Acta Cryst. (1993). D**49**, 129–147

Automated Refinement of Protein Models

BY VICTOR S. LAMZIN* AND KEITH S. WILSON

European Molecular Biology Laboratory (EMBL), c/o DESY, Notkestrasse 85, D-2000 Hamburg 52, Germany

(*Received* 1 *June* 1992; *accepted* 18 *August* 1992)

Abstract

An automated refinement procedure (ARP) for protein models is proposed, and its convergence properties discussed. It is comparable to the iterative least-squares minimisation/difference Fourier synthesis approach for small molecules. ARP has been successfully applied to three proteins, and for two of them resulted in models very similar to those obtained by conventional least-squares refinement and rebuilding with *FRODO*. In real time ARP is about ten times faster than conventional refinement. In its present form ARP requires high (2.0 Å or better) resolution data, which should be of high quality and a starting protein model having about 75% of the atoms in roughly the correct position. For the third protein at 2.4 Å resolution, ARP was significantly less powerful but nevertheless gave definite improvement, in the density map at least.

Introduction

Crystallography involves the determination of three-dimensional structure from experimental X-ray diffraction data. The amplitudes of the structure factors are available from the experiments but, unfortunately, this is insufficient for direct calculation of the electron density. The phases are needed for this. A highly simplified flow chart of a crystal structure determination is shown in Fig. 1. In the left-hand column the procedure for small-molecule structures is shown. The phase problem for such structures is generally solved by either direct or Patterson methods. These assume that the amplitudes have been measured to approximately atomic resolution (at least 1.2–1.0 Å). The initial phases are implicitly refined in direct methods, or extended from a partial atomic fragment in direct or Patterson methods. The resulting atomic coordinates are refined by least-squares minimisation of the residuals between the observed and calculated amplitudes. At 1.0 Å resolution the observations exceed the parameters by more than five to one, even with anisotropic atomic temperature factors. The refinement of the model through iterative cycles of least squares, with difference Fourier syntheses to update the model, proceeds essentially automatically through programs such as *SHELX* (Sheldrick, 1976).

In protein crystallography the situation is severely complicated by the lack of atomic resolution data for both structure solution and refinement. The structure of rubredoxin, with an FeS_4 cluster, has been solved recently by direct methods (Sheldrick, Dauter, Wilson, Hope & Sieker, 1993) but data to such high resolution are rarely available for proteins. Hence the use of direct or Patterson methods for *ab initio* phase determination is precluded, as is the use of unrestrained least squares: at resolutions below about 2.5 Å the number of parameters actually exceeds the number of observations even with isotropic atomic temperature factors. The procedures used for proteins are shown in the two right-hand columns of Fig. 1. The phase problem is initially solved by obtaining a model from an electron-density map calculated using multiple isomorphous replacement with anomalous scattering (MIR) phases or from molecular replacement (MR) using a known related structure. Usually the amplitudes computed from this model give relatively poor agreement with the observed structure-factor amplitudes, and corresponding large errors in the starting phases. There follows a difficult and time consuming stage in proceeding from this initial model to the final one.

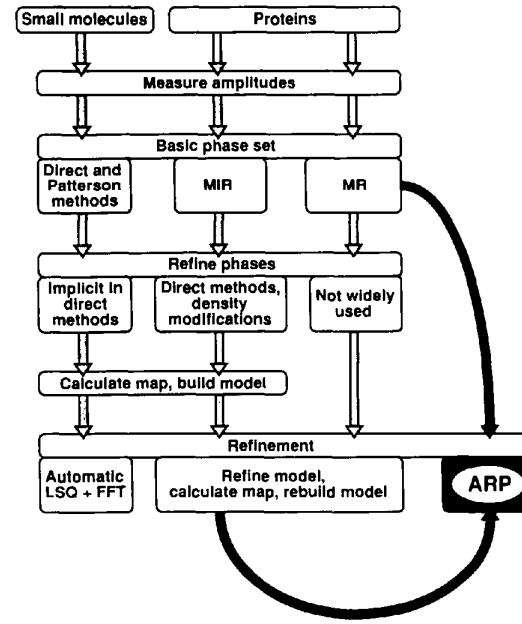


Fig. 1. A highly simplified flow chart of the steps in small-molecule and protein crystallographic analyses.

* Permanent address: Institute of Biochemistry, Russian Academy of Sciences, Leninsky pr. 33, Moscow 117071, Russia.

0907-4449/93/010129-19$06.00

From an MIR model (Fig. 1, central column), the phases can often be improved before interpretation of the map, *inter alia* through density modification including methods such as non-crystallographic symmetry averaging, solvent flattening or histogram matching [reviewed by Podjarny, Bhat & Zwick (1987)]. The map is then interpreted in terms of a (usually) partial atomic model. For MR the model is obtained directly from the results of the rotation and translation function, but is in general chemically and/or spatially different from the protein being studied.

In the standard approach many iterations are required involving manual rebuilding using computer graphics alternating with stereochemically restrained least-squares refinement. A combination of calculated and initial phases may be used in these iterations. The need for the restraints arises from the paucity of observations in almost all protein structure determinations. The X-ray data are complemented by a set of stereochemical restraints based on the known structures of small-molecule models: this effectively increases the number of observations. The method works well in the later stages of refinement. However, the radius of convergence, *e.g.* of the *PROLSQ* package (Hendrickson & Konnert, 1981; Agarwal, 1978; Baker & Dodson, 1980) seems to be at best one-third of the resolution and the program cannot automatically move atoms into new features of the map which are far from the current atomic positions. This necessitates iterative manual rebuilding of the refined model for the large errors inherent in initial protein models.

Attempts to increase the radius of convergence of conventional least-squares minimisation resulted in the incorporation of molecular dynamics. Refinement using simulated annealing [*e.g.* the *X-PLOR* package (Brünger, Kuriyan & Karplus, 1987; Brünger, 1988)] allows atoms to cross barriers between the minima of the multiparameter target function and can sometimes flip, *e.g.* the main-chain carbonyl group, or considerably rotate the side chain. However, simulated annealing cannot move atoms through other atoms (*i.e.* it cannot refine a structure having wrongly traced fragments of polypeptide fold) and requires a large amount of computer time.

In summary, for solving protein structures at less than atomic resolution, the application of direct methods or methods manipulating the electron density for phase improvement does not give phases sufficiently close to the phases calculated from the refined model to produce a density map showing essentially all the atoms of the structure. Restrained least-squares refinement procedures do not provide an automatic way of proceeding to such a final model.

We propose in this paper an automated refinement procedure (ARP) for proteins, as indicated in Fig. 1. ARP is comparable to the iterative least-squares/Fourier synthesis approach for small molecules. In its present form it requires high (ideally 2.0 Å or better) resolution data, which should be of high quality, and a starting protein model which has about 75% of the atoms in roughly the correct position. Three applications of ARP are described and its convergence properties discussed.

Automated refinement procedure

An initial protein model from molecular replacement, or built into an isomorphous density synthesis and preliminarily refined with constraints, can have a large number of regions which need to be substantially corrected. A typical example, where an aspartate side chain and water molecule are incorrectly placed, is shown in Fig. 2(*a*). Such regions increase the *R* factor and make the local geometry worse. The correct structure is shown in Fig. 2(*b*). The question is how to pass from the incorrect to the correct model. Usually such corrections require a large amount of time, especially for big proteins, both in the identification and correction of these sections of the structure.

The solution seems very simple in principle. Assume the side-chain oxygen atom is relabelled as the water, and the water atom as the side-chain oxygen, and some cycles of unrestrained refinement are performed. The *CG* atom which is out of the density should be removed from the model and the peak in positive density identified as a new atom. Thus the solution involves searching of Fourier syntheses for removing atoms, for placing new atoms in roughly the correct position and least squares for the optimisation of their parameters. Such a procedure should be iterated and all stages can be carried out automatically. This is possible if the X-ray data extend to a resolution where the atomic positions can be roughly estimated from the density map. The nominal resolution should be at least 2.0 Å, where features separated by about 1.4 Å (2.0 × 0.7 Å) can be expected to be resolved on the basis of James theorem (James, 1957). There are many possibilities of how the updated model can be extracted from the Fourier syntheses, and only our first attempts are described here. Additional criteria for updating the model can also be envisaged in the future.

ARP is such a combination of least-squares refinement with automatic updating of the model on the basis of the calculated Fourier syntheses. ARP is, in essence, comparable to the procedure used for the refinement of small-molecule structures, where a partial model is developed through alternating rounds of least squares and inspection

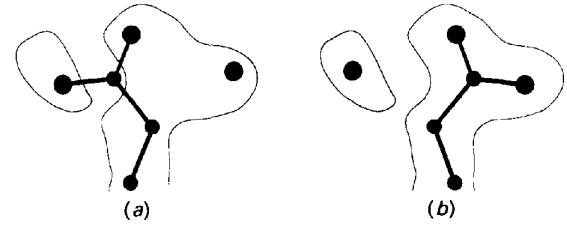


Fig. 2. A schematic representation of a side chain and a water molecule with a rough electron-density contour. (*a*) Incorrect model. (*b*) Correct model.

of difference maps. The use of fast Fourier transforms in both refinement and map calculation allows a comparable procedure to be tractable in terms of computing time for proteins.

Practical implementation

ARP is based on the 'atomicity' of the protein structure. It involves three main steps as shown in Fig. 3. Firstly all atoms in the initial protein model are reset to atoms of the same type, to water for instance. This can be done in different ways, *e.g.* assigning an equal number of electrons and the same temperature factor to each atom, or retaining the number of electrons or temperature factors, or retaining both at the values for the atom in the initial model. The different protocols are described in more detail below. Renaming is not an essential part of the procedure and was applied here for the convenience of using the program *PROLSQ*. In future versions it will not be required.

In the second step the relabelled ARP model is subjected to unrestrained least-squares refinement of the positional and thermal atomic parameters against the X-ray data. The refinement is performed using data in the whole range of resolution from the start.

During the third step the ARP model is updated either after one or after several cycles of refinement (different protocols are shown below). Updating consists of the following.

(i) Calculation of $(3F_o - 2F_c)$ and $(2F_o - 2F_c)$ electron density maps using phases calculated from the model. The difference map is calculated with amplitudes $(2F_o - 2F_c)$ to be on the same scale as the $(3F_o - 2F_c)$ map. The $(3F_o - 2F_c)$ map is used as it represents a point-by-point summation of the F_o map and difference Fourier synthesis, the latter been accorded double weight as is appropriate for the predominantly acentric data (Luzzati, 1953). A grid spacing of approximately one-fifth of the resolution was found to be fine enough.

(ii) Rejection of a small percentage of atoms if they are in low $(3F_o - 2F_c)$ density. Atoms are only considered for rejection if the interpolated value of the electron density at the atomic centre is less than 1σ above the mean density. The percentage of atoms rejected depends on the resolution. If the rejection is performed after several cycles of unrestrained refinement it appears to be reasonable to reject between 1 and 10% of the total number of atoms. However convergence can be achieved more rapidly if the rejection is performed after each cycle of least squares, when only the 0.1-1.0% of atoms in lowest density should be rejected.

(iii) Addition of new atoms found in positive difference density. The grid points where the amplitude of the difference density exceeded the upper value of the density used for removal, and which were not too close to the existing atoms, were analysed. A 'not too close' value of 1.2 Å, slightly less then the average distance between neighbouring protein atoms, was found to be most useful but the absolute value seems to be of low importance. The process of interpretation of a set of grid points at which to add new atoms is simple and may be described as follows. The grid point with highest electron-density value is assigned as a new atom. All grid points located closer to this new atom than the distance specified are now rejected. Various values from 1.2 to 2.5 Å were tried with approximately the same results. However, the minimum distance between new atoms should not be smaller than say 70% of the resolution, otherwise it could lead to undesirable placing of several new atoms in one peak where there should be only one atom. The grid point with the next highest amplitude is now checked and this is iterated until the number of new atoms equals the target specified.

