Published in final edited form as:
Nat Protoc. 2008 ; 3(7): 1171–1179. doi:10.1038/nprot.2008.91.

Automated macromolecular model building for X-ray crystallography using ARP/wARP version 7

Gerrit G Langer[a,#], **Serge X Cohen**[b,#], **Victor S Lamzin**[a,*], and **Anastassis Perrakis**[b,*]

[a]*European Molecular Biology Laboratory, c/o DESY, Notkestrasse 85, 22607, Hamburg, Germany.*
[b]*Department of Molecular Carcinogenesis, Netherlands Cancer Institute, Plesmanlaan 121, 1066CX Amsterdam, The Netherlands*

Abstract

ARP/wARP is a software suite to build macromolecular models in X-ray crystallography electron density maps. Structural genomics initiatives and the study of complex macromolecular assemblies and membrane proteins all rely on advanced methods for 3D structure determination. ARP/wARP meets these needs by providing the tools to obtain a macromolecular model automatically, with a reproducible computational procedure. ARP/wARP 7.0 tackles several tasks: iterative protein model building including a high-level decision-making control module; fast construction of the secondary structure of a protein; building flexible loops in alternate conformations; fully automated placement of ligands, including a choice of the best fitting ligand from a "cocktail"; and finding ordered water molecules. All protocols are easy to handle by a non-expert user through a graphical user interface or a command line. The time required is typically a few minutes although iterative model building may take a few hours.

Keywords

Structural genomics; X-ray crystallography; software; model building; ligand placement

INTRODUCTION

Overview

The advent of structural genomics initiatives[1] and medically oriented high-throughput structure determination projects[2], emphasized the need for advanced methods for structure determination[3]. In X-ray macromolecular crystallography, availability of comprehensive software packages like CCP4[4], CNS[5] and PHENIX[6] has had a major impact on structural biology research. Crystallographic model building has been traditionally done by expert users, with the aid of specialized interactive graphics software such as O[7] and more recently Coot[8]; a recent trend has been the automation of this process. First exemplified in the ARP/wARP package[9, 10], it was followed promptly by significant developments, e.g. in RESOLVE[11], TEXTAL[12], Bucanneer[13] and ACMI[14].

ARP/wARP has been used extensively in the past ten years, for thousands of structure determination experiments in macromolecular crystallography. These cover a diversity of studies of, e.g. a three-protein complex that is crucial in chromosome segregation[15]; complexes

*Corresponding authors.
#These authors contributed equally to the work presented

in spindle assembly checkpoint formation[16]; ubiquitin conjugation[17]; transcriptional regulation of mRNA[18]; cargo transport along microtubules[19]; studies of bioluminescence[20]; the structural dissection of an enzyme involved in synthesis of inflammatory mediators[21]; ligand recognition by lipoprotein receptors[22]; and investigations of membrane-binding proteins in signal transduction in photo-response[23], the plant aquaporin mechanism[24] and the functional characterisation of a prokaryotic Ca2+-gated K+ channel[25]. Moreover, ARP/wARP is often used as a standard benchmark to evaluate the quality of electron density maps produced by new methods, as exemplified in[26]. Finally, ARP/wARP has been integrated in many crystallography automated pipelines as the default model building engine[27-34].

Initial implementations of the ARP/wARP protein backbone (main-chain) tracing algorithms were specific for high-resolution structures[9] but subsequent developments[35, 36] have allowed successful automated building of a considerable part of the model at a resolution of as low as 2.7 Å[37] or 2.8 Å[38]. Although ARP/wARP has begun as a tool to build protein chains, it now provides a much wider spectrum of functionalities, Figure 1. All ARP/wARP modules perform better at higher crystallographic resolution. However, the ligand and loop building modules are applicable in many projects where automated model building does not succeed. Finally, the secondary structure recognition module extends the applicability of the procedure to data at only 4.5 Å resolution.

Automated model building

Free atoms and hybrid models—A crystallographic electron density map is always sampled to a regular grid. Essentially the goal of model building is to condense the information of the electron density map to a crystallographic molecular model, made by atoms with known chemical identity. As a first step towards building a model, ARP/wARP condenses the map information to a set of 'free atoms'[39] that have no chemical identity: these atoms are carefully chosen to represent as good as possible the electron density, but still resemble in their distribution a protein-like model. As model building and refinement proceed, some free atoms gain chemical identity (they are recognized as part of a protein chain) while others remain free. This mixture is an ARP/wARP hybrid model that combines two sources of information: it incorporates chemical knowledge from the partially built protein model, while its free atoms continue to interpret the electron density in areas where no model is yet available. Finally, use of the atomic positions in the hybrid model as guides for model building in the electron density maps allows implementing computationally more efficient algorithms.

Main chain—Main chain tracing in ARP/wARP uses all available atoms of the hybrid model (containing both free atoms and atoms from a partial protein model) as potential C_α atoms[9]. Peptides between potential C_α pairs are recognized by matching the electron density that surrounds each potential C_α pair [40] to that precomputed for true C_α pairs from known structures. The recognized peptides are subsequently assembled into linear polypeptide chain fragments using a limited depth-first graph search algorithm, a procedure we have previously described [35], where the main chain is built up from overlapping sets of four C_α fragments that are selected to match conformations observed in the Protein Data Bank (PDB). Chain fragments including partial 'guessed' side chains are refined to fit the electron density using the steepest descent algorithm.

Side chains—The protein chains are subsequently docked in sequence with side chains built in the best rotamer configuration[41] and refined in real space using an implementation of the downhill simplex algorithm. This allows the torsion angles of each side chain to be gradually changed in a step-wise manner, so as to fit atoms to the electron density, while keeping the side chain bonded atom distances and angles intact.

Loop building—After sequence docking, the missing parts of the model can be easily identified. Using this knowledge and a distribution of five C_α fragments that we derived from known structures, many structurally likely conformations are constructed and the ones that fit best the electron density are chosen. Incorporating prior information allows building in low-density regions. An exact description of the algorithms has been recently published[42].

