Breaking good resolutions with ARP/wARP

Richard J. Morris[a], Petrus H. Zwart[b], Serge Cohen[c], Francisco J. Fernandez[b,c], Mattheos Kakaris[c], Olga Kirillova[b], Clemens Vonrhein[d], Anastassis Perrakis[c*] and Victor S. Lamzin[b]

[a] *European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD,UK*
[b] *EMBL Hamburg, c/o DESY, Notkestraße 85, D-22603 Hamburg, Germany*
[c] *NKI, Department of Molecular Carcinogenesis, Plesmanlaan 121, 1066 CX Amsterdam NL. E-mail: perrrakis@nki.nl*
[d] *Global Phasing Ltd., Sheraton House, Castle Park, Cambridge CB3 0AX, UK*

New procedures are outlined that enable ARP/wARP to automatically build protein models with diffraction data extending to about 2.5 Å. An overview of ongoing research is given and possible future advances are discussed.

Keywords: protein crystallography, model building, ARP/wARP, structural genomics.

1. Introduction

X-ray crystallography has become a routine tool to assist the investigation into biological phenomena by providing the researcher with detailed atomic models of the bio-molecules of interest. Advances in the field of synchrotron radiation enable diffraction data to be recorded at an ever-increasing rate at dedicated beam-lines. The fraction of structures solved using synchrotron data collection is rapidly increasing (Minor *et al*., 2000). More and more emphasis is now given to the need of robust, easy-to-use and efficient software pipelines that would allow rapid, preferably on-site, determination of protein crystal structures (e.g. Brunzelle *et al*., 2002).

There are well-established computational and experimental techniques for recovering the lost phase information in a crystallographic X-ray diffraction experiment. Once initial phase information has been made available, a three-dimensional image of the electron density can be computed. At this stage it is desired to construct a chemically sensible model of the macromolecule, which represents the experimental electron density distribution with a set of labelled atoms and their corresponding coordinates.

For any Structural Genomics project or other high-throughput structure determination initiative to deliver macromolecular models at the expected rates, the traditionally time consuming and labour intensive step of model building has to be made fast, reliable and highly automated. ARP/wARP (Perrakis *et al*., 1999; Lamzin *et al*., 2001; Morris *et al*., 2002) has successfully tackled this problem but with the limitation of requiring high resolution, good quality data. The previous ARP/wARP version 5.1 required data to 2.0 Å (in some cases 2.3 Å was sufficient). In this contribution, we present new procedures and methods that have enabled these conditions to be relaxed in the current software release (version 6.0 from July 2002). Successful model building can now be carried out with diffraction data extending to about 2.5 Å, thus providing a major advance in terms of the number of structures that could move into the reach of auto-building by ARP/wARP. Judging from the PDB (Bernstein *et al*., 1977; Berman *et al*., 2000) statistics, Figure 1, ARP/wARP is now applicable to about 75 % of the structures. However, the remaining 25 % are generally harder to solve – both because of limited resolution and their larger average size, Figure 2. Ongoing developments, that are briefly outlined, give promises for future applications at even lower resolution.

1.1 The Automated Refinement Procedure (ARP)

The basic idea of ARP is to couple density interpretation with refinement of the atomic parameters (Lamzin & Wilson, 1993; Lamzin & Wilson, 1997). The approach of allowing the macromolecular model to consist only of what is found in the electron density map and to provide the flexibility of having atoms removed or added to account for density features that emerge in the course of refinement, proved an extremely powerful tool in overcoming limitations of convergence radius of, especially pre-likelihood, refinement programs. Whereas conventional refinement might attempt and fail to move a restrained atom over an energy barrier to a new position, ARP would simply take the atom out of the model, thereby ignoring the restraints, and then place it back somewhere else.

Significant phase improvement has furthermore been obtained by the wARP concept (Perrakis *et al*., 1997), where several different atomic models were used to interpret the same electron density. Subsequent refinement of these models and combination of phases of the individual models, resulted in maps of higher quality compared to maps calculated from the individual contributors. Although this method has been largely superseded by the novel model building

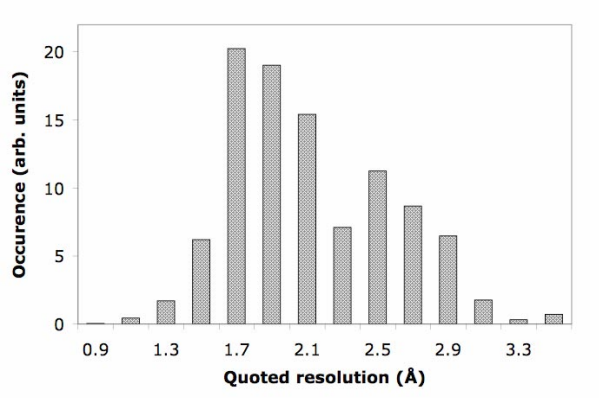

Figure 1

Distribution of quoted resolution for macromolecular entries in the PDB.

