Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

ARP/wARP and molecular replacement

Anastassis Perrakis,[a]* Maria Harkiolaki,[b] Keith S. Wilson[b] and Victor S. Lamzin[c]

[a]Department of Molecular Carcinogenesis, Plesmanlaan 121, 1066 CX Amsterdam, The Netherlands, [b]York Structural Biology Laboratory, Department of Chemistry, University of York, York YO10 5DD, England, and [c]EMBL, c/o DESY, Notkestrasse 85, 22603 Hamburg, Germany

Correspondence e-mail: perrakis@nki.nl

Received 15 August 2001
Accepted 21 August 2001

The aim of *ARP/wARP* is improved automation of model building and refinement in macromolecular crystallography. Once a molecular-replacement solution has been obtained, it is often tedious to refine and rebuild the initial (search) model. *ARP/wARP* offers three options to automate that task to varying extents: (i) autobuilding of a completely new model based on phases calculated from the molecular-replacement solution, (ii) updating of the initial model by atom addition and deletion to obtain an improved map and (iii) docking of a structure onto a new (or mutated) sequence, followed by rebuilding and refining the side chains in real space. A few examples are presented where *ARP/wARP* made a considerable difference in the speed of structure solution and/or made possible refinement of otherwise difficult or uninterpretable maps. The resolution range allowing complete autobuilding of protein structures is currently 2.0 Å, but for map improvement considerable advances over more conventional refinement techniques are evident even at 3.2 Å spacing.

1. Introduction

The *ARP/wARP* (Lamzin *et al.*, 1999) software suite is designed for automated model building and refinement, with the aim of delivering essentially complete models starting from electron-density maps alone (Perrakis, Morris *et al.*, 1999) or from partial models. In this manuscript, we will deal with its applicability to the refinement of models that have been correctly placed in the cell by molecular replacement.

Over the last few years, considerable efforts have been made in the development of automated model-building proecedures. Most notable is that encoded in the commercial software (Molecular Simulation Inc.) *QUANTA* (Oldfield, 1996), which automates the procedures to a considerable extent, but still requires interactive decision making by the user. The algorithms implemented in the software *O* (Jones *et al.*, 1991) also facilitate automation, but still rely on interactive decision making in front of a graphics workstation. Similar tools are implemented in the popular *XtalView* (McRee, 1999) package, while programs such as *TURBO-FRODO* (Roussel & Cambillau, 1991) and *MAIN* (Turk, 1992) also provide alternative methods. Other approaches aim to identify secondary-structure features in electron-density maps, with the primary goal of assisting the user in the decision-making process. The *ESSENS* (Kleywegt & Jones, 1997) software was the first attempt, while a more efficient implementation of the same principle (utilizing fast Fourier transformations instead

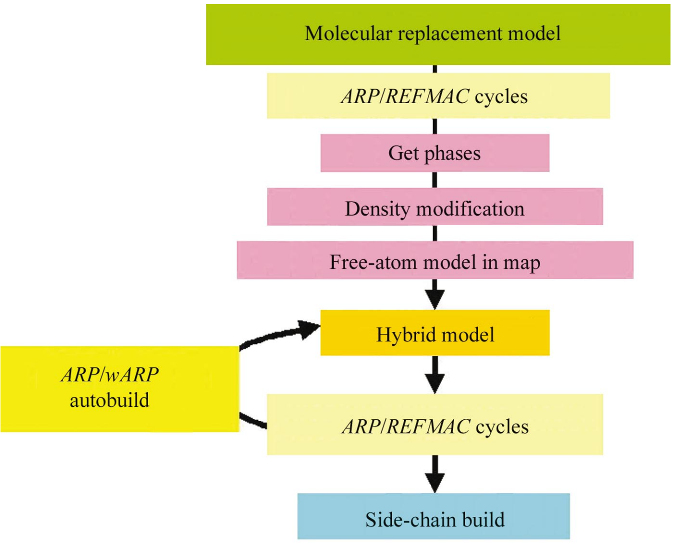

Figure 1
The generalized *ARP*/*wARP* flow scheme in the case of molecular replacement.

of real-space searches) is implemented in *FFFEAR* (Cowtan, 1998). The increasing number of publications on the use of databases for the positioning of structural modules and polypeptide chain fragments into the electron-density map, such as *DADI* (Diller *et al.*, 1999), *TEXAL* (Holton *et al.*, 2000), *ConfMatch* (Wang, 2000) and *MAID* (Levitt, 2001) reflects additional advances in this field.