The percentage of atoms to be added depends on the resolution, on the percentage rejected and also on the number of atoms in the initial model. In each of the examples described here the initial coordinate set was incomplete and did not include waters. The best results were obtained if the number of atoms was constantly increased during the refinement to a final value ~120% of the number of expected protein atoms. Usually 10% of the excess atoms corresponded to real water molecules bound to the protein and 10% compensated for lack of electrons in the places where the protein model has sulfur, phosphorus or other heavy atoms, or to imitate pseudo solvent flattening. The rationale of including 10% extra atoms is explained below.

As the percentage of the atoms to be updated is relatively small there is no need to adjust the temperature factors for new atoms to correspond closely to the shape of the electron density. Phase combination is not required because at all stages phases calculated from the atomic model are used without additional phase information from density modification. In all the protocols no weighting was used in calculating the maps.

All steps in ARP can be iterated in a completely automated manner until convergence is achieved. The crystallographic R factor and values of high moments of the $(3F_o - 2F_c)$ map can be used as criteria of convergence. In addition the ARP model can be visually inspected at any point

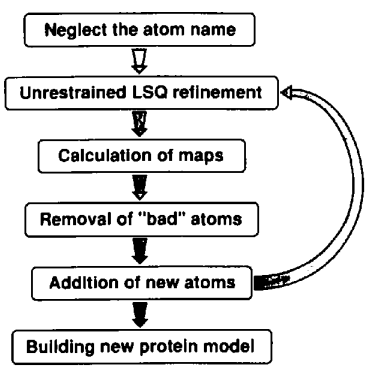


Fig. 3. A flow chart of the automated refinement procedure (ARP).

during the procedure to ensure it continues to resemble the protein in terms of stereochemistry *etc.* Several applications of the procedure showed that if the starting model was good enough and the resolution was higher than 2.0 Å it rapidly converged to an R factor below 20%, close to the value expected after refinement was complete and led to an ARP model similar to the final model. Relabelling of the ARP model to the correct protein atoms can be carried out automatically or manually using computer graphics. A preliminary version of a relabelling program has been developed and is briefly described in the section on formate dehydrogenase. It will be improved in the near future.

Software

For unrestrained refinement of the model and calculation of density maps the *SFKH*, *PROLSQ*, *RSTATS* and fast Fourier transform programs from the *CCP*4 suite (SERC Daresbury Laboratory, 1979) were used. Analysis of electron-density maps, rejection of the atoms located in low density, addition of new atoms according to difference density maps, resetting of the atom name and keeping the model within the asymmetric unit were all performed by the *ARP* program, specially written for this purpose. The user only needs to specify the number of atoms of the ARP model to reject, the number to add, the minimum distance between new and old atoms, the minimum distance between new atoms, and the temperature factor for new atoms. The *ARP* program is in FORTRAN77, does not use computer-specific codes and is compatible with the *CCP*4 map format. The code is available from the authors on request.

Map characteristics

Several characteristics of electron-density syntheses are used in the following sections as criteria for the quality of the models/phases used in their computation. They are briefly summarised here.

The correlation coefficient between two maps is calculated as:

$$\text{Correlation} = (1/N)\Sigma[(\rho_{1i} - \langle\rho_1\rangle)(\rho_{2i} - \langle\rho_2\rangle)]/(\sigma_1 + \sigma_2)$$

where $\langle\rho_1\rangle$ and $\langle\rho_2\rangle$ are the mean densities of the two maps to be correlated, σ_1 and σ_2 are the r.m.s. densities and N is the number of grid points.

Electron-density distributions, in common with other distributions, can be characterised by the values of their moments. The first moment is the mean value ($\langle\rho\rangle$) of the density and the square root of the second moment is the r.m.s. density (σ). Both $\langle\rho\rangle$ and σ have units of e Å^{-3} and do not depend on the phases. The mean density of the maps presented below is 0, as no F_{000} term was included in the Fourier summation. The third and fourth moments, the skewness and kurtosis respectively, are dimensionless pure numbers characterising only the shape of the distribution. They are dependent on the phases. The skewness characterises the asymmetry of the density distribution about its mean and is defined as:

$$\text{Skewness} = (1/N)\Sigma[(\rho_i - \langle\rho\rangle)/\sigma]^3,$$

where $\langle\rho\rangle$ is the mean density, σ the r.m.s. density and N the number of points. The kurtosis measures the relative sharpness (or flatness) of the density distribution compared to a normal distribution and is defined as:

$$\text{Kurtosis} = (1/N)\Sigma[(\rho_i - \langle\rho\rangle)/\sigma]^4 - 3,$$

where the factor -3 makes the kurtosis zero for a normal distribution. The significance of skewness and kurtosis is shown schematically in Fig. 4. We have found the skewness and kurtosis excellent indicators of map quality. Details of the analysis will be published elsewhere.

Applications

The application of ARP to three proteins is described here. Throughout this section we define the initial model as that input to ARP, the ARP model as that refined during ARP and the final model as that obtained by conventional restrained least-squares refinement. The maps calculated with phases from these models are described as initial, ARP and final maps, respectively. Similar terminology is used for atoms and for phases calculated from the models.

For all three examples data were collected on an imaging plate scanner, built in-house, using synchrotron radiation from beamline X31 at EMBL Hamburg. In each case the data were more than 90% complete overall, more than 98% at low resolution and more than 66% of the theoretical unique intensities were greater than 3σ in the outer resolution shell.

Example 1: apo formate dehydrogenase

NAD-dependent formate dehydrogenase (E.C. 1.2.1.2., FDH) from the methylotrophic bacterium *Pseudomonas sp.* 101 catalyses the oxidation of the formate anion with concomitant reduction of NAD to NADH (Egorov, Avilova, Dikov, Popov, Rodionov & Berezin, 1979). FDH is a dimer with two subunits of 393 residues and the primary structure has been determined (Popov, Shumilin, Ustinnikova, Lamzin & Egorov, 1990). The holo FDH structure has been determined by MIR and refined by restrained least-squares minimisation (Lamzin, Aleshin,

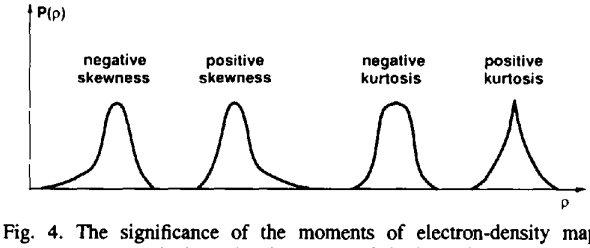


Fig. 4. The significance of the moments of electron-density maps, particularly the skewness and the kurtosis.

Strokopytov, Yukhnevich, Popov, Harutunyan & Wilson, 1992).

Apo formate dehydrogenase (apo FDH) crystallises in space group $P2_1$, with cell dimensions a = 110.5, b = 54.5, c = 70.3 Å, β = 101.9°. There is one dimeric molecule per asymmetric unit and each subunit consists of two domains. X-ray data to 1.8 Å have been collected. The crystal structure of apo FDH was solved by molecular replacement using the refined holo model (Lamzin, Popov, Harutunyan & Wilson, 1993).

Here the apo model from molecular replacement was refined by three cycles of constrained rigid-body minimisation in resolution range from 8.0 to 4.0 Å using the program *CORELS* (Sussman, Holbrook, Church & Kim, 1977). The apo and holo forms of FDH mainly differ by the relative orientation of the two domains as is found in other dehydrogenases (Eklund & Brändén, 1987). Thus constrained refinement was performed using each of the four FDH domains in the dimer as a separate rigid body. The r.m.s. deviation from the initial model after rigid-body refinement to the final model was about 0.5 Å for C*A* atoms and about 0.9 Å for all protein atoms. The resulting R factor was 37.4% for data between 10.0 and 1.8 Å resolution.

ARP protocol. The model of apo FDH after constrained rigid-body refinement was used as the initial model for the development of the automated refinement procedure (ARP).

All atoms were renamed as water atoms with an equal number of electrons and equal atomic temperature factors. All data in the range 10.0 to 1.8 Å were used from the start. The starting R factor was 42.6%. 24 cycles of ARP were run. Each consisted of one cycle of unrestrained refinement in the whole resolution range, followed by removal of the 0.3% of atoms in the weakest ($3F_o$ - $2F_c$) density and the addition of the 1.0% of peaks in the strongest ($2F_o$ - $2F_c$) density. The R factor at the end of ARP fell to 13.8%.

The initial and ARP models were inspected in the ARP ($3F_o$ - $2F_c$) density using computer graphics. Most ARP atoms were located rather close to initial protein atoms and the interatomic distances in the ARP model were similar to the corresponding distances in the protein. 90% of the new protein model was automatically constructed using a rebuilding program. The program checked if ARP atoms were at ($3F_o$ - $2F_c$) density greater than the mean value by 1σ and had temperature factors less 50 Å^2. The x, y, z coordinates of each protein atom in the initial model were changed to those of the nearest ARP atom if there was one within 1 Å. The temperature factors were adjusted according to the number of electrons in the protein atom to keep the peak density the same. The remaining 10% of the structure corresponded to regions with large shifts from the initial model and these were rebuilt manually from the ARP model and ($3F_o$ - $2F_c$) density. 514 water molecules were automatically assigned from the ARP model. The new protein model, with rather poor geometry (r.m.s. deviation in bond lengths of 0.18 Å), was subjected to restrained refinement using *PROLSQ*, mainly to improve the stereochemistry. After ten cycles the R factor dropped to 17.5% and the r.m.s. deviation in bond lengths to 0.021 Å.

Table 1. *Formate dehydrogenase: a comparison of conventional restrained refinement (Lamzin et al., 1992) and ARP*

For both approaches, the initial model is from molecular replacement using holo FDH followed by three cycles of refinement with four rigid bodies (see text)

Methods used	R factor (%) 10.0–1.8 Å	R.m.s. deviation in bond length (Å)	Real time
51 cycles of restrained least-squares refinement (*PROLSQ*), one cycle of simulated annealing (*X-PLOR*), plus much manual rebuilding	16.9	0.022	2 months
Automated refinement procedure, automatic rebuilding (90% of the model), manual rebuilding (10% of the model), ten cycles of restrained refinement	17.5	0.021	1 week

A comparison of ARP and conventional restrained refinement is shown in Table 1. In both cases the initial model was that from MR after rigid-body refinement with *CORELS* at 8.0–4.0 Å resolution with each of the four protein domains treated as a rigid body. The conventional refinement of apo FDH gave a final model with an R factor of 16.9%, r.m.s. deviation in bond distances of 0.022 Å and took about two months. The complete protocols of the conventional restrained and simulated annealing refinement will be published elsewhere (Lamzin *et al.*, 1993). In brief, the first 20 cycles of restrained refinement were run with the program *PROLSQ*. *X-PLOR* was then used with a heat stage followed by slow cooling at 2.4 Å resolution. This gave an R factor of 21.6% at 2.4 Å and an average discrepancy for all protein atoms from the final model of 0.6 Å. We did not widely investigate the best parameters for *X-PLOR* and these could almost certainly be optimised further. Substantial errors in the model remained, for example the end of the side chain Arg 85 was more than 6 Å from its final position. A further 31 cycles of restrained refinement interspaced with several graphics sessions were required to complete the refinement.

In contrast a model with nearly the same characteristics was obtained after only two days for ARP plus three to four days to rebuild the 10% of the structure which could not be automatically assigned and to tidy up the stereochemistry. The ARP model is essentially identical to that from standard methods. The r.m.s. deviation between C*A* atoms of the two models is 0.09 Å and about 0.19 Å for all protein atoms with temperature factors less than 50 Å^2.

The final FDH model refined by conventional least squares contained 5814 protein atoms with non-zero occupancy and 500 water molecules. The ARP model con-

tained 6844 atoms. From these 5814 protein plus 514 water atoms were included for the restrained refinement after ARP. The other 516 ARP atoms were rejected.

