Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Ab initio phasing of a 4189-atom protein structure at 1.2 Å resolution

Jeremy R. H. Tame

Protonic Nanomachine Project, ERATO, Kyoto 619-0237, Japan

Correspondence e-mail: jtame@npn.jst.go.jp

Received 27 April 2000
Accepted 30 August 2000

The phase problem remains a key rate-limiting step in the determination of macromolecular X-ray structures. Direct methods, applying probability theory to the native data set, can routinely solve structures of up to about 200 non-H atoms, although much larger structures have been solved given sufficiently high resolution data and the presence of heavy atoms. Here it is shown that maximum-likelihood refinement of free-atom models with *ARP*/*wARP* can solve *ab initio* a much larger metalloprotein structure than the largest so far solved by conventional direct methods. The protein, OppA, is not naturally associated with metal ions but was co-crystallized with uranium.

1. Introduction

Despite considerable research into methods for phasing X-ray crystal structures *ab initio*, conventional direct methods continue to struggle with macromolecular structures larger than a few hundred non-H atoms (Uson & Sheldrick, 1999; Hauptman, 1997).[1] The most successful programs are *Shake-and-Bake* (Weeks & Miller, 1998) and *SHELXD* (Sheldrick, 1998) which uses a similar dual-space iteration method. Currently, only three structures larger then 1000 non-H atoms have been solved from native data alone. Hen egg-white lysozyme (1001 non-H atoms) was solved in 1997, and high-potential iron protein HiPIP (1264 atoms) and cytochrome c_3 (2024 atoms) were solved in 1999 (Deacon *et al.*, 1998; Parisini *et al.*, 1999; Frazao *et al.*, 1999). Only a few years earlier, structures of around 400 atoms represented significant challenges (Weeks *et al.*, 1995; Schafer *et al.*, 1996). Phasing is rather easier if heavy atoms are present in the structure, as these can be located by Patterson or simple search methods and provide useful phase information. The presence of eight Fe atoms in cytochrome c_3 was essential in the *ab initio* determination of the structure. Recent developments in the *ARP*/*wARP* package have allowed spectacular improvements in phasing proteins (Perrakis *et al.*, 1997, 1999), which prompted me to try the direct phasing of a 4189-atom (59 kDa) protein complex, the oligopeptide-binding protein OppA bound to lysyl-alanyl-lysine and several uranyl ions. OppA is a periplasmic binding protein involved in bacterial uptake of short peptides and is the largest known member of the binding-protein family (Hiles *et al.*, 1987). The protein was crystallized with uranyl acetate, leading to a crystal lattice held together by uranyl ions bonding to neighbouring protein

[1] The meaning of the term '*ab initio*' has been the subject of much discussion. The term is defined by Uson & Sheldrick (1999).

Table 1
The completeness of the data in resolution shells.

Overall, 162 955 reflections were measured of a possible 182 213 (89.4%). $I/\sigma(I)$ in the outermost resolution shell (1.20–1.21 Å) was 3.11. The overall temperature factor (B) of the data set is 10.7 Å^2. The crystals are in space group $P2_12_12_1$, with unit-cell parameters $a = 109.19$, $b = 76.04$, $c = 70.28$ Å. The asymmetric unit contains one copy of mature OppA from *Salmonella typhimurium* (SwissProt entry P06202) and eight uranyl (VI) ions. 572 solvent sites were identified in the model deposited in PDB in 1996 (1jet). The overall ratio of reflections to parameters (x, y, z, B) is approximately 8.5 given this level of solvation.

High-resolution limit (Å)	No. possible reflections	No. observed reflections	Observed (%)
3.23	9422	9383	99.6
2.35	15225	14569	95.7
1.94	19256	17543	91.1
1.68	22491	20850	92.7
1.51	25352	23453	92.5
1.38	27841	24935	89.6
1.28	30247	26353	87.1
1.20	32379	25869	79.9

Table 2
Phasing OppA using *ARP/wARP* and *SHELX*.