The basic concept differentiating and underlying the *ARP*/*wARP* package is the unified view of the iterative building and refinement steps. Model building is, in this sense, an integral process that alternates between real-space model adjustment and reciprocal-space parameter refinement. To implement this principle, novel concepts have been introduced allowing the macromolecular model itself, not only the parameters of the model, to change *on the fly*.

Thus, while other approaches keep the model fixed and only adjust its parameters, *ARP* allows the model to change both in the number and position of atoms (Lamzin & Wilson, 1993), according to the density. The key to this is the use of *hybrid models*: 'free' atoms can be used to describe almost every feature of an electron-density map, but this interpretation rarely resembles a conventional conception of a protein. Nevertheless, parts of the free-atom model can be automatically recognized as elements of a protein and by applying model-building algorithms an atomic model can be constructed for the protein. Combination of this partial macromolecular model with a free-atom set (a hybrid model) allows a considerably better overall description of the current map. The model provides additional information (in the form of stereochemical restraints), while free atoms describe prominent features in the electron density (unaccounted for by the current model).

These features allow *ARP*/*wARP* to represent experimental electron-density maps as sets of free atoms, which by iteration of model building and refinement converge to hybrid models which contain most of the protein model automatically built. Here, we describe the applicability of *ARP*/*wARP* to what at first sight is a different case; namely, the refinement and rebuilding of models from molecular-replacement solutions. Although this might be presumed to be more simple (because a model is already available), model bias and other considerations in practice challenge the capabilities and stretch the limitations of the *ARP*/*wARP* software.

2. *ARP*/*wARP* protocols for refining and autobuilding models from molecular-replacement solutions

The protocols presume that you have obtained a molecular-replacement solution with an associated PDB file. This model, together with the native diffraction data, is the mandatory input for *ARP*/*wARP*. The general flowchart in which these data can be utilized is depicted in Fig. 1, while Fig. 2 depicts detailed protocols appropriate to more specific subsets of problems as demonstrated by the subsequent examples.

2.1. Automatically building a new model – *warpNtrace*

This mode requires the native diffraction data to extend to a spacing better than 2.0 Å. If the resolution is between 2.0 and 2.3 Å and the starting model is good and/or the solvent content is high, then this approach is worth a try. If the model is particularly bad (*i.e.* it was very difficult or even unexpected to find a molecular-replacement solution) or incomplete (less than 2/3 of the final model), data to a resolution higher than 2.0 Å might be necessary.

2.1.1. Direct use of the molecular-replacement model in *warpNtrace* cycles. The available model is here fed directly into the *warpNtrace* procedure. A new map is calculated after a single refinement cycle that is performed mainly to obtain reliable σ_A weights (Read, 1986) and then a new model is built automatically. An important element is that the stereochemistry of the molecular-replacement model is completely ignored. The atoms of the model are used solely as guides for the autotracing, but the autotracing algorithm does *not* compare ambiguous areas against the existing model. This might appear to be a limitation of the algorithm, since prior knowledge is ignored, but in reality presents a vital means of minimizing model bias. This simplistic procedure (Fig. 2*b*) can yield impressive results, as depicted in §3.1.

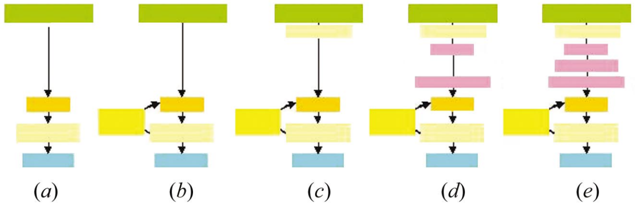

Figure 2
A scheme of the *ARP*/*wARP* protocols presented in §§2.1.1–2.1.4 (*b*)–(*e*) and in §2.2 (*a*). The protocols are sorted in levels of increasing complexity. The action in each box is that listed at the same position and in the same colour in Fig. 1.