Results. Fig. 5 shows the *R*-factor distribution for the initial model, ARP model, ARP model after rebuilding plus restrained refinement (*PROLSQ*), and the final model. There is a dramatic reduction in *R* factor from initial to ARP model in all resolution shells. For the final and ARP model the variation of *R* factor with resolution is very similar.

Average phase differences from the final phase set are shown in Fig. 6 and phase statistics in Table 2. The initial phases differed by 37.9° from the final phases, 28.7% of them by more than 45°. For the 10% strongest reflections the mean phase difference was 15.2° and 4.3% of these had phases differing by more than 45°. Here we define the strongest reflections as those with highest normalised structure-factor amplitudes (*E* values). Large structure factors with poor agreement make it difficult to improve the model by conventional refinement. ARP phases were considerably better and differed only by 18.9° on average from final phases. None of the 10% strongest reflections had phases differing more than 45°. Thus ARP improved the phases overall, especially for the strongest reflections resulting in their assignment to the correct quadrant. This substantially improves the calculated density and makes it much easier to rebuild the model.

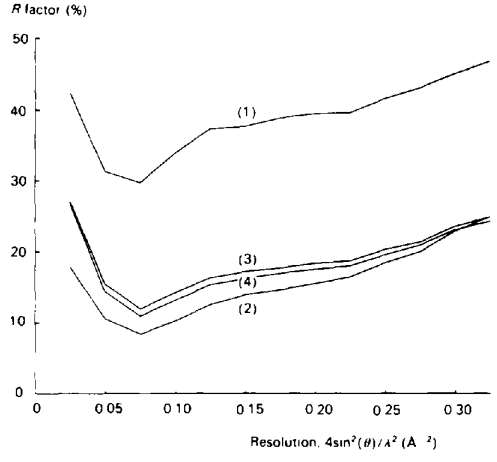


Fig. 5. Formate dehydrogenase: *R*-factor distribution. (1) Initial model (overall *R* factor 37.4%). (2) ARP model (13.8%). (3) Model after rebuilding and ten cycles of restrained refinement with *PROLSQ* (17.5%). (4) Final model, solved by standard methods (16.9%).

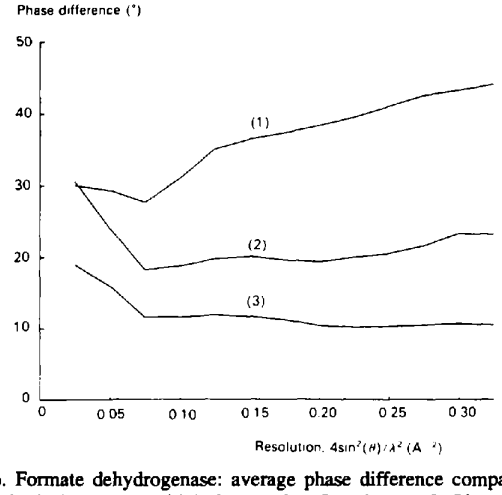


Fig. 6. Formate dehydrogenase: average phase difference compared to the final phases. (1) Initial phases. (2) ARP phases. (3) Phases from ARP model after rebuilding and ten cycles of restrained refinement.

Table 2. *Formate dehydrogenase: phase statistics in the* 10.0–1.8 Å *resolution range for the various models*

$\Delta\varphi$ (°) is the average phase difference from the final phases.

	All reflections		10% strongest reflections	
Model	$\Delta\varphi$	% with $\Delta\varphi > 45°$	$\Delta\varphi$	% with $\Delta\varphi > 45°$
Initial	37.9	28.7	15.2	4.3
ARP	18.9	9.3	6.2	0.0
ARP after rebuilding and ten cycles of restrained refinement	11.2	3.6	3.6	0.0

The *R* factor overall and in resolution shells for the ARP model is lower than for the final model especially at low resolution (Fig. 5). The former property is due to the lack of stereochemical restraints in ARP and the latter to the effect of including more 'water' molecules with high *B* factors in the ARP compared to the final model. Similarly ARP phases in the lowest resolution shell look as though they became worse than the initial phases, Fig. 6. Again this is misleading. The ARP phases include extra information about solvent and due to this differ from the final phases at low resolution. After rebuilding and restrained refinement the model gave phases closer to the final model than the ARP model because the former were refined using geometrical restraints but the ARP model was not restrained.

Histograms of the density distributions are shown in Fig. 7 and some characteristics of the electron densities in Table 3. The initial map has a relatively high correlation, 0.77, to the final map. However, the correlation coefficient can sometimes exaggerate the agreement. The initial map is very noisy with r.m.s. density of 0.64 e $Å^{-3}$, and has skewness and kurtosis substantially lower than the final map. The skewness and kurtosis clearly show that the initial FDH map had a density distribution substantially different from the final map. In contrast the distribution of ARP density is in essence identical to that for the final map and the r.m.s. density is similar for both maps. Indeed the skewness and kurtosis of the ARP map are 'better', *i.e.* greater, than those for the final map. The correlation coefficient of the ARP map to the final map is 0.93.

The overall r.m.s. deviation in C*A* atoms between the initial and final models is only 0.5 Å but there are several places where these models differ significantly. An example using ARP electron density is shown in Fig. 8. There is a movement of a large loop between the initial and final

models, with a systematic shift of about 2 Å. Fig. 8(*a*) shows the initial model and a set of ARP atoms: the ARP atoms look like a protein. The ARP density is certainly good enough to allow the correct model to be built. The same region of map in the final model is shown in Fig. 8(*b*). This indicates that ARP converges on the correct solution with a model similar to the final one.

In summary, application of ARP to FDH at 1.8 Å resolution gave a density map with parameters almost identical to the final map. 90% of the protein model was rebuilt automatically and only 10% of the atoms needed to be rebuilt manually according to the ARP map. The resulting model was refined by ten cycles of restrained refinement and proved to be almost identical to the final model from conventional methods. In real time use of ARP for apo FDH made the refinement approximately ten times faster.

Example 2: narbonin

This example shows the application of ARP to a partial (85%) complete MIR model with X-ray data measured to 1.8 Å but without complete amino-acid sequence information.

Narbonin is a monomeric seed globulin from *Vicia narbonesis* L. It has been crystallised (Hennig, Schlesier, Pfeffer & Höhne, 1990) in space group $P2_1$, a = 46.9, b = 75.5, c = 50.9 Å, β = 120.5°, with one molecule per asymmetric unit and has a molecular mass of 33 kdalton. Only partial (about 30%) primary structure information is available. A model was built into a 2.2 Å resolution MIR map and preliminarily refined with restraints to an R factor of 29.9% to 1.8 Å by Michael Hennig and coworkers (to be published). The refinement needed a large number of steps of alternating least-squares minimisation, MIR and calculated phase combination and rebuilding with *FRODO*. This partial model had several breaks in the polypeptide chain and contained 85% of the protein atoms. The ($3F_o$ - $2F_c$) and ($2F_o$ - $2F_c$) density maps did not clearly show which regions of the model should be corrected or how the missing regions should be built, especially difficult in the absence of chemical sequence. It was used as the starting model for ARP.

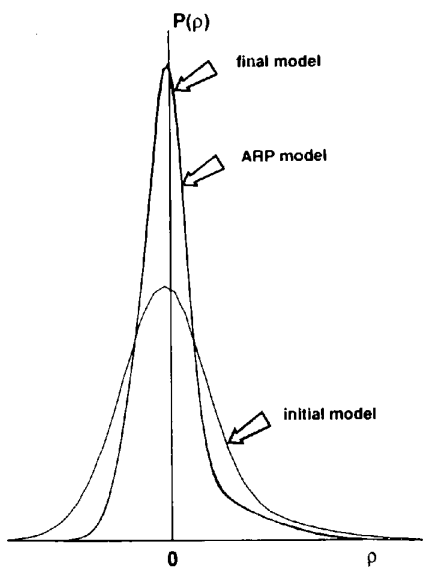


Fig. 7. Formate dehydrogenase: electron-density histograms for ($3F_o$ - $2F_c$) maps on the absolute scale with F_{000} set to zero for initial model, ARP model and final model. Maps in all later figures are calculated with the same coefficients unless otherwise stated.

Table 3. *Formate dehydrogenase: characteristics of the* 1.8 Å ($3F_o - 2F_c$) *electron-density distributions*

Map	Correlation to final map	R.m.s. (e Å $^{-3}$)	Skewness	Kurtosis
Initial	0.77	0.64	0.82	2.22
ARP	0.93	0.49	1.55	4.89
Final	--	0.42	1.42	4.42

ARP protocol. Because of the absence of primary structural information all atoms in the initial model were assigned the same number of electrons before ARP: this increased the starting R factor to 31.2%. Six steps of ARP were carried out (Table 4, left-hand side). In each step several cycles of unrestrained refinement of x, y, z and B parameters for each atom were carried out with all data in the resolution range 10.0–1.8 Å followed by updating of the model by rejection of the 2–10% atoms in the lowest ($3F_o$ - $2F_c$) density and addition of the 10–15% in the highest ($2F_o$ - $2F_c$) density. Thus the total number of atoms was gradually increased to 2796. The R factor dropped to 14.9%. The ARP map proved to be much better than the initial one.

The initial protein model was corrected manually using the ARP electron density with *FRODO*. As for the FDH case, ARP atoms were used as guide points in the rebuilding as most of them lay in the correct positions for real atoms. The polypeptide fold could be traced in places where the initial model had breaks, and side chains built in places where the initial map was not good enough to identify the amino acid. The resulting model, with 2171 protein atoms and 273 water molecules, was refined with stereochemical restraints and the R factor dropped from 28.8 to 20.5%.

ARP was applied for a second time with the protocol shown on the right-hand side of Table 4. The R factor dropped from 22.8 to 13.2%. The ARP density map was inspected, the model rebuilt in several parts and refined with geometrical constraints. The R factor fell from 23.1 to 16.9%. The final model had 2269 protein and 234 water atoms.

Results. Fig. 9 shows how the R-factor distribution changed during the refinement. There are four curves corresponding to the initial model, two ARP models and the final model. During the first ARP the R factor fell in all resolution shells. The difference between the first and second ARP models is most significant at high resolution. The R factor overall and in resolution shells for the final

model are somewhat higher than for the ARP models. This is to be expected: firstly because of the restraints used for the refinement of the real protein model and secondly because of additional waters in the ARP model (as seen for FDH).

Fig. 10 shows how the average phase difference to the final phases was reduced during ARP. The first ARP gave a model with greatly improved phases for the high-resolution shells. The second ARP resulted in phases very close to those for the final model. The improved phases obtained in the second application are due to the superior starting model used for ARP.

Phase statistics are summarised in Table 5. The initial model had phases differing by 39.2° from the final model. 30.1% of the phases differed by more than 45°, *i.e.* vectors of the calculated structure factors for those reflections were in the incorrect quadrant. For the 10% strongest (highest E value) reflections the average phase difference was 16.2% and 4.6% of the strongest reflections were in the wrong quadrant. Phases calculated from the first ARP model were substantially improved. In spite of the ARP phases differing by 33.2° on average and 23.4% of the phases deviating by more than 45°, only 1.6% of the 10% strongest reflections were in the wrong quadrant. Phases after the second ARP differed by 9.6° for the 10% strongest reflections and had an overall average difference of 19.4° from the final phases.

Electron-density histograms for the initial, first ARP and final maps are shown in Fig. 11. The distribution for the ARP map differs dramatically from that for the initial map, is very similar to that for the final map and is less diffuse. Characteristics for these density maps are presented in Table 6. During the first ARP the correlation to the final map improved from 0.74 to 0.83. The skewness and kurtosis clearly show that the initial narbonin model gave a density distribution substantially different from the final map. After the first ARP these parameters were 'better' than for the final map.

