Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Molecular replacement by evolutionary search

Charles R. Kissinger,*† Daniel K. Gehlhaar, Bradley A. Smith and Djamal Bouzida

Pfizer Inc., 10350 North Torrey Pines Road, San Diego, CA 92037, USA

† Current address: Structural GenomiX Inc., 10505 Roselle Street, San Diego, CA 92121, USA.

Correspondence e-mail: chuck_kissinger@stromix.com

Received 29 March 2001
Accepted 18 July 2001

Stochastic search algorithms can be used to perform rapid six-dimensional molecular-replacement searches. A molecular-replacement procedure has been developed that uses an evolutionary algorithm to simultaneously optimize the orientation and position of a search model in a unit cell. Here, the performance of this algorithm and its dependence on search model quality and choice of target function are examined. Although the evolutionary search procedure is capable of finding solutions with search models that represent only a small fraction of the total scattering matter of the target molecule, the efficiency of the search procedure is highly dependent on the quality of the search model. Polyalanine models frequently provide better search efficiency than all-atom models, even in cases where the side-chain positions are known with high accuracy. Although the success of the search procedure is not highly dependent on the statistic used as the target function, the correlation coefficient between observed and calculated structure-factor amplitudes generally results in better search efficiency than does the *R* factor. An alternative stochastic search procedure, simulated annealing, provides similar overall performance to evolutionary search. Methods of extending the evolutionary search algorithm to include internal optimization, selection and construction of the search model are now beginning to be investigated.

1. Introduction

New molecular-replacement methods have been introduced recently that use stochastic search algorithms to optimize the orientation and position of a molecular model in a unit cell. Unlike traditional approaches, these methods do not divide the molecular-replacement procedure into separate rotation and translation searches. Instead, the three rotational and three translational parameters for a search model are optimized simultaneously. Several stochastic search techniques have been applied successfully to the molecular-replacement problem, including a genetic algorithm (Chang & Lewis, 1997), evolutionary programming (Kissinger *et al.*, 1999) and simulated annealing (Glykos & Kokkinidis, 2000). These methods are several orders of magnitude faster than a comprehensive six-dimensional search and have proven to be effective tools for molecular replacement.

Six-dimensional searches have potential advantages over traditional molecular-replacement methods that use separate rotation and translation searches. Although significant advances have been made in increasing the effectiveness of rotation-search methods (Navaza & Saludjian, 1997; Brünger, 1997; Tong & Rossmann, 1997), the signal-to-noise ratio in a rotation search is relatively low because the correlation with

the observed diffraction data is modest for a search model that is correctly rotated but mistranslated. Furthermore, the subsequent translation search can be highly sensitive to small errors in the orientation obtained from the rotation search. Conversely, when the orientation and position of a search model are optimized simultaneously the maximum achievable correlation between observed and calculated diffraction data can be obtained.

While advances in computing power have made systematic six-dimensional searches increasingly practical, particularly when spread over multiple processors (Sheriff *et al.*, 1999), these generally remain very lengthy calculations. Stochastic search methods are dramatically faster than a complete systematic six-dimensional search. They do have some disadvantages, however. These algorithms are non-deterministic and the correct solution will not be found on every attempt. Multiple search attempts are typically required in order to ensure that the global optimum has been found. The usefulness of this type of algorithm, therefore, depends critically on its efficiency and reliability across the broad range of molecular-replacement problems. We have developed a program, *EPMR*, which uses an evolutionary algorithm to perform rapid six-dimensional molecular-replacement searches (Kissinger *et al.*, 1999). Here, we examine the performance of this algorithm and the effect of the quality of the search model and the choice of statistic used as the target function. We also compare the performance of this algorithm with that of another stochastic search algorithm, simulated annealing.

2. Molecular replacement using *EPMR*

In an evolutionary algorithm (Fogel *et al.*, 1966), the global optimum of a function is found through the iterative optimization of a population of initially random trial solutions. During each cycle of optimization, the solutions are ranked and the best ones are retained to serve as 'parents' of the next generation. Random variations are applied to the values of these parent solutions to generate a new population and this process is repeated for a number of generations. The size of the 'mutations' applied to the parent solutions is typically adjusted to provide broad sampling of the entire search space in the initial stages of the search while gradually focusing in on only the most promising areas. Evolutionary algorithms have been applied successfully to a wide variety of optimization problems involving multidimensional non-linear search spaces (Fogel, 1995).