2.1.2. Pre-refinement of the molecular-replacement model in *ARP*/*REFMAC* cycles. Here, the starting model is pre-refined for a number of *REFMAC* (Murshudov *et al.*, 1997) and *ARP* cycles with maximum-likelihood refinement combined with addition and deletion of atoms. After a given number of cycles a $2mF_o - DF_c$ map is calculated and the *warpNtrace* procedure is initiated. This may seem a better protocol, since prior information from the molecular-replacement model is exploited to a larger extent. However, if the model is retained for too many refinement cycles, there is a danger of introducing increasing model bias by forcing wrong areas of the model into density. Nevertheless, in our experience this is the best procedure (Fig. 2*c*) and is now the default in *ARP*/*wARP* 5.2.

2.1.3. Use of the molecular-replacement model as a source of phase information only. Although similar to that described in §2.1.2, this procedure possesses a fundamental difference: after the first cycle of refinement, the σ_A-weighted map is used as an 'experimental' map. The model is completely discarded, in contrast to the procedure in §2.1.1 where the atoms of the model were used as guides for the autotracing. Here, a new free-atoms model is constructed, as for MIR or MAD maps. The advantage is that the starting model might be too inaccurate and lots of atoms may be out of density or far from their true positions; thus, it would be misleading to use them for tracing and harder to remove the bias arising from their presence. By discarding them in a single step and reinterpreting the map as a free-atoms model, the ensuing more accurately placed free atoms facilitate the tracing and reduce model bias. This procedure (Fig. 2*d*) is particularly powerful when a significant part of the structure is missing from the molecular-replacement model, when it is crucial to start by initially placing a more-or-less correct number of atoms in density rather than slowly increasing the number of atoms through *ARP*/*wARP* cycles. The disadvantage of this protocol is that in practice the free-atoms placement tends to be less accurate than the molecular-replacement model, unless the model is either very different from the structure or very incomplete.

2.1.4. Use of the molecular-replacement model as a source of phases for density modification. This can be seen as an extension of §2.1.3 with the phases obtained from the model being used for density modification. The calculated phases ($\varphi_{\rm calc}$) are used as 'experimental phases' ($\varphi_{\rm best}$) with the σ_A weights as the FOMs. We typically use *DM* (Cowtan & Main, 1998) for the density modification. However, initial trials with *RESOLVE* (Terwilliger, 1999), which is designed to take into account the existence of a partial model and reduce the bias, have recently proved to be extremely powerful (data not shown). This protocol (Fig. 2*e*) was most useful when non-crystallographic symmetry (NCS) was available for the density-modification step and has yielded impressive results (as shown in §3.3 and §3.4) which could not be obtained by any of the other protocols. It is important to note that in the case presented in §3.3 it was essential to downweight the model contribution in phase recombination during density modification.

2.2. Model update and refinement to enhance map quality

This procedure (Fig. 2*a*) is recommended whenever the resolution is lower than 2.3 Å. Here, the aim is not to autobuild a model but rather by combining model update and refinement to deliver a map that is much better than that produced by any conventional refinement program alone. The final model is partially deleted (atoms set to zero occupancy) and has free atoms added as 'waters'. Manual rebuilding is required at the end of this process. Building of new side chains (*e.g.* for point mutants or close homologues) can be carried out automatically as described in the following session. If the autobuilding fails, this mode should nevertheless be carried as it will improve the map and assist the graphics rebuilding session. This mode has yielded impressive results, even at 3.2 Å resolution, as demonstrated in §3.5.

2.3. Automatic sequence docking and side-chain building

Sequence docking followed by side-chain building, best rotamer assignment and real-space torsional refinement can be applied at any resolution. It can also be applied after any of the protocols described above. This method can be an integral part of the auto-building protocol, but its use is not limited to this. The algorithms used and their results will be described in a different publication.

3. Examples

There are several successful examples of *ARP*/*wARP* application in the refinement of models from molecular replacement. There are also examples were *ARP*/*wARP* failed to automatically build a model. There are two factors that affect the success of *ARP*/*wARP* autobuilding in the case of molecular replacement: resolution and quality of the starting model. In general, the better the resolution and the better the starting model is, the more chances *ARP*/*wARP* has to succeed. If resolution is better that approximately 1.7 Å, a correct molecular-replacement solution is available and the molecular-replacement model is more than half of the final model, *ARP*/*wARP* autobuilding will succeed. If the data are between 1.7 and 2.2 Å it is hard to predict what will happen, but for complete good-quality models there are very high chances for the autobuilding to succeed. For data worse than 2.2 Å, autobuilding will not work regardless of the quality of the starting model. The *ARP*/*wARP* protocols for map improvement should be generally applicable regardless of resolution limits and should deliver better results than *REFMAC* refinement alone. In the few cases that we could not obtain any improvement with *ARP*/*wARP*, it proved that conventional refinement was not applicable as well. The most typical pathology of these cases (other than an incorrect solution of the molecular replacement) was badly measured data, missing or incomplete low-resolution reflections being the predominant cause.

