SOLVE and RESOLVE: automated structure solution, density modification, and model building

Thomas Terwilliger

Los Alamos National Laboratory, USA.
E-mail: terwilliger@lanl.gov

The software SOLVE and RESOLVE can carry out all the steps in macromolecular structure solution, from scaling and heavy-atom location through phasing, density modification, and model-building in the MAD, SAD, and MIR cases. SOLVE uses scoring scheme to convert the decision-making in macromolecular structure solution to an optimization problem. RESOLVE carries out the identification of NCS, density modification, and automated model-building. The procedure is fully automated and can function at resolutions as low as 3 Å.

Keywords: **MIR/MAD; SAD; SAS; software.**

1. Introduction

During the past decade there have been profound changes in the approaches used to carry out macromolecular structure determination in cases where a closely-related structure is not available. An obvious change is in the overall method used. A decade ago the MIR method was dominant, today it is the MAD and SAD methods using synchrotron radiation and often using selenomethionine as a phasing tool (Hendrickson, 2000) that dominate the field. Another change is in the automation of structure solution. A decade ago many new structures were "solved" by inspection of Patterson maps. Today most are solved by automatic interpretation of Patterson maps or by direct methods (Grosse-Kunstleve & Brunger, 1999; Sheldrick, 1998; Terwilliger & Berendzen, 1999; Weeks & Miller, 1999). A third change is the increased use of Bayesian statistical methods for analysis and interpretation of crystallographic data (Bricogne, 1997; Fourme et al., 1999; McCoy, 2002). A fourth change is the introduction of the powerful idea of iterative model-building and refinement as an approach for improving crystallographic phases (Lamzin, 1993; Perrakis et al, 1999). A fifth change is the introduction of automated model-building at moderate resolution (Oldfield, 1997; Levitt 2001; Ioerger and Sacchettini, 2002; Terwilliger, 2001). These changes and many others have had a substantial effect on the field of macromolecular crystallography by making the process of structure determination far easier and faster than it was previously. Even more importantly, these changes have widened the visions of structural biologists, making ideas such as structural genomics potentially feasible.

2. SOLVE-automated structure solution

Automating a process such as structure solution requires several things to be in place. First, each of the component steps have to be worked out. Next, these steps need to be linked together in a seamless way so that the output from one step can be readily used as the input to the next. Finally, a means of making decisions has to be implemented. In the MIR or MAD methods, the key decision to be made consists of the identification of the sites of the heavy or anomalously-scattering atoms in the structure. This really consists of many smaller decisions, such as which of two possible enantiomorphic heavy-atom sets is correct, or whether an additional site is to be included. Before this main decision can be even approached, many smaller decisions, such as the choice of resolution cutoff, the rejection of implausible measurements, and the choice of optimal scaling procedures need to be made.

2.1. Scaling a SAD dataset

In this work the methods used to solve structures by the SAD method will be used as an example of how SOLVE and RESOLVE work. The overall process for structure solution for SAD data using SOLVE has several steps (Terwilliger & Berendzen, 1999). The data consist of measurements of F+ and F- for most or all reflections to a given resolution. The first step is to scale this SAD dataset. Optimally, the original indices of these measurements have been preserved so that local scaling can be applied to minimize systematic errors such as those introduced by absorption. A reference dataset is created by merging all the measurements into the asymmetric unit of the crystal. Then this reference dataset is used to scale all the measurements using their original indices. Finally matched pairs of F+ and F- measurements are identified and the mean F and the anomalous differences Δ_{ANO} are obtained. This procedure is designed to minimize systematic errors by scaling F+ and F- observations of the same reflection to the same reference and by keeping measurements of F+ or F- that are in different regions of reciprocal space separate. If multiple measurements of a given anomalous difference are available, they are averaged.

2.2. Possible solutions to the anomalous difference Patterson function for a SAD dataset

The second overall step for SAD structure solution is to generate a large number (typically 10-30) of plausible 2-site solutions to the anomalous difference Patterson function. This is carried out using the HASSP automated superposition method (Terwilliger, 1987). It might seem that finding 2-site solutions to Patterson functions that may have as many as 60 or 70 sites would not be productive or even possible, but this step is found to be quite reliable, both for structures with just a few sites and for structures with many sites. Once a 2-site solution is found, it is used as the basis for generating additional potential sites using difference Fourier methods (Terwilliger & Berendzen, 1999).