In our procedure, the three rotational and three positional parameters that describe the orientation and position of the search model within the unit cell are optimized using the correlation coefficient between observed and calculated structure-factor amplitudes to rank the solutions. The procedure is described in detail in a previous publication (Kissinger *et al.*, 1999). Briefly, an initial population of molecular-replacement solutions is created by assigning a random orientation and position to each population member. The correlation coefficient for each solution is calculated and compared with those of a small number of other randomly chosen solutions. The number of these competitions that each solution wins is used to rank the solutions and determine which ones will survive to the next generation. Surviving solutions are retained in the next generation without modification and are also used to regenerate the rest of the population by applying normally distributed random variations to the values for the orientation and position. Structure factors for each solution are calculated rapidly during the course of the optimization by pre-calculating the molecular transform of the search model and applying the rotations and translations in reciprocal space (Lattman & Love, 1970; Huber & Schneider, 1985; Castellano *et al.*, 1992). A population size of 300 solutions, evolving over 50 generations, has proven to provide a good compromise between the speed and probability of success of an individual search. At the end of 50 generations, the population member with the best correlation coefficient is chosen for a local conjugate-gradient optimization. Searches for multiple copies of a molecule in the asymmetric unit can be carried out sequentially. (Simultaneous searches for multiple copies of a molecule are discussed below.) In the sequential procedure, when each copy of the molecule is found its static contribution to the structure factors is calculated and included in the correlation-coefficient calculations during subsequent searches.

Table 1
Number of positions evaluated using three alternative molecular-replacement procedures.

The systematic six-dimensional search and rotation/translation search examples assume angular intervals of 5° and translational intervals of 2 Å. Values for the evolutionary search procedure represent a single attempt using a population size of 300 and 50 generations. (Surviving solutions from previous generations do not need to be re-evaluated.)

Test case	Systematic six-dimensional search	Rotation + translation search	Evolutionary six-dimensional search
RVP†	1.4×10^8	4.9×10^4	1.0×10^4
CNFK‡	2.0×10^9	1.1×10^5	1.0×10^4

† Rhinovirus protease–inhibitor complex, space group $P2_12_12$, a = 62.3, b = 77.6, c = 34.1 Å. ‡ Calcineurin–FKBP12–FK506 complex, space group $P4_32_12$, $a = b$ = 105, c = 162 Å.

Evolutionary six-dimensional molecular-replacement searches are computationally efficient. As illustrated in Table 1 for two representative test cases, the use of an evolutionary algorithm is typically four or five orders of magnitude more efficient than a complete systematic six-dimensional search. In fact, a six-dimensional evolutionary search frequently requires fewer structure-factor calculations than would an equivalent combination of systematic rotation and translation searches [although separate rotation/translation searches can still be performed more quickly using the highly efficient 'fast rotation' (Crowther, 1972; Navaza, 1993) and 'fast translation' functions (Navaza & Vernoslova, 1995)]. A typical evolutionary molecular-replacement search can be completed in several minutes on a modern workstation. The true efficiency of this approach depends ultimately, however, not only on the speed of a single search attempt but also on how many

attempts are required to ensure that the optimal solution has been found. For this reason, we have examined the effectiveness of the search procedure under a variety of conditions. In the following sections we explore two aspects of the performance of the algorithm: the search efficiency, defined as the probability of success in finding the correct solution on a single search attempt, and the sensitivity, defined as the maximum reduction in search model quality that still produces the correct solution. Our results are illustrated using a single test case, but are representative of results obtained on a wide variety of molecular-replacement problems.

3. Search efficiency and model accuracy

To determine the effect of model quality on the effectiveness of the evolutionary search procedure, we examined the search efficiency using models of systematically decreasing accuracy or completeness. Fig. 1 shows the success rate of the procedure with search models that were incrementally truncated by removing residues from the carboxyl terminus. For both a polyalanine model and a model including side chains, the search efficiency decreased in a roughly linear manner as the completeness of the search model was decreased. We obtained very similar results on a variety of other test cases. The same behaviour also was seen for search models with increasing amounts of coordinate error (data not shown). Clearly, in molecular-replacement problems where the search model is marginal, a large number of search attempts are required to ensure that the optimal solution is obtained. Although this is a drawback of the procedure, it may be compensated by the potential in some cases to use less complete or less accurate search models than is possible with conventional methods (Kissinger *et al.*, 1999).