The user should relate specific cases for which she/he intends to apply *ARP*/*wARP* to the following examples that show the application of the protocols we discussed in §§2.1–

2.3. It is not possible to give at this stage of the program development more exact 'recipes' or tabulate program parameters that should be used for specific cases. The examples are presented in increasing complexity and users are advised, if defaults fail, to try the protocols in a similar order to the presented examples. If resolution is lower than 2.3 Å, the protocol in §2.2 should be tried immediately, as exemplified in §3.5.

3.1. Equine infectious anaemia virus dUTPase

During the original structure solution for this protein (Dauter *et al.*, 1999), an *ARP/wARP* protocol in which the model was converted to free atoms and then refined against the 2.0 Å data was pioneered (*unrestrained ARP* or *uARP*). This convincingly showed the power of *ARP* in refinement of molecular-replacement solutions. Automated model building was omitted from the *ARP* cycling at the time, simply because the corresponding algorithms had not been developed. To demonstrate the power of the new protocols, the molecular-replacement model was used to initiate the simple protocol (§2.1.1). After 15 *ARP/wARP* autobuilding cycles a new model had been created, lacking only one N- and one C-terminal residue. $R_{\rm free}$ decreased from 54.1 to 27.6% without user intervention. Side chains were built using the protocol described in §2.3. A comparison of the initial and *ARP/wARP* models with the final structure from the PDB is depicted in Fig. 3.

3.2. Ankyrin repeat domain of Bcl-3

The search model for this structure was an edited model of human IκBα (Jacobs & Harrison, 1998), which was trimmed based on sequence alignments and trial-and-error methods (F. Michel *et al.*, to be published). The initial model contained five and a half ankyrin repeats with non-conserved residues mutated to alanine (Fig. 4). The model from the molecular-replacement solution by *AMoRe* (Navaza, 1994) was first refined by simulated annealing in *CNS* (Brunger *et al.*, 1998), which reduced the $R_{\rm free}$ to 46.5%. This model was subjected to *ARP/wARP* refinement against the 1.86 Å data. The initial refinement step proved in practice to be vital for the subsequent autobuilding to work. Later tests showed that it could be substituted by 20 *ARP/wARP* cycles before the start of autobuilding, as described in the protocol in §2.1.2. The *warpNtrace* process provided an essentially complete model. The $R_{\rm free}$ fell to 30.2%, yielding an excellent map and a model that lacked only ten N-terminal, 21 C-terminal and 15 internal residues from the total 228 residues. This example demonstrates convincingly the additional improvement that *ARP/wARP* can yield compared with conventional refinement strategies, provided that the resolution of the data is sufficient.

3.3. FprA

The structure of this *Mycobacterium tuberculosis* iron-metabolism enzyme was solved (R. T. Bossi and A. Mattevi, to be published) using as a search model bovine adrenodoxin reductase with 40% sequence identity. There were two FprA molecules in the asymmetric unit and molecular replacement using the program *Beast* (Read, 2001) located both. Although the map quality clearly supported a correct solution, all efforts to refine and autobuild with *ARP/wARP* failed. The only protocol that worked was a possibly generally applicable modification of that presented in §2.1.4. The *DM* program was used to extend the phases from 4.5 to 2.5 Å (*DM* keyword: SCHEME FROM 4.5) in a total of 400 cycles (*DM* keyword: NCYCLE 400) using the twofold NCS. During phase recombination the model phases were downweighted (*DM* keyword: *COMBINE* RPERT WEIGHT 0.2). The combination of these