The phase problem is essentially solved by the *warp_solve.sh* step using the default parameters for phasing from heavy atoms alone. This step is the most computationally demanding and required 36 h on a 300 MHz Pentium PC running Linux (Red Hat 6.1). Different runs produced very similar models with a final R factor of around 23%. The numbers shown here are typical. The map produced is remarkably clear and side chains are instantly recognisable. Several cycles of *warpNtrace* give a clear indication of the polypeptide chain, but do not reduce the R factor significantly. The final R factor for the model 1jet in the PDB is 22.9%, close to that reached by *warp_solve*. The relatively high R factor is a consequence of disorder and anisotropic movement of the uranium ions.

Starting model	Step
5 uranium ions (from *SHELX*)	*warp_solve*
5 uranium ions, 5024 water molecules	*warpNtrace*
5 uranium ions, 2840 protein atoms, 1788 water molecules	*side_dock*
5 uranium ions, 475 residues, 995 water molecules	

molecules. The structure was solved in 1994 using a combination of MAD phasing in one crystal form, partly soaking out weakly occupied uranium sites in another form and then averaging the two (Tame *et al.*, 1994). The protein was the first case of treating MAD data as a special case of MIR (Glover *et al.*, 1993), although MAD data good enough to phase the protein directly were only collected after adequate phasing for solving the structure had already been obtained (Glover *et al.*, 1995). The strong lattice contacts lead to highly ordered crystals which diffract to atomic resolution. The structure is readily soluble with *SHELX* and *ARP/wARP* using the native data extending to 1.2 Å resolution. *SHELX* is required only to find the uranyl sites, which provide enough phase information for *ARP/wARP* to calculate almost perfect phases for the structure, which is more than twice the size of cytochrome c_3, the largest structure solved *ab initio* to date.

2. Methods

OppA complexed with the tripeptide Lys-Ala-Lys was crystallized by the hanging-drop method at 293 K using 10% PEG 4000, 50 mM sodium acetate pH 5.5, 1 mM uranyl acetate. X-ray data were collected at Daresbury SRS PX 9.6 from a single crystal cryocooled to 120 K (Tame *et al.*, 1996). The data were processed with *DENZO* (Otwinowski & Minor, 1997) and reduced to structure factors with the *CCP*4 suite (Collaborative Computational Project, Number 4, 1994). The

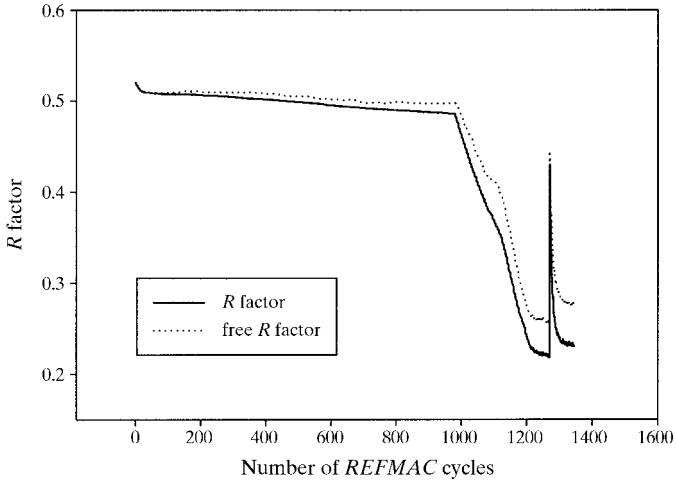

Figure 1
A plot of the R factor and free R factor of the model against *REFMAC* cycle number running *warp_solve.sh*. In this particular experiment, the five refinement cycles required 975, 145, 75, 75 and 75 *REFMAC* cycles. The number of atoms in the models output by each cycle was 200, 1948, 3956, 4681 and 4954. The sharp jump in R factor corresponds to the start of the last cycle, the model being 'shaken' at this point.