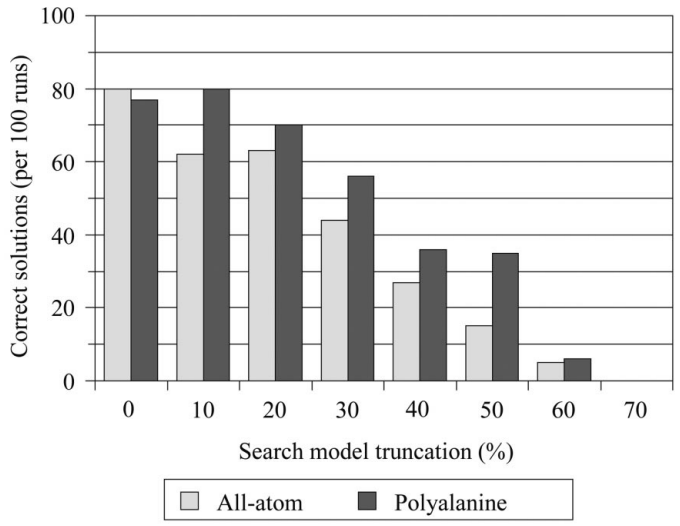

Figure 1
Effect of model completeness on evolutionary search efficiency. Results are shown for polyalanine and all-atom search models that were incrementally truncated at the carboxy-terminal end by roughly 10% of the total residues (11–12 residues) at a time. The evolutionary search used a population size of 300 evolving over 50 generations and data between 15 and 4 Å resolution. The diffraction data used in this test was from PDB entry 6rhn, histidine triad nuclear binding protein (Brenner *et al.*, 1997). The search models were derived from the refined coordinates of the same protein determined in a different crystal form (PDB entry 4rhn, 115 residues).

Table 2
Efficiency of alternative evolutionary search protocols for a marginal search model.

The success rate using the 60% truncated search model from Fig. 1 was measured for the three population sizes and the probability of success after the specified number of search attempts was calculated.

Population size	No. of attempts	Probability of success (%)
300	30	77
600	15	74
900	10	61

The chance of success in a given search attempt can be improved by increasing the population size or number of iterations of optimization, at the expense of more computation time. There is not a linear increase in search efficiency with increases in population size or number of iterations, however. The probability of success asymptotically approaches but never reaches 100%, with each incremental increase in population size providing a smaller improvement in search efficiency. It is usually more advantageous to increase the number of search attempts rather than the population size, as illustrated in Table 2 for three protocols that required roughly the same amount of computation time.

Interestingly, we have found that the search efficiency is frequently higher for polyalanine search models than for models that include side chains. This is true even for models in which the side-chain positions are quite accurate. We see two possible reasons for this. One is that a polyalanine model, although less complete, is frequently significantly more accurate. For the test case shown in Fig. 1, the RMS differences between the search-model coordinates and final refined coordinates are 0.37 Å for the polyalanine model and 0.71 Å for the all-atom model. It has been a common observation that molecular-replacement models that are less complete but more accurate provide better results than more complete less accurate ones. The increased search efficiency occurs in this case despite the fact that the correlation coefficients are significantly lower for the polyalanine models than for the all-atom models. We are investigating the possibility that use of polyalanine search models may be advantageous because this results in a smoother variation in the value of the correlation coefficient throughout the search space, thus providing an easier landscape for the optimization algorithm to traverse.

We also compared the maximum search-model truncation that was possible using both polyalanine and all-atom search models. The search models were truncated one residue at a time until the solution with the highest correlation coefficient that was obtained after 100 search attempts no longer corresponded to the correct solution. Surprisingly, nearly as many residues could be removed from the polyalanine model as from the all-atom model. The minimally successful all-atom search model, comprising 41 residues, represented roughly 35% of the total number of atoms in the molecule, while the

minimal polyalanine model, 44 residues, represented only about 24% of the complete molecule.

These results were obtained using a high-quality search model in which most side-chain positions were very close to those in the final refined structure (RMS coordinate differences from the final model for side-chain atoms only were 1.1 Å). In molecular-replacement problems involving homology models derived from distantly related proteins, the results would be expected to be much less favourable for an all-atom model. Our experiments suggest that there is rarely an advantage in including side chains in a model for molecular replacement, at least when using the evolutionary search procedure.