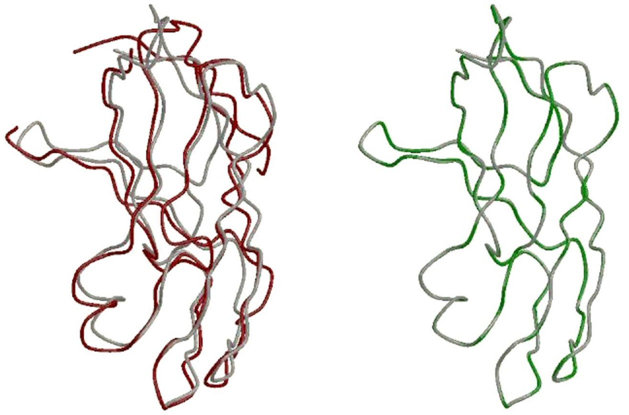

Figure 3
Left: a superposition of the search model (in red) for the eIAV dUTPase (§3.1) with the final structure (light grey). Right: the *warpNtrace* model in green superposed on the final structure (light grey).

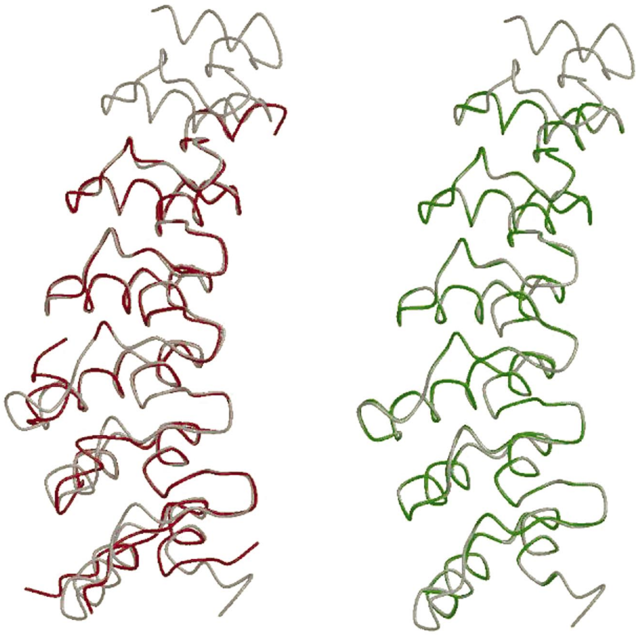

Figure 4
Left: a superposition of the search model (in red) for the ankyrin repeat domain of Bcl-3 (§3.3) with the final structure (light grey). Right: the *warpNtrace* model in green is superposed on the final structure (light grey).

parameters proved essential. Subsequently, *warpNtrace* was initiated from the resulting phase set and the 1.9 Å native data. After 15 building cycles, 900 of the expected 922 peptides were traced and the *R* factor converged to 18%, with a map of outstanding quality. Side chains were built as described in §2.3. In this example, where the 'usual' protocols did not work, effectively spending some time to figure out proper parameters for *DM* saved lots of tedious refinement and manual rebuilding.

