parent solution. Through this process, evolutionary algorithms can provide broad comprehensive stochastic sampling that gradually focuses on the most promising regions of the search space.

Evolutionary programming algorithms share several characteristics with genetic algorithms (Fogel, 1995), which have also recently been applied to six-dimensional MR searches (Chang & Lewis, 1997). However, the two search methods differ in several critical respects. Most importantly, in evolutionary programming the parameters are represented as a real-valued vector, instead of the bit-string used in a standard genetic algorithm. The parameters are thus allowed to vary continuously, eliminating the need to choose a sampling interval for problems involving real-valued parameters. In addition, evolutionary programming is based on simultaneous modification of all parameters during generation of offspring. Traditional genetic algorithms rely primarily on crossover, in which parts of two parents are combined, which tends to leave the values of some parameters unchanged. Thus, evolutionary programming is likely to be more effective in searching spaces where the variable parameters are correlated (Salomon, 1996).

The evolutionary programming search procedure for MR is summarized in Fig. 1. A starting population of trial MR solutions is generated by assigning random values to the six rigid-body parameters describing the orientation and position of the search model in the unit cell. A stochastic tournament is used to determine which solutions survive into the next generation. Surviving members of the population are retained in the next generation without modification and are used to produce offspring to regenerate the population. Offspring are generated by applying normally distributed random mutations to the orientation and translation of the parent solution. The process is repeated for a fixed number of generations, after which the solution with the highest correlation coefficient is chosen for a conjugate-gradient optimization procedure (Powell, 1977).

The standard deviations for the mutation sizes are set at the start of the process, but are allowed to vary in a self-adaptive manner, with selection pressure determining the optimal mutation sizes as the search progresses. In this procedure (Schwefel, 1981; Fogel, 1995), the standard deviation of the Gaussian mutation σ_i' for each variable i is generated from the value σ_i used to generate the parent solution by the following equation,

$$\sigma_i' = \sigma_i \exp[\alpha N(0,1) + \beta N_i(0,1)], \tag{1}$$

where $\alpha = 1/(2^{1/2}n)$, $\beta = 1/(2n)^{1/2}$, $N(0, 1)$ is a Gaussian random number with zero mean and unit variance, $N_i(0,\ 1)$ is a different random number for each variable and n is the number of variables. Thus, α affects the size of all variables for a particular offspring, while β affects the individual variables.

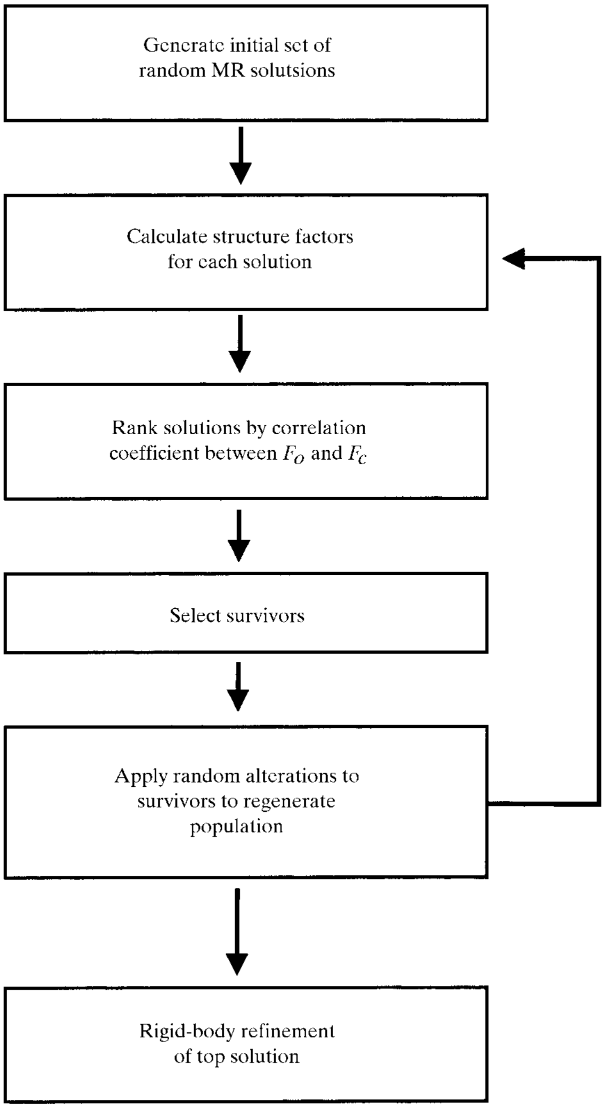
Generate initial set of random MR solutsions
Calculate structure factors for each solution
Rank solutions by correlation coefficient between F_O and F_C
Select survivors
Apply random alterations to survivors to regenerate population
Rigid-body refinement of top solution