4. Alternative scoring methods

We next investigated whether the search efficiency could be improved by using a statistic other than the standard correlation coefficient between observed and calculated structure-factor amplitudes as the target function. Fig. 2 shows the search efficiency for a polyalanine search model using three different targets: the correlation coefficient using standard structure factors, the correlation coefficient using normalized structure factors and the *R* factor. We have found little difference in search efficiency between correlation coefficients based on either unnormalized or normalized structure factors. In the test case shown, the use of normalized structure factors results in generally better search efficiency. Because our procedure does not attempt to optimize an overall temperature factor for the search model, the use of normalized structure factors might be advantageous in some cases because it minimizes the effect of errors in the temperature factors assigned to the atoms in the search model. The *R* factor consistently provided lower efficiency in our search procedure.

The maximum amount of search-model truncation that was possible using the different target statistics is shown in Fig. 3. Use of normalized structure factors allowed truncation of one additional residue from the search model in this case. Using the *R* factor, significantly less truncation was possible before the best-scoring solution no longer corresponded to the correct solution.

We are continuing to investigate alternative scoring methods. The latest version of *EPMR* allows the user to choose the correlation coefficient between standard or normalized structure-factor amplitudes (or the squares of the structure factors) or the *R* factor as the target function. More sophisticated scoring methods, such as those based on maximum likelihood (Bricogne, 1992; Read, 2001), might lead to further improvements in search efficiency and sensitivity.

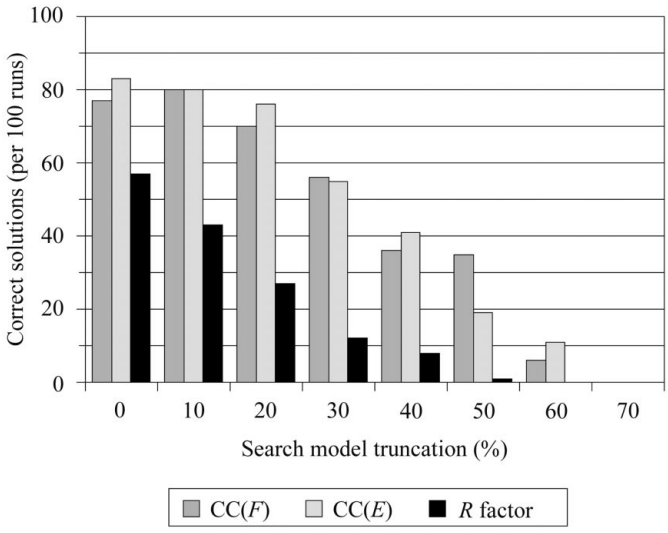

Figure 2
Effects of alternative scoring methods on search efficiency. The polyalanine search model was truncated incrementally as described in Fig. 1. Solutions were scored using the correlation coefficients between observed and calculated structure-factor amplitudes, labelled CC(*F*), normalized structure-factor amplitudes, CC(*E*), or the *R* factor.

5. Alternative stochastic search algorithms

We also investigated whether the search efficiency, particularly with low-accuracy models, could be improved using an alternative stochastic search technique. Simulated annealing (Kirkpatrick *et al.*, 1983; Kirkpatrick & Swendsen, 1985) is another widely used stochastic method for finding solutions to non-linear optimization problems. This technique has been used for crystallographic refinement (Brünger *et al.*, 1987) and for protein structure determination from NMR data (Nilges *et al.*, 1988); it has also recently been applied to the molecular-replacement problem (Glykos & Kokkinidis, 2000).

