Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

An evaluation of automated model-building procedures for protein crystallography

John Badger

Structural GenomiX, Inc., 10505 Roselle Street, San Diego, CA 92121, USA

Correspondence e-mail:
john_badger@stromix.com

Received 22 January 2003
Accepted 20 February 2003

The computer programs *ARP*/*wARP*, *MAID* and *RESOLVE* are designed to build protein structures into experimentally phased electron-density maps without any user intervention, requiring only diffraction data and sequence information. However, the *MAID* and *RESOLVE* systems, which seek to extend the range of automated model-building to ∼3 Å resolution, have yet to receive significant testing outside the small numbers of data sets used in their development. Since these two systems employ a large number of scoring functions and decision-making heuristics, additional tests are required to establish their usefulness to the crystallographic community. To independently evaluate these programs, their performance was tested using a database containing 41 experimentally phased maps between 1.3 and 2.9 Å resolution from a diverse set of protein structures. At resolutions higher than 2.3 Å the most successful program was *ARP*/*wARP* 6.0, which accurately built an average of 90% of the main chain. This system builds somewhat larger fractions of the model than the previous version *ARP*/*wARP* 5.1, which accurately built an average of 87% of the main chain. Although not specifically designed for model building into high-resolution maps, *MAID* and *RESOLVE* were also quite successful in this resolution regime, typically building ∼80% of the main chain. At 2.3–2.7 Å resolution the *MAID* and *RESOLVE* programs automatically built ∼75% of the main-chain atoms in the protein structures used in these tests, which would significantly accelerate the model-building process. Data sets at lower resolution proved more problematic for these programs, although many of the secondary-structure elements were correctly identified and fitted.

1. Introduction

Recent years have seen the widespread adoption of the *ARP*/*wARP* 5.1 software (Perrakis *et al.*, 1999) to trace large portions of protein structures from experimentally phased maps, subject to the limitation that the diffraction data extend to better than ∼2.3 Å resolution. More recently, two new systems have appeared, *MAID* (Levitt, 2001) and *RESOLVE* (Terwilliger, 2001), which are intended to provide significant automated model-building capabilities down to ∼3 Å resolution. In addition, version 6.0 of *ARP*/*wARP* seeks to extend the range of applicability of the software to ∼2.5 Å resolution. Successful automation of the initial model-building step in the crystal structure determination process would be a very significant advance. For example, ∼39% of structures in the Protein Data Bank (Berman *et al.*, 2000) were solved using data below 2.3 Å resolution and these structures include many of the largest and most biologically interesting structures. Despite the development of efficient GUI-driven model-building tools within the framework of computer graphics programs (for example, Jones *et al.*, 1991; McRee, 1999; Turk, 2001; Oldfield, 2002), full automation of protein model building at moderate resolution represents the critical challenge to high-throughput structure determination since interactive model building of large structures can be both time-consuming and error-prone.

The *ARP*/*wARP*, *MAID* and *RESOLVE* programs use distinctly different algorithms for automated model building. The *ARP*/*wARP* software uses an iterative peak-picking approach in which dummy atoms are placed in or removed from electron-density maps and the resulting model subjected to conventional crystallographic refinement (Perrakis *et al.*, 1999). By detecting patterns in the sets of atomic sites that resemble polypeptide acid chains, dummy atoms are subsequently replaced by polyalanine fragments (Morris *et al.*, 2002). Cycles of refinement of the mixed polypeptide/dummy atom model, with density-map calculations, peak-picking and polypeptide-chain detection are continued until as complete a main-chain model as possible has been built. The general approach encoded in the *MAID* program (Levitt, 2001) is to automate some of the model-building strategies employed by humans when facing a new electron-density map. Following a map-skeletonization procedure, regular α-helical and β-sheet structures are identified and placed. The program then attempts to identify the amino-acid sequence based on fitting side-chain densities. These fits are subsequently extended by attempting to add residues from a library of allowed main-chain φ–ψ values that fit the electron density and constraints imposed by connectivity with the previously fitted structure. The *MAID* model-building process does not include conventional crystallographic refinement, but employs simulated-annealing real-space torsion dynamics over small structural regions to remove stereochemical anomalies, assist in side-chain placement and optimize backbone conformations. The *RESOLVE* system (Terwilliger, 2001) uses an FFT-based convolution approach to recognize and place secondary-structural elements in the electron-density map. Subsequent main-chain extension, sequence alignment and side-chain placement employs matching features in the map to libraries of fragments and densities with steps to assemble the fragments into a single model.