Figure 1
Flowchart of molecular replacement by evolutionary search.

Proper sampling of the search space is essential for an evolutionary algorithm to be generally effective. Although it is convenient to express rotations for MR searches in terms of Eulerian angles (Rossmann & Blow, 1962), the use of Eulerian angles (or spherical polar angles) results in non-uniform sampling of the angular space. Values near the poles [$\theta_2 = 0°$ or 180°, according to the convention of Rossmann & Blow (1962)] are much more finely sampled than at the equator ($\theta_2 = 90°$). Uniform sampling in Euler space thus leads to a highly biased search. To circumvent this problem, we employed a quaternion representation (Minkler & Minkler, 1990) for the search-model rotation. The quaternions for the initial population were calculated so as to evenly sample rotation space, and were mutated in subsequent generations through application of Gaussian random rotations about a random axis. As a result, the searches are unbiased and equally efficient regardless of the location of the solution in rotation space.

The choice of the objective function used to evaluate the fitness of each member of the population is also critical to the success of an evolutionary programming algorithm. We examined several functions, including the conventional crystallographic R factor and several forms of correlation coefficient between observed and calculated structure factors in either standard or normalized form. The objective function that proved most effective was the linear correlation coefficient between observed and calculated (non-normalized) structure factors,

$$C = \frac{\sum(|F_o| - \overline{|F_o|})(|F_c| - \overline{|F_c|})}{[\sum(|F_o| - \overline{|F_o|})^2]^{1/2}[\sum(|F_c| - \overline{|F_c|})^2]^{1/2}}. \tag{2}$$

Determining this correlation coefficient requires the calculation of a complete set of structure factors for each new member of the population in each generation. In a typical search procedure, structure factors are calculated approximately 10000 times. The speed of the procedure thus depends critically on the efficiency of the structure-factor calculations. The computation times would be prohibitively long if structure factors were calculated by fast Fourier transform (FFT). However, a more rapid method of structure-factor calculation can be applied to a molecule undergoing rigid-body transformation. If the Fourier transform of the molecular electron density is sampled on a sufficiently fine grid, structure factors for that molecule in any orientation and position in the unit cell can be calculated by appropriate transformation of reflection indices, interpolation and application of appropriate phase shifts (Lattman & Love, 1970; Huber & Schneider, 1985). Using this method, structure factors are calculated once by FFT for the search model at the origin of an artificial *P*1 cell. The dimensions of the cell are chosen to allow sufficiently fine sampling of the transform so that interpolation errors are minimized in subsequent calculations. A cell of approximately four times the extent of the search model in each direction yields an average error of less than 1% in the structure-factor magnitudes. Subsequent structure-factor calculations are performed as follows.

(i) The indices of each observed reflection are transformed into the lattice of the molecular transform and rotated according to the current rotation of the search model.

(ii) The structure factor at the non-integral indices is calculated using linear interpolation into the table of *P*1 structure factors.

(iii) Phase shifts corresponding to the current translation of the search model in the unit cell are applied.

(iv) The contributions from all symmetry-related molecules are summed. This procedure provides a dramatic speed increase over FFT calculation and allows the evolutionary search procedure to be quite rapid.