Secondary structure recognition—At resolution of 3.0 Å and lower, where electron density maps lack atomic features, ARP/wARP uses a different algorithm to build protein helices and strands. Sparse map grid points with about 1 Å spacing are selected as potential C_α atoms on the basis of their density. They are then fed into a complex scheme of successive filtering steps that yield fragments of appropriate helical or stranded conformations. These are used to generate candidate trace ensembles that then undergo averaging. Finally, peptide backbone and C_β atoms are added, the secondary structural chain fragments are subject to real space refinement, and the most likely chain direction is selected. The procedure has been designed to work at resolution down to 4.5 Å.

Ligand building

When the protein structure is (nearly) completed, smaller compounds - ligands, cofactors - bound to the protein are modeled in the difference electron density map. First, regions of difference density that have approximately the same volume as the ligand are identified. Subsequently, we use numeric features of the density region and its sparse representation (similar to the one used to build free atoms for chain tracing), to produce an ensemble of putative ligand structures to best fit the local density. The single best model is chosen after restrained (steepest descent) real space refinement of all candidates in the ensemble[43, 44].

Cocktail screening—The technique that compares the shapes of difference electron density blobs to the shape of the ligand to be built, is used to distinguish compounds from a list (cocktail) of ligand candidates. The ligand that fits best is selected for further construction of the ensemble and subsequent restrained refinement.

Solvent building

After the protein part of the model is complete, either manually or using automated software, a solvent structure can be constructed in a difference electron density map. The protein part of the model is not rebuilt. Apart from van der Waals repulsion, no restraints are applied to the refinement of solvent even if the protein part is highly restrained. Therefore ordered solvent comprises on average about 10% of the model, improvement of solvent indirectly improves the density corresponding to the protein part. The output is the protein model with the solvent molecules transformed with symmetry operations to lie around the protein[40, 45].

Iterations

Building protein chains or solvent with ARP/wARP proceeds in an iterative fashion. When the quality of the (partially built) model is sufficiently high, the phases improve overall and result in an enhanced electron density where a more accurate and more complete model may be built. In essence, ARP/wARP, like human crystallographers, links model building and refinement together into a unified process that iteratively proceeds towards the final macromolecular model. An important component within iterations is the model update. Parts of the existing model located in weak density can be removed and new atoms added where the density acquired pronounced features.

MATERIALS

Equipment setup

System requirements—ARP/wARP 7.0 runs on any Linux platform we are aware of, including machines with Intel Itanium2 processors and under Mac OSX (both Power PC and Intel based machines). Alpha True64 Unix and SGI Irix distributions are also available. There is no implementation available for Windows.

AMD processors manufactured before about 2003 and Intel processors before 2001, which do not support the SSE2 instruction set, will not run programs for sequence docking and loop building. A corresponding warning message is printed during ARP/wARP installation. Some computational intensive parts of ARP/wARP are optimised using a data-parallel approach (taking advantage of SIMD capabilities of recent processors, namely SSE and AltiVec) and the ARP/wARP decision-making 'Expert System' can make use of any number of processor cores that are available in a specific machine to execute parallel sub processes.

ARP/wARP software and its availability—ARP/wARP is being developed using Fortran, C, C++, Python, Tcl/Tk, and C-shell. Some software modules use the Clipper [46] and ATLAS libraries for specific crystallographic and mathematical computations respectively. All software is being compiled with the Intel, IBM and GCC (Gnu Compiler Collection) compilers, and distributed as executables. ARP/wARP is freely available to all users and free of charge for academic usage. It can be downloaded from this site: http://www.arp-warp.org

Installation of the ARP/wARP package is made user friendly through an *install.sh* script. A User Guide contained within the ARP/wARP package can be referred to for additional information. There are no software dependencies other than the crystallographic CCP4 package[4] and particularly the program REFMAC[47], which ARP/wARP uses. Installation also implements the Graphical User Interface (GUI) which is connected to the GUI of CCP4, also known as CCP4i[48].

If a computer has access to the Internet and the *curl* utility is available, provided the data are available in a suitable format, a protein model building task can be submitted from the GUI to a 64-processor Linux cluster located at the EMBL Hamburg. The results can be viewed via an Internet browser. There is also a possibility to submit model building for remote execution directly through the web link http://cluster.embl-hamburg.de/ARPwARP/remote-http.html but this offers limited customization.

The data

ARP/wARP automated model building can be invoked after experimental phasing, molecular replacement or subsequent phase improvement and density modification; in crystallographic jargon, 'after solving the structure' or 'after solving the phase problem'.

- Diffraction data amplitudes to a resolution of about 2.7 Å (or higher) should be recorded and provided in an MTZ format. ARP/wARP examines the Wilson plot of the data to check data quality; a warning message is issued even if there are minor concerns. It is recommended that the user selects a 'test' dataset of reflections that can be used for cross-validation, often referred to as 'Free R flag'[49].
- Structure factor phases (if available) should be provided along with the structure factor amplitudes in the same MTZ file, together with their associated figures of merit; or a molecular model should be available to calculate the phases.

Detailed instructions for preparing data available are outlined in the first steps of the procedure.

The most important output of ARP/wARP is in all cases a standard coordinates file in the PDB format. This file should be manually inspected by the user for occasional errors, as indicated in detail in the Procedure.

PROCEDURE

Preparation of the reflections file

1| Unless the MTZ file is already available, convert your 'hkl' file to the MTZ format. CCP4i has all the necessary tools under 'Reflection Utilities'.

2| If the structure was solved with a heavy atom method, include the best available phases and their figures of merit.

3| If phases were determined by a heavy atom method, include the Hendrickson-Lattman coefficients (HL), which can be used as restraints.

<CRITICAL STEP> If Hendrickson-Lattman coefficients are available, including them in the procedure (see steps 19 and 34) can sometimes make the difference between a successful model building and failure. It is recommended that for phase restraints you use phase information that comes directly from the phasing procedure and not from a post-processing tool such as solvent flattening or averaging.

4| Check that the MTZ file has all possible reflections for your structure and the resolution you work with. Even if a certain hkl was not measured it should be present with a flag that it has not been measured. It is good practice to always use the 'uniqueify' script from the CCP4 collection.

5| Assign an Rfree column for about 1,000-1,500 reflections. The script *uniqueify* that can be run from the command line and comes together with CCP4 can also take care of that; to invoke it simply type *uniqueify* in the command line.