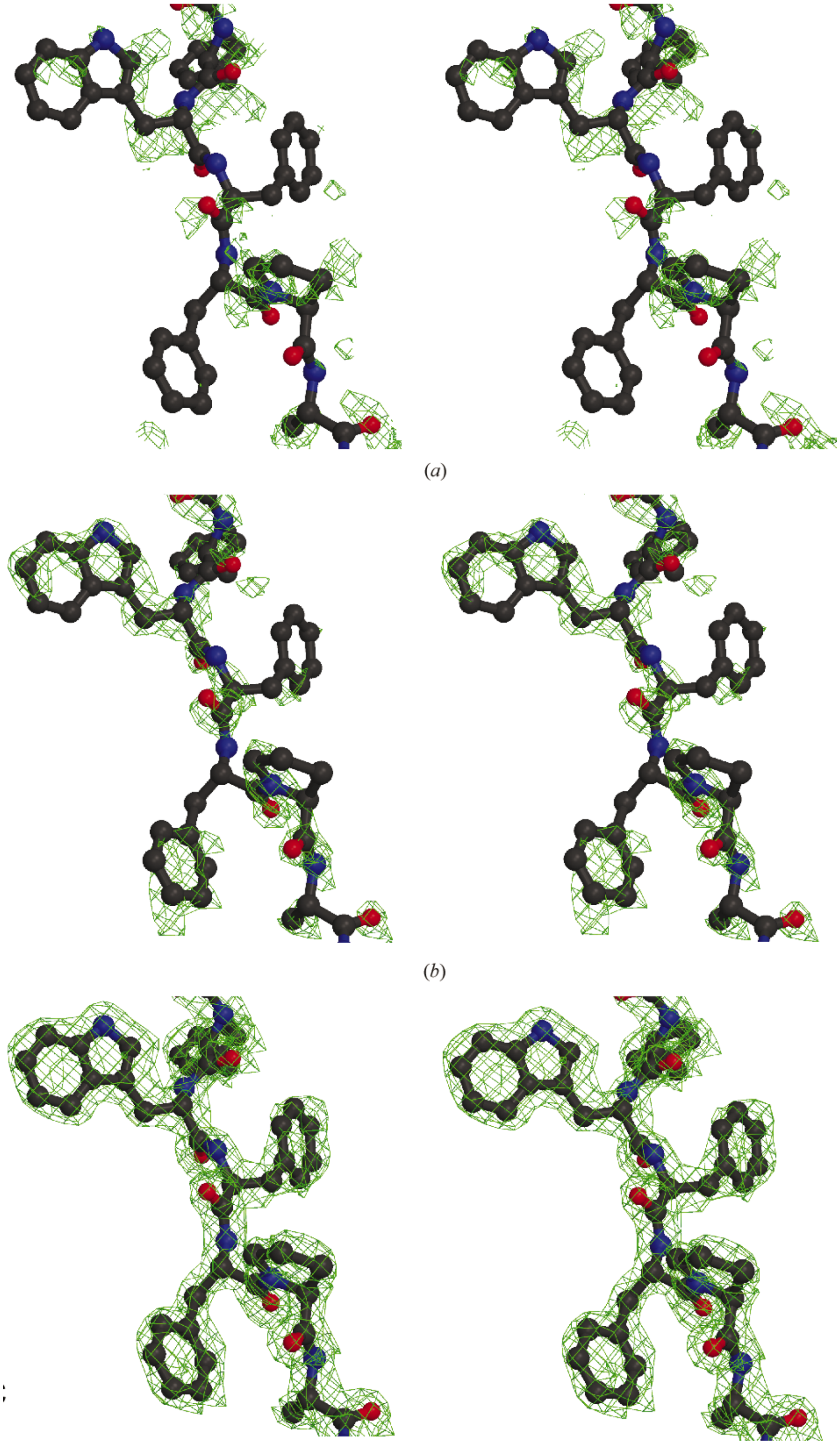

Figure 5
Stereo pairs of the final model for *B. subtilis* dUTPase with the electron-density maps (*a*) after molecular replacement and rigid-body refinement, (*b*) after density modification and NCS averaging and (*c*) after *ARP*/*wARP* refinement and automated model building. All maps were contoured at 1.2 r.m.s. from the mean density. The region shown is from the fourth molecule, which was not included in the molecular-replacement starting model.

3.4. *Bacillus subtilis* dUTPase

The structure of *B. subtilis* dUTPase (Persson *et al.*, 2001 and unpublished results) was solved by molecular replacement using *AMoRe*. The search model was the feline immunodeficiency virus dUTPase. However, it was impossible to refine, by *ARP*/*wARP* or other conventional means, the resulting model containing a trimer in the asymmetric unit. Therefore, the protocol described in §2.1.4 was tried. The density-modification step included solvent flattening, histogram matching and threefold NCS averaging with *DM*. The *warpNtrace* procedure was initiated with the 1.7 Å data. Surprisingly, it quickly became apparent that more residues than expected in the trimer (3 × 142 = 426) were being traced (450–480), despite the free-atoms model being tuned to only contain enough atoms for a trimer. The crystallographic *R* factor fell to 18.5% and the map was of excellent quality (Fig. 5). Visual inspection of the model clearly showed (Fig. 5) that a new fourth molecule was present in the asymmetric unit and had been partially (∼50%) traced by *ARP*/*wARP*. This subunit formed a trimer with symmetry-equivalent subunits in the $P6_3$ cell. Thus, there were four subunits per asymmetric unit, which was not expected from the self-rotation or the molecular-replacement solution. What in our opinion is remarkable about this particular example is that with good quality diffraction data, even with *a posteriori* wrong input (incomplete model, wrong solvent content for *DM*, wrong number of residues to

ARP/wARP), the crystallographic software is sufficiently robust to yield a correct solution. It must be noted that the NCS information was not used within *ARP/wARP*, since these options were not implemented at the time. A protocol that incorporates NCS is now implemented and will be available for testing in the next release of *ARP/wARP*.