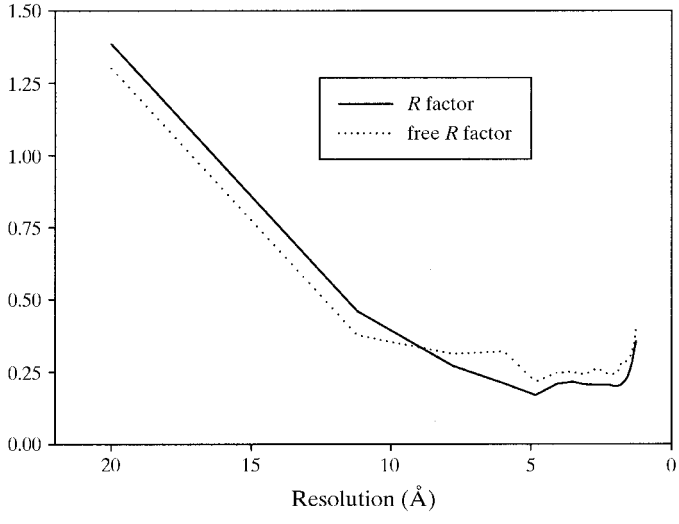

Figure 2
A plot of R factor *versus* resolution for the model produced by *warp-solve.sh*.

uranium positions were found by Patterson search using *SHELXS* with squared structure factors. Anomalous data were not used. The fractional coordinates were converted to PDB format and placed in a file 'heavy.warp'. In the *warp_solve* step, five refinement cycles were carried out, each of 15 iterations. Multiple model averaging was not used; all jobs were performed with a single processor (Pentium PC running Linux or Silicon Graphics R10000). Given the quality of the model produced by *warp_solve*, many cycles of *warpNtrace* were not necessary. A total of 100 cycles were run with ten cycles between rebuilding, however, in order to try to connect the chain as much as possible. Maximum-likelihood refinement using the *CCP*4 program *REFMAC* (Murshudov *et al.*, 1997) was used throughout. Initially, default settings were used which switched off the use of $R_{\rm free}$. Excellent phases could also be determined using 5% of the reflections to calculate $R_{\rm free}$ throughout the *warp_solve* procedure.

3. Results and discussion

OppA is currently the second largest structure in the PDB refined to 1.2 Å or better (PDB code 1jet). The X-ray data are summarized in Table 1. Using the 1.2 Å native data, five uranyl positions can be found readily with *SHELX*. Eight or more uranyl ions are found on refining the structure (currently, over 30 OppA coordinate sets have been deposited in the PDB), but only five of these appear fully occupied, with low temperature factors. The *SHELX*-derived uranium positions were fed into the shell script *warp_solve.sh*, part of the *ARP*/*wARP* package. This has been written (in the words of the document) to 'take care of everything'. In the case of OppA, it produces a very good model for *warpNtrace* to derive excellent phases with no manual intervention whatsoever.

The complete procedure is outlined in Table 2. The starting model of five U atoms (each given an occupancy of 0.8 and a temperature factor of 11) has an *R* factor of 52.1% and a free *R* factor of 52.0%. The mean phase error is 90.1°. *warp_solve.sh* gave a model (consisting of 5024 water molecules and the original U atoms only) with an overall *R* factor of 23.0% and free *R* factor of 27.6% (Figs. 1 and 2). In the initial experiment, the heavy-atom positions used were (by chance) the wrong hand. Running *warpNtrace* therefore gave a beautiful map with clear left-handed α-helices of glycine residues, but otherwise extremely good though inverted