An important attribute of the *ARP*/*wARP*, *MAID* and *RESOLVE* systems is that they may be run in genuinely automated modes, *i.e.* they may be driven by simple scripts requiring minimal user inputs (sets of experimental phases and sequence information), and produce relatively complete atomic models as output. Thus, these systems could provide the technology needed for scaleable high-throughput structure-determination platforms capable of producing large numbers of new protein structures across the entire resolution range. Furthermore, these fully automated procedures are much more amenable to consistent testing and protocol development than their GUI-driven counterparts since the capabilities of fully automated systems can be repeatedly and systematically evaluated using crystallographic databases containing large quantities of structure data. This approach to evaluation and development of software for protein crystallography does not appear to been taken in any previous studies, presumably because the Protein Data Bank (Berman *et al.*, 2000) contains almost no diffraction data files that record experimental phase information and individual structure-determination laboratories do not usually capture this data in accessible databases with standardized annotation and formats (Badger, 2001). Version 6.0 of the *ARP*/*wARP* program, which contains a new protein-tracing algorithm (Morris *et al.*, 2002) and utilizes likelihood-weighted refinement technology (Murshudov *et al.*, 1997), and the recently developed *MAID* (Levitt, 2001) and *RESOLVE* (Terwilliger, 2001) model-building systems have yet to be subjected to many tests outside of the training data used by their developers. Since all of these programs encode complex sets of decision-making criteria, many additional tests with structure data outside these training examples are needed to establish their usefulness to the crystallographic community. The studies described in this paper independently evaluate the capabilities and limitations of these three automated model-building systems using a relatively large and diverse test set of experimentally phased maps.

2. Methods

2.1. Test data

The automated model-building systems were evaluated using a database of 41 phase sets that had been previously obtained by SAD, MAD and SIRAS SeMet phase determination techniques. This database constitutes the complete corpus of phased diffraction data for structures solved during the course of a novel bacterial structure-determination program extending to December 2001, *i.e.* these are genuinely 'real-world' examples without any pre-selection to exclude data sets that might be unfavorable for automated model-building trials. The experimental phasing procedures had used *SHARP* (de La Fortelle & Bricogne, 1997) or *CCP*4/*MLPHARE* (Otwinowski, 1991) for Se-site refinement with subsequent density modification using either *CCP*4/*SOLOMON* (Abrahams & Leslie, 1996) or *CCP*4/*DM* (Cowtan, 1994). The resulting 41 phase sets span the resolution range 1.3–2.9 Å (Fig. 1), with phase differences from the final refined models lying in the range 28.9–68.0° and a mean phase difference of 42.3° (Fig. 2). The refined models contained between 101 and 1234 amino acids in the crystal asymmetric unit. The models span a wide range of structure types, including predominantly α-helical structures and predominantly β-sheet structures as well as mixed α/β structures.