3.5. *Alu* ribonucleoprotein particle

The structure of *Alu*RNA (5′ domain, 50 nucleotides) in complex with the protein heterodimer SRP9/14 from the signal recognition particle (Weichenrieder *et al.*, 2000) was solved using as a molecular-replacement model the structure of the free SRP9/14 protein (Birse *et al.*, 1997). The search model, *i.e.* the protein part of the ribonucleoprotein particle, was approximately 60% of the scattering volume. Data from an extremely small flat plate were collected at the microfocus beamline (Perrakis, Cypriani *et al.*, 1999) to the diffraction limit of that crystal at 3.2 Å. In an attempt to obtain a better map, since autobuilding is not possible at that resolution, the starting model was subjected to the protocol described in §2.2. This procedure increased the quality of the map around the protein model and led to a dramatic improvement of the electron density in the region occupied by the RNA, which could then be built manually from scratch without difficulty. The $R_{\rm free}$ fell from 46.5 to 38.3%, much lower than by conventional refinement or simulated-annealing protocols in Cartesian or torsional parameterizations.

We would like to thank all *ARP/wARP* users for their help and criticism, especially those whose data we present in this paper, in particular Andrea Mattevi and R. Persson for making data available prior to publication. Many thanks also to Zbyszek Dauter for the *uARP* applications that inspired much of this work. *ARP/wARP* developers participate in the EU-funded AUTOSTRUCT network (QLRT-1999-30398). The software can be downloaded from http://www.arp-warp.org.

References

Birse, D. E., Kapp, U., Strub, K., Cusack, S. & Aberg, A. (1997). *EMBO J.* **16**, 3757–3766.

Brunger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J. S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.

Cowtan, K. (1998). *Acta Cryst.* D**54**, 750–756.

Cowtan, K. & Main, P. (1998). *Acta Cryst.* D**54**, 487–493.

Dauter, Z., Persson, R., Rosengren, A. M., Nyman, P. O., Wilson, K. S. & Cedergren-Zeppezauer, E. S. (1999). *J. Mol. Biol.* **285**, 655–673.

Diller, D. J., Redinbo, M. R., Pohl, E. & Hol, W. G. (1999). *Proteins*, **36**, 526–541.

Holton, T., Ioerger, T. R., Christopher, J. A. & Sacchettini, J. C. (2000). *Acta Cryst.* D**56**, 722–734.

Jacobs, M. D. & Harrison, S. C. (1998). *Cell*, **95**, 749–758.

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

Kleywegt, G. J. & Jones, T. A. (1997). *Acta Cryst.* D**53**, 179–185.

Lamzin, V. S., Perrakis, A. & Wilson, K. S. (1999). *International Tables for Crystallography, Crystallography of Biological Macromolecules*, edited by M. Rossmann & E. Arnold, pp. 720–722. Dordrecht: Kluwer Academic Publishers.

Lamzin, V. S. & Wilson, K. S. (1993). *Acta Cryst.* D**49**, 129–147.

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

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

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

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

Oldfield, T. J. (1996). In *Crystallographic Computing 7. Proceedings from the Macromolecular Crystallography Computing School*, edited by P. E. Bourne & K. Watenpaugh.

Perrakis, A., Cypriani, F., Castagna, J.-C., Claustre, L., Burghammer, M., Riekel, C. & Cusack, S. (1999). *Acta Cryst.* D**55**, 1765–1770.

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

Persson, R., Harkiolaki, M., McGeehan, J. & Wilson, K. S. (2001). *Acta Cryst.* D**57**, 876–878.

Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.

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

Roussel, A. & Cambillau, C. (1991). *Silicon Graphics Geometry Partners Directory*, p. 81. Mountain View, CA, USA: Silicon Graphics.

Terwilliger, T. C. (1999). *Acta Cryst.* D**55**, 1863–1871.

Turk, D. (1992). *Weiterentwicklung eines Programms fuer Molekuelgraphik und Elektrondichte-Manipulation und seine Anwendung auf verschiedene Protein-Strukturaufklaerungen.* Technische Universitaet, Muenchen, Germany.

Wang, C. E. (2000). *Acta Cryst.* D**56**, 1591–1611.

Weichenrieder, O., Wild, K., Strub, K. & Cusack, S. (2000). *Nature (London)*, **408**, 167–173.