Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Automated structure solution, density modification and model building

Thomas C. Terwilliger

Bioscience Division, Mail Stop M888, Los Alamos National Laboratory, Los Alamos, NM 87545, USA

Correspondence e-mail: terwilliger@lanl.gov

Received 31 January 2002
Accepted 10 September 2002

The approaches that form the basis of automated structure solution in *SOLVE* and *RESOLVE* are described. The use of a scoring scheme to convert decision making in macromolecular structure solution to an optimization problem has proven very useful and in many cases a single clear heavy-atom solution can be obtained and used for phasing. Statistical density modification is well suited to an automated approach to structure solution because the method is relatively insensitive to choices of numbers of cycles and solvent content. The detection of non-crystallographic symmetry (NCS) in heavy-atom sites and checking of potential NCS operations against the electron-density map has proven to be a reliable method for identification of NCS in most cases. Automated model building beginning with an FFT-based search for helices and sheets has been successful in automated model building for maps with resolutions as low as 3 Å. The entire process can be carried out in a fully automatic fashion in many cases.

1. Introduction

Macromolecular X-ray crystallography is undergoing a period of rapid change. The technologies needed for the steps of structure determination are becoming reliable and powerful enough that they can be linked together into an automatic sequence that can yield a structure in an automatic or nearly automatic fashion. The clear potential of fully automated structure determination has supplied much of the impetus for the vision of large-scale structure determination in structural genomics. Here, some of the approaches that form the basis of the automated structure-solution approach implemented in *SOLVE* and *RESOLVE* are described. Many of the details of these methods have been described in earlier papers on *SOLVE* (Terwilliger & Berendzen, 1999*a*) and on *RESOLVE* (Terwilliger, 2000) and here the focus will be on the general approaches and on the aspects that are key to the automation of structure solution.

2. *SOLVE* – automated structure solution

Automation of steps in macromolecular X-ray crystallography has two principal requirements. Firstly, all the individual steps in the process need to be made seamless and individually reliable and, secondly, a process for decision making needs to be developed. These two steps are not completely independent, as it is rather common to find a need to insert steps in a process as it is being readied for automation.

The development of *SOLVE* rather closely followed the division of steps described above. Nearly all the routines that form the core procedures in *SOLVE* were developed before *SOLVE* integration was carried out. This includes, for

example, the origin-removed difference Patterson refinement of heavy-atom parameters in *HEAVY* (Terwilliger & Eisenberg, 1983) that allows rapid refinement during the heavy-atom search procedure. Similarly, the estimation of heavy-atom F_A structure factors from MAD data (*MADBST*; Terwilliger, 1994*a*) and the conversion of MAD data to a pseudo-SIRAS form (*MADMRG*; Terwilliger, 1994*b*) which set up the Patterson search for heavy-atom trial solutions (Terwilliger *et al.*, 1987), the iterative heavy-atom refinement/difference Fourier approach to building up heavy-atom solutions in *SOLVE* and the Bayesian MAD phasing algorithm (Terwilliger & Berendzen, 1996) used for final phase calculation in MAD structure determination were developed prior to integration. Routines for scaling (using local scaling; Matthews & Czerwinski, 1975), peak searching, data input and output and other housekeeping routines were also largely developed before the full integration of the structure-solution process.

Once all the core procedures in *SOLVE* were developed and placed into one program unit, it became realistic to think of automating the process of structure solution (finding heavy-atom sites through phasing). The linking together of routines in *SOLVE* and the decision-making process were incorporated together. These two aspects required a surprisingly large amount of software and most of the code in the *SOLVE* program is for these purposes.

The decision-making process in *SOLVE* is very simple. It uses a scoring system to make all the important decisions, the principal one of which is to decide which of two heavy-atom solutions is the better one. Once a way of evaluating the quality of heavy-atom solutions is decided upon, then the decision as to which one to pursue can be as simple as choosing the one with the higher score (see Fig. 1).

SOLVE uses four criteria to score solutions. These are the agreement of the observed and calculated Patterson functions, the internal consistency of the heavy-atom model as measured by the figure of merit, the internal consistency of the model as measured by cross-validation difference Fourier analysis and the believability of the electron-density map as measured by the definition of clear solvent and macromolecule-containing regions. For each criterion, the scoring is carried out in two steps. Firstly, a raw score that measures something related to the quality of the solution is calculated. This is then typically converted to a Z score by subtracting the mean value of scores for a number of trial solutions (most of which are incorrect) and dividing by the standard deviation of those scores. Such a Z score is, roughly speaking, a measure of the log probability that a solution with that score would be obtained by chance. The overall score for a solution is simply the sum of Z scores for all the criteria.