The density characteristics shown in Table 6 as well as changes in average phase difference or R factor reflect the overall improvement after ARP. Fig. 12 shows local examples of density improvement. As incomplete primary structure information was available the side chains of residues 91 and 92 were incorrectly identified in the initial model (Fig. 12*a*). The density map from the initial model did not clearly show the correct structure. The density was considerably improved after the first ARP (Fig. 12*b*). The ARP atoms were connected by interatomic distances typical for amino-acid side chains. Another example is shown in Figs. 12(*c*) and 12(*d*). There are two regions of the polypeptide fold containing two phenylalanines. The initial density (*c*) is relatively poor and it is difficult to recognise side chains or to trace the polypeptide fold. The ARP atoms and density clearly show the correct structure in this region (*d*). The initial density does not indicate phenylalanine side chains.

The application of ARP to narbonin at 1.8 Å resolution resulted in substantial improvement of the density especially in places where the initial map was poor. The distribution of ARP electron density was close to that of the final density. The positions of ARP atoms generally corresponded to real atoms and the model was rebuilt from the new features in the ARP map. The second ARP application which started from the improved protein model resulted in further improvement. The model after the second ARP was successfully rebuilt and refined by Michael Hennig. In retrospect much time would have been saved in the analysis by use of ARP at an earlier stage.

Example 3: pyrophosphatase

The refinement of pyrophosphatase (PP) is an example of ARP at medium resolution. Mn-dependent yeast inorganic PP from *Saccharomyces cerevisiae* catalyses the hydrolysis of inorganic pyrophosphate (Cooperman, 1982). PP is a dimer with two chemically identical subunits, each having a molecular mass of 32.0 kdalton. The primary structure is known (Cohen, Sterner, Keim & Heinrikson, 1978). PP has been crystallised in space group $P2_12_12_1$, cell dimensions a = 116.5, b = 106.3, c = 56.1 Å, with one dimer per asymmetric unit (Chirgadze, Kuranova, Nevskaya, Teplyakov, Wilson, Strokopytov, Harutunyan & Höhne, 1991). X-ray data to 2.4 Å were recorded.

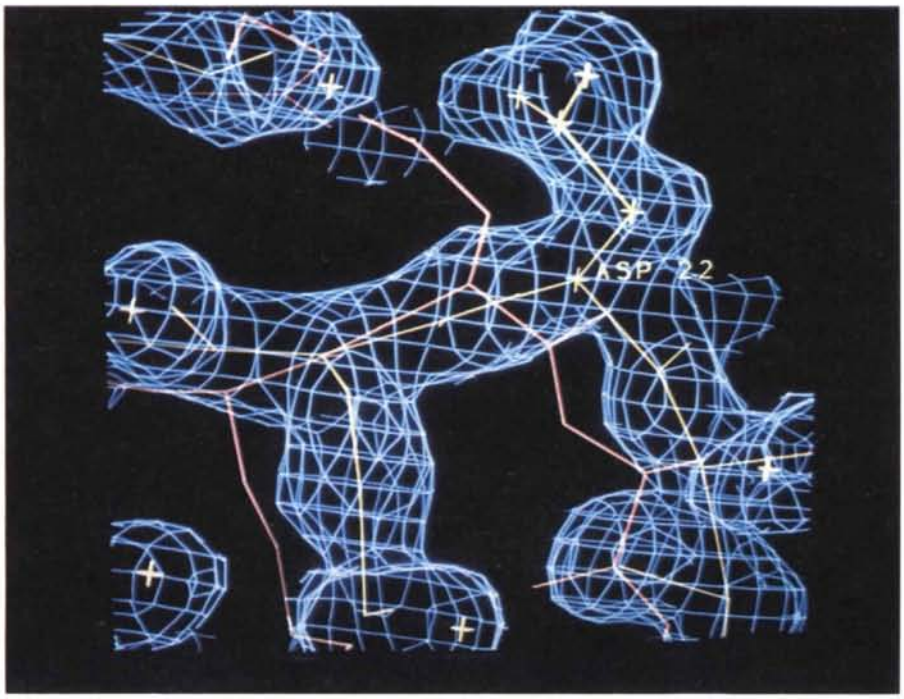
ARP protocol. The MR model of PP, preliminarily refined with stereochemical restraints to an R factor of 27.3% in the resolution range from 7.0 to 2.4 Å (E. Harutunyan and coworkers, to be published), was used as the initial model for ARP.

Two different refinement protocols were tried from this initial model. The first protocol is shown in Table 7. In each step several cycles of unrestrained refinement of x, y, z and B factors were run in the resolution range 7.0–2.4 Å, followed by updating of the model through rejection and addition of atoms. The fraction of the model updated in each iteration was gradually decreased from about 10% at step one to 2% at step four. After five steps the R factor for this first model had dropped to 12.5%.

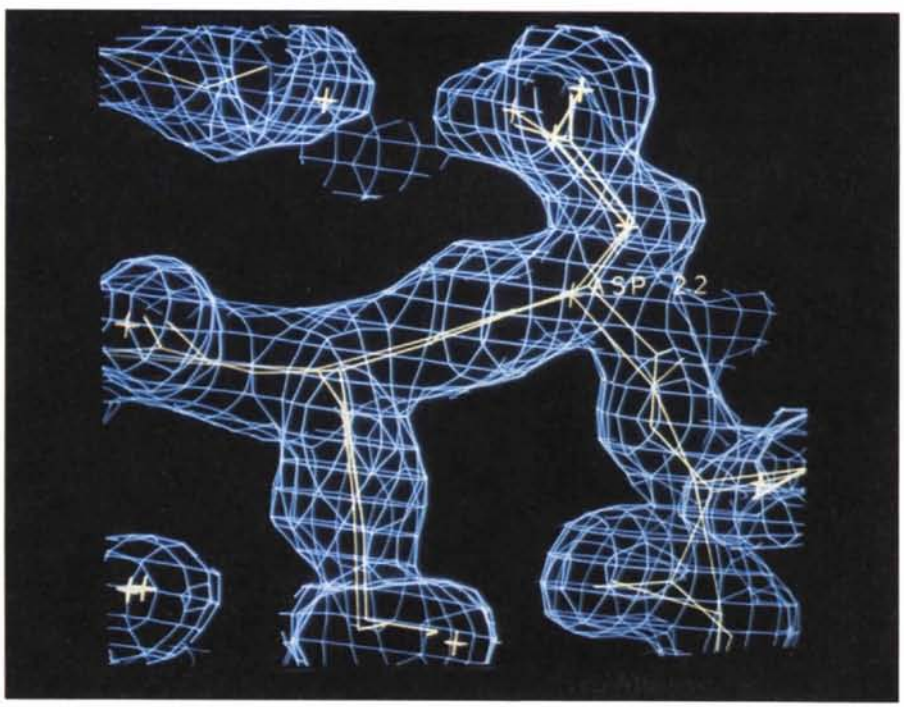
The second ARP was performed in a 'smooth' manner as for FDH. 50 cycles of unrestrained refinement were carried out, but the model was updated from the $(3F_o - 2F_c)$ and $(2F_o - 2F_c)$ maps after each cycle: the nine 'weakest' atoms (0.2% of the total number of atoms) were removed and nine new atoms were added. All new atoms were assigned temperature factors of 30 Å^2, which is approximately equal to the average temperature factor. After 50 cycles the R factor dropped to 12.6%. The total number of atoms in both ARP refinements was kept constant because the limited resolution of the data does not allow the introduction of a large number of waters to the model.

Regions of the $(3F_o - 2F_c)$ electron-density map calculated from the initial model and the two maps from the first and second ARP models are shown in Fig. 13. The ARP maps are a considerable improvement upon and are

less noisy than the initial map. New features and connectivities appeared in several places where the initial map was relatively poor. The density characteristics for the PP maps are shown in Table 8. Both ARP maps have lower r.m.s. values than the initial map which indicates they are less noisy. Greater values for the skewness and the kurtosis show that the ARP maps are more asymmetric as is typical for maps from refined models.



(*a*)



(*b*)

Fig. 8. Formate dehydrogenase: approximately 2 Å shift of the loop from the initial to the final model. The electron density shown is from the ARP model, again the $(3F_o - 2F_c)$ map. The contour level is 1σ above mean density. (*a*) Initial model (red) and ARP atoms connected according to protein interatomic distances (green). (*b*) Same place with ARP and final model.

Table 4. *Narbonin: ARP protocols at* 10.0–1.8 Å *resolution*

	First application						Second application		
Step	1	2	3	4	5	6	1	2	3
R factor, start (%)	31.2	25.1	21.9	19.8	18.3	17.3	22.8	17.9	15.6
R factor, end (%)	25.9	21.0	19.0	16.9	15.8	14.9	15.9	14.7	13.2
No. of atoms	1924	2181	2360	2622	2733	2796	2444	2405	2597
No. of cycles of unrestrained refinement	9	7	5	6	6	6	6	6	6
Cutoff ($3F_o - 2F_c$) density (σ)	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
No. of atoms rejected	42	31	45	203	260		184	82	
Threshold ($2F_o - 2F_c$) density (σ*)	2.0	1.7	1.4	1.2	1.1		1.5	1.3	
No. of atoms added	299	210	307	299	310		145	275	

* The threshold values are shown in units of σ for the ($F_o - 2F_c$) map.

Table 5. *Narbonin: phase statistics in the* 10.0–1.8 Å *resolution range*

$\Delta\varphi$ is the average phase difference (°) from the final phases.

	All reflections		10% strongest reflections	
Model	$\Delta\varphi$	% with $\Delta\varphi > 45°$	$\Delta\varphi$	% with $\Delta\varphi > 45°$
Initial	39.2	30.1	16.2	4.6
1st ARP	33.2	23.4	12.0	1.6
2st ARP	19.4	9.6	6.5	0.0

Table 6. *Narbonin: characteristics of the* 1.8 Å ($3F_o - 2F_c$) *electron-density distributions*

Model	Correlation to final map	Skewness	Kurtosis
Initial	0.74	0.84	2.23
1st ARP	0.83	1.37	3.87
Final	—	1.27	3.34

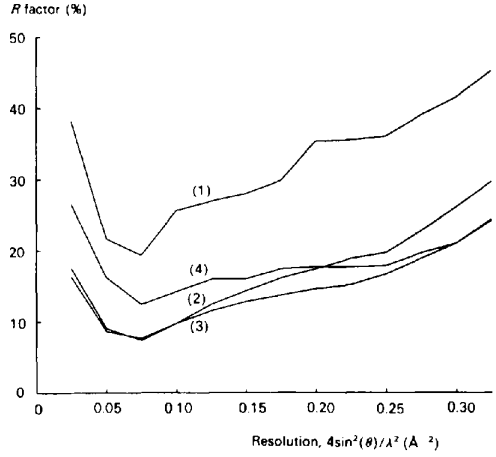


Fig. 9. Narbonin: *R*-factor distribution. (1) Initial model (total *R* factor is 29.9%). (2) First ARP model (14.9%). (3) Second ARP model (13.2%). (4) Final model (16.9%).

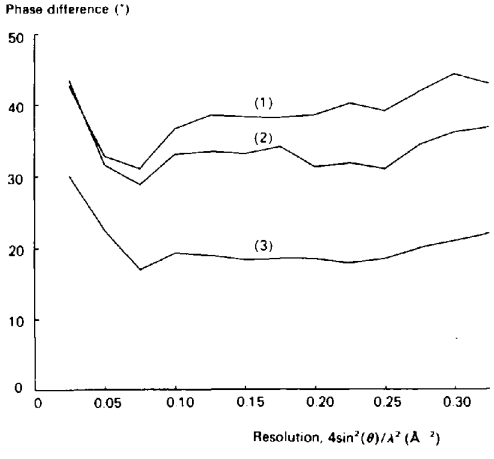


Fig. 10. Narbonin: average phase difference compared to the final phases. (1) Initial model. (2) First ARP phases. (3) Second ARP phases.