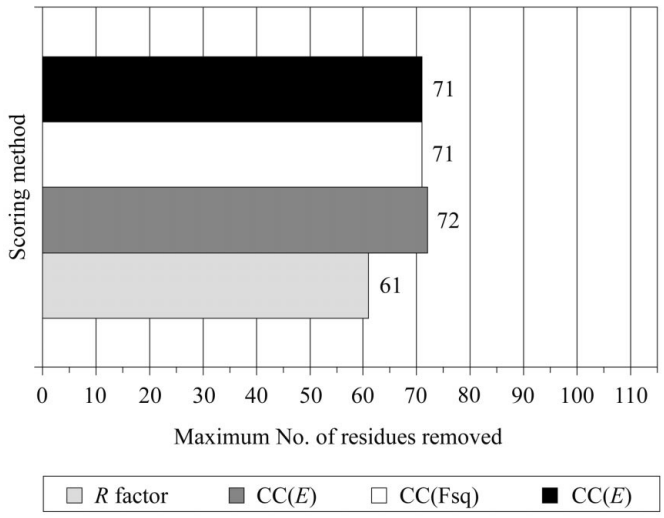

Figure 3
Maximum search-model truncation using four alternative scoring methods. The polyalanine search model was truncated from the carboxyl terminus one residue at a time. Shown is the maximum number of residues that could be removed from the search model before the highest scoring solution obtained in 100 runs of the evolutionary search procedure no longer corresponded to the correct solution. Solutions were scored using the correlation coefficients between observed and calculated structure-factor amplitudes, labelled CC(*F*), the squares of the structure-factor amplitudes, CC(Fsq), normalized structure-factor amplitudes, CC(*E*), or the *R* factor.

We implemented a six-dimensional molecular-replacement search procedure using a simulated-annealing algorithm following the recommendations from Kirkpatrick (1984). The temperature was decreased exponentially in 75 steps, with 300 trial moves at each temperature. To facilitate efficient sampling of the search space, we used a dynamically optimized Monte Carlo method based on the acceptance ratio of the trial moves (Bouzida & Kumar, 1992). Using this method, the acceptance ratio (the ratio of accepted moves to the total number of trial moves) is optimized by adjusting the maximum move size during each temperature step.

We compared the performance of this algorithm to that of our evolutionary search procedure. The results are shown in Fig. 4. The efficiency of the simulated-annealing algorithm was lower than that of the evolutionary algorithm, but also showed a roughly linear decrease in search efficiency with decreasing model completeness. This appears to be an inherent characteristic of these kinds of search methods when applied to the molecular-replacement problem. We believe the lower overall efficiency of the simulated-annealing algorithm primarily reflects a difference in the amount of tuning that we have performed to the two algorithms rather than an inherent superiority of the evolutionary algorithm. As expected, the maximum possible search-model truncation was identical for the two methods (data not shown).

6. The future of stochastic search methods in molecular replacement

Stochastic search algorithms have proven to be reliable and efficient tools for solving molecular-replacement problems. The ability to easily extend searches to more than six dimensions suggests a number of interesting ways in which these methods can be enhanced. For instance, when there are multiple copies of a molecule in the asymmetric unit, the standard application of our method is to search for each copy sequentially, adding a static contribution to subsequent structure-factor calculations as each solution is found. However, procedures have recently been described for searching simultaneously for multiple copies of a molecule (Glykos & Kokkinidis, 2000, 2001) and we have recently added this capability to *EPMR*. A simultaneous search has the same potential advantage over a sequential search that a six-dimensional search has over sequential rotation and translation searches; the signal-to-noise ratio may be increased, but at the expense of increased computation time. The increase in the number of variables being optimized requires a much larger population size for the evolutionary search to be effective. There is currently a practical limit of two or three molecules that can be found simultaneously using our procedure without requiring impracticably long search times. (Problems involving larger numbers of molecules in the asymmetric unit can be attempted using a series of two-molecule searches, however.)

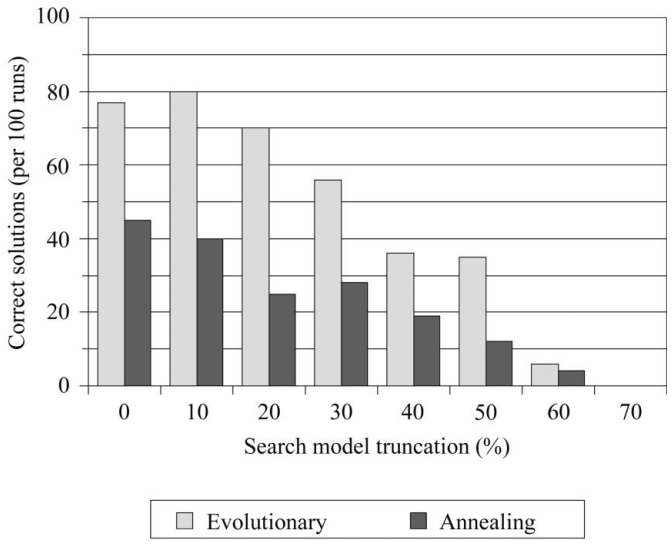

Figure 4
Effect of model completeness on the efficiency of two stochastic search procedures. The polyalanine search model was truncated incrementally as described in Fig. 1 and the success rates of our standard evolutionary algorithm and a simulated-annealing algorithm were measured. The correlation coefficient using unnormalized structure factors was used as the target statistic for both methods.