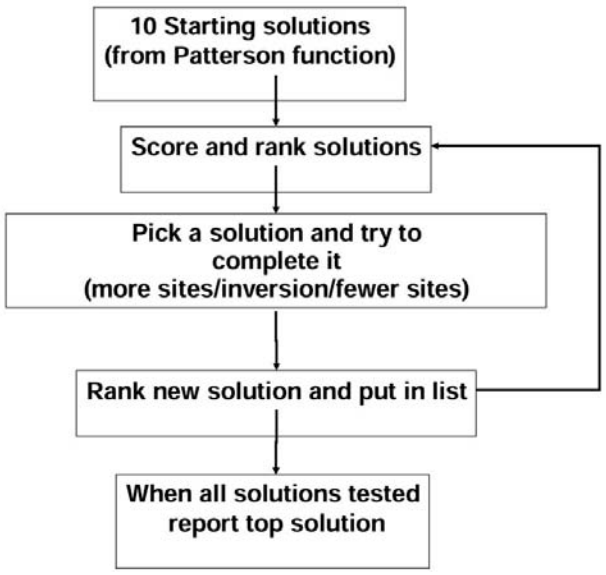

Figure 1
Flowchart of *SOLVE* decision-making process.

The raw score for agreement of calculated and observed Patterson functions is obtained from the mean peak height (normalized to the r.m.s. of the Patterson) at positions expected for the heavy-atom solution. In order to express the expectation that a particular mean peak height is much less likely to be found for a solution with many sites, the raw score is multiplied by the square root of the number of sites. Additionally, peaks that fall near the origin or on very high noise peaks are identified as having heights higher than expected based on the other peak heights. These peak heights are limited to the expected height plus one standard deviation.

The score for the figure of merit is simply the figure of merit itself. *SOLVE* uses origin-removed difference Patterson refinement, which yields a relatively unbiased figure of merit. Consequently, although there is some uncertainty in the figure of merit, it is a good measure of the quality of the solution.

The cross-validation difference Fourier score is obtained by using a variation on the long-established method of evaluating heavy-atom derivatives by removing a site or derivative, using all the other sites or derivatives to calculate phases, calculating a difference Fourier and evaluating whether there are peaks for the sites or derivatives that had been removed (Dickerson *et al.*, 1961). To convert this process to a scoring scheme, a raw score for the cross-validation difference Fourier is the mean peak height (normalized to the r.m.s. of the map) at the coordinates of a heavy-atom site after removing it and carrying out this process. Once again, the score is multiplied by the square root of the number of sites to weight solutions with many sites more strongly than those that have few.

The scoring of the native Fourier is based on the identification of solvent and macromolecule-containing regions that are contiguous and clearly separated from each other. The algorithm used is to divide the asymmetric unit into boxes 5–10 Å on a side and to calculate the standard deviation of electron density within each box. For boxes in the macromolecule-containing region this standard deviation will typically be high and for boxes in the solvent-containing region it will typically be small. Furthermore, as the solvent-containing regions and the macromolecule-containing regions are typically larger in extent than 5–10 Å, the standard deviations of electron density in neighboring boxes tend to be strongly correlated. The score for the native Fourier is therefore just the mean correlation of standard deviation of

electron density for adjacent boxes. This turns out to be quite a good indicator of the quality of an electron-density map, particularly for identifying small improvements in an already good-quality map (Terwilliger & Berendzen, 1999*b*).

The algorithm in *SOLVE* for finding, evaluating and selecting heavy-atom solutions is iterative (Fig. 1). Firstly, the Patterson function is used to identify plausible pairs of heavy-atom sites (even if there are far more sites actually present). Each set of about ten such pairs of trial sites is generally considered as a starting point for generating the entire set of heavy-atom sites. Each of these trial solutions is scored using the four criteria described above and sorted from high to low score. The highest-scoring trial solution is tested as a seed for generating additional sites and completing the solution. This process consists of calculating a difference Fourier, picking the top peaks that are not already part of the solution and generating additional trial solutions by combining some of these peaks with the sites already part of the seed. Additional solutions are generated by removing sites and by inversion. Each additional trial solution is scored and added to the list of scored solutions in the appropriate place. Beginning with a single seed, this process is repeated until no further improvement in Z score is obtained from any additions or deletions of sites or by inversion. If the solution found using the first seed satisfies standard criteria for quality (typically a figure of merit > 0.5 and a Z score above 10), then further solutions are generally not pursued.

This iterative algorithm for identifying heavy-atom solutions in MAD, MIR and SAD experiments has proven useful in the solution of many structures, including proteins as large as the small subunit of the ribosome (Wimberly *et al.*, 1999) and with as many as 56 selenium sites (W. W. Smith and C. Janson, unpublished data).