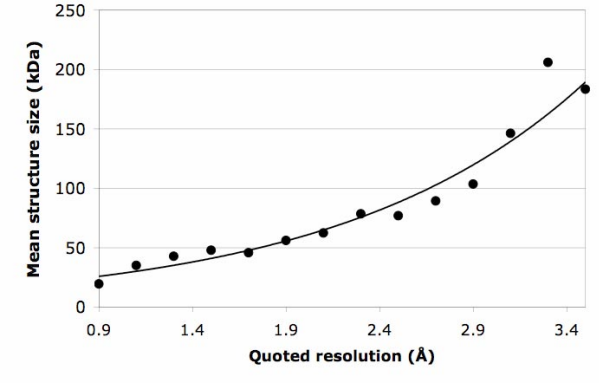

Figure 2

Average molecular mass of a structure as a function of resolution in the PDB. An exponential trend line is also shown.

routines, the wARP principle is now being reconsidered in the light of advances in Bayesian statistical methods as outlined by Read (2001).

1.2 ARP combined with pattern recognition and refinement automates model building

An ARP model consists of a set of atoms that approximately reproduce the density calculated from the measured structure factor amplitudes and the current set of phase estimates. There are no atom type or bond assignments and therefore no stereo-chemical restraints (Konnert & Hendrickson, 1980) on these atoms. Following a nomenclature suggested by Isaacs & Agarwal (1985) we therefore refer to them as "free atoms". To go from this set of free atoms to a valid chemical model of the molecule requires a further layer of interpretation. For accurately placed atoms, this chemical interpretation reduces to an exercise in connecting points to produce well-known covalent geometry. However, the initial phases are often of poor quality and substantial improvement of these phases is required to allow the atom placement to be sufficiently accurate for automated model building. ARP/wARP aims at providing the best possible density for model building and this goal is achieved *via* coupling of model building with restrained refinement. The central concept constitutes the use of the *hybrid* models. A hybrid model contains a proportion of atoms belonging to a protein structure (providing stereo-chemical restraints) and free atoms in the area of the density where a stereo-chemically sound model has not yet been built. If the density is not of sufficient quality to immediately build a full protein model, the correspondence between atomic positions and phases is employed: the improvement of atomic parameters in real space will provide better phase estimates in reciprocal space. If even small parts of the macromolecular model can be correctly built, subsequent refinement utilising restraints will give rise to phases that produce an improved electron density map. This iterative procedure is outlined in Perrakis *et al*. (1999).

1.3 ARP/wARP as a versatile model building package

ARP/wARP is a software package for the interpretation of electron density maps and automated protein model building combined with refinement. It is not a standalone package but rather a large set of numerical routines for density analysis, density features extraction and object classification through probabilistic reasoning. Moreover it encompasses a data management layer in the form of utilities, scripts and, more recently, a GUI. ARP/wARP depends on the use of a number of CCP4 programs (Collaborative Computational Project, Number 4, 1994) including the state-of-the-art program REFMAC (Murshudov *et al*., 1997) for maximum likelihood refinement.

2. The ARP/wARP Version 6.0

The latest release of ARP/wARP, Version 6.0, shows a number of key improvements over previous versions. In this section the major developments will be highlighted.

2.1 Graphical User Interface

The ARP/wARP interface has been designed based on the CCP4i principles and libraries. The main feature of the interface is what can be called "single button tracing". All the user has to do is to provide essential input information (the file with the structure factors, the size of the protein in amino-acid and – optionally – the sequence) and press the *run* button. There is a simple choice between diverse protocols. Although the system has sensible defaults, almost all parameters are customisable. It is a central point in the philosophy of the interface that nearly user-free operation must be possible, while retaining all the functionality that is desired by an expert user. All output files are accessible from the main interface window. Graph files can be visualised.