In many molecular-replacement problems, it would be valuable to be able to specify particular internal degrees of freedom within the search model; for instance, to allow the angle between two domains to vary. Brünger (1990) has demonstrated the efficacy, when using separate rotation and translation searches, of optimizing the molecular-replacement model after the rotation search by 'Patterson correlation' refinement to improve the performance of the subsequent translation search. We have found that the addition of a small number of internal degrees of freedom is possible in the evolutionary search procedure without a significant decrease in search efficiency. In this type of procedure, it is necessary to calculate a separate molecular transform for each independently varying part of the search model, however, and the resulting high memory requirements can limit the amount of internal optimization that can be attempted.

We are also extending the search algorithm to include not only optimization but also selection of the best molecular model. This is accomplished by including an additional parameter in the optimization that specifies which one of a set of search models is used. Initial experiments suggest that it is possible to select the optimal model from a set of similar aligned search models while simultaneously optimizing its position and orientation. This may be possible with larger databases of more distantly related structures as well if a suitable alignment is possible. The memory requirements for holding the molecular transforms of a large number of search models again represent a practical impediment to this approach, however. For this reason, screening of multiple search models might best be accomplished by evaluating the alternative models in parallel on multiple computers. Evolutionary algorithms are ideally suited for implementing in parallel across a large number of processors. We have developed a highly parallel version of *EPMR* to run on multiprocessor computers or computer clusters. This will facilitate

the rapid screening of large databases of search models. The ultimate realization of this approach might be the use of a database of structures derived from the entire Protein Data Bank.

The extensible nature of stochastic search algorithms also makes them well suited for implementing procedures in which the search model is not simply optimized but instead actually constructed from structural fragments during the search process. A theoretical basis for such an approach has been discussed by Bricogne (1997). This kind of search-model 'construction' clearly can be accomplished using a limited number of large domains, where the problem is essentially the same as that of finding multiple molecules in the asymmetric unit. We are investigating methods of extending the procedure to much smaller fragments.

The program *EPMR* can be downloaded free of charge from ftp://ftp.agouron.com/pub/epmr/.

References

Bouzida, D., Kumar, S. & Swendsen, R. H. (1992). *Phys. Rev. A*, **45**, 8894–8901.

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

Bricogne, G. (1992). *Proceedings of the CCP4 Study Weekend. Molecular Replacement*, edited by E. J. Dodson, S. Gover & W. Wolf, pp. 62–75. Warrington: Daresbury Laboratory.

Bricogne, G. (1997). *Methods Enzymol.* **277**, 14–18.

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

Brünger, A. T. (1997). *Methods Enzymol.* **276**, 558–580.

Brünger, A. T., Kuriyan, J. & Karplus, M. (1987). *Science*, **235**, 458–460.

Castellano, E. E., Oliva, G. & Navaza, J. (1992). *J. Appl. Cryst.* **25**, 281–284.

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

Crowther, R. A. (1972). *The Molecular Replacement Method*, edited by M. G. Rossmann, pp. 173–178. New York: Gordon & Breach.

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.

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

Glykos, N. M. & Kokkinidis, M. (2001). *Acta Cryst.* D**57**, 1462–1473.

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

Kirkpatrick, S. (1984). *J. Stat. Phys.* **34**, 975–986.

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

Kirkpatrick, S. & Swendsen, R. H. (1985). *Commun. ACM*, **28**, 363–373.

Kissinger, C. R., Gehlhaar, D. K. & Fogel, D. B. (1999). *Acta Cryst.* D**55**, 484–491.

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

Navaza, J. (1993). *Acta Cryst.* D**49**, 588–591.

Navaza, J. & Saludjian, P. (1997). *Methods Enzymol.* **277**, 581–594.

Navaza, J. & Vernoslova, E. (1995). *Acta Cryst.* A**51**, 445–449.

Nilges, M., Marius, G. & Gronenborn, A. M. (1988). *FEBS Lett.* **239**, 129–136.

Read, R. J. (2001). *Acta Cryst.* D**57**, 1373–1382.

Sheriff, S., Klei, H. E. & Davis, M. E. (1999). *J. Appl. Cryst.* **32**, 98–101.

Tong, L. & Rossmann, M. G. (1997). *Methods Enzymol.* **276**, 594–610.