2.2. Automated model building

Each of the programs *ARP*/*wARP* 5.1, *ARP*/*wARP* 6.0, *MAID*, *RESOLVE* 2.02 and *RESOLVE* 2.04 were run with default settings *via* automated scripts on a Linux computer cluster. The *MAID* and *RESOLVE* programs were run on maps computed from all 41 data sets and *ARP*/*wARP* was run on the 28 test cases for which the resolution exceeded 2.3 Å. The run times for all of these programs were typically several hours. The only input information provided to these programs was the set of experimental phases and either a sequence file (for *MAID* and *RESOLVE*) or the expected number of amino acids in the crystal asymmetric unit (for *ARP*/*wARP*). Although several of these test structures contained more than one copy of the molecule in the asymmetric unit, no knowledge of non-crystallographic symmetry was employed by any of programs in these trials. For calculations with *ARP*/*wARP*, 200 cycles were run with model reconstruction every ten cycles. The calculations with *ARP*/*wARP* 5.1 used the 'F protocol', which employs least-squares refinement of the atomic model with *CCP*4/*REFMAC*4. Calculations with *ARP*/*wARP* 6.0 employed the script-driven 'R protocol' with the new main-chain tracing algorithm (Morris *et al.*, 2002) and maximum-likelihood structure refinement using *CCP*4/*REFMAC*5 (Murshudov *et al.*, 1997). For the calculations with *MAID*, the density σ-cutoff of 1.2 was used for the initial skeletonization and the final structures were screened to ensure that only a unique set of molecular fragments (*i.e.* there were no fragments repeated through application of crystallographic symmetry) remained in the final coordinate file. For calculations with *RESOLVE* the correct number of copies of the protein was provided to the program.

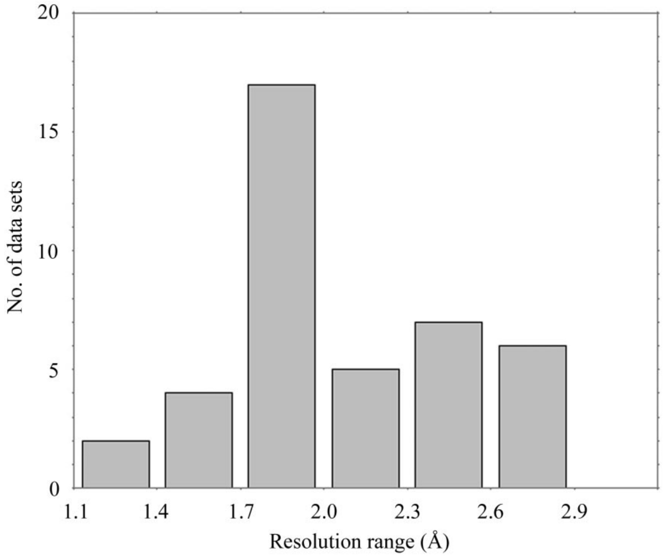

Figure 1
Distribution of resolution for the 41 phased data sets used in automated model-building tests.

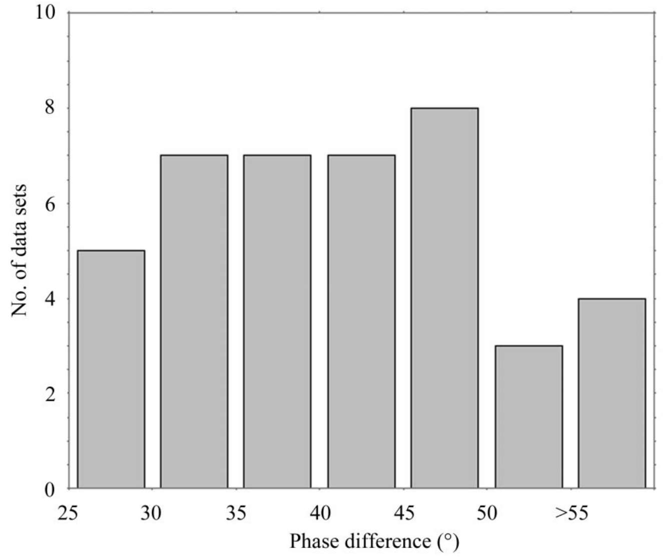

Figure 2
Distribution of phase differences between experimental and model phases for the 41 phased data sets used in automated model-building tests.

Table 1
Mean fractions of main-chain traces built and correctly built for the programs tested at high and medium–low resolution.