2.3. Scoring a heavy-atom solution

The third overall step is to evaluate and rank the current set of heavy-atom solutions. This is the critical step for automated structure solution for the SAD (or MAD or MIR) methods. In the SOLVE software the scoring of heavy-atom solutions serves as the principal decision-making tool: the solution with the higher score is "better". Of course such an approach requires that the scoring system be reliable. In practice, a scoring system that combines information from several sources can be quite reliable.

The SOLVE scoring system has four components (Terwilliger & Berendzen, 1999). For each component a numerical score is calculated. The first is the quality of the electron density map that is obtained using a particular heavy-atom solution to calculate phases. This criteria is very powerful for identifying the correct hand of the heavy-atom solution and for discriminating between a solution that gives a very good map and one that gives a mediocre map. It is less useful for distinguishing between two solutions that are both very poor. The property used to evaluate the quality of an electron density map is the presence of contiguous regions of relatively flat

solvent and of contiguous regions of protein density which are not at all flat.

The second component of the SOLVE scoring system is the agreement between the anomalous difference Patterson function and the function predicted from the heavy-atom solution. The third component is the cross-validation anomalous difference Fourier. In this component, all the sites except one are used to calculate phases, and these phases are used with the measured anomalous differences to calculate a map that should show peaks at the sites of all anomalously-scattering atoms. The peak height at the position of the omitted site is a measure of the reliability of that site. The fourth component of scoring is simply the figure of merit of the phasing calculation. The figures of merit calculated by SOLVE are relatively unbiased and therefore are a reasonable indication of the actual quality of the phases. Consequently solutions that lead to higher figures of merit are often better than those that lead to lower ones.

All four components of the SOLVE scoring procedure are combined together using a "Z-score" approach. In this approach, the Z-score for a particular component and heavy-atom solution describes how high the numerical score for this solution is, normalized to the scores for all the 2-site solutions that SOLVE considered at the beginning of the structure solution process. The Z-scores for the four components of the SOLVE scoring system are then added together to yield an overall score. Finally, this overall score is corrected to reduce any very large contributions from any one component, by subtracting half the difference between the largest contribution and the average of all the others from the final score.

2.4. Phase calculation for SAD data

SOLVE uses a simple framework for calculating phase probability distributions for SAD data. First the parameters describing the heavy-atom solution are refined using a Patterson-based approach. An origin-removed anomalous difference Patterson function is calculated from the measured anomalous differences. Then the heavy-atom parameters are refined so as to lead to a predicted origin-removed anomalous difference Patterson function that matches the observed one as closely as possible. This method yields unbiased and generally quite accurate estimates of the occupancies and positions of the anomalously-scattering atoms.

Once the heavy atom parameters are refined, the basic phase calculation for SAD data is straightforward: for a particular reflection, the probability of phase ϕ is proportional to the probability of measuring the observed value of the anomalous difference Δ_{ANO} given the best estimates available of the anomalous-scattering part of the heavy-atom structure factors (δ_H^+, δ_H^- ; calculated from the heavy-atom model; see Terwilliger, 1994),) and of the mean structure factor amplitude ($|F| = |F^+ + F^{-*}|/2 \approx [|F^+| + |F^-|]/2$; obtained from the measured data):

$$p(\phi) \propto p(\Delta_{ANO} \mid \delta_H^+, \delta_H^-, |F|, \phi) . \qquad (1)$$

For given values of the anomalous-scattering part of the heavy-atom structure factors (δ_H^+, δ_H^-), and values of the mean structure factor amplitude ($|F|$) and of the phase (ϕ) of F , a value of the anomalous difference Δ^C_{ANO} can be calculated. The probability in Eq. (1) is then just the probability of measuring the value Δ_{ANO} if the true value were Δ^C_{ANO}. For centric reflections, Eq. (1) yields no phase information because the anomalous difference is zero, independent of the phase. It is useful to include a Sim-based phase probability (Sim 1959) in the phase estimate as well so that there is some phase information for centric reflections. In SOLVE this additional phase probability information is of the form,

$$p(\phi) \propto e^{-2w|F||F_H|\cos(\phi - \phi_H)/<F^2>} \qquad (2)$$

where ϕ_H is the phase of the heavy-atom structure factor F_H and w is a weighting factor included in this expression simply to scale the phase probability information from the heavy-atoms structure factor in an approximate way to the phase probabilities from the anomalous differences. Empirically it is found that although the most accurate phases are obtained with $w=1$, this lead to very large peaks at the sites of the heavy-atoms, and the best maps after statistical density modification are obtained with smaller values of w. Typically the weighting factor is set to $w=<m>/2$, where $<m>$ is the mean figure of merit of phasing.