Building secondary structure elements

6| Start this procedure that you can use to e.g. validate an initial density map for its interpretability by starting the 'ARP/wARP Quick Fold' interface from CCP4i.

7| Choose the name of the MTZ file you will use, created in steps 1-2. An MTZ file with phases is required.

8| Assign the correct labels for the amplitudes and their associated uncertainties (sigmas).

9| Assign the correct labels for the best phases and their associated uncertainties (figures of merit).

10| Enter the approximate total number of residues expected in the asymmetric unit.

11| Run the job. The output will be a PDB file with the constructed helices and strands. This model is a partial main-chain trace that can be used as an indication of the interpretability of the map and can provide a quick impression about the fold of the molecule. The constructed helices and strands may contain some local errors but can be used for subsequent manual model extension.

Automated protein model building using the 'ARP/wARP Classic' interface

12| Prepare the sequence file, unless you do not know the sequence of the protein in question. The first line of the sequence file must start with the ">" character and contain any title afterwards while the second line must be empty. Subsequent lines should have the amino acid sequence in one letter code. If you are working with a complex containing different polypeptide chains, separate each chain with about ten alanine residues.

<CAUTION> Even if you have multiple copies of the same protein in the asymmetric unit, please put in the file only the sequence for one copy (monomer). You can define the multiplicity later at step 17. Do not duplicate a sequence in the file, as this may cause errors during sequence docking.

13| Proceed using Option A if the initial structure was solved by Molecular Replacement or a partial model is available. Proceed using Option B if the initial map was obtained by a heavy atom phasing method such as SAD, MAD, SIR, MIR or a combination thereof.

- **A.** Molecular Replacement or a partial model
  - **i.** Prepare the PDB file for your model. The file should contain the CRYST and SCALE information (you can use the CCP4 program PDBSET to create these if needed). It is important that the cell present in the PDB file is identical to that in the MTZ file.
  - **ii.** Inspect the PDB file for non-standard residues, e.g. ligands. These may not be properly handled during automated model building and may cause the job to fail.
  - **iii.** Choose the protocol 'Automated model building starting from existing model'.
  - **iv.** Choose the name of the reflection file obtained in steps 1 and 5.
  - **v.** Assign the correct labels for the amplitudes and their associated uncertainties (sigmas).
  - **vi.** Choose the name of the coordinates (PDB) file from step i above.
- **B.** Initial map obtained by a heavy atom replacement method such as SAD, MAD, SIR, MIR or a combination thereof
  - **i.** Choose the protocol 'Automated model building starting from experimental phases'.
  - **ii.** Choose the name of the MTZ reflection file obtained in steps 1-5.
  - **iii.** Assign the correct labels for the amplitudes and their associated uncertainties (sigmas).
  - **iv.** Assign the correct labels for the best phases and their associated uncertainties (figures of merit) obtained in step 2.

14| Choose the name of the sequence file from step 13.

15| Provide the expected total number of residues in the asymmetric unit.

16| Define how many copies of each chain you expect in the asymmetric unit if non-crystallographic symmetry is present.

<CAUTION> In this mode it is not possible to define complexes that have for example two copies of one molecule and one of another in this interface. In this case please use the 'Expert System'; specific instructions for how to prepare the sequence file are available in step 29.

17| Inspect the number of autobuilding cycles (default is 10). Although the default will do in most cases, a larger number of cycles should not worsen the results. At the end of the run the software may print a suggestion to run additional cycles. This may result in higher completeness of the automatically built model.

18| If experimental phases in the MTZ file were accompanied by the HL coefficients (see step 3), use them as phase restraints. You can do that by choosing 'Phased ML' instead of 'Maximum Likelihood' in the choice of the target function.

<CRITICAL STEP> If data to about 2.3 Å resolution or lower are only available, it is especially advised to add phase restraints.

19| Run the job. Wait for the output PDB file with coordinates of the model, an MTZ file with phases and electron density maps. For models that appear complete the number of incorrectly built residues is expected to be very very small. As an indication, at every cycle, as well as at the end of the job, the estimated correctness of the model is printed. A successful run is also indicated when most of the chains docked into sequence and R factors that are typically in the low twenties. If a free R factor is used, it should by the end of the job be not much higher than plain R factor. In any case the output PDB should be visually inspected according to standard crystallographic practice.

20| To run the job remotely check the button "Submit the job for remote execution at the Hamburg cluster". Enter your email address and run the job as normal. The log file will contain submission information and will guide you for how to examine the results.

21| If you wish to build missing loops, launch the 'ARP/wARP Loops' interface from CCP4i.

22| Choose the name of the PDB file from step 19.

23| Choose the MTZ file with the phases from step 19.

24| The interface will detect automatically which loops are missing from your structure. Choose the loop to build and also indicate the sequence in the available box.

25| Optionally you can output multiple loop conformations.

26| Run the job that will output coordinate files in the PDB format.

<CAUTION> You should be aware that the program will try to build the loop through low density, as long as it conforms to geometrical criteria. It is emphasized that the user should inspect the loops visually and make a scientific decision if the electron density map supports the specific loop conformation or not.

27| If more than one loop needs to be built, use the output of step 26 as input to step 22. Repeat this as many times as there are loops to build.

Automated protein model building using the ARP/wARP 'Expert System' interface

28| Prepare the sequence file(s) with the format described in step 12. If you are working with a complex containing different polypeptide chains, prepare a separate file for each chain, each having a unique amino-acid sequence.

<CAUTION> It is important to put each sequence in a separate file. Even if you have multiple copies of the same protein in the asymmetric unit, please put the sequence in the file only once. You can define the multiplicity of each sequence later at step 32.

29| If a model from molecular replacement or a partial model is available, proceed as in step 14 A above. However, here you can choose how the PDB file should be handled: 'as it is' and proceed to the model building; reset the chemical identity of all atoms to free atoms first; or use it to extract phases and make a map in which to build a free atoms model. If you have enough computational power available try all three options above. For a more informed decision please read the comments provided in a separate publication[50]. If the initial map was obtained by a heavy atom replacement method such as SAD, MAD, MIR or a combination thereof, proceed as in step 14 B above.

30| Choose the names of the sequence files; and use the 'Add Input PIR file' button to define chains with different amino acid sequence.