To optimize the performance of the evolutionary programming algorithm, we systematically varied the population size, the number of generations, the number of competitors in the stochastic tournament, the starting mutation sizes and other parameters, while evaluating the procedure on a variety of test cases. Because the search procedure is non-deterministic, the correct solution will not be found on every run, even if the population size is made very large. We found that a population size of 300 evolving over 50 generations provided a good compromise between computation time and search efficiency. Starting mutation sizes were set to 6° for rotations and 3 Å for translations. Six competitors were used in the stochastic tournament, with the top 50% of solutions kept as survivors at the start of the search, decreasing linearly to 20% at the end of the search. No crossover between surviving solutions was used.

3. Examples of use

The procedures described above have been incorporated into the program *EPMR*, which was designed to provide MR solutions in a highly automated manner. The only input files required by the program are those containing the search-model coordinates, the observed structure-factor amplitudes and the cell parameters. A set of command-line options allows the user to control the program's operation. In its default mode, *EPMR* will search for a single molecule in the asymmetric unit, running the evolutionary search procedure up to ten times or until a correlation coefficient of 0.5 or higher is obtained. By default, data in the resolution range 15–4 Å are used in the search, and the coordinates of the solution having the highest correlation coefficient at the end of the specified number of runs are written to a file.

Searches for multiple molecules or domains in the asymmetric unit are conducted sequentially. After a solution is found, its static contribution to the structure factors is included in the correlation-coefficient calculations on subsequent searches. These sequential searches require smaller population sizes and thus shorter run times than would be necessary for higher dimensional simultaneous searches.

The following example structure determinations illustrate the application of *EPMR*.

3.1. Type 2 rhinovirus protease

A complex of the protease from type 2 rhinovirus with a bound inhibitor was crystallized in space group $P2_12_12$, a = 62.28, b = 77.63, c = 34.10 Å with one complex in the asymmetric unit (Ferre & Matthews, 1998). The structure was determined using as a search model the protease from type 14 rhinovirus (Matthews *et al.*, 1994), which shares 48% sequence identity with type 2 rhinovirus protease. The evolutionary search procedure was applied using all data in the resolution range 15–4 Å (1552 reflections), and was carried out over 50 generations using a population size of 300. A solution with a correlation coefficient of 0.42 (*R* factor = 0.53) was found on the first run of the search procedure, and was later verified by refinement and identification of the bound ligand in difference electron-density maps. The progress of the search is illustrated in Fig. 2.

The total run time was 255 CPU s on a Silicon Graphics Octane workstation with a 175 MHz MIPS R10000 processor. The procedure required approximately 10000 structure-factor

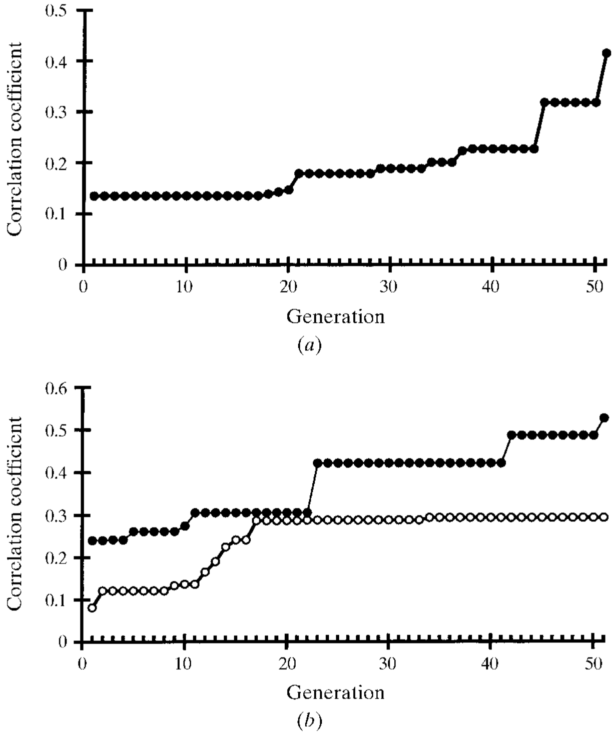

Figure 2
Evolution of MR solutions for (*a*) a rhinovirus protease–inhibitor complex, (*b*) an FKBP12.6–inhibitor complex, showing solutions for the first (open circles) and second (filled circles) molecule in the asymmetric unit. The correlation coefficient of the top solution in each generation is shown. Generation 51 represents the final conjugate-gradient minimization step.