3. RESOLVE-statistical density modification

The initial electron density maps obtained using SAD data are typically not of very high quality due to the inherent ambiguity in the crystallographic phases calculated from anomalous differences alone. Figure 1A shows a section through a SAD electron density map obtained by using the peak wavelength data from a selenomethionine-containing initiation factor 5A from *P. aerophilum* (Peat et al., 1998). The map shows correct features, but is quite noisy.

Statistical density modification is an approach to density modification that maintains independence of different sources of phase information. The fundamental information that is used in density modification procedures (Rossmann, 1972; Bricogne, G. 1976; Wang, 1985; Xiang et al., 1993; Cowtan & Main, 1993; Szoke, 1993; Abrahams & Leslie, 1996; van der Plas & Millane, 1996) is that phases which lead to maps that are plausible are more likely to be correct than phases which lead to implausible maps. In statistical density modification (Terwilliger, 2001), the plausibility of an electron density map is quantified using a "map probability function". In essence, this is a function that has a high value if all the values of electron density in the map are consistent with expectations about the map, and a low value if they are not. For example, if a solvent region can be reliably identified, then if most of the values of density in the solvent region are close to the mean of the density in this region, then the map is plausible, but if many are not, then it is less plausible. Similarly, if the non-crystallographic symmetry is present within a defined envelope and most of the points within this envelope have values of electron density matching the values at NCS-related points, then the map is plausible. The statistical density modification procedure allows the probability of each possible value of each crystallographic phase to be estimated (given the current values of all the other phases). This phase information can then be combined with the experimental phase information to yield an improved electron density map. Figure 1B shows the same region of the IF5A crystal structure as Fig. 1A, except that statistical density modification has been applied. This structure has a solvent content of about 60% but no non-crystallographic symmetry. The density-modified SAD map is considerably improved over the original SAD-phased map.

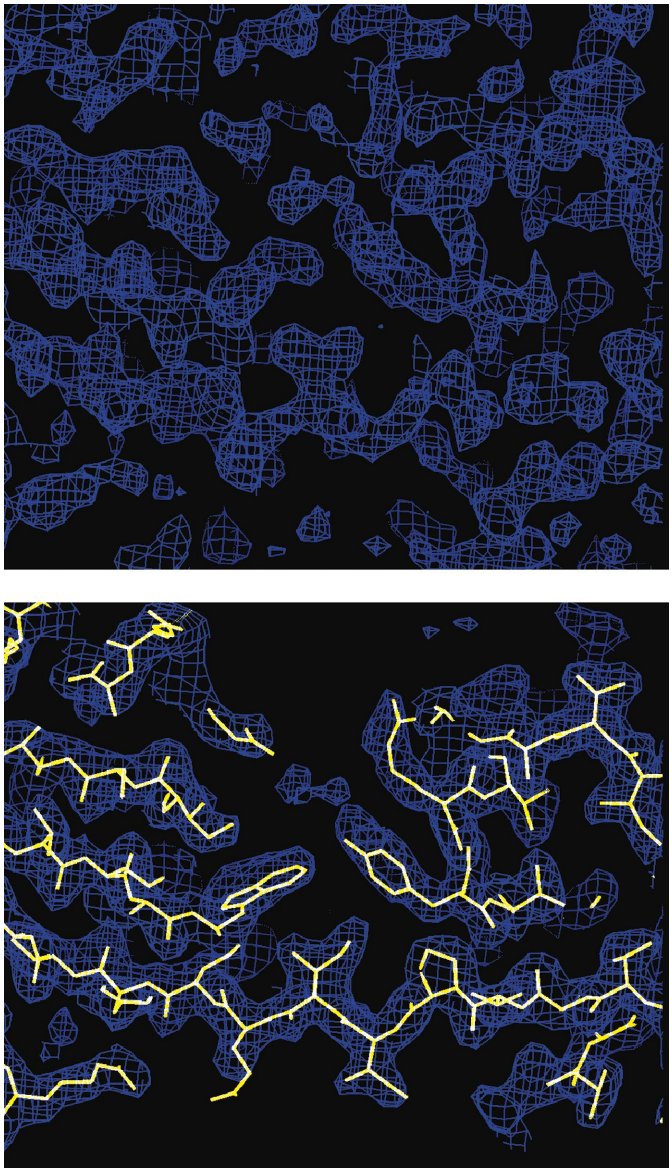

Figure 1

SOLVE and RESOLVE electron density maps and model using SAD data from initiation factor 5A. Top: SAD SOLVE electron density map. Bottom: RESOLVE density-modified electron density map, with superimposed RESOLVE model.