31| Define how many copies of each chain you have in the asymmetric unit

32| If your diffraction data comes from a Se-Met substituted protein, choose this option.

33| If experimental phases were accompanied in the MTZ file by the HL coefficients (see step 3) you may use them for phase restraints in refinement. You can do that by selecting 'HL' phase restraints

<CRITICAL STEP> If data to about 2.3 Å resolution or lower are only available, it is especially advised to use phase restraints.

34| Run the job. The 'Expert system' will decide automatically when to stop. When a fairly complete model is obtained (default 80%, under 'Decision Parameters') a separate process is started: this model will be cleaned from free atoms, loops and the solvent structure will be built, the model will be validated for its fit to the density, and the coordinate and the MTZ files will be written and registered in the job output. Every time a new model is better than the best so far, the process described above will be iterated; the job will only stop if more than 95% of the model is built or if 40 autobuilding attempts have been made. These settings can be modified in the 'Decision parameters' panel of the interface. At the end of the job a short density-based validation procedure is run[51] and particular model regions that are likely to contain errors are highlighted in the log file. Although these regions should be inspected first, it is not impossible to have additional problematic regions and the model should be inspected manually using standard validation tools. R factors should be as discussed in step 20.

Building ligands

35| Launch the 'ARP/wARP Ligands' interface from CCP4I.

36| Prepare an MTZ file with amplitudes and the associated uncertainties (you can use the same MTZ file as was used for the automated model building) and select it in the interface.

37| Assign the correct labels for the observed amplitudes and their associated uncertainties (sigmas).

38| Prepare the PDB coordinates file of your protein that is expected to have the bound ligand (you can use the PDB file from the automated model building) and select it in the interface.

<CRITICAL STEP>. It is advised that all water molecules (e.g. from a partially built solvent structure) are removed prior to ligand building. Should this not be done and a solvent site occupies the ligand binding site, the shape of the difference electron density, which ARP/wARP computes for ligand search, may become adversely affected.

39| Prepare the PDB coordinates file for every ligand you plan to build. You can use services such as HICUP[52] or PRODRG[53]. Proceed to option A if the ligand binding pocket is unknown, to option B if the ligand bindng pocket is known through another ligand that has been modelled previously, or to option C if the approximate XYZ coordinates of the region that the ligand should be built into are known.

- **A.** The ligand binding pocket is unknown
  - **i.** Choose the default protocol that will build the ligand 'in the most likely place of the complete asymmetric unit'.
- **B.** The ligand binding pocket is known

i. Choose the protocol that allows building the ligand 'around the same approximate place as a previous ligand'.

ii. Choose the PDB coordinates file that contains the previously modelled ligand.

C. The approximate XYZ coordinates of the region that the ligand should be built into are known

i. Choose the protocol for building the ligand 'around an approximate XYZ position'.

ii. Enter the XYZ center coordinates and the radius of the sphere that is likely to contain the ligand.

40| Depending on the type of ligand experiment you performed, proceed to Option A if a known ligand needs to be modelled, or to option B if there are many likely ligands present after soaking the crystals in a cocktail of different ligands.

A. Building a single ligand

i. Select the PDB file for this ligand in the interface.

B. Finding the most likely bound ligand from a cocktail

i. Prepare a PDB file that contains the coordinate sections of all respective ligands concatenated with the residue names or residue numbers distinguishing them from one another. A 'CRYST' line is not needed.

ii. Select the file you created above in place of the PDB coordinates file for the single ligand.

41| Run the job. The short log file concludes with the real space map correlation for the built ligand, which is typically of an order of 0.8 or higher. It also prints a self-validation statement whether the ligand is built successfully or not. A PDB file with the modelled ligand will be the output. Additionally, a difference map and the detailed log file are created for inspection.

Build the solvent structure

42| Launch the 'ARP/wARP Solvent Building' interface from CCP4I.

43| Prepare an MTZ file with amplitudes and the associated uncertainties as above. The use of Rfree is advised.

44| Prepare the PDB coordinates file of your protein for adding solvent atoms.

45| Select the number of solvent building cycles (default is 20).

46| Run the job. A PDB file with the modelled water molecules will be the output.

CAUTION. The solvent sites are built by ARP/wARP iteratively. At each cycle the density maps are re-computed using the observed and the current calculated structure factors. Towards the end of solvent building some of the previously built waters may be automatically removed. When the procedure converges, at each cycle the number of added waters is about equal to those removed. This should normally occur within the default 20 cycles, but sometimes further iterations may be required. The procedure of adding waters should lower the Rfree factor from the start to the end of the job. The built solvent sites should be visually inspected.

Automated command line scripts

All of the actions that result in an automated model and have been explained in steps 6-20 and 29-46, can also be performed without the CCP4i interface, by issuing single commands from a terminal window, or from within another software script:

auto_albe.sh (helices and strands)

auto_ligand.sh (ligand building)

auto_tracing.sh (model building 'Classic')

auto_flex_warp.sh (model building 'Expert System')

auto_solvent.sh (building solvent)

The input files to these automatic scripts are the same as for the interface. Executing the automatic scripts without arguments provides a short on-line help. More detailed instructions are given in the ARP/wARP User Guide.

Web service for model building using the 'Classic' protocol

1. Navigate your browser to http://cluster.embl-hamburg.de/ARPwARP/remote-http.html
2. View the Disclaimer as well as the ARP/wARP and the CCP4 licensing conditions and press the 'Continue' button.
3. Enter your Email address to which instructions on how to view the results will be sent.
4. Choose the model building protocol (start from experimental phases or existing model).
5. Choose the MTZ file.
6. Click 'Submit' to proceed to Step 2.
7. Choose how to enter or upload your sequence file. Please refer to step 13 in Procedure for sequence file preparation.
8. Enter the total number of residues in the asymmetric unit.
9. Choose the number of chemically identical molecules in the asymmetric unit.
10. Choose the PDB file for the starting model, if applicable.
11. Choose the number of automated model building cycles.
12. Choose the MTZ labels. Those for the observed structure factor amplitudes and their associated uncertainties are compulsory for model building starting from the existing model. The labels for phases and their figures of merit are additionally needed if starting from experimental phases.
13. Choose the dissemination level (World, ARP/wARP Developers, or Confidential).
14. Submit the job and follow the instructions for viewing the results.