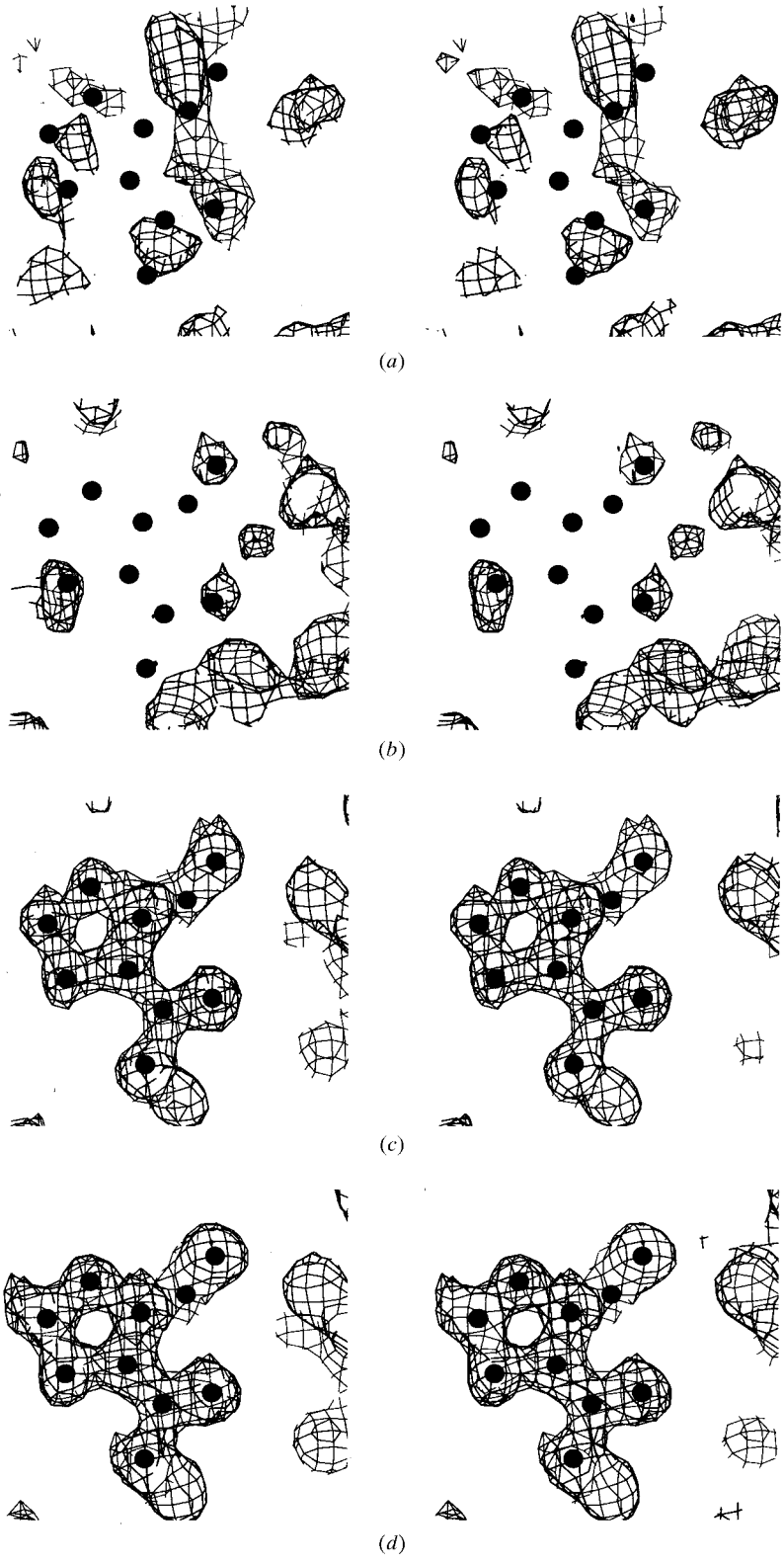

Figure 3
Stereoviews of the electron-density maps at various stages of phasing and the atomic model generated by *warpNtrace*. The black spheres are pseudo-water molecules built into the density by *ARP*/*wARP*, seen from an identical view in each panel. The initial map, calculated from phases from the five uranium positions alone, is shown in (*a*) and (*e*). (*b*) and (*f*) show the map after the first cycle (975 rounds of *REFMAC*). 145 additional rounds of *REFMAC* clearly improve the map greatly in places, (*c*) and (*g*). A proline and a tryptophan residue can be seen very clearly in (*d*) and (*h*), respectively, showing the map produced by *warpNtrace*.