2.2 Data Analysis

Prior to launching computationally intense automated building modules, ARP/wARP now executes minimal checks on the general quality of the diffraction data. The Wilson plot of the dataset provides a simple but powerful means to identify clearly suspicious data. We have implemented an "expected" Wilson plot derived by Popov & Bourenkov (2003) from 72 randomly chosen, good quality data sets collected at EMBL Hamburg and Max Planck beamlines at DESY. A related representation, constructed using normalised structure factor amplitudes, has recently been given by Morris & Bricogne (2003). ARP/wARP checks the agreement between the observed and the expected Wilson plot and reports strong deviations. Figure 3 shows the leishmanolysin example (courtesy of Peter Metcalf, distributed within the ARP/wARP package as "psp") where the lowest resolution shell has a mean intensity of 70 % compared to the expected value – possibly due to missing a few strong, overloaded reflections. Model building (with the same default parameters) against all data resulted in 453 residues (out of 475) in 7 fragments. When the data from the inner shell were excluded, the tracing produced 467 residues in 3 fragments. More tests are needed to conclude whether a truncation of poorly measured data is beneficial for a general case, but an inspection of the Wilson plot seems to be a good thing to do.

2.3 Graph searching strategies enhance main chain tracing

In cases where the initial electron density map does not allow all atoms to be placed with confidence and accuracy, a decision-making process during the model building becomes necessary to "guess" the most likely interpretation. This situation can be shown by randomising the atomic coordinates of a refined structure and attempting to find the original connectivity between the atoms. Based only on local bonding geometry this becomes increasingly difficult with the inaccuracy of the atomic positions as many pairs of originally non-bonded atoms fall within valid bonded distance limits. In such situations, one has to rely on the experience of a crystallographer and interactive graphics software to find the most

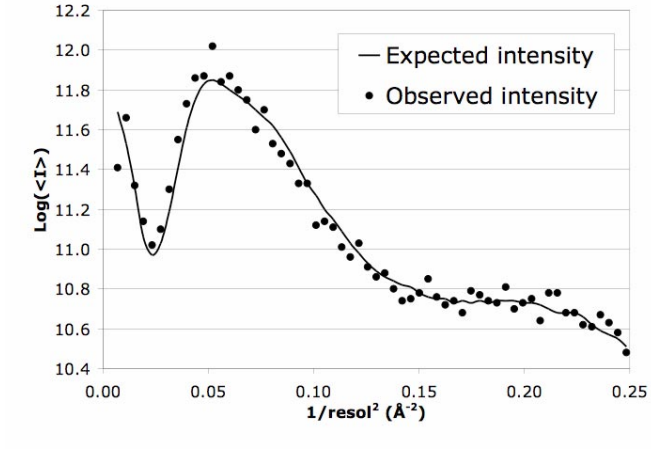

Figure 3

Observed and expected intensities as a function of the resolution for the *psp* example.

plausible solution or to attempt to formulate some heuristic that mimics this process. In Morris *et al*. (2002) we demonstrated that this problem may be formulated as a constrained integer optimisation problem and outlined efficient graph searching techniques similar to well-established methods of Artificial Intelligence to find good approximations to an NP-hard problem. In short, a density-weighted match between found and expected protein Cα geometry is computed and the best set of highest scoring main chain fragments is sought. The search technique is a modified depth-limited search algorithm (Russel & Norvig, 1995). An implementation of this method has been shown to cope better with inaccurate free atom positions.

2.4 Sequence docking and side chain building

For the sequence docking, a feature vector is used that represents the possible connectivity between the free atoms in the vicinity of each Cα atom. Each observed feature vector is compared to all twenty possible side chain connectivity vectors. This way each main chain fragment is represented as an array of probability vectors and slid across the given protein sequence. After the fragment with the best confidence score is docked, each of its residue is assigned to a specific side chain. For defining rotamers ARP/wARP uses the "Penultimate Rotamer Library" (Lovell *et al*., 2000). We express the density of a side chain as a function of its torsion angles and restrained coordinates of the Cα position. A non-derivative minimisation method is used for real space torsion angle refinement.
Sequence docking and side chain building are incorporated within the automated scripts. This procedure is typically carried out during the last cycles of auto-building but is also available as a standalone application. A limitation of the side chain module is its inability to automatically handle non-crystallographic symmetry (NCS).

2.5 Examples

The capabilities of ARP/wARP have recently been reviewed by Badger (2003). Examples were mentioned where the use of the default ARP/wARP parameters gave reasonable results.
The ARP/wARP scripts are meant to take care of most standard cases in the best manner, but adding extra knowledge and fine-tuning various parameters can enhance the auto-building capabilities. The following examples, Table 1, were taken from the autoSHARP structure solution software suite, version 3.0.16 (Vonrhein & Bricogne, 2003), which used case-dependent adaptation of the ARP/wARP scripts. Solvent flipping was carried out by SOLOMON (Abrahams & Leslie, 1996). The structure of RNAse at 2.5 Å, starting from good phases, was traced with ARP/wARP straightforwardly. 96 % of the main chain was built and all side chains were constructed and fit into the density. The structure of GerE at 2.66 Å is rather an exotic example. Though being well beyond the resolution limit to which ARP/wARP 6.0 was designed to work, a good half of the main chain (but no side chains) was automatically constructed.