calculations. (Structure factors for surviving solutions do not need to be recalculated.) By comparison, a systematic six-dimensional search using an angular increment of 5° and translational steps of 2 Å would require over 140 million structure-factor calculations. The evolutionary search was thus over 14000 times more efficient than a systematic search in this case. In fact, the evolutionary algorithm is also several times more computationally efficient than a traditional rotation and translation search, which would have required approximately 49000 calculations. The total CPU time for all equivalent steps using *AMoRe*, which has highly efficient implementations of the rotation and translation functions, on the same problem was approximately 58 s. The total CPU time using the direct-rotation function, Patterson correlation refinement, translation function and rigid-body refinement in *X-PLOR* (Brünger, 1992) was 55 min. Thus, *EPMR* can carry out a six-dimensional MR search on a time-scale comparable to that currently required for separate rotation and translation searches.

With traditional MR methods, an indication that a correct answer has been found is given by the peak height relative to the overall background. This is not possible using *EPMR* because a complete search is not performed. A good indication that the correct solution is found using *EPMR* is that the solution is obtained repeatedly in multiple runs. In this example, the correct solution was obtained in 64 out of 100 runs of the search procedure. The highest correlation coefficient for an incorrect solution was 0.312.

3.2. FK506-binding protein 12.6

A complex of FK506-binding protein (FKBP) 12.6 with a bound inhibitor was crystallized in space group $P1$, $a = 32.80$, $b = 35.27$, $c = 47.68$ Å, $\alpha = 85.73$, $\beta = 72.16$, $\gamma = 66.32°$ with two molecules in the asymmetric unit. A search model was constructed using the FKBP12 protein from the FKBP12–rapamycin complex (Protein Data Bank entry 1FKB; Van Duyne *et al.*, 1991). 18 residues (out of 107) that were non-identical in the two proteins were truncated to alanine. A sequential search was performed for the two molecules in the asymmetric unit using *EPMR*. 1480 reflections in the resolution range 15–4 Å were used. Each evolutionary search was carried out over 50 generations using a population size of 300. On the first run, a solution was found for one of the molecules in the asymmetric unit with a correlation coefficient of 0.292 (R factor = 0.502). The partial structure contribution of the first solution was calculated and added to the structure-factor calculations during the search for the second molecule. A second solution was found with a correlation coefficient for the combined solutions of 0.526 (R factor = 0.398). The total run time for the two searches was 138 CPU s on the workstation described in §3.1. The procedure required approximately 20000 structure-factor calculations. By comparison, a rotation search to find the first molecule and systematic six-dimensional search for the second molecule using increments of 5° and 2 Å would require over one billion structure-factor calculations. Separate rotation and translation searches would require roughly 190000 structure-factor calculations. In this case, the evolutionary search was over 50000 times more efficient than a systematic six-dimensional search and roughly nine times more efficient than a rotation/translation search.

3.3. Cytochrome c'

The crystal structure of cytochrome c' from *Alcaligenes denitrificans* was originally determined using a combination of MR and phase information from the anomalous scattering of the heme iron (Baker *et al.*, 1995). The protein crystallized in space group $P6_522$, $a = b = 54.6$, $c = 180.4$ Å, with one molecule in the asymmetric unit. According to the authors, initial attempts at finding a MR solution using *ALMN* (Collaborative Computational Project, Number 4, 1994) and *X-PLOR* with Patterson correlation refinement (Brünger, 1990) were unsuccessful. The structure was eventually determined after a large number of alternative solutions from *AMoRe* (Navaza, 1994) using several different search models were analyzed for consistency with the iron position obtained from an anomalous difference Patterson map and correlation with electron-density maps phased by anomalous scattering (Baker *et al.*, 1995).