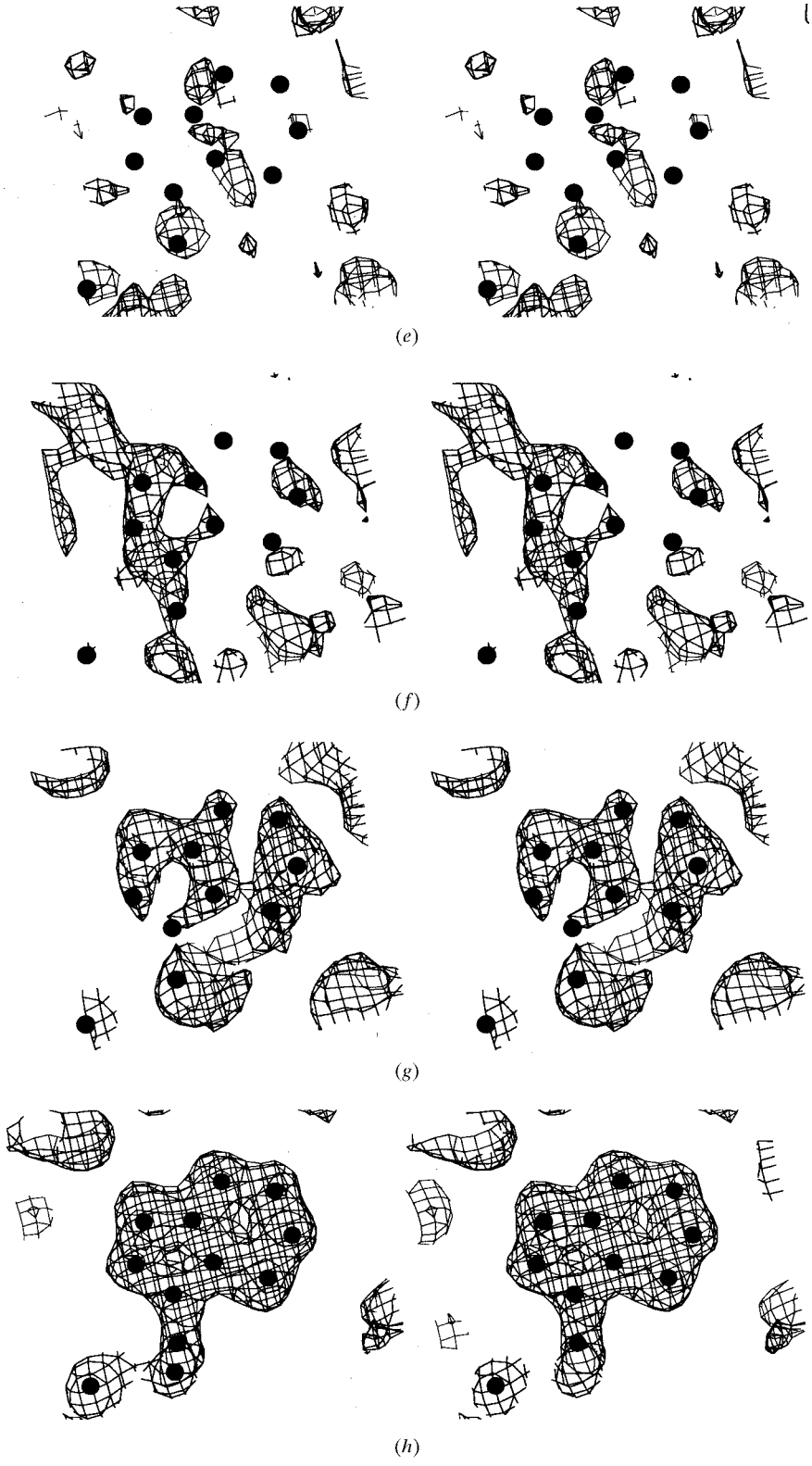

Figure 3 (continued)
The program did not recognize these residues, as the starting model was in the wrong hand and therefore all the atomic positions of the derived model are inverted through the origin. Changing hand can be carried out either by inverting the uranium positions and beginning again with *warp_solve* or (more quickly) inverting the positions of the output model before running *warpNtrace*. The quality of the phases is apparent from the positioning of water molecules close to the expected positions of C and N atoms in the side chains and the holes found in some proline residues. Electron density is contoured at 1σ for each map. This figure was produced with *BOBSCRIPT* (Esnouf, 1997).

phases. The map shows prolines and aromatic side chains with clear holes in the rings and all side chains are immediately recognisable, though in the D rather than L configuration (Fig. 3). To build a model of the correct hand, the coordinates from *warp_solve.sh* were inverted through the origin and *warpNtrace* was rerun. The script could build 444 of the 517 residues immediately, broken into 18 chains and with a connectivity index of 0.92. The final 1.2 Å map has a correlation coefficient of 0.9359 with the map derived from the deposited coordinates (1jet), the mean phase error being 21.1° (Fig. 4). To show that the correct phases could be obtained directly, the uranium positions derived from *SHELX* were also inverted through the origin and *warp_solve* rerun, again giving an excellent model, this time with an *R* factor of 23.6% and containing 4945 water molecules.