4. RESOLVE-automated model-building

Automated model-building is carried out in the RESOLVE software using a sequential process (Terwilliger, 2002a; Terwilliger, 2002b). In the first stage, helices and strands are identified by matching templates to the density in a map. In the next stage, fragments of helices or strands from a library built from refined protein structures are matched to this density and extended in both directions using tripeptide fragment libraries. In the third stage, side chains are identified, once again using libraries from refined protein structures. In the final stage, the molecule is assembled, making use of non-crystallographic symmetry if available.

4.1. Identification of helices and strands

RESOLVE uses an FFT-based procedure to identify helices and strands in an electron density map. A template for helices (6 amino acids) and a similar template 4 amino acids long for strands were constructed. Then to identify helices or strands in a map, many rotations of each template are carried out, and for each rotation an FFT-based convolution search was carried out to identify locations where the density in the map was correlated with the density in the template. These positions and orientations are refined to maximize this correlation, resulting in a sorted list of positions and orientations of helices and of strands.

4.2. Matching and extension with fragment libraries

The positions and orientations of helical and strand fragments are used as a starting point for placing fragments of structure from refined proteins into the electron density. Each member of a set of 17 helical fragments from 6 to 24 residues long is compared with the electron density using the position and orientation identified in the FFT-based search. A similar procedure is applied for strand fragments, using a library of 17 strands from 4 to 9 amino acids long. Each segment is scored based on the mean density at the coordinates of main-chain atoms in the segment and its length, and the segment with the highest score is retained.

These helices and strands are then extended in both directions, this time using libraries of 3-amino acid fragments derived from refined protein structures. To extend, the first amino acid of a fragment to be tested is superimposed on the last amino acid that has already been placed, and the density at the coordinates of all the other atoms in the test fragment is examined. RESOLVE uses a look-ahead procedure for extending the main chain: the score for a test fragment is a combination of the density at the coordinates of atoms in the fragment, and the density at the coordinates of the best fragment that can be added on to this fragment. In this way, a fragment will not usually be added unless it can be extended again.

This process of identifying helices and strands and extending them leads to many overlapping fragments of main-chain. To build a single "best" main-chain, an iterative procedure is used. The longest chain is identified. Then this chain is extended using whichever chain leads to the longest extension. The process is repeated until that chain cannot be extended further, and all chains that occupy the same space as the growing main chain are eliminated. This process is then repeated starting with the next longest remaining chain until no more chains are available.

4.3. Side-chain identification

Once the main chain has been built, the possible locations of side chains are partly determined, but there are several possibilities for the orientations for most side chains (Ponder, 1987). RESOLVE uses a library of side chain templates to match side chain density in a map with side chain types and rotamers in a probabilistic fashion. This approach yields probabilities for each side chain type at each position in the main chain model. The amino acid sequence can then be compared with these probabilities and an alignment found that maximizes the correspondence between the known sequence and the side chain probabilities at each position. In some cases this analysis can identify errors in the main chain model in which the wrong number of amino acids is present in a loop. These cases can be identified because an alignment will be found for the side chains in two adjacent sections of main chain, but the alignment one group of side chains will be different than the alignment for the other. RESOLVE will then break the main chain between the two alignments.

4.4. Molecular assembly

The model-building procedure described above leads normally to several chains, most assigned to the amino acid sequence of the protein, but located in arbitrary asymmetric units of the crystal. It is most useful to have a compact (but crystallographically equivalent) version of the molecule for purposes of interpretation. The assembly of fragments of a molecular model into a compact model is accomplished in RESOLVE using a scoring procedure to evaluate possible assemblies. The scoring includes non-crystallographic symmetry, if present, the compactness of the entire assembly, and the plausibility of distances between the end of one chain and the beginning of the next, given the number of amino acids separating them in the sequence. RESOLVE first brings all chains as close together as possible, then tries to increase the overall score of the assembly by iteratively taking one chain from the assembly and placing it in all plausible and crystallographically available locations. This procedure normally yields an assembly with the correct non-crystallographic symmetry and a high degree of compactness and connectivity.