The fractions built are calculated using the number of amino acids observed in the final refined models. A 'correctly built' amino acid is an amino acid in which a CA atom is less than 1 Å from a CA position in the final refined model.

Program	Resolution (Å)	Fraction built	Fraction correctly built
ARP/*wARP* 5.1	1.3–2.3	0.88	0.87
ARP/*wARP* 6.0	1.3–2.3	0.91	0.90
MAID	1.3–2.3	0.84	0.82
RESOLVE 2.02	1.3–2.3	0.79	0.77
RESOLVE 2.04	1.3–2.3	0.79	0.77
MAID	2.3–2.9	0.72	0.60
RESOLVE 2.02	2.3–2.9	0.67	0.55
RESOLVE 2.04	2.3–2.9	0.66	0.57

To evaluate the quality and completeness of the main-chain traces produced by these programs, statistics of the total numbers of amino acids built and the numbers of CA atoms placed within 1 Å of a correct CA position were collected from each run. Although somewhat arbitrary, this cutoff was chosen because it has been used in assessments by other authors (for example, Ioerger & Sacchetti, 2002) and because it should usually be possible to refine or refit a model built with this degree of accuracy by currently available crystallographic methods. Although all three model-building systems used in the tests described in this report construct complete main-chain traces, powerful energy-optimization procedures (Correa, 1990) would probably also be able to reconstruct complete main-chain traces from these sets of CA positions.

3. Results and discussion

3.1. Automated model building at high resolution (d < 2.3 Å)

The results of these tests (Tables 1 and 2) show that for the types of high-resolution maps that can be routinely obtained by SeMet phasing, refined by effective density-modification procedures, the available automated model-building procedures should be expected to provide over 90% of the main chain of the final model in most cases and over 75% of the main chain of the final model in almost all cases.

These results demonstrate that the current version of *ARP*/*wARP* (version 6.0) usually builds somewhat more of the model than *ARP*/*wARP* 5.1 and is the most successful

Table 2
Distribution of numbers of automatically built models according to the fraction of the model chain correctly built as a function of the program used to build the model.

The results are in two batches, corresponding to high and medium–low resolution. The fractions built are calculated using the number of amino acids observed in the final refined models. A 'correctly built' amino acid is an amino acid in which a CA atom is less than 1 Å from a CA position in the final refined model.

		Fraction correctly built			
Program	Resolution (Å)	1.00–0.90	0.90–0.75	0.75–0.50	0.50–0.00
ARP/*wARP* 5.1	1.3–2.3	14	9	4	0
ARP/*wARP* 6.0	1.3–2.3	18	8	1	0
MAID	1.3–2.3	10	12	6	0
RESOLVE 2.02	1.3–2.3	2	13	12	0
RESOLVE 2.04	1.3–2.3	3	12	12	0
MAID	2.3–2.9	1	3	5	4
RESOLVE 2.02	2.3–2.9	0	1	7	5
RESOLVE 2.04	2.3–2.9	0	1	8	4

program overall. It is worth noting that *ARP*/*wARP*'s superior results in terms of model accuracy are not severely biased by the inclusion of refinement runs within the model-building cycles; in fact, the number of correctly placed CAs by *ARP*/*wARP* is usually greater than the total numbers of CAs placed by the *MAID* and *RESOLVE*.

Although the main intention behind the development of *MAID* and *RESOLVE* systems is to provide model-building methods that will be useful at medium to low resolution, these programs do work effectively in the high-resolution regime, building the majority of the protein structure in all cases.

None of these software systems created significant regions of inaccurate or erroneous structure when used with high-resolution diffraction data; the amount of incorrect structure that was built corresponded to just 1–2% of the total chain trace (Table 1).