Results. The average difference between the initial phases and the first ARP phases is 36.2° for all reflections and 13.0° for the 10% reflections with highest *E* values. Between the initial phases and the second ARP phases the values are 32.2 and 12.5° respectively. Between the two ARP models the differences are 29.0 and 10.4° respectively. The average phase difference is only an overall characteristic. Phases from the two ARP models deviate from each other as much as they both deviate from the initial model. Nevertheless the two ARP maps look approximately the same (Figs. 13*b* and 13*c*). However, the second ARP map (Fig. 13*c*) showed improvement in several places. Electron-density histograms are shown in Fig. 14. The histograms for the two ARP maps are nearly identical. Both are considerably 'better' than the histogram for the initial protein map, in that they are much more similar to the expected histogram at this resolution.

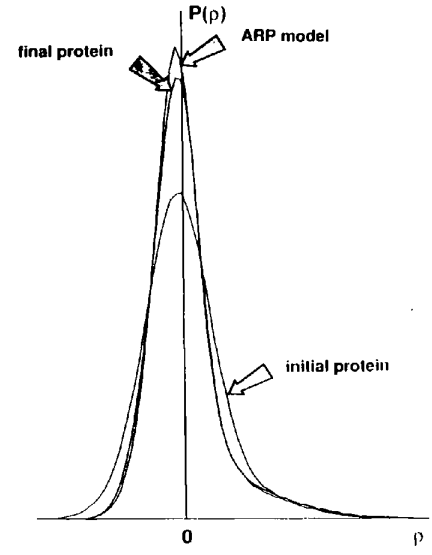
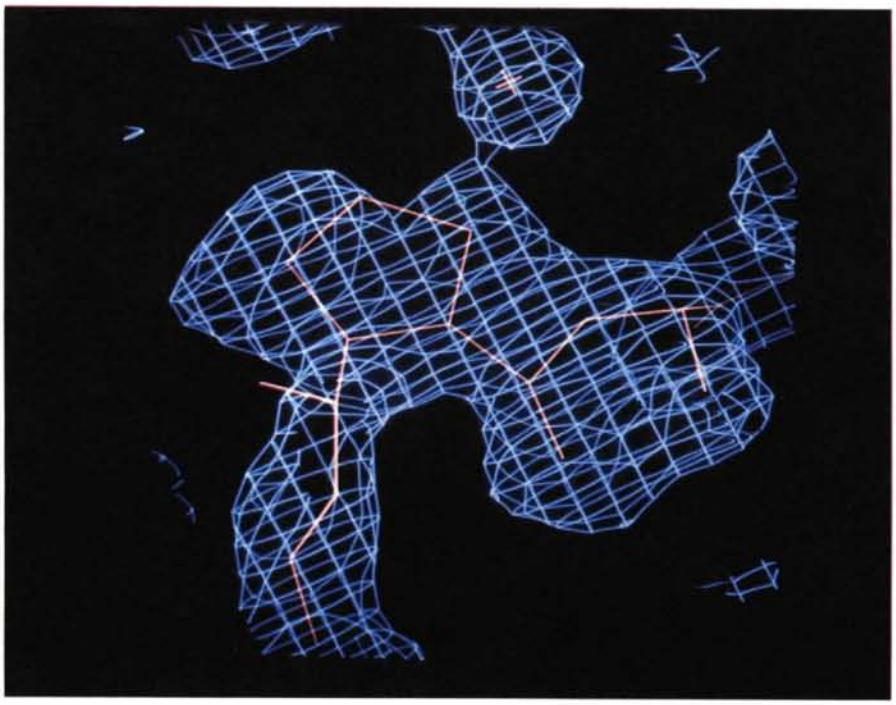
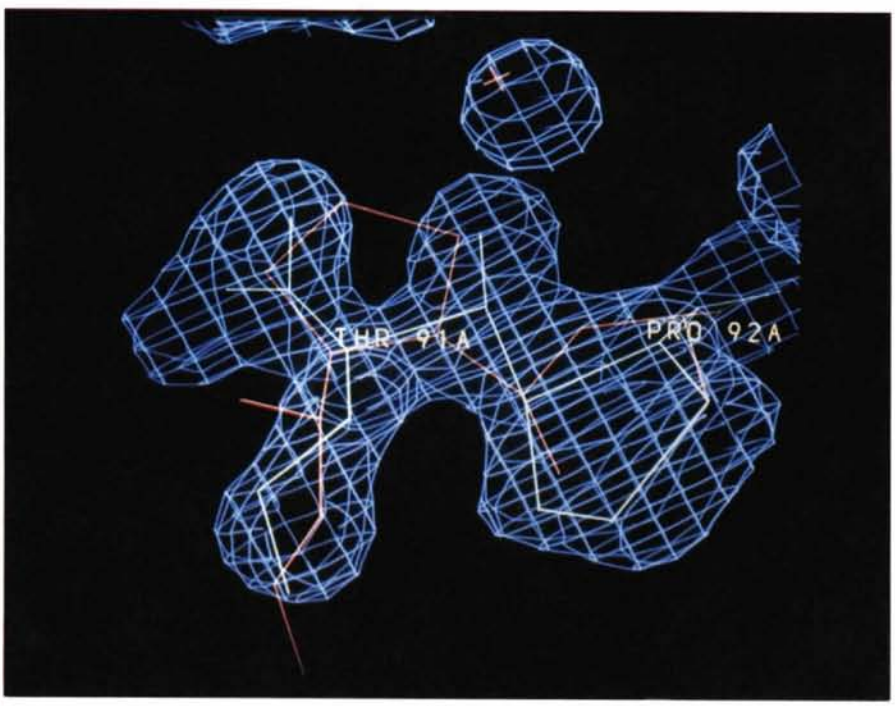


Fig. 11. Narbonin: electron-density histograms for maps from initial model, after first ARP and final model.

Thus the application of ARP to PP at 2.4 Å resolution gave an improved ($3F_o$ - $2F_c$) map. The noise has been substantially reduced and several new features have appeared, which make it easier to correct the model. ARP effectively resulted in solvent flattening in spite of the fact that the molecular boundaries were not specially set. The distribution of the new electron density is considerably better according to histogram-matching criteria. Two



(*a*)

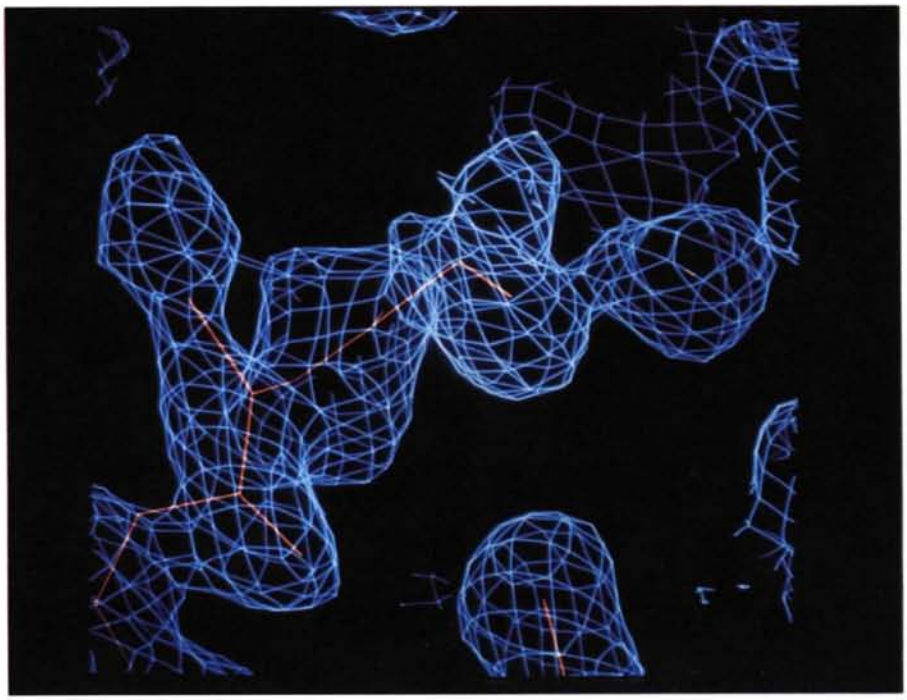


(*b*)

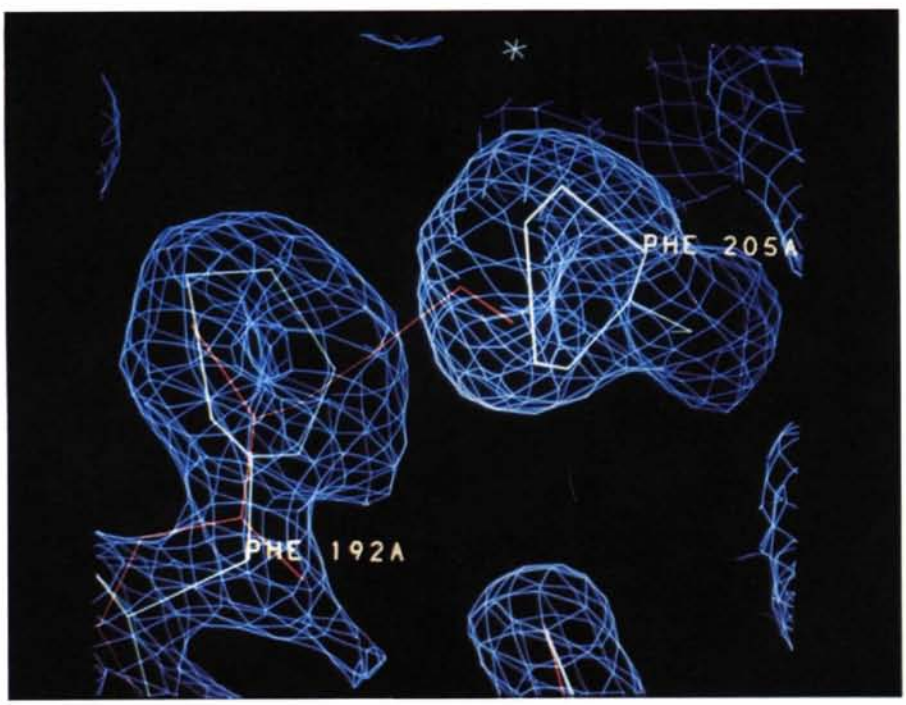
Fig. 12. Narbonin: (*a*) Badly fitting initial model and initial density around residues 91 and 92. (*b*) Density and atoms after first ARP.

different refinement protocols resulted in similar electron-density maps. However, the performance of 'smooth' ARP is much more convenient and makes the procedure completely automatic: this ARP was completed within a single computer operation with constant parameters.

The results were much less impressive, however, than for the previous examples when 1.8 Å data were used. The ARP model was not adequate to allow automatic reconstruction of the protein from the ARP model to proceed with confidence. Rebuilding of the protein model



(*c*)



(*d*)

Fig. 12 (*cont.*) Narbonin: (*c*) Incompletely built initial model and the initial density around residues 192 and 205. (*d*) Same region with electron density improved by first ARP and with ARP atoms.

Table 7. *Pyrophosphatase: first ARP protocol at 7.0–2.4 Å resolution*

26 177 reflections were used.

Step	1	2	3	4	5
R factor, start (%)	28.4	24.2	22.1	18.4	16.5
R factor, end (%)	18.5	16.3	14.3	13.1	12.5
No. of atoms	4479	4482	4479	4480	4463
No. of cycles of unrestrained refinement	10	7	10	10	8
Cutoff ($3F_o - 2F_c$) density (σ)	1.0	1.0	1.0	1.0	
No. of atoms rejected	469	250	99	88	
Threshold ($2F_o - 2F_c$) density (σ^*)	1.0	1.0	1.0	1.0	
No. of atoms added	472	247	100	71	

*The threshold values are shown in units of σ for the ($3F_o - 2F_c$) map.