The structure was redetermined with *EPMR* using the observed structure factors deposited with the structure (entry 1CGN) in the Protein Data Bank (PDB; Bernstein *et al.*, 1977) and the deposited coordinates for one of the search models used in the original structure determination, residues 3–125 of the cytochrome c' from *Rhodospirillum molischianum* (PDB entry 2CCY). The evolutionary search procedure was applied using a search model consisting of the polyalanine backbone and heme group of the *R. molischianum* protein, a model that was not sufficient to yield the correct solution in the original structure determination. All data in the resolution range 15–4 Å (1527 reflections) were used. The correct solution, with a correlation coefficient of 0.515 (R factor = 0.598) was found on the fifteenth run of the search procedure. The highest correlation coefficient obtained for an incorrect solution was 0.452 after 100 runs of the search procedure, indicating that the correct solution could clearly be identified above the background. The highest correlation obtained when the procedure was carried out in space group $P6_122$ was 0.461, indicating that the correct space group could also be clearly identified. Thus, *EPMR* was able to unambiguously identify the correct solution in this case without additional phase information from other techniques, while using a non-optimal polyalanine search model.

4. Comparison with existing MR methods

To determine if our method offers consistent advantages in signal-to-noise over existing MR methods, we compared the performance of *EPMR* with two widely used programs, *X-PLOR* (Brünger, 1992) and *AMoRe* (Navaza, 1994). We chose four test cases that had both coordinates and structure factors deposited in the PDB. The test cases, summarized in

Table 1
Test cases.

PDB entry code	Protein name	Crystal parameters	Reference
1UBQ	Ubiquitin	$P2_12_12_1$; a = 50.84, b = 42.77, c = 28.95 Å	Vijay-Kumar *et al.* (1987)
6RHN	Histidine-triad nuclear binding protein	$P4_32_12$; a = b = 40.34, c = 143.03 Å	Brenner *et al.* (1997)
1CBY	CytB delta-endotoxin	$P6_122$; a = b = 66.81, c = 170.79 Å	Li *et al.* (1996)
1VIP	Phospholipase A_2	$I4_132$; a = b = c = 122.78 Å	Carredano, E. Westerlund, B., Persson, B., Saarinen, M., Ramaswami, S., Eaker, D. & Eklund, H. Unpublished work.

Table 1, represent a range of space groups of moderate to very high symmetry.

We first evaluated the ability of each program to identify the correct MR solution as increasing amounts of coordinate error were introduced into the search model. The refined protein coordinates deposited in the PDB were taken as the ideal search-model coordinates. Both polyalanine and all-atom search models were constructed, and random normally distributed errors with a defined root-mean-square deviation (RMSD) were introduced into the coordinates of each search model. The RMSD was increased in 0.1 Å increments. Although this type of coordinate error is not representative of that occurring in real-world search models, it does yield models with systematically decreasing correlation between observed and calculated structure factors. The randomized coordinates were submitted to each of the three programs using the experimental conditions described in Fig. 3. An MR search was considered successful if the solution providing the highest correlation coefficient in the final step in each procedure was within 5.0° and 2.0 Å of the deposited coordinates. (The exact choice of cutoffs did not significantly affect the results.)

Fig. 3 compares the maximum amount of coordinate error that could be introduced into the search model when using each of the three programs. The maximum coordinate error varied widely between the different test cases. In four of the eight test conditions, a larger amount of coordinate error could be introduced when using *EPMR* than either of the other two methods. In the remaining four test cases, an equal amount of coordinate error could be introduced when using *EPMR* or *AMoRe*, with somewhat less error allowed using *X-PLOR*.

With all three programs, more error could be introduced when using the polyalanine models than with the all-atom models. We attribute this to the fact that the positions of the randomly displaced atoms in the polyalanine models would sometimes approximate the positions of side-chain atoms not present in the search models.