The *ARP/wARP* package also allows very rapid fitting of side chains using the script *side_dock.sh* (Perrakis *et al.*, 1999). In this case, *side_dock.sh* was able to build almost the entire structure (475 residues) across two asymmetric units, with nine breaks in the polypeptide chain. These breaks correspond to the positions of uranyl ions which cause large ripples in the map. A stereo figure of autobuild_all.brk created by *side_dock.sh* is shown in Fig. 5. Moving two chains by simple crystallographic symmetry to the same asymmetric unit as the other eight allowed the r.m.s. deviation from the PDB model 1jet to be calculated with *LSQKAB* (Kabsch, 1976; Collaborative Computational Project, Number 4, 1994). The r.m.s. deviation from 1jet of the 475 C^{α} atoms is 0.687 Å; over all the main-chain atoms of these residues it is 0.644 Å (with an average displacement of 0.206 Å). Other than building in the missing 42 residues, almost no manual adjustments would be required to finish with a model with geometry and *R* factor comparable with the structure refined much more laboriously which was deposited in the PDB in 1996. The Ramachandran plot of the automatically generated model is shown in Fig. 6.

As noted in the paper discussing the 1.4 Å resolution refinement of OppA (Tame *et al.*, 1995), co-crystallization with

metals is perhaps an underused technique in protein crystallization. While the results may not always be as successful as in this case, the rapid phasing and refinement of a structure incorporating heavy metals may prove an incentive in the future for crystal screens including such atoms.

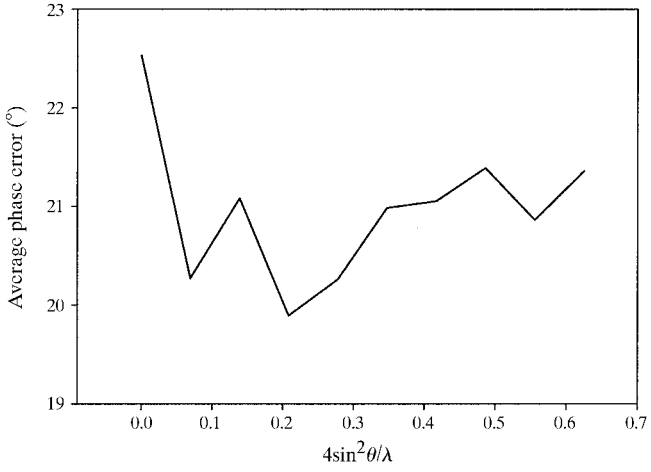

Figure 4
A plot of phase error (weighted by figure of merit) of the model produced by *warpNtrace*. This plot was produced using *PHISTATS*, part of the *CCP*4 package.