Table 8. *Pyrophosphatase: characteristics of the* 2.4 Å *($3F_o - 2F_c$) electron-density distributions*

Map	R.m.s. (e Å $^{-3}$)	Skewness	Kurtosis
Initial	0.30	0.43	1.49
1st ARP	0.24	0.76	2.51
2nd ARP	0.22	0.73	2.54

according to the ARP electron density has been carried out and the *R* factor is currently 19.0%.

Properties of the procedure

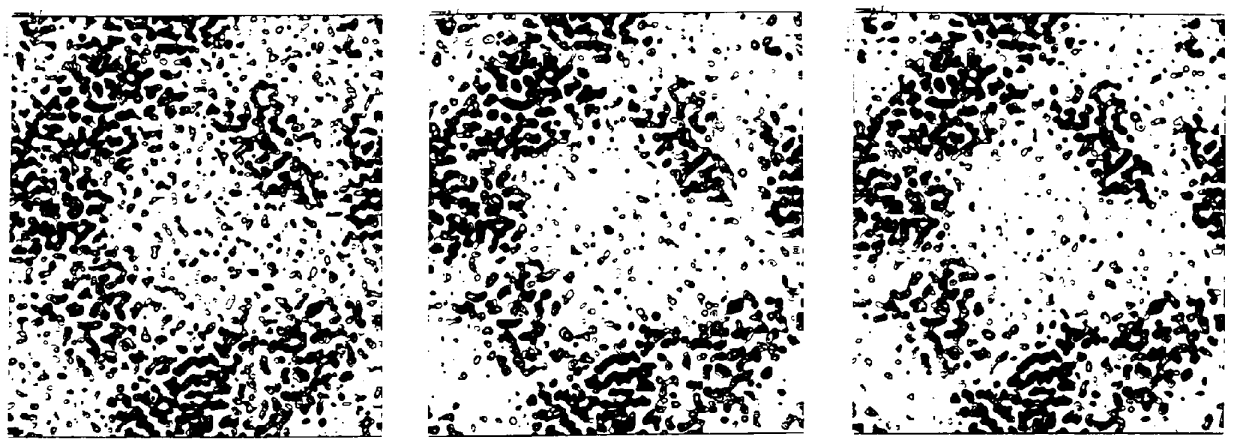
Trajectories of atoms during ARP

Application of the ARP to refinement of several proteins shows its ability to produce a correct model and thus to improve the calculated density by introducing atoms in 'new' places where the initial model had no, or not enough, atoms. How does the set of atoms in the initial model converge to the new set of atoms? The power and mechanism of improvement of the model using ARP are described here.

The course of ARP for Ala 20 in FDH is shown schematically in Fig. 15, and with the improved electron density from the ARP model in Fig. 16, where the points indicate intermediate positions of the atoms during the refinement. C*B* in the initial model moved out of density and was removed when the model was updated: this trajectory is coloured blue in Fig. 16. C*A* in the initial model moved to the true position of the C*B* atom. The true place for the C*A* atom was occupied by a new atom picked up in positive difference density after several cycles of ARP. This trajectory is shown in red. The initial main-chain N and O atoms moved to what were, in reality, water positions during ARP and their places were taken up by C and C*A* atoms from neighbouring residues.

To refine the FDH model a set of atoms with the same number of electrons was used. Such atoms can satisfactorily describe the electron-density distribution of atoms with a similar number of electrons (C, N, O). Fig. 17 shows what happened for a typical S-containing residue. During ARP the atom initially located near the final S position remained essentially where it was. To compensate for the lack of electrons at the S position the initial C*E* atom moved to the S position. Thus in the ARP model there were two atoms at the S position. A new atom, with trajectory shown in red, was picked up and moved to the final place of the C*E* atom. The resulting ARP electron density is typical of a methionine residue.

The trajectories shown in Figs. 16 and 17 are generally not linear. The atoms continually change direction and can sharply change after the model has been updated. These pictures clearly show the difference between ARP, with automatic rejection and addition of atoms, from the use of unrestrained refinement alone. When a big enough positive peak in difference density appears it is immediately included as a new atom, provided it satisfies certain positional requirements. If the peaks roughly correspond to the true position of atoms in the structure then such addition



(*a*) (*b*) (*c*)

Fig. 13. Pyrophosphatase: (*a*) Electron-density map calculated for initial model (*R* factor 27.3% at 2.4 Å resolution). (*b*) Map after first ARP (*R* factor 12.5%). (*c*) Map after ARP performed with the 'smooth algorithm', see text (*R* factor 12.6%). Contour levels from 1.0 to 5.0 in steps of 1.0 of r.m.s density. Fractional map limits are x 0-1, y 0-1, z 0-0.08.

immediately improves the local density and model. This local improvement then promotes more accurate positioning of neighbouring atoms. If any atom moves out of ($3F_o$ - $2F_c$) density it becomes a potential source of noise and is immediately taken out of the model. In general, if an atom is not placed in the correct position it is much more simple and powerful to remove it and pick it up in a new position from the difference map than to try to refine it by restrained least squares with manual rebuilding.

Convergence

FDH was successfully refined using ARP starting from a model obtained by molecular replacement and preliminary rigid-body refinement at low resolution, where each domain was refined as a rigid body. This resulted in an ARP model deviating by an average of 0.5 Å in C*A*-atom position compared to the initial model. This ARP model was excellent and corresponded closely to the final refined model of FDH.

The behaviour of the ARP was then tested by deliberately using weaker preliminary models. Firstly only two rigid bodies were used in the constrained refinement, the two independent subunits of FDH, *i.e.* the subunits were not refined in two domains as above. After four cycles of refinement in the resolution range 8.0–4.0 Å the model deviated on average by 1.4 Å in C*A*-atom positions compared to the final model. This model is referred to as the 'complete' initial model because it included all 5996 protein atoms corresponding to the 383 residues (from a total of 393 in the sequence) which were ordered in the holo FDH structure. Only 373 residues were ordered in the final apo model. Three systematically incomplete initial models were obtained using methods similar to rigid-body refinement. The 'polyglycine' model included only main-chain atoms (about 51% of the total number). The 'polyalanine' model included main-chain and C*B* atoms (63%). The 'polyserine' model included main-chain, C*B* and C*G*, O*G*, S*G* atoms of the 'complete' model (76%). The incomplete models thus correspond approximately to starting models with increasingly different sequences but with no deletions or insertions.

The four models were each refined using a similar ARP protocol. 50 cycles of ARP were carried out. Each cycle consisted of: one cycle of unrestrained refinement at resolution 10.0–1.8 Å; rejection of approximately 0.4% of the atoms located in the lowest ($3F_o$ - $2F_c$) density; addition of 0.7–1.7% of new atoms picked in positive ($2F_o$ - $2F_c$) density to increase the total number of atoms after the 50 cycles to be approximately 20% more than the true number of protein atoms. All new atoms were given temperature factors of 20 Å^2 during the first 20 cycles of ARP and 30 Å^2 during the last 30 cycles.

The results of the refinement are presented in Table 9. After two rigid-body refinements all models, even the 'complete' one, had an initial *R* factor more than 50%. For comparison the initial model obtained after refinement with four rigid bodies had an *R* factor of 37.4%, (see above). After ARP all models had a number of atoms equal to 117% of the number of protein atoms in the final model. *R*-factor values for the ARP refined models vary from 22.0% for the 'polyglycine' model to 18.2% for the 'complete' model but all are higher than the *R* factor of 16.9% for the final model. All initial models gave calculated phases deviating on average by more than 60° from the final phases. After ARP the phase differences were reduced for all models. For the 'polyglycine' the phase deviation fell from 71.7 to 66.9°, but for the 'complete' model much more, from 63.2 to 43.9°. If only the 10% strongest reflections are considered the phase improvement is even better. The 'polyglycine' phase deviation fell from 55.3 to 43.7°, for the 'complete' model from 42.3 to 20.1°. Thus ARP generally improved phases, but was most effective

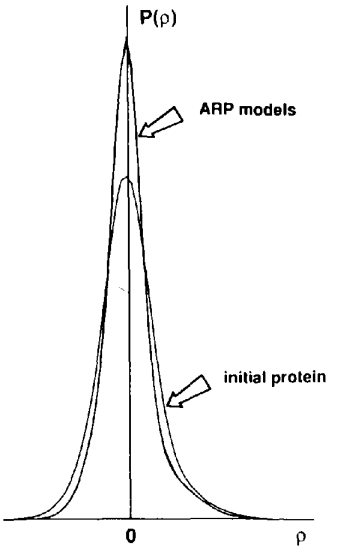


Fig. 14. Pyrophosphatase: electron-density histograms for the initial and two ARP maps obtained using different protocols.

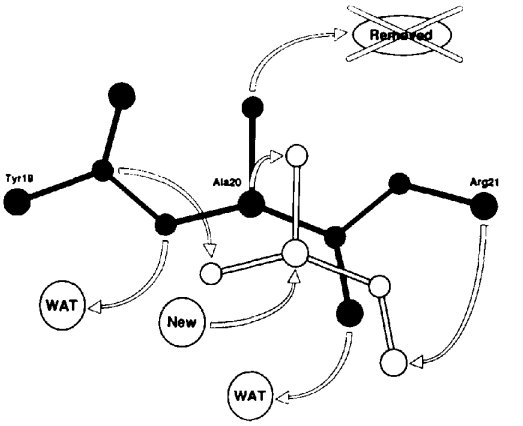


Fig. 15. A schematic representation of ARP for Ala 20 in FDH. The initial model is shown in black and the final model in white. Pointers indicate the movement of atoms during ARP.

for the strongest reflections, which are essential for the high quality of electron-density maps.

The same regions of the ($3F_o$ - $2F_c$) maps are shown in Fig. 18. All the initial maps are very difficult to interpret. The ARP maps look much clearer. The maps for 'polyglycine' and 'polyalanine' ARP models still have gaps in several places of the polypeptide fold and in many places the side-chain density is absent. The map for the refined 'polyserine' model has most features present in the final protein map. The map for the refined 'complete' model is quite similar to the final map. The latter two maps can be interpreted easily. Overall characteristics of the ($3F_o$

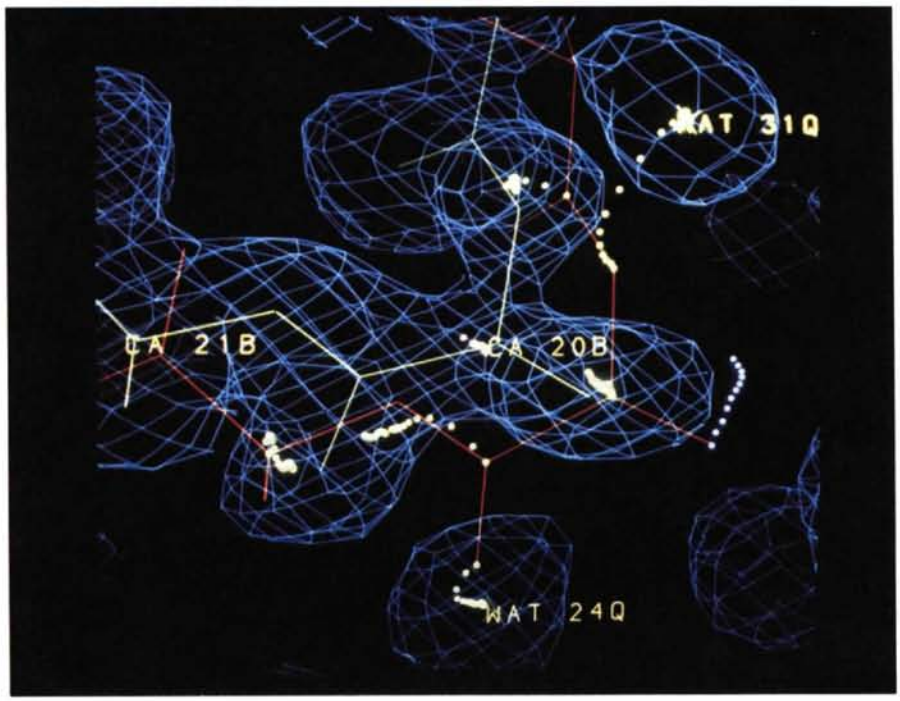


Fig. 16. A region of FDH model near Ala 20. The ($3F_o$ - $2F_c$) electron density after ARP is contoured at 1σ above the mean density. The initial model is shown in red, final model in green. Points indicate intermediate positions of atoms during refinement. Trajectories of movement of atoms which are retained from the initial model are coloured in green. Intermediate positions of the C*B* atom which moved out of density and was automatically removed are shown in blue. The final position of the C*A* atom was occupied by a new atom picked up in positive difference density, of which the trajectory is shown in red.