3. *RESOLVE* – NCS identification and statistical density modification

Density modification is a well established and very powerful method for improving the quality of electron-density maps (Rossmann, 1972; Bricogne, 1976; Wang, 1985; Xiang *et al.*, 1993; Cowtan & Main, 1993; Szöke, 1993; Abrahams & Leslie, 1996; van der Plas & Millane, 2000). The foundation of the method is that a crystallographer has substantial prior knowledge about what an electron-density map ought to look like. For example, most crystals of macromolecules have substantial regions containing disordered solvent that appear very flat in a good map. Similarly, many crystals show non-crystallographic symmetry (NCS). In many cases, the solvent regions or NCS can be identified fairly accurately from an experimental map. In these cases, the plausibility of the electron-density map can be evaluated based on the flatness of the solvent region or the similarity of NCS-related regions.

In the statistical density modification method developed for *RESOLVE*, the agreement of the electron-density map with the experimental map and the agreement of the map with expectation are simultaneously maximized. The procedure is conceptually straightforward. In concept (though not exactly in practice), a cycle of density modification consists of an updating of the phase-probability distribution for each reflection, considering each reflection one at a time. The phase-probability distribution for a reflection is the product of the distribution obtained from experiment (MAD, MIR, SAD *etc.*) and the distribution obtained from the way a 'map-probability' function depends on the phase of that reflection. Statistical density modification is particularly well suited for automated procedures because the approach is relatively insensitive to both choice of solvent content and number of cycles (Terwilliger, 1999).

The map-probability function is a mathematical description of the believability of an electron-density map (Terwilliger, 2001). It is the sum, over all positions in the asymmetric unit, of the believability of the electron density at that point, given the knowledge of whether that point is in the solvent or macromolecule-containing regions. To use the map-probability function to estimate the phase probability for a particular reflection k, the following process is used (in concept). Firstly, a map is calculated omitting this reflection. A series of maps are then calculated by adding in reflection k with each possible phase. Each map is evaluated using the map-probability function, yielding an estimate of the relative probability that this map is correct and, consequently, an estimate of the probability that this phase for reflection k is correct.

In practice, the map-probability function and the way it changes with the phases of reflections is evaluated using an FFT-based method (Terwilliger, 1999). Consequently, the probabilities for all the phases can be calculated in one pass and the process is quite rapid.

This statistical density-modification procedure, like other methods based on the same fundamental information, can make use of many types of information. In particular, non-crystallographic symmetry is an exceptionally powerful source of phase information. When supplied with a file listing heavy-atom sites (such as that written by *SOLVE*), *RESOLVE* identifies potential NCS in the sites by looking for symmetry operators that relate most of the sites to each other. The search is made rapid by eliminating implausible operators that relate sets of sites that do not have matching interatomic distances. *RESOLVE* then evaluates whether the NCS actually exists by examining the correlation of electron density in the experimental map in NCS-related regions. If there is substantial correlation, the NCS operations are refined and the NCS is used as an additional source of prior knowledge about the electron-density map, along with the presence of solvent-containing regions.

4. *RESOLVE* – automated protein model building

RESOLVE carries out automated model building of protein chains in a three-step procedure. Firstly, the locations of helical or sheet-containing regions are identified with a convolution-based method. Next, a tripeptide-fragment library is used to extend the helices and sheets in each direction and the resulting fragments are joined. Thirdly, side chains are added and the sequence aligned to the model.

Helical and sheet-containing regions are identified in *RESOLVE* using an FFT-based method similar to that described by Cowtan (1998). Templates for helices and sheets were created from the average electron density in a library of segments. These templates are rotated in all appropriate orientations and a convolution of the templates is then carried out with the electron-density map. The peaks in this convolution search correspond to plausible locations and orientations of a helix or sheet. These preliminary parameters are then refined and the resulting orientations and positions of helices are sorted according to the final correlation of density with the map. Next, a library of helices or sheets from a database of high-resolution protein structures is oriented using the refined parameters and compared one by one with the density in the map. The ends of each possible fit to the map are trimmed back until only atoms in positive density are left and the longest such fragment obtained for each helix or sheet location is kept.

The second stage in model building is to extend the helices or sheets using a library of tripeptide fragments. The libraries of tripeptides used contain about 10 000 entries and cover about 99% of the tripeptides in a database of 600 proteins within about 0.5 Å r.m.s.d. A tripeptide is oriented by overlapping its first (or last) residue with the last (or first) residue of a helix or sheet. The match to the electron density is then evaluated and the best matches are kept. The process is then repeated by extending the end of the tripeptide further until there is no tripeptide that matches the density. This process results in many overlapping fragments, all containing a helix or sheet at some point and extending in one or both directions from there.