We next evaluated the ability of the three MR methods to identify the correct solution with increasingly truncated search models. For these tests we chose a test case, 6RHN, which had coordinates available for the search model used in the original structure determination. This protein, histidine-triad nuclear binding protein, was originally determined using the coordinates of the protein with bound adenosine (PDB entry 4RHN; Brenner *et al.*, 1997). The RMSD between the 4RHN coordinates and the final 6RHN coordinates was 0.71 Å for all protein atoms and 0.30 Å for polyalanine atoms. We progressively truncated the protein or polyalanine coordinates of 4RHN by removal of C-terminal residues. The models were truncated in increments of approximately 5% of the total number of residues in the protein, corresponding to five or six

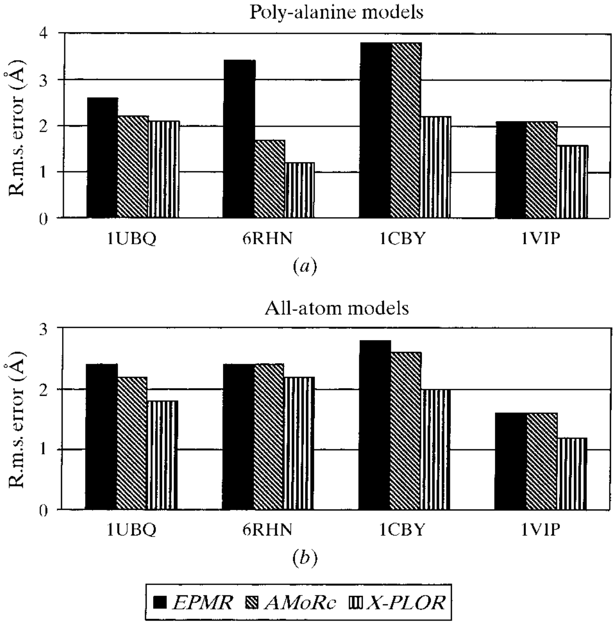

Figure 3
Maximum coordinate error for three MR methods. The tests were carried out with *X-PLOR* version 3.1 (Brünger, 1992) using the direct-rotation function, Patterson correlation (PC) refinement and translation functions and *AMoRe* from version 3.2 of the *CCP4* package (Collaborative Computational Project, Number 4, 1994). In *X-PLOR*, the top 500 peaks from each rotation search were each subjected to ten cycles of PC refinement. The solution having the highest correlation coefficient after PC refinement was submitted to the translation function. For the *AMoRe* rotation function, the integration radius was set to the largest value allowed by the program that was also less than or equal to the maximal extent of the search model from its center of mass. The top 99 rotation-function peaks were submitted to the translation search. The angular increments for the rotation searches were 2.5° in *AMoRe* and 2° in *X-PLOR*. *EPMR* was allowed a maximum of 100 runs to identify the correct solution, using a population size of 300 and 50 generations. Data in the resolution range 15–4 Å were used in all cases. The temperature factors of the deposited coordinates were retained. Each search model was rotated away from the correct orientation by an arbitrary amount before use.

of the 104 residues. The truncated coordinates were submitted to each of the three MR procedures using the same test conditions described in Fig. 3. The results are shown in Fig. 4. For both the all-atom and polyalanine models, the largest degree of truncation was possible when using *EPMR*. The all-atom model could be truncated by approximately 60% using *EPMR*, compared with 55% using *AMoRe* and 40% using *X-PLOR*. The polyalanine model could be truncated by approximately 55% using *EPMR*, 40% using *X-PLOR* and 35% using *AMoRe*.