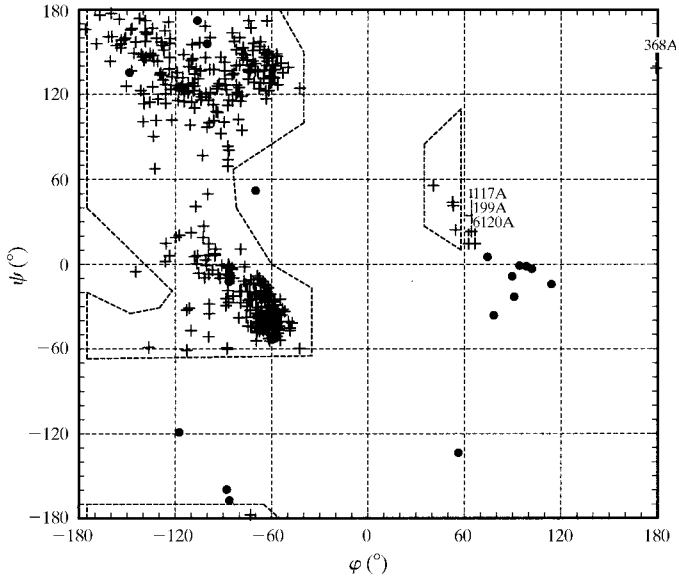

Figure 6
The Ramachandran plot of the model built automatically by *side_dock.sh*. This figure was produced with *Xtalview* (McRee, 1999).

The author thanks the Royal Society for a University Research Fellowship.