Table 1
Examples of structure determination with autoSHARP and ARP/wARP at medium resolution (MCC is the map correlation to the map calculated from the final model)

	RNAse (CCP4 provided example)	GerE (CCP4 provided example)
Highest resolution	Native 2.5 Å	Low energy remote 2.7 Å
Wilson B factor (Å^2)	30	53
Initial phases	MIRAS with 3 derivatives	Se MAD at 4 wavelengths
FOM after SHARP / solvent flipping	0.66 / 0.92	0.55 / 0.92
Residues traced with ARP/wARP	185 in 4 fragments	254 in 15 fragments
Total number of Residues	192	444

3. Ongoing developments

3.1 Bayesian approach for side chain placement

The sequence docking and side chain fitting algorithms have been implemented again in a series of new object oriented software modules (Cohen *et al*., 2003). Even in favourable cases of automated model building ARP/wARP rarely succeeds in tracing a single continuous main chain, but rather identifies a set of main chain fragments, typically a few longer stretches of tens of residues and many other pieces of shorter length. A sliding algorithm is employed to assign the main chain fragments to the correct place in the protein sequence space and to assemble a macromolecular model. This procedure is now performed in a probabilistic manner. This new module includes an efficient scheme for real space refinement that combines an exhaustive, discrete global search followed by a continuous local minimisation and also handles NCS.

3.2 Exploitation of non-crystallographic symmetry

The degree to which ARP/wARP traces the main chain may change according to the quality of density in a particular area. In cases where NCS is present, this additional information may be used to deliver more complete models by extending polypeptide fragments generated during the main-chain tracing step. This could be particularly valuable in cases of similar conformation of the NCS-related fragments but different quality of the electron density resulting from e.g. poor phases, model bias or disorder. Powerful methods to detect structural similarities are available (e.g. Levitt & Gerstein, 1998; Singh & Brutlag, 1997) which exploit orientation-independent distance and/or vector-based measures coupled with suitable scoring schemes. Modifications of these methods are currently being investigated for identification of putative NCS relations in order to provide a robust way of exploiting the geometrical redundancy during main chain tracing.

3.3 Utilisation of the secondary structure

Identification of secondary structural elements in an electron density map based on prior knowledge of their motifs and stereo-chemistry should considerably enhance model building in general and, particularly, provide the extension to the lower (e.g. 3.0 Å) resolution of the X-ray data. We currently exploit a discriminant analysis pattern recognition technique for location of helical fragments. A helical structural motif is parameterised by a set of small overlapping fragments, which fulfil a number of stereo-chemical conditions including interatomic distances, valence and dihedral angles. Extension to the location of β-stranded fragments is currently being investigated.

3.4 Estimation of coordinate error

We have developed a novel approach for estimation of the positional error on a set of ARP/wARP free atoms. The method utilises known geometrical features of protein models and estimates the parameters of the error model on the basis of derived generalised Rice distributions for erroneous positions. The obtained coordinate error correlates well with the map quality. Both the real and reciprocal space variants are being implemented (Zwart & Lamzin, 2003a).

3.5 Building of non-protein fragments

Using an approach that resembles the Conditional Dynamics proposed by Scheres & Gros (2002) and the formalisms used in the coordinate error estimation procedure, a good prediction of the possible chemical nature of a particular area in the unit cell can be obtained (Zwart & Lamzin, 2003b). Although the procedure is being designed for the automatic recognition and building of bound ligands and small molecular fragments, its extension to an interpretation of parts of protein structure is trivial.

4. Discussion

The ARP/wARP software suite (© European Molecular Biology Laboratory) is based on the paradigm of treating model building and refinement as one unified procedure for optimising phase estimates. ARP/wARP is routinely used to automatically build protein models in experimental maps, where good quality data are available to sufficient resolution. It has proved to be a powerful tool for removing bias and subsequent model building starting from molecular replacement solutions. ARP/wARP has been a key part of many structure solution pipelines in both academic and industrial laboratories.