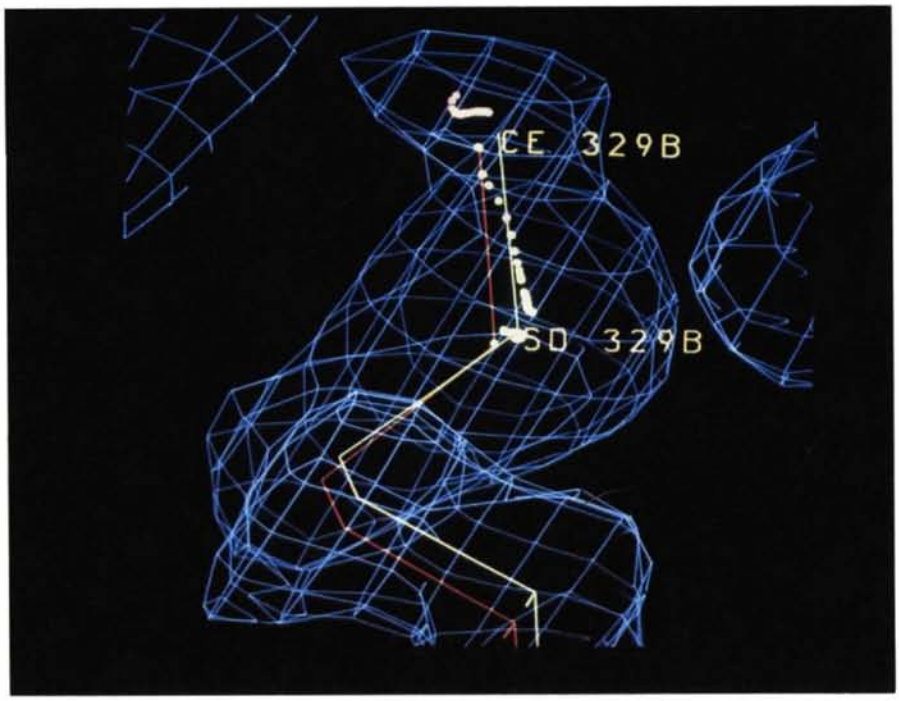


Fig. 17. A region of the FDH model near Met 329. The map and contour level are as in Fig. 16. The initial model is shown in red, the final model is shown in green. The points indicate intermediate positions of the atoms during the refinement. Trajectories of movement of atoms which are retained from the initial model are coloured in green. To compensate for the lack of electrons at the S atom, the C*E* atom moved from the original place to the S position. The trajectory for the new atom picked up and moved to the final place of the C*E* atom is shown in red.

Table 9. *Application of the ARP to incomplete and poorly prerefined FDH models*

Completeness is estimated as the ratio of total number of atoms to the number of protein atoms in the final model. $\Delta\varphi$ is the average phase differerence (°) to the final phases. The models were as follows: 'polyglycine' – main-chain atoms only (total 3064); 'polyalanine' – main-chain and C*B* atoms (total 3772); 'Polyserine' – main-chain, C*B* and C*G*,•O*G*, S*G* atoms (total 4552); 'complete' – 5996 atoms of 2 × 383 residues. Refinement protocol: the initial models are those after molecular replacement followed by four cycles of refinement as two rigid bodies at 8.0–4.0 Å resolution, where each monomer was treated as a single unit. This resulted in r.m.s. deviation in C*A* positions for all initial models compared to the final protein model about 1.4 Å. 50 cycles of ARP were carried out. Each cycle consisted of one cycle of unrestrained refinement in the resolution range 10–1.8 Å, rejection of 0.4% 'worst' atoms and addition of 0.7–1.7% new atoms.

Model	Completeness (%)	*R* factor (%)	$\Delta\varphi$ (°) All reflections	$\Delta\varphi$ (°) 10% strongest reflections	Correlation to final map	$(3F_o - 2F_c)$ density map R.m.s. (e Å^{-3})	Skewness	Kurtosis
'Polyglycine'								
Initial → ARP	51 → 117	55.2 → 22.0	71.7 → 66.9	55.3 → 43.7	0.35 → 0.44	0.87 → 0.45	0.23 → 1.08	0.63 → 3.03
'Polyalanine'								
Initial → ARP	63 → 117	54.3 → 20.8	70.4 → 64.6	53.3 → 40.9	0.38 → 0.48	0.86 → 0.42	0.22 → 1.06	0.54 → 2.88
'Polyserine'								
Initial → ARP	76 → 117	52.6 → 19.8	66.8 → 53.4	47.4 → 27.9	0.44 → 0.62	0.84 → 0.40	0.27 → 1.17	0.66 → 3.54
'Complete'								
Initial → ARP	100 → 117	50.5 → 18.2	63.2 → 43.9	42.3 → 20.1	0.49 → 0.72	0.86 → 0.41	0.34 → 1.27	0.75 → 3.92
Final	108	16.7	—	—	—	0.42	1.42	4.42

$- 2F_c$) maps are also presented in Table 9. The correlation of all initial maps to the final map was very poor. After ARP the correlation coefficient increased, especially for the 'polyserine' model from 0.44 to 0.62 and for the 'complete' model from 0.49 to 0.72. All maps for the refined models have reasonable skewness and kurtosis parameters, but these are significantly lower compared to the final map. The electron-density histograms plotted in Fig. 19 show the real improvement of the density distribution for all the models.

Table 10 shows the results of analysis of $(3F_o - 2F_c)$ maps on the basis of the final protein model. The electron density for each protein atom of the final model was interpolated from the maps. The number of residues of the final protein model having all atoms in density greater than 1σ above the mean value appears to be a good criterion for estimating the quality of the improved density. If all atoms of a residue are at the required density it is highly probable that this residue can be easily built manually or automatically. The final FDH map has 640 such residues, 86% of the number of residues with non-zero occupancy. All initial maps had less than one-third such residues. There was no substantial increase in the number of such residues in the 'polyglycine' and 'polyalanine' ARP maps. However, the 'polyserine' ARP map has 42% such residues and the 'complete' ARP map about 65%. Returning to the FDH model after refinement with four rigid bodies (see above) the map after ARP had 635 such residues, 85% relative to 86% in the final map.

The density maps, their characteristics, *R*-factor values and average phase difference clearly demonstrate the power, possibilities and limitations of the ARP. The better the initial model, the better the resulting ARP improvement. Two easily monitored parameters can be used as convergence criteria, the *R* factor and the characteristics of the $(3F_o - 2F_c)$ density map. Using X-ray data of good quality and completeness the *R*-factor value for an ARP model below 20% at 1.8 Å resolution indicates a substantially correct model, and at about 15% a model and a map which are nearly identical to the final map. Other criteria which can be used to monitor the ARP are the skewness and kurtosis of the density distribution. The 'expected' values of these depend on the resolution, percentage of solvent in the cell and the overall temperature factor. Thus they can be estimated roughly from any other protein density map at comparable resolution [as suggested for histogram matching by Lunin (1988) and Zhang & Main (1990)]. However, the values will be perturbed by the presence of heavy atoms such as metals in the protein and even to some extent by sulfur atoms. The values of the root-mean-square density, presented in Table 9, show that this parameter is not a useful indicator of the success or failure of the refinement.

Radius of convergence

The 'radius of convergence' of ARP can be estimated roughly in terms of completeness of the initial model and how different it is to the final model, namely the root-mean-square deviation in atomic position. At 1.8 Å resolution a molecular replacement model, including all the protein atoms with root-mean-square deviation in C*A* positions of about 0.5 Å compared to the final model, can be straightforwardly refined using ARP (as shown for FDH). This remains true if most of the model is correct and the model is about 85% complete but still has several unresolved regions (as shown for narbonin). If the final model is significantly displaced from the initial model, as for the poorly preliminarily refined FDH with root-mean-square deviations in C*A* positions of 1.4 Å, but is approximately complete, ARP results in a model giving a high-resolution density map quite similar to the final map. Even in the case of 75% completeness of such a poor model ARP gives substantial improvement and makes the density map much more interpretable.

Concluding remarks

ARP has been successfully applied to three proteins. It is clearly more powerful when high (better than 2.0 Å) data are available, but nevertheless gives definite improvement,

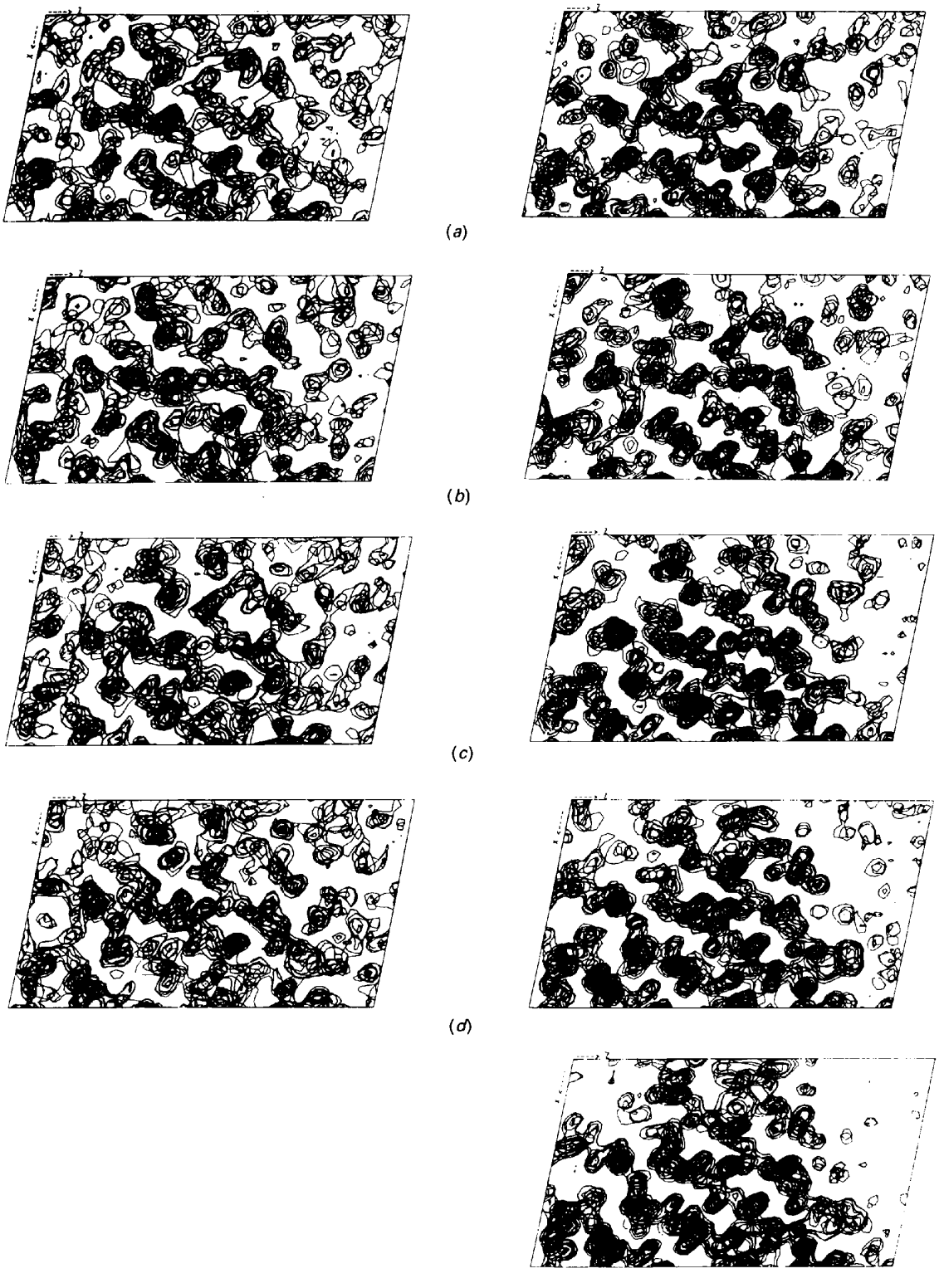


Fig. 18. Convergence properties of ARP. (*a*)-(*d*) The initial (left) and after ARP (right) density maps for incomplete and poorly preliminary refined FDH models (see text). The initial models used: (*a*) 'polyglycine', (*b*) 'polyalanine', (*c*) 'polyserine', (*d*) 'complete'. (*e*) Final map. The contour levels are 1, 2, 3, 4 and 5σ above mean density. Maps limits are: x 102/192-137/192, y 1/96-11/96, z 16/120-75/120.

at least in the density map, even at 2.4 Å. ARP requires as input a protein model which is more than 75% complete, *i.e.* which has 75% of the atoms within the least-squares radius of convergence of a true atomic position. The better the initial model, the better the result, at least in the present implementation.