3.2. Automated model building at medium–low resolution ($2.3 < d < 2.9$ Å)

In the medium–low resolution regime both the *MAID* and *RESOLVE* systems provide similar proportions of correct structure (Tables 1 and 2), building ~66% of the main chain in most cases. Since the model-building results at medium–low resolution are much less homogenous than results at high resolution, they are tabulated in more detail in Table 3. If the examples for which the resolution is lower than 2.7 Å, as well as the example with the 68° phase difference (essentially a low-resolution map, where the phases are very poorly determined at medium-high resolution), are excluded from consideration, a much greater degree of model-building success is realised. In the remaining nine test cases, *MAID* built 57–90% of the main chain accurately and *RESOLVE* 2.04 built 51–82% of the main chain accurately. In the medium–low resolution regime there is clearly a significant dependence in the quality of the results on the precise resolution of the data and accuracy of the experimental phase set.

Table 3
Detailed breakdown of automated model-building results for 13 structures at medium–low resolution.

For each phased data set the results are expressed for each program as the fraction of model correctly built/fraction of model built.

Resolution (Å)	Phase difference (°)	*MAID*	*RESOLVE* 2.02	*RESOLVE* 2.04
2.35	35.6	0.72/0.75	0.83/0.89	0.82/0.87
2.35	68.1	0.39/0.52	0.39/0.49	0.41/0.50
2.36	35.0	0.78/0.86	0.67/0.73	0.73/0.75
2.40	59.2	0.64/0.77	0.58/0.68	0.57/0.66
2.46	44.0	0.77/0.85	0.63/0.72	0.61/0.68
2.50	45.7	0.90/0.91	0.60/0.64	0.61/0.64
2.60	40.4	0.67/0.86	0.65/0.75	0.65/0.75
2.65	38.3	0.75/0.83	0.65/0.71	0.72/0.79
2.70	38.7	0.57/0.64	0.48/0.66	0.51/0.65
2.70	36.1	0.74/0.78	0.72/0.76	0.72/0.76
2.71	61.8	0.28/0.67	0.28/0.62	0.32/0.63
2.80	45.0	0.47/0.66	0.39/0.67	0.40/0.55
2.90	54.3	0.14/0.27	0.20/0.38	0.22/0.35

Although use of *ARP*/*wARP* 5.1 in *ab initio* model-building applications was generally considered to be restricted to data extending to better than 2.3 Å, this limitation appears to be relaxed with *ARP*/*wARP* 6.0. Model-building trials with the six structures between 2.3 and 2.5 Å (Table 3) using the protocol described in §2.2 yielded significant success (87 and 76% of the main chain) for the first and third structures on the list. Both of these structures were obtained from initial experimental maps with low phase errors. Almost none of the main-chain residues were located in the other four cases. Automated model building with *ARP*/*wARP* 6.0 would clearly be worth attempting for structures in this resolution regime. Alternative *ARP*/*wARP* protocols might also be more successful in this resolution range.

The main result from this study is that for medium-resolution maps (2.3–2.7 Å) of good but not unusually high quality it should usually be possible to obtain ~75% of the structure prior to interactive model building. Although both the *MAID* and *RESOLVE* procedures begin by identifying and fitting sections of regular secondary-structure fragments and then proceed by building out from these fragments, they use completely different algorithms for these purposes. Another emerging system, *CAPRA* (Ioerger & Sacchettini, 2002), uses pattern-recognition techniques to build CA traces and may also provide comparable results when full main-chain building techniques are implemented. Given the early stage of development of all of these systems, it would seem likely that one or all of them will be capable of further improvement in the near future. Nevertheless, a very significant challenge is posed by the next step in the development of these systems: completing the structure in map regions where the electron density is too poor for the current automated procedures to make reliable choices. As currently coded, these programs appear to rely on reasonably unambiguous electron density and simple stereochemical concepts (the Ramachandran plot, secondary-structure identification and dimensions of protein atomic groups) to build structure with enough certainty to maintain acceptably low error rates (Table 1). For example, some test calculations with calculated data from final refined models (*i.e.* structures completed by interactive model building) show that parts of electron-density maps corresponding to loop and random coil regions of the structure are eroded in *ARP*/*wARP* runs if the atoms are poorly resolved and are often not built in map interpretations by *MAID*. More sophisticated sets of stereochemical rules will be needed to interpret map regions where the electron-density data is unclear.