Timing

Setting up the ARP/wARP protocols is straightforward and does not take more than a few minutes. No interactive decision of any kind is needed until completion of the jobs. Execution time depends on the size of the molecule in question and hereafter is referred to the currently available computers with CPU clock frequencies around 2 GHz. In most cases the secondary structure tracing and ligand building jobs run within a few minutes. Solvent building is iterative but should complete within tenths of minutes for most proteins under 1,000 residues. Automated model building 'Classic' jobs for a protein of about 500 residues should take well below an hour to complete; for larger or smaller models the time needed scales approximately linear with the size of the structure. 'Expert System' completion time is comparable to 'Classic': for straightforward jobs 'Expert System' runs for a shorter time while for difficult cases it can take much longer.

Troubleshooting

The ARP/wARP suite has been tested extensively and is used in well over 1,000 laboratories by more than 3,000 users. In case of installation problems, or premature exits or detected bugs the users are encouraged to contact the corresponding authors (see http://www.arp-warp.org for more information).

In cases of insufficient model completeness after a 'Classic' automated model building, more cycles can be run using the protocol described in Procedure from step 13 onwards. 'Expert system' will continue automatically until the best completeness has been reached. Sometimes, better results can be obtained by increasing the number of building cycles.

In cases when automated ligand building fails, it is advised to give as input the approximate coordinates for the binding site to make sure that ligand building is attempted in the correct place. In cases where only a part of the ligand is observed in the density, while another part is not visible because it e.g. occupies multiple conformations and is thus disordered, it might be worth trying to model only the ordered part. An example where the density was insufficient to model the whole ligand, but truncating the ligand produced a much more realistic fit is illustrated in Figure 2.

It should be emphasized that all automatically built models may contain a few errors (particularly at medium-to-low resolution or in case of poor data quality). Thus we advise manual inspection and validation according to standard crystallographic practices.

ANTICIPATED RESULTS

Automated model building

ARP/wARP has been commonly (mis)conceived as a software suite for automated building of protein structures at high resolution. Although ARP/wARP rarely fails if initial phases are reasonable and the resolution of the data is higher that 2.0 Å, it can very often produce fairly complete models when the resolution is as low as 2.7 Å, Figure 3. At lower resolution obtaining most of the model with ARP/wARP 7.0 is rather a surprise than the rule, but occasional useful traces are not uncommon.

Compared to other approaches we are aware of, ARP/wARP would most often produce a more complete and accurate model at resolutions of 2.0 Å or higher, and in lesser time. Conversely, it will produce a less complete model compared to other approaches at resolution 2.5 Å or lower, even if it will be at least as accurate and still faster than other approaches. Exceptions to the above rules - either way - are rather common like most exceptions, and we would encourage all users to run ARP/wARP in all cases if time and resources permit. The worse case

scenario is that ARP/wARP will not produce a meaningful and helpful result; however ARP/wARP will never produce erroneous results and build the wrong structure.

Building secondary structure elements

ARP/wARP can produce within minutes a secondary structure trace of the protein (the maximum estimated performance for identifying helices being around 80% and for strands about 50%) even if the X-ray data extend to as low as 4.5 Å resolution. That makes it a first choice solution for fast evaluation of the success of a data collection experiment immediately after (or even before) all the data have been collected at a synchrotron beamline or a home source.

Ligand building

Building bound ligands and solvent constitute important steps of model completion in macromolecular crystal structure determination. The ARP/wARP ligand building module was tested on several thousand protein-ligand complexes from the Protein Data Bank with X-ray data spanning 1.0 to 3.0 Å resolution. The overall success (defined as the automatically 'reproduced' ligand model with an r.m.s.d. from the 'truth' of 1.0 Å or lower) is about 70% overall, and in over 80% of the cases the binding site could be correctly identified. In addition, there is a novel option for 'cocktail screening' where the most likely ligand is automatically identified and built from a list of potential candidates, whose full potential is yet to be assessed.

References

1. Stevens RC, Yokoyama S, Wilson IA. Global efforts in structural genomics. Science 2001;294:89–92. [PubMed: 11588249]
2. Banci L, et al. First steps towards effective methods in exploiting high-throughput technologies for the determination of human protein structures of high biomedical value. Acta Crystallogr D Biol Crystallogr 2006;62:1208–17. [PubMed: 17001097]
3. Lamzin VS, Perrakis A. Current state of automated crystallographic data analysis. Nat Struct Biol 2000;7(Suppl):978–81. [PubMed: 11104005]
4. C.C.P.N. The CCP4 suite: programs for protein crystallography. Acta Crystallogr D 1994;50:760–763. [PubMed: 15299374]
5. Brunger AT. Version 1.2 of the Crystallography and NMR system. Nat Protoc 2007;2:2728–33. [PubMed: 18007608]
6. Adams PD, et al. PHENIX: building new software for automated crystallographic structure determination. Acta Crystallogr D Biol Crystallogr 2002;58:1948–54. [PubMed: 12393927]
7. Jones TA, Zou J-Y, Cowan SW, Kjeldgaard M. Improved methods for the building of protein models in electron density maps and the location of errors in these models. Acta Crystallogr A 1991;47:110–119. [PubMed: 2025413]
8. Emsley P, Cowtan K. Coot: model-building tools for molecular graphics. Acta Crystallogr D Biol Crystallogr 2004;60:2126–32. [PubMed: 15572765]
9. Perrakis A, Morris R, Lamzin VS. Automated protein model building combined with iterative structure refinement. Nat Struct Biol 1999;6:458–63. [PubMed: 10331874]
10. Morris RJ, Perrakis A, Lamzin VS. ARP/wARP and automatic interpretation of protein electron density maps. Methods Enzymol 2003;374:229–44. [PubMed: 14696376]
11. Terwilliger T. SOLVE and RESOLVE: automated structure solution, density modification and model building. J Synchrotron Radiat 2004;11:49–52. [PubMed: 14646132]
12. Ioerger TR, Sacchettini JC. TEXTAL system: artificial intelligence techniques for automated protein model building. Methods Enzymol 2003;374:244–70. [PubMed: 14696377]
13. Cowtan K. The Buccaneer software for automated model building. 1. Tracing protein chains. Acta Crystallogr D Biol Crystallogr 2006;62:1002–11. [PubMed: 16929101]