ARP resembles the use of alternating cycles of least squares and difference Fourier syntheses used in small-molecule crystallography where atomic resolution data are available. The fast Fourier transform is essential if the calculations for proteins, both for the diagonal approximation least-squares refinement and the maps, are to be carried out in a tractable time. The restricted resolution of data is the cause of the poorer convergence properties in the case of proteins. It is expected that the method would be much more powerful with 1.0 Å data, and this will be tested in the near future.

As can be seen from Fig. 1, ARP is concerned with the refinement of a model starting with a set of atoms most of which are in essentially the correct position. The property it is based on is atomicity. ARP differs completely from *e.g.* direct methods which are also based on atomicity and use statistical relationships between the structure factors as in those expressed by Sayre's equation (Sayre, 1972). They are only valid for structures containing resolved equal atoms, and for *ab initio* phase generation from amplitudes alone they normally require data to atomic (about 1 Å) resolution, which is the most serious limitation of direct methods. At lower resolution direct methods can be used, in principle, for the extension and refinement of an initial set of phases and several attempts have been made for proteins *e.g.* Sayre (1974) and Agarwal & Isaacs (1977). These have met with limited success and the methods have not been applied generally.

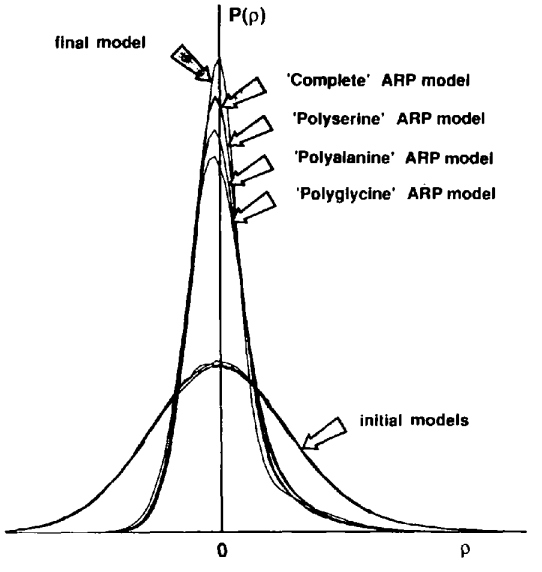


Fig. 19. Convergence properties of ARP. Electron-density histograms for incomplete and poorly preliminary refined FDH models (see text). There are four initial models, after ARP and final model.

Table 10. *Number and percentage of residues in the final FDH model having all atoms in* $(3F_o - 2F_c)$ *density greater than* 1σ *above the mean density*

The percentage is given relative to the 746 residues which are ordered in the final apo FDH model.

Model used to calculate density	Initial density No.	%	Density after the ARP No.	%
'Polyglycine'	49	7	78	10
'Polyalanine'	105	14	124	17
'Polyserine'	162	22	270	36
'Complete'	222	30	419	56
Final protein	—	—	640	86

Another approach to refine and extend an initial phase set at less than atomic resolution is density modification, in a variety of implementations. The map is modified according a number of possible restraints, and the phases resulting from the inversion of the modified map used in a cyclic manner. The most popular and widely used restraint is solvent flattening as described by Wang (1985). The procedure requires the definition of the molecular boundary and the density outside this boundary is set to zero. No model is required; however, this means that atomicity is not imposed on the structure.

Exploitation of non-crystallographic symmetry through molecular averaging as an additional restraint (*e.g.* Bricogne, 1974) is clearly more powerful than solvent flattening alone, especially in the case of high symmetry such as in viruses. The restraint of histogram matching, where the current electron-density distribution is adjusted to the expected distribution, has also been reported to be quite powerful (Lunin, 1988; Zhang & Main, 1990). However, it does not implicitly impose real atomicity on the density either.

Several density-modification methods are based on the representation of the initial density by dummy atoms, *i.e.* using atomicity as a restraint. In the method suggested by Agarwal & Isaacs (1977) a dummy model is built into an MIR map and the atoms refined to get a new set of calculated phases, which are then combined with the previous phases and used to extend the resolution. This was tested on the refinement of insulin with phase extension from 3.0 to 1.5 Å and resulted in some improvement in the map obtained, but the phases still differed by 70° from the final phases. However, the dummy atoms were only used to represent the MIR density and were not a real protein structure. A similar approach was reported by Lunin & Urzhumtsev (1984) for γ-crystallin at 2.7 Å resolution.

In a modification of this approach (Lunin, Urzhumtsev, Vernoslova, Chirgadze, Nevskaya & Fomenkova, 1985) dummy atoms were put around existing protein atoms to represent the 'difference' map calculated from the partial model. This is intermediate between restrained refinement of a protein model and density modification with the restraint of atomicity but without the imposition of 'protein-like' structure. It was developed on γ-crystallin refinement

at 2.7 Å resolution. Urzhumtsev and coworkers (Urzhumtsev, Lunin & Vernoslova, 1989) briefly reported the application of this approach to the refinement of pea lectin with phase extension from 3.0 to 2.4 Å.

In ARP the initial set of atoms comes from a protein model and the initial map is calculated with phases from this model. The initial protein phases are not used again. The refinement imposes atomicity and real protein structure, arising implicitly from the high-resolution data coupled with the reasonable starting model.

ARP has been shown to work very successfully when used with accurate high-resolution data and a reasonable starting model. In its present form ARP should not be used indiscriminately at lower resolution with poor models until considerably more experience and improved criteria have been obtained as such incorrect applications could potentially lead to incorrect models.

The authors thank Zbigniew Dauter for helpful discussion, and Emil Harutunyan and Michael Hennig for providing data and models prior to publication. This work was supported in part by an EMBL postdoctoral fellowship to VSL.

References

AGARWAL, R. C. (1978). *Acta Cryst.* A**34**, 791-809.

AGARWAL, R. C. & ISAACS, N. W. (1977). *Proc. Natl Acad. Sci. USA*, **74**, 2835-2839.

BAKER, E. N. & DODSON, E. J. (1980). *Acta Cryst.* A**36**, 559-572.

BRICOGNE, G. (1974). *Acta Cryst.* A**30**, 395-405.

BRÜNGER, A. T. (1988). *X-PLOR Manual*. Version 1.5. Yale Univ., USA.

BRÜNGER, A. T., KURIYAN, J. & KARPLUS, M. (1987). *Science*, **235**, 458-460.

CHIRGADZE, N. Y., KURANOVA, I. P., NEVSKAYA, N. A., TEPLYAKOV, A. V., WILSON, K. S., STROKOPYTOV, B. V., HARUTUNYAN, E. H. & HÖHNE, W. (1991). *Kristallografiya*, **36**, 128-132.

COHEN, S. A., STERNER, R., KEIM, P. S., & HEINRIKSON, R. L. (1978). *J. Biol. Chem.* **253**, 889-897.

COOPERMAN, B. S. (1982). *Methods Enzymol.* **87**, 526-548.

EGOROV, A. M., AVILOVA, T. V., DIKOV, M. M., POPOV, V. O., RODIONOV, Y. V. & BEREZIN, I. V. (1979). *Eur. J. Biochem.* **99**, 569-576.

EKLUND, H. & BRÄNDÉN, C.-I. (1987). In *Pyridine Nucleotide Coenzymes*, edited by D. N. Y. DOLPHIN, pp. 51-98. New York: Wiley.

HENDRICKSON, W. A. & KONNERT, J. H. (1981). In *Biomolecular Structure Conformation, Function and Evolution*, Vol. 1, edited by R. SRINIVASAN, pp. 43-57. Oxford: Pergamon Press.

HENNIG, M., SCHLESIER, B., PFEFFER, S. & HÖHNE, W. E. (1990). *J. Mol. Biol.* **215**, 339-340.

JAMES, R. W. (1957). In *The Optical Properties of the Diffraction of X-rays*, Vol. 2, *The Crystalline State*. London: Bell.

LAMZIN, V. S., ALESHIN, A. E., STROKOPYTOV, B. V., YUKHNEVICH, M. G., POPOV, V. O., HARUTUNYAN, E. H. & WILSON, K. S. (1992). *Eur. J. Biochem.* **206**, 441-452.

LAMZIN, V. S., POPOV, V. O., HARUTUNYAN, E. H. & WILSON, K. S. (1993). In preparation.

LUNIN, V. YU. (1988). *Acta Cryst.* A**44**, 144-150.

LUNIN, V. YU. & URZHUMTSEV, A. G. (1984). *Acta Cryst.* A**40**, 269-277.

LUNIN, V. YU., URZHUMTSEV, A. G., VERNOSLOVA, E. A., CHIRGADZE, YU. N., NEVSKAYA, N. A. & FOMENKOVA, N. P. (1985). *Acta Cryst.* A**41**, 166-171.

LUZZATI, V. (1953). *Acta Cryst.* **6**, 142-152.

PODJARNY, A. D., BHAT, T. N. & ZWICK, M. (1987). *Annu. Rev. Biophys. Biophys. Chem.* **16**, 351-373

POPOV, V. O., SHUMILIN, I. A., USTINNIKOVA, T. B., LAMZIN, V. S. & EGOROV, TS. A. (1990). *Bioorg. Chem.* **16**, 324-335.

SAYRE, D. (1972). *Acta Cryst.* A**28**, 210-212.

SAYRE, D. (1974). *Acta Cryst.* A**30**, 180-184.

SERC Daresbury Laboratory (1979). *CCP4. A Suite of Programs for Protein Crystallography*. SERC Daresbury Laboratory, Warrington, England.

SHELDRICK, G. M. (1976). *SHELX*. Program for crystal structure determination. Univ. of Göttingen, Germany.

SHELDRICK, G. M., DAUTER, Z., WILSON, K. S., HOPE, H. & SIEKER, L. C. (1993). *Acta Cryst.* D**49**, 18-23.

SUSSMAN, J. L., HOLBROOK, S. R., CHURCH, G. M. & KIM, S. H. (1977). *Acta Cryst.* A**33**, 800-804.

URZHUMTSEV, A. G., LUNIN, V. YU. & VERNOSLOVA, E. A. (1989). *J. Appl. Cryst.* **22**, 500-506.

WANG, B. C. (1985). *Methods Enzymol.* **115**, 90-112.

ZHANG, K. Y. J. & MAIN, P. (1990). *Acta Cryst.* A**46**, 41-46.