5. Conclusions

The use of an evolutionary algorithm allows six-dimensional MR searches to be carried out rapidly and offers several advantages over traditional MR methods. Most importantly, the signal-to-noise limitations of sequential rotation and translation searches are avoided. We believe that increased signal-to-noise is the primary explanation for the superior ability of *EPMR* to identify solutions with highly truncated or inaccurate search models. The signal-to-noise ratio in our method is further improved owing to the fact that the rotational and translational parameters for the search model are continuously variable rather than sampled at fixed intervals as is the case in both traditional rotation/translation searches and in the genetic algorithm approach of Chang & Lewis (1997). Although optima are unlikely to be missed entirely if sampling intervals are carefully chosen, the signal-to-noise is invariably compromised unless a solution occurs exactly at a sampled position. Uniform sampling of rotation space in our method further ensures that the search procedure is equally efficient regardless of the orientation of the solution.

MR searches using *EPMR* are rapid and highly automated. The program requires very simple inputs and can produce solution coordinates for problems involving single or multiple molecules in the asymmetric unit in a single step, often within minutes, without user intervention. Our results show that, in some cases, *EPMR* can find solutions using significantly less accurate or less complete search models than is possible with two conventional methods that employ separate rotation and translation searches. These results reflect only a single specific set of test conditions. Judicious selection of resolution ranges, rotation-function integration radius in *AMoRe*, or other parameters could yield improved results with the other two programs. We do not expect that *EPMR* will give superior results on all MR problems. Nevertheless, the program has now been used in a number of laboratories to solve novel crystal structures, including several for which other MR methods had failed (Behnke *et al.*, 1999; Segelke *et al.*, 1999; Stanfield, 1998; C. R. Kissinger, unpublished results).

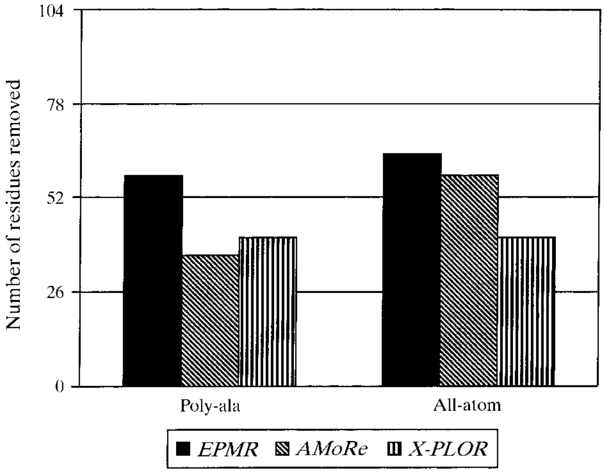

Figure 4
Maximum search-model truncation for three MR methods. The diffraction data were obtained from PDB entry 6RHN. The search models were derived from the protein coordinates of entry 4RHN (a total of 104 residues). The test conditions were as described in Fig. 3.

EPMR could potentially be enhanced in a number of ways. For instance, although we have found that searching sequentially for multiple molecules or domains in the asymmetric unit is an effective approach, simultaneous searches could be a useful alternative. The increased number of search parameters will require larger population sizes in the search, which will increase the computation time, but there could be offsetting advantages in search efficiency. We are investigating the relative efficacy of this approach. Such multi-body searches could be particularly effective when combined with the use of non-crystallographic symmetry restraints.

A shortcoming of our method is that the search efficiency drops as the quality of the search model decreases. When a search model is highly inaccurate or incomplete, it can take a large number of attempts before the correct solution is obtained. However, it should be pointed out that this often may be a search model for which existing methods are not able to identify the solution at all. We are exploring the use of other scoring functions in order to maximize the search efficiency when the model is marginal. It might be more useful in these cases to optimize the internal geometry of the model during the search process. We have recently incorporated the ability to optimize individual segments of a search model into the conjugate-gradient optimization procedure that follows the evolutionary search. This can be particularly valuable in increasing the effectiveness of automated sequential searches for multiple molecules in the asymmetric unit in those cases where the search model requires significant internal adjustment. We are now experimenting with methods of incorporating internal optimization of the search model into the evolutionary search itself.

Ultimately, it should be possible to incorporate not only optimization, but also selection of the search model into the procedure. Instead of a single search model, a set of structural models would be allowed to compete in the evolutionary search process. Although this will undoubtedly necessitate much larger population sizes and much longer computing times, early experiments suggest that this approach is feasible. When combined with a comprehensive database of protein structures, such a procedure could greatly expand the range of applicability of MR.