14. DiMaio F, et al. Creating protein models from electron-density maps using particle-filtering methods. Bioinformatics 2007;23:2851–8. [PubMed: 17933855]
15. Jeyaprakash AA, et al. Structure of a Survivin-Borealin-INCENP Core Complex Reveals How Chromosomal Passengers Travel Together. Cell 2007;131:271. [PubMed: 17956729]
16. Mapelli M, Massimiliano L, Santaguida S, Musacchio A. The Mad2 Conformational Dimer: Structure and Implications for the Spindle Assembly Checkpoint. Cell 2007;131:730. [PubMed: 18022367]
17. Penengo L, et al. Crystal Structure of the Ubiquitin Binding Domains of Rabex-5 Reveals Two Modes of Interaction with Ubiquitin. Cell 2006;124:1183. [PubMed: 16499958]
18. Bono F, Ebert J, Lorentzen E, Conti E. The Crystal Structure of the Exon Junction Complex Reveals How It Maintains a Stable Grip on mRNA. Cell 2006;126:713. [PubMed: 16923391]
19. Allingham JS, Sproul LR, Rayment I, Gilbert SP. Vik1 Modulates Microtubule-Kar3 Interactions through a Motor Domain that Lacks an Active Site. Cell 2007;128:1161. [PubMed: 17382884]
20. Nakatsu T, et al. Structural basis for the spectral difference in luciferase bioluminescence. Nature 2006;440:372. [PubMed: 16541080]
21. Molina DM, et al. Structural basis for synthesis of inflammatory mediators by human leukotriene C4 synthase. Nature 2007;448:613–616. [PubMed: 17632546]
22. Fisher C, Beglova N, Blacklow SC. Structure of an LDLR-RAP Complex Reveals a General Mode for Ligand Recognition by Lipoprotein Receptors. Molecular Cell 2006;22:277. [PubMed: 16630895]
23. Moukhametzianov R, et al. Development of the signal in sensory rhodopsin and its transfer to the cognate transducer. Nature 2006;440:115. [PubMed: 16452929]
24. Törnroth-Horsefield S, et al. Structural mechanism of plant aquaporin gating. Nature 2006;439:688. [PubMed: 16340961]
25. Ye S, Li Y, Chen L, Jiang Y. Crystal Structures of a Ligand-free MthK Gating Ring: Insights into the Ligand Gating Mechanism of K+ Channels. Cell 2006;126:1161. [PubMed: 16990139]
26. Qian B, et al. High-resolution structure prediction and the crystallographic phase problem. Nature 2007;450:259. [PubMed: 17934447]
27. Brunzelle JS, et al. Automated crystallographic system for high-throughput protein structure determination. Acta Crystallogr D Biol Crystallogr 2003;59:1138–44. [PubMed: 12832756]
28. Fu ZQ, Rose J, Wang BC. SGXPro: a parallel workflow engine enabling optimization of program performance and automation of structure determination. Acta Crystallogr D Biol Crystallogr 2005;61:951–9. [PubMed: 15983418]
29. Holton J, Alber T. Automated protein crystal structure determination using ELVES. Proc Natl Acad Sci U S A 2004;101:1537–42. [PubMed: 14752198]
30. Liu ZJ, et al. Parameter-space screening: a powerful tool for high-throughput crystal structure determination. Acta Crystallogr D Biol Crystallogr 2005;61:520–7. [PubMed: 15858261]
31. Minor W, Cymborowski M, Otwinowski Z, Chruszcz M. HKL-3000: the integration of data reduction and structure solution--from diffraction images to an initial model in minutes. Acta Crystallogr D Biol Crystallogr 2006;62:859–66. [PubMed: 16855301]
32. Ness SR, de Graaff RA, Abrahams JP, Pannu NS. CRANK: new methods for automated macromolecular crystal structure solution. Structure 2004;12:1753–61. [PubMed: 15458625]
33. Panjikar S, Parthasarathy V, Lamzin VS, Weiss MS, Tucker PA. Auto-Rickshaw: an automated crystal structure determination platform as an efficient tool for the validation of an X-ray diffraction experiment. Acta Crystallogr D Biol Crystallogr 2005;61:449–57. [PubMed: 15805600]
34. Vonrhein, C.; Blanc, E.; Roversi, P.; Bricogne, G. in Crystallographic methods. S, D., editor. Humana Press; Totowa, NJ: 2006.
35. Morris RJ, Perrakis A, Lamzin VS. ARP/wARP's model-building algorithms. I. The main chain. Acta Crystallogr D Biol Crystallogr 2002;58:968–75. [PubMed: 12037299]
36. Morris RJ, et al. Breaking good resolutions with ARP/wARP. J Synchrotron Radiat 2004;11:56–9. [PubMed: 14646134]
37. Colf LA, Juo ZS, Garcia KC. Structure of the measles virus hemagglutinin. Nat Struct Mol Biol 2007;14:1227–8. [PubMed: 18026116]