The fragments of backbone structure are then connected using variations on a simple algorithm: a pair is connected if they overlap over at least two C^{α} atoms. This eliminates most incorrect fragments, including most of those tracing the chain backwards, and yields segments of connected polypeptide backbone. In some cases however (particularly at resolutions lower than 3 Å), the chain still can be traced backwards.

The final step in the automated model building carried out by *RESOLVE* is the assignment of side chains and their alignment to the protein sequence. Once main-chain coordinates have been estimated, the expected position of the side chain is known and density in this region can be directly compared with templates of the 20 side chains in their common conformations. *RESOLVE* uses a library of side-chain templates consisting of average electron-density maps for each of the common side-chain conformations. Side-chain assignment is carried out in two steps. Firstly, the electron density at each side-chain position is compared with each of the side-chain templates and the best-matching conformation for each side chain is noted. The correlation coefficient obtained from this match is used to estimate the probability that each side chain is the correct one at this position. The sequence of the protein is then aligned in each possible register with this series of side-chain matches and the most probable alignment is chosen.

5. Conclusions

Automation of macromolecular structure solution has required a substantial investment in software for linking together routines for all the steps in the process and in the development of a reliable decision-making process. The use of a scoring scheme to convert the decision making in macromolecular structure solution to an optimization problem has proven very useful. In many cases, a single clear heavy-atom solution can be obtained and used for phasing. The statistical density-modification procedure is well suited to an automated approach to structure solution because the method is relatively insensitive to choices of solvent content and number of cycles. The detection of NCS in heavy-atom sites and checking of potential NCS operations against the electron-density map has proven to be a reliable method for identification of NCS in most cases. Automated model building beginning with an FFT-based search for helices and sheets has been successful for maps with resolutions as low as 3 Å. Although the *SOLVE* and *RESOLVE* programs are separate program units, the output of *SOLVE* (the structure-factor amplitudes and phases in solve.mtz and heavy-atom sites in ha.pdb) are default inputs to *RESOLVE*, so that the two programs can be run sequentially without any user decisions or input between the two. Together, they provide a fully automatic procedure for structure solution, phase improvement and preliminary model building.

The author is grateful to Joel Berendzen for help in the development of *SOLVE* and for continuing discussions, and to the NIH for generous support.

References

Abrahams, J. P. & Leslie, A. G. W. (1996). *Acta Cryst.* D**52**, 30–42.
Bricogne, G. (1976). *Acta Cryst.* A**32**, 832–847.
Cowtan, K. D. (1998). *Acta Cryst.* D**54**, 750–756.
Cowtan, K. D. & Main, P. (1993). *Acta Cryst.* D**49**, 148–157.
Dickerson, R. E., Kendrew, J. C. & Strandberg, B. E. (1961). *Acta Cryst.* **14**, 1188–1195.
Matthews, B. W. & Czerwinski, E. W. (1975). *Acta Cryst.* A**31**, 480–487.
Plas, J. L. van der & Millane, R. P. (2000). *Proc. SPIE*, **4123**, 249–260.
Rossmann, M. G. (1972). *The Molecular Replacement Method*. New York: Gordon & Breach.
Szöke, A. (1993). *Acta Cryst.* A**49**, 853–866.
Terwilliger, T. C. (1994*a*). *Acta Cryst.* D**50**, 11–16.
Terwilliger, T. C. (1994*b*). *Acta Cryst.* D**50**, 17–23.
Terwilliger, T. C. (1999). *Acta Cryst.* D**55**, 1863–1871.
Terwilliger, T. C. (2000). *Acta Cryst.* D**56**, 965–972.
Terwilliger, T. C. (2001). *Acta Cryst.* D**57**, 1763–1775.
Terwilliger, T. C. & Berendzen, J. (1996). *Acta Cryst.* D**53**, 571–579.
Terwilliger, T. C. & Berendzen, J. (1999*a*). *Acta Cryst.* D**55**, 849–861.
Terwilliger, T. C. & Berendzen, J. (1999*b*). *Acta Cryst.* D**55**, 1872–1877.
Terwilliger, T. C & Eisenberg, D. (1983). *Acta Cryst.* A**39**, 813–817.
Terwilliger, T. C., Kim, S.-H. & Eisenberg, D. (1987). *Acta Cryst.* A**43**, 1–5.
Wang, B.-C. (1985). *Methods Enzymol.* **115**, 90–112.
Wimberly, B. T., Brodersen, D. E., Clemons, W. M., Morgan-Warren, R. J., Carter, A. P., Vonrhein, C., Hartsch, T. & Ramakrishnan, V. (1999). *Nature (London)*, **407**, 332–339.
Xiang, S., Carter, C. W. Jr, Bricogne, G. & Gilmore, C. J. (1993). *Acta Cryst.* D**49**, 193–212.