EPMR is available from the authors upon request.

The authors thank Steve Freer, Ernie Villafranca, Robert Love, Karen Maegley and Dave Matthews (Agouron Pharmaceuticals, Inc.) for valuable discussions and critical reading of this manuscript, and Duncan McRee (The Scripps Research Institute) for making source code for FFT routines available.

References

Baker, E. N., Anderson, B. F., Dobbs, A. J. & Dodson, E. J. (1995). *Acta Cryst.* D**51**, 282–289.

Behnke, C. A., Yee, V. C., Le Trong, I., Pedersen, L. C., Stenkamp, R. E., Kim, S.-S., Reeck, G. R. & Teller, D. C. (1999). In preparation.

Bernstein, B. E. & Hol, W. E. J. (1997). *Acta Cryst.* D**53**, 756–764.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Bowie, J. U. & Eisenberg, D. (1994). *Proc. Natl Acad. Sci. USA*, **91**, 4436–4440.

Brenner, C., Garrison, P., Gilmour, J., Peisach, D., Ringe, D., Petsko, G. A. & Lowenstein, J. M. (1997). *Nature Struct. Biol.* **4**, 231–238.

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.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Chang, G. & Lewis, M. (1997). *Acta Cryst.* D**53**, 279–289.

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

DeLano, W. L. & Brünger, A. T. (1995). *Acta Cryst.* D**51**, 740–748.

Ferre, R. A. & Matthews, D. (1998). Personal communication.

Fogel, D. B. (1995). *Evolutionary Computation: Towards a New Philosophy of Machine Intelligence.* Piscataway, NJ: IEEE Press.

Fogel, L. J., Owens, A. J. & Walsh, M. J. (1966). *Artificial Intelligence Through Simulated Evolution.* New York: John Wiley.

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

Gehlhaar, D. K., Verkhivker, G. M., Rejto, P. A., Sherman, C. J., Fogel, D. B., Fogel, L. J. & Freer, S. T. (1995). *Chem. Biol.* **2**, 317–324.

Huber, R. (1965). *Acta Cryst.* **19**, 353–356.

Huber, R. & Schneider, M. (1985). *Acta Cryst.* **18**, 165–169.

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

Li, J., Koni, P. A. & Ellar, D. J. (1996). *J. Mol. Biol.* **257**, 129–152.

Matthews, D., Smith, W. W., Ferre, R. A., Condon, B., Budahazi, G., Sisson, W., Villafranca, J. E., Janson, C. A., McElroy, H. E., Gribskov, C. L. & Worland, S. (1994). *Cell*, **77**, 761–771.

Minkler, G. & Minkler, J. (1990). *Aerospace Coordinate Systems and Transformations.* Baltimore: Magellan.

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

Oh, B.-H. (1995). *Acta Cryst.* A**51**, 140–144.

Powell, M. J. D. (1977). *Math. Programming*, **12**, 241–254.

Rabinovich, D. & Shakked, Z. (1984). *Acta Cryst.* A**40**, 195–200.

Rossmann, M. G. & Blow, D. M. (1962). *Acta Cryst.* **15**, 24–31.

Salomon, R. (1996). *BioSystems*, **39**, 263–278.

Schwefel, H.-P. (1981). *Numerical Optimization of Computer Models.* Chichester: John Wiley.

Segelke, B. W., Trakhanov, S., Knapp, M., Newhouse, Y. M., Weisgraber, K. & Rupp, B. (1999). In preparation.

Stanfield, R. (1998). Personal communication.

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

Van Duyne, G. D., Standaert, R. F., Schreiber, S. L. & Clardy, J. C. (1991). *J. Am. Chem. Soc.* **113**, 7433–7434.

Vijay-Kumar, S., Bugg, C. E. & Cook, W. J. (1987). *J. Mol. Biol.* **194**, 531–544.