38. Wuerges J, et al. Structural basis for mammalian vitamin B12 transport by transcobalamin. Proc Natl Acad Sci U S A 2006;103:4386–91. [PubMed: 16537422]
39. Agarwal RC, Isaacs G. Method for obtaining a high resolution protein map starting from a low resolution map. Proceedings of the National Academy of Sciences 1977;74:2835–2839.
40. Lamzin VS, Wilson KS. Automated refinement for protein crystallography. Methods in enzymology 1997;277:269–305. [PubMed: 18488314]
41. Lovell SC, Word JM, Richardson JS, Richardson DC. The penultimate rotamer library. Proteins 2000;40:389–408. [PubMed: 10861930]
42. Joosten K, et al. A knowledge-driven approach for crystallographic protein model completion. Acta Crystallogr D Biol Crystallogr 2008;64:416–24. [PubMed: 18391408]
43. Zwart PH, Langer GG, Lamzin VS. Modelling bound ligands in protein crystal structures. Acta Crystallogr D Biol Crystallogr 2004;60:2230–9. [PubMed: 15572776]
44. Evrard GX, Langer GG, Perrakis A, Lamzin VS. Assessment of automatic ligand building in ARP/wARP. Acta Crystallogr D Biol Crystallogr 2007;63:108–17. [PubMed: 17164533]
45. Lamzin VS, Wilson KS. Automated refinement of protein models. Acta Crystallogr D Biol Crystallogr 1993;49:129–147. [PubMed: 15299554]
46. Cowtan K. The Clipper C++ libraries for x-ray crystallography. IUCr computing comission newsletter 2003;2:4–9.
47. Murshudov GN, Vagin AA, Dodson EJ. Refinement of macromolecular structures by the maximum-likelihood method. Acta Crystallogr D Biol Crystallogr 1997;53:240–255. [PubMed: 15299926]
48. Potterton E, Briggs P, Turkenburg M, Dodson E. A graphical user interface to the CCP4 program suite. Acta Crystallogr D Biol Crystallogr 2003;59:1131–7. [PubMed: 12832755]
49. Brunger AT. Free R value: A novel statistical quantity for assessing the accuracy of crystal structures. Nature 1992;355:472–475. [PubMed: 18481394]
50. Cohen SX, et al. ARP/wARP and molecular replacement: the next generation. Acta Crystallogr D Biol Crystallogr 2008;64:49–60. [PubMed: 18094467]
51. Cohen SX, et al. Towards complete validated models in the next generation of ARP/wARP. Acta Crystallogr D Biol Crystallogr 2004;60:2222–9. [PubMed: 15572775]
52. Kleywegt GJ, Henrick K, Dodson EJ, van Aalten DM. Pound-wise but penny-foolish: How well do micromolecules fare in macromolecular refinement. Structure 2003;11:1051–9. [PubMed: 12962624]
53. Schuttelkopf AW, van Aalten DM. PRODRG: a tool for high-throughput crystallography of protein-ligand complexes. Acta Crystallogr D Biol Crystallogr 2004;60:1355–63. [PubMed: 15272157]

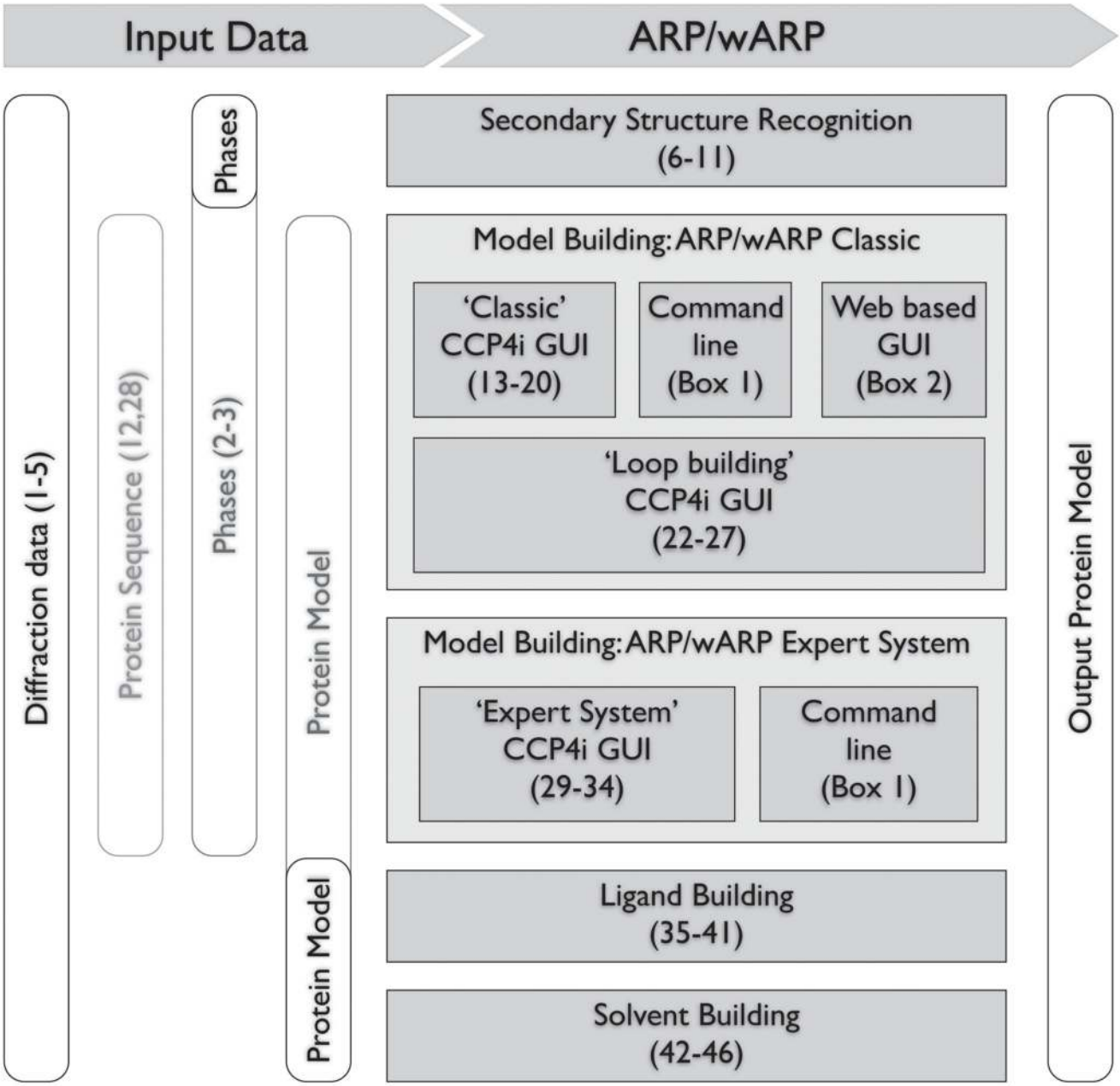

Figure 1.

A flowchart of the ARP/wARP procedure. The arrow on top indicates the flow of the data. ARP/wARP modules are labelled in the middle in grey shaded boxes; the numbers in parentheses refer to the steps in Procedure that describe them. The rounded rectangular boxes to the left represent input data (black for required data, light grey for optional input - the sequence - and medium grey for alternative input - the phases or a model) and to the right to the output data. The vertical span of the input / output boxes refers to the procedures they are connected to in the middle.