A related question is to ask whether the proportions of the models built by *MAID* and *RESOLVE* could be increased by computing new maps from the initial partial model and repeating the model building, analogous to the recycling procedures carried out by the *ARP*/*wARP* system. This procedure might be expected to be useful if the phase error in the initial partial model was less than the phase error in the experimental map. In a sample calculation with the first structure from Table 3, the phases calculated from the partial model built by *MAID* (75% correctly traced) were found to be almost exactly equidistant (49° in both cases) from the experimental map and the final refined model. Building a new model, a likelihood-weight map calculated from this partial model gave a model with 74% of main-chain residues built, which is a marginally lower completeness than that of the initial partial model built from the experimental map. If the initial model built by *MAID* had greater phasing power (*i.e.* it were either more accurate or more complete), it might be possible to iteratively develop larger models by automated procedures.

Besides their obvious utility for providing large partial models, these automated model-building procedures also provide a direct operational test of the usefulness of experimentally phased maps for structure determination. In the 2–3 cases where automated model-building procedures failed to produce a useful result, visual inspections of the maps at the outset of the structure determination indicated that full structure interpretation would be problematic and might not succeed. It seems likely that if these automated procedures fail with phased data in the high–medium resolution range, a crystallographer using interactive model-building tools will also find it difficult or impossible to provide a successful structure determination.

I wish to thank the developers of the *MAID* and *RESOLVE* systems, David Levitt and Tom Terwilliger, for providing access to software and comments on the results, Tassos Perrakis, Victor Lamzin and Clemens Vonrhein for information regarding the range of applicability of the *ARP*/*wARP* program, Jorg Hendle for maintaining the crystallographic database at Structural GenomiX, Inc., Chuck Kissinger and Tom Peat for support and discussion, and the protein crystallography group at SGX for providing the structures on which this work depended.

References

Abrahams, J. P. & Leslie, A. G. W. (1996). *Acta Cryst.* D**52**, 30–42.

Badger, J. (2001). *CCP4 Newsl. Protein Crystallogr.* **39**. http://www.ccp4.ac.uk/newsletter39/05_sgx.html.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N. & Bourne, P. E. (2000). *Nucleic Acids Res.* **28**, 235–242.

Correa, P. E. (1990). *Proteins Struct. Funct. Genet.* **7**, 366–377.

Cowtan, K. (1994). *Jnt CCP4/ESF–EACBM Newsl. Protein Crystallogr.* **31**, 34–38.

Ioerger, T. R. & Sacchettini, J. C. (2002). *Acta Cryst.* D**58**, 2043–2054.

Jones, T. A., Zou, J.-Y., Cowan, S.-W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.

La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–494.

Levitt, D. G. (2001). *Acta Cryst.* D**57**, 1013–1019.

McRee, D. E. (1999). *J. Struct. Biol.* **125**, 156–165.

Morris, R. J., Perrakis, A. & Lamzin, V. S. (2002). *Acta Cryst.* D**58**, 968–975.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.

Oldfield, T. (2002). *Acta Cryst.* D**58**, 487–493.

Otwinowski, Z. (1991). *Proceedings of the CCP4 Study Weekend. Isomorphous Replacement and Anomalous Scattering*, edited by W. Wolf, P. R. Evans & A. G. W. Leslie, pp. 80–86. Warrington: Daresbury Laboratory.

Perrakis, A., Morris, R. J. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.

Terwilliger, T. C. (2001). *Acta Cryst.* D**57**, 1755–1762.

Turk, D. (2001). *Methods in Macromolecular Crystallography*, edited by D. Turk & L. Johnson, pp. 148–158. Amsterdam: IOS Press.