The current release, ARP/wARP Version 6.0, works with density driven procedures for placing and removing atoms but is complemented by geometrical pattern recognition algorithms for the model building steps. This has allowed for the limit of applicability for diffraction data resolution to be extended to about 2.5 Å. Iterative cycles of density modelling by placing atoms, unrestrained refinement of their parameters, automated model building and restrained refinement of the built fragments provide a powerful means of phase refinement and produce an almost complete protein model. Initial phase estimates may be provided in the form of a molecular replacement solution, MIR/SIRAS/MAD/SAD phases or other experimental measurements. Pattern recognition techniques play a key role and the development of more robust algorithms for medium resolution is currently underway. ARP/wARP is an experimental hypothesis-generating and testing procedure for placing atoms in the most likely positions, using artificial intelligence search techniques combined with geometrical comparisons against stereo-chemical expectation values obtained from the PDB to construct the polypeptide chain. The iterative approach with maximum likelihood refinement of the current model at every stage has proved to be powerful tool for overcoming the insufficient robustness of the map interpretation routines for poor phases.

The authors thank A. Popov and G. Bourenkov for provision of the empirical Wilson plot and G. Murshudov for fruitful discussions.

References

Abrahams, J.P. & Leslie, A.W.G. (1996). *Acta Cryst*. **D52**, 30-42.
Badger, J. (2003). *Acta Cryst.* **D59**, 823-827
Brunzelle, J.S., Shafaee, P., Weigand, S., Yang, X., Ren, Z. & Anderson, W.F. (2002). In Proceedings of *International Conference of Structural Genomics*, Berlin, October 2002, pp. 75.
Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, Jr., E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shomanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535-542.
Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N. & Bourne, P.E. (2000). *Nucleic Acids Research* **28**, 235-242.
Cohen, S., Morris, R.J., Lamzin, V.S. & Perrakis, A. (2003). In preparation.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst*. **D50**, 760-764.
Cowtan, K. (1998). *Acta Cryst*. **D54**, 750-756.
Isaacs, N.W. & Agarwal, R.C. (1985). In *Meth. Enzymol*. **115**, 112-117.
Konnert, J.H. & Hendrickson, W.A. (1980). *Acta Cryst*. **A36**, 344-350.
Lamzin, V.S. & Wilson, K.S. (1993). *Acta Cryst.* **D49**, 129-149.
Lamzin, V.S. & Wilson, K.S. (1997). *Meth. Enzymol*. **277**, 269-305.
Lamzin, V.S., Perrakis, A. & Wilson, K.S. (2001). International Tables for Crystallography, Crystallography of Biological Macromolecules, edited by M. Rossmann & E. Arnold, pp. 720-722, Dordrecht: Kluwer Academic Publishers.
Levitt, M & Gerstein, M. (1998). *Proc. Natl. Acad. Sci*. **95**, 5913-5920.
Lovell, S.C, Word, J.M., Richardson, J.S., and Richardson D.C. (2000). *Proteins: Structure, Function and Genetics* **40**, 389-408.
Minor, W., Tomchick, D.R. & Otwinowski, Z. (2000). *Structure* **8**, R105-R110.
Morris, R. J., Perrakis, A. & Lamzin, V.S (2002). *Acta Cryst*. **D58**, 968-975.
Morris, R. J., & Bricogne, G. (2003). *Acta Cryst.* **D59**, 615-617.
Murshudov, G.N., Vagin, A.A. & Dodson, E.J. (1997). *Acta Cryst*. **D53**, 240-255.
Perrakis, A., Sixma, T.K., Wilson, K.S. & Lamzin, V.S. (1997), *Acta Crys* **D53**, 448-455.
Perrakis, A., Morris, R.J. & Lamzin, V.S. (1999). *Nature Structural Biology* **6**, 458-463.
Popov, A. & Bourenkov, G. (2003). *Acta Cryst. Section D, in press.*
Read, R.J. (2001). Acta Cryst. **D57**, 1373-1382
Russel, S. & Norvig, P. (1995). Artificial Intelligence. pp. 77-80, Prentice Hall, ISBN 0-13-360124-2.
Scheres, S.H.W. & Gros, P. (2001). *Acta Cryst.* **D57**, 1820-1828
Singh, A.P. & Brutlag, D.L. (1997). In the proceedings of the *Fifth International Conference on Intelligent Systems for Molecular Biology*, pp. 284-293. Menlo Park, Calif: AAAI Press.
Vonrhein, C. & Bricogne, G. (2003). In preparation.
Zwart, P.H. & Lamzin, V.S. (2003a). In preparation.
Zwart, P.H. & Lamzin, V.S. (2003b). In preparation.