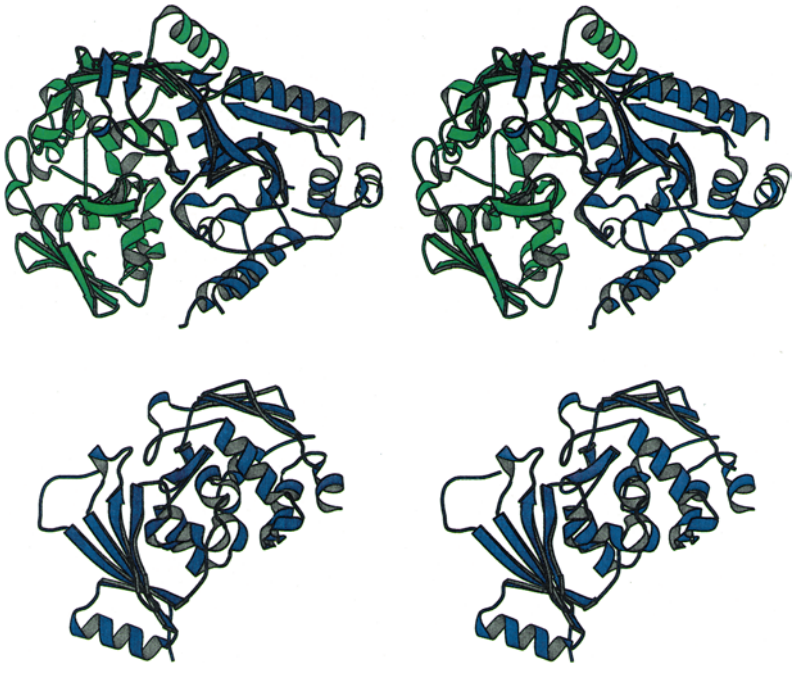

Figure 5
A stereoview of the overall final polypeptide model from *ARP/wARP*, autobuild_all.brk. Water molecules have been omitted for clarity. The model consists of ten polypeptide chains, shown in blue, broken at points close to U atoms in the structure. Two chains were built into a neighbouring molecule. The symmetry equivalents of these two chains (related by $\frac{1}{2} - x, -y, -\frac{1}{2} + z$) are shown in green, forming an almost complete monomer with the other eight chains. This figure was created with *MOLSCRIPT* (Kraulis, 1991). Secondary-structure assignments were made automatically using *MOLAUTO*, part of the *MOLSCRIPT* package.

References

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Deacon, A. M., Weeks, C. M., Miller, R. & Ealick, S. E. (1998). *Proc. Natl Acad. Sci. USA*, **95**, 9284–9289.
Esnouf, R. M. (1997). *J. Mol. Graph.* **15**, 132–134.
Frazao, C., Sieker, L., Sheldrick, G. M., Lamzin, V. S., LeGall, J. & Carrondo, M. A. (1999). *J. Biol. Inorg. Chem.* **4**, 162–165.
Glover, I. D., Denny, R. C., Nguti, N. D., McSweeney, S. M., Kinder, S. H., Thompson, A. W., Dodson, E. J., Wilkinson, A. J. & Tame, J. R. H. (1995). *Acta Cryst.* D**51**, 39–47.
Glover, I. D., Denny, R. C., Nguti, N. D., McSweeney, S. M., Thompson, A. W., Dodson, E. J., Wilkinson, A. J. & Tame, J. R. H. (1993). *Jnt CCP4/ESF–EACBM Newslett. Protein Crystallogr.* **29**.
Hauptman, H. A. (1997). *Curr. Opin. Struct. Biol.* **7**, 672–680.
Hiles, I. D., Gallagher, M. P., Jamieson, D. J. & Higgins, C. F. (1987). *J. Mol. Biol.* **195**, 125–142.
Kabsch, W. (1976). *Acta Cryst.* A**32**, 922–923.
Kraulis, P. J. (1991). *J. Appl. Cryst.* **24**, 946–950.
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.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Parisini, E., Capozzi, F., Lubini, P., Lamzin, V. S., Luchinat, C. & Sheldrick, G. M. (1999). *Acta Cryst.* D**55**, 1773–1784.
Perrakis, A., Morris, R. J. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.
Perrakis, A., Sixma, T. A., Wilson, K. S. & Lamzin, V. S. (1997). *Acta Cryst.* D**53**, 448–455.

Schafer, M., Schneider, T. R. & Sheldrick, G. M. (1996). *Structure*, **4**, 1509–1515.
Sheldrick, G. M. (1998). *Direct Methods for Solving Macromolecular Structures*, edited by S. Fortier, pp. 401–411. Dordrecht: Kluwer.
Tame, J. R. H., Dodson, E. J., Murshudov, G., Higgins, C. F. & Wilkinson, A. J. (1995). *Structure*, **3**, 1395–1406.
Tame, J. R. H., Murshudov, G., Dodson, E. J., Neil, T. K., Dodson, G. G., Higgins, C. F. & Wilkinson, A. J. (1994). *Science*, **264**, 1578–1581.
Tame, J. R. H., Sleigh, S., Wilkinson, A. J. & Ladbury, J. E. (1996). *Nature Struct. Biol.* **3**, 998–1001.
Uson, I. & Sheldrick, G. M. (1999). *Curr. Opin. Struct. Biol.* **9**, 643–648.
Weeks, C. M., Hauptman, H. A., Smith, D., Blessing, R. H., Teeter, M. M. & Miller, R. (1995). *Acta Cryst.* D**51**, 33–38.
Weeks, C. M. & Miller, R. J. (1998). *J. Appl. Cryst.* **32**, 120–124.