5. Conclusions and prospects

In cases where high-quality MAD, MIR, or SAD crystallographic data are available, the SOLVE and RESOLVE software can very often carry out the entire process of structure solution, phase calculation, density modification, non-crytstallographic symmetry identification, model-building, and molecular assembly in a completely automated fashion. At present the models obtained are preliminary models, requiring both further building by an expert crystallographer and identification and correction of errors. Recently scripts have been developed for RESOLVE model building that allow iteration of the model-building, refinement and density modification process, resulting in more complete models. It seems possible that over the next few years iterative model-building, refinement and density modification, combined with more thorough model-building in loop regions and with identification of ligands bound to protein, could lead to nearly-complete models.

One important outcome of automation of the structure determination process is that it speeds up the process of structure solution. Another is that it allows experienced crystallographers to test many ideas about how to solve a particularly difficult structure. Perhaps the most important outcome is still in its infancy. This is that automation can allow far more systematic error checking and error analysis than can be done manually. This may ultimately provide a sound basis for error analysis in the interpretation of protein structures.

The author would like to thank Joel Berendzen and Li-Wei Hung and the members of the PHENIX consortium for many discussions and ideas during the development of SOLVE and RESOLVE, the NIH for support, and many SOLVE and RESOLVE users for valuable feedback on the software and ideas for future developments.

References

Abrahams, J. P. & Leslie, A. G. W. (1996) *Acta Cryst*. D52, 30-42.
Bricogne, G. (1976) *Acta Cryst.* A32, 832-847.
Bricogne, G. (1997). *Methods Enzymol.* 276, 361-423.
Cowtan, K. D. & Main, P. (1993) *Acta Cryst.* D49, 148-157.
Fourme, R., Shepard, W., Schilta, M., Prange, T., Ramin, M., Kahn, R., de la Fortelle, E. & Bricogne, G. (1999) *J. Synchrotron Rad.* 6, 834-844 .
Grosse-Kunstleve, R. W. &A.T. Brunger, A. T. (1999) *Acta Cryst*. D55, 1568-1577.
Hendrickson, W. A. (2000). Trends Biochem. Sci. 25, 647-653.
Ioerger, T. R. & Sacchettini, J. C. (2002). *Acta Cryst.* D58, 2043-2054.
Lamzin, V. S. & Wilson, K. S. (1993) *Acta Cryst* D49, 129-147.
Levitt, D.G. (2001). *Acta Cryst*. D57, 1013-1019.
McCoy, A. J. (2002). *Curr. Opinion Struct. Biol*. 12, 670-673.
Oldfield, T. (2002). Acta Cryst. D58, 487-493.
Peat T. S., Newman J., Waldo G. S, Berendzen J., Terwilliger T. C. (1998). *Structure* 6, 1207-1214.
Perrakis, A., Morris, R.M. and Lamzin, V.S. (1999) *Nature Structural Biology* 6, 458-463.
Ponder, J. W. & Richards, F. M. (1987). J. Mol. Biol. 193, 775-791.
Rossmann, M. G. (1972) *The molecular replacement method.* New York, Gordon & Breach.
Sheldrick, G. M. *in* "Direct Methods for Solving Macromolecular Structures" (S. Fortier, ed.), p. 401. Kluwer Academic Publishers, Dordrecht, 1998.
Sim, G. A. (1959). Acta Cryst. 12, 813-815.
Szoke, A. (1993) *Acta Cryst.* A49, 853-866.
Terwilliger, T. C. & . Berendzen, J. (1999) *Acta Cryst*. D55, 849-861.
Terwilliger, T. C., Kim, S.-H., and D. Eisenberg. (1987) *Acta Cryst.*. A43, 1-5
Terwilliger, T. C. (1994). *Acta Cryst.* D50, 17-23.
Terwilliger, T. C. (2000). Acta Cryst. D56, 965-972.
Terwilliger, T. C. (2001). *Acta Cryst.* D57, 1755-1762.
Terwilliger, T. C. (2002a) *Acta Cryst.* D59, 34-44.
Terwilliger, T. C. (2002b) *Acta Cryst* D59, 45-49.
van der Plas, J. L & Millane, R. P. (2000) *Proceedings of SPIEE* 4123, 249-260.
Wang, B.-C. (1985) *Methods Enzymol.* 115, 90-112.
Weeks, C. M. & Miller, R. (1999) *Acta Crystallogr*. D55, 492-500.
Xiang, S., Carter, C. W., Jr., Bricogne, G. & Gilmore, C. J. (1993) *Acta Cryst.* D49, 193-212.