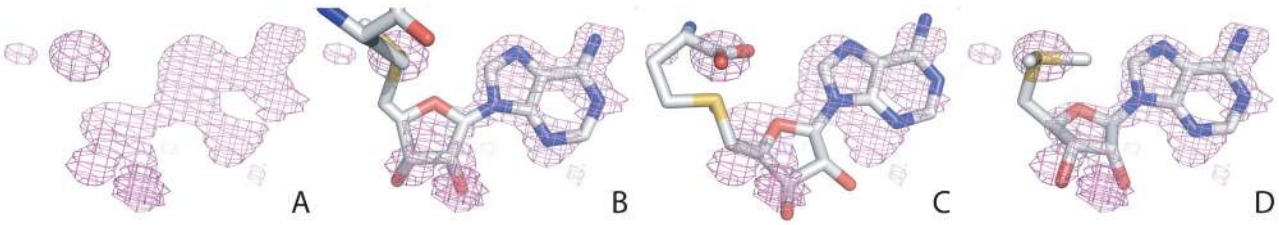

Figure 2.
Trouble shooting for partially ordered ligands. (A) The difference electron density map in the area of the ligand. (B) The ligand adenosylmethionine as modeled in the deposited structure in the PDB (1v2x). (C) A full ligand as built with ARP/wARP. (D) A partial ligand (with the methionine moiety up to its C_β atom removed) built with ARP/wARP, which matches the density better.

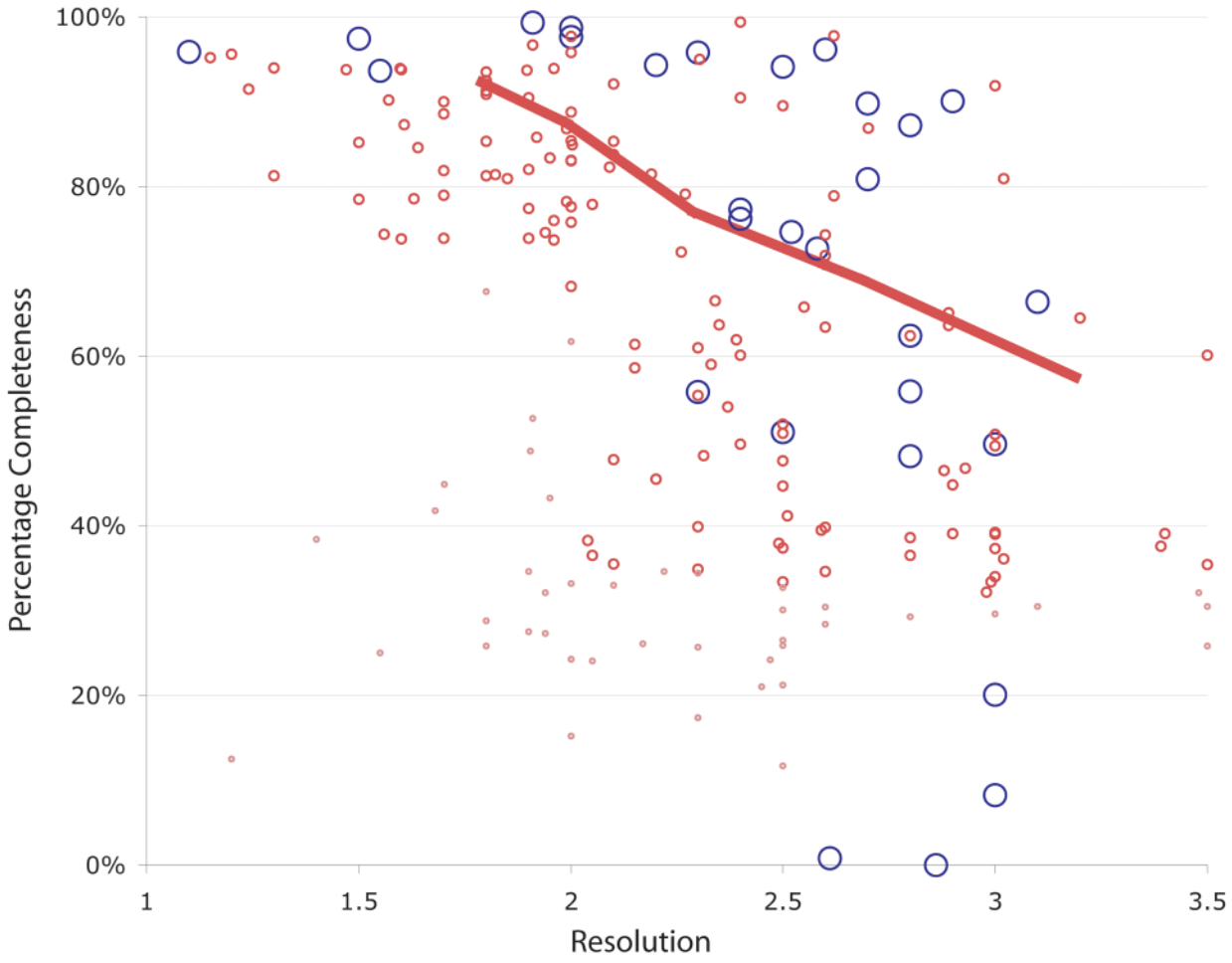

Figure 3.

Model completeness achieved by ARP/wARP 7.0 'Classic' as a function of resolution. The blue circles represent a set of diverse structures donated by users in http://xtal.nki.nl/Depot. For these the starting map and the final structure are both known and the model completeness is calculated as the percentage of correct residues compared to the residues in the final structure. The red circles represent a set of diverse structures submitted by users to the ARP/wARP web server (http://cluster.embl-hamburg.de/ARPwARP/remote-http.html). For these structures the starting map or model is known but no final structure is available; thus model completeness is calculated as the percentage of traced residues compared to the anticipated number of residues in the final structure as input by the user. A trend line (counting only the average top 50% of jobs) is also shown for these structures. The lower model completeness even at high resolution for the remotely submitted jobs compared to the Depot structures is likely due to the fact that very rarely all residues in a sequence can ever be built in a crystallographic structure and thus the ARP/wARP performance could be underestimated. We also cannot exclude that many cases that end up in our web server could be the 'hopeless' local cases that users attempt to run remotely after failure in their lab.