Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

ACORN: a review

J.-X. Yao,[a]* E. J. Dodson,[a]* K. S. Wilson[a] and M. M. Woolfson[b]

[a]York Structural Biology Laboratory, Department of Chemistry, University of York, Heslington, York YO10 5YW, England, and [b]Department of Physics, University of York, Heslington, York YO10 5YW, England

Correspondence e-mail: yao@ysbl.york.ac.uk

Received 7 February 2006
Accepted 7 March 2006

The *ACORN* system was originally developed as a means of *ab initio* solution of protein structures when atomic resolution data were available. The first step is to obtain a starting set of phases, which must be at least slightly better than random. These may be calculated from a fragment of the structure, which can be anything from a single metal atom to a complete molecular-replacement model. A number of standard procedures are available in *ACORN* to orientate and position such a fragment. The fragment provides initial phases that give the first of a series of maps that are iteratively refined by a dynamic density-modification (DDM) process. Another FFT-based procedure is Sayre-equation refinement (SER), which modifies phases better to satisfy the Sayre equation. With good-quality atomic resolution data, the final outcome of applying DDM and SER is a map similar in appearance to that found from a refined structure, which is readily interpreted by automated procedures. Further development of *ACORN* now enables structures to be solved with less than atomic resolution data. A critical part of this development is the artificial extension of the data from the observed limit to 1 Å resolution. These extended reflections are allocated unit normalized structure amplitudes and then treated in a similar way to observed reflections except that they are down-weighted in the calculation of maps. *ACORN* maps, especially at low resolution, tend to show C atoms less well, in particular C^{α} atoms which fall within the first diffraction minimum of their three neighbours. Two new density-modification procedures (DDM1 and DDM2) and a density-enhancement procedure (ENH) have been devised to counter this problem. It is demonstrated that high-quality maps showing individual atoms can be produced with the new *ACORN*. *ACORN* has also been demonstrated to be very effective in refining phase sets derived from physical processes such as those using anomalous scattering or isomorphous derivative data. Future work will be directed towards applying *ACORN* to resolutions down to 2 Å.

1. Direct methods and protein

In the period from about 1970 to 1990 automated direct methods were developed to the stage where, for all practical purposes, it could be claimed that the problem of phasing small-molecule structures had been solved, as long as data of reasonable quality were available. Structural protein crystallography also had its successes based on physical methods, such as anomalous scattering and isomorphous replacement, and on molecular replacement, which became more and more applicable as the protein database grew.

The improvement of both the quality and the resolution of protein data sets, brought about by the introduction of better instrumentation and experimental procedures, and the ready availability of high-power high-capacity computers prompted attempts to extend direct methods into the protein regime. Woolfson & Yao (1990) demonstrated that a small protein, avian pancreatic polypeptide, could be solved by the standard direct-methods program *SAYTAN*, but this structure, with 302 independent atoms including one zinc, was not very much larger than the small-molecule structures that had already been solved. Other work by Mukherjee and coworkers (Mukherjee & Woolfson, 1995; Mukherjee *et al.*, 1999, 2000, 2001) showed that direct methods could give a starting point for the solution of proteins with up to about 1000 atoms, although it required considerable painstaking effort to go from the maps provided by direct methods to the final solution. It was notable that some of the trial structures used by Mukherjee and coworkers had less than atomic resolution data. Procedures that are incorporated in the *Shake-and-Bake* method (Weeks & Miller, 1999; Xu *et al.*, 2000) and in *SHELXD* (Sheldrick & Gould, 1995; Sheldrick, 1997) extend direct methods to proteins with up to 1000 independent non-H atoms as long as atomic resolution data are available.

2. The *ACORN* concept

To make progress with creating an automated procedure for solving protein structures, it seemed sensible to move away from reciprocal-space methods that increased in scale as the square of the number of reflections to be phased. In any case, for larger proteins, as the work of Mukherjee and coworkers showed, the best that could be expected is a rather imprecise set of phases that had to be developed somehow into a refined structure. The *ACORN* concept grew first from the idea of finding other ways to obtain a trial set of phases and then later from consideration of how to go, by automatic processes, from approximate phases to a refined structure. In the first instance it was assumed that atomic resolution (better than 1.2 Å) data were available.

Although the *ab initio* solution of proteins seems to be a formidable task when just considered in terms of the size of the structures, there are some compensating factors. There are just 20 amino acids that constitute the main part of a protein, differing only in their side chains. More significantly, perhaps, many proteins contain characteristic substructures of amino acids, such as α-helices, the presence of which can be deduced either from the known structures of related proteins or from features of the Patterson function. In addition, the positions of the heavy atoms in metalloproteins are often easy to determine and even substructures of lighter atoms, such as sulfur, can be determined from anomalous difference data.

These considerations led to the first idea of developing a protein structure from initial phases provided by a small part of the scattering matter in the cell. The fragment providing this starting point can be of several different types.

2.1. Experimentally positioned anomalous scatterers

When moderately heavy atoms are present in a structure, *e.g.* Zn, Mg, Se, they can usually be located by inserting anomalous difference magnitudes into a direct-methods procedure. With very precise data, even lighter atoms such as S and Cl can be picked up in this way (Dauter *et al.*, 1999, 2000). The separations of the anomalous scatterers are usually large enough for them to be located with less than atomic resolution data.

2.2. Molecular replacement

The whole, or a large part of, a molecule from a solved related structure can be used as a starting model. This can be located in orientation and position by a standard molecular-replacement program.

2.3. A small fragment

Usable fragments can be as little as 1–8% of the total structure and consist of a single α-helix or a small part of a β-sheet. In favourable cases, the use of *AMoRe* (Navaza, 1994) may correctly locate the fragment. A slower but more effective way to locate the fragment is to find the orientation and position that gives the best correlation coefficient

$$\mathrm{CC} = \frac{\langle |E_{\mathrm{frag}} E_{\mathrm{o}}| \rangle - \langle |E_{\mathrm{frag}}| \rangle \langle |E_{\mathrm{o}}| \rangle}{\sigma_{\mathrm{frag}} \sigma_{\mathrm{o}}}, \qquad (1)$$

where E_{frag} is the normalized structure factor for the fragment, E_{o} is the observed normalized structure factor and the σ is the corresponding standard deviations.

The process is factorized into first finding the orientation and then the translation to position the fragment correctly; the details of this process have been described by Foadi *et al.* (2000).

3. Refining the phases

The initial mean phase error derived from the fragment, whatever its origin, is usually of the order of 75°. An electron-density map calculated with such phases will show the fragment which gave the phases, but elsewhere the map is usually impossible to interpret in structural terms. In the process of phase refinement, the maps are calculated just with terms corresponding to the largest E values, typically with $|E| > 1.2$, but with a lower limit if more E values are needed. Three processes are used to refine the phases.

3.1. Patterson superposition *via* the sum function

In principle, the Patterson function contains all the information about the structure. It is found that applying the superposition Patterson sum function right at the beginning of the phase-refinement process provides some benefit. The Fourier coefficients of the map are $|E_{\mathrm{o}}||F_{\mathrm{o}}|E_{\mathrm{frag}}$, which have the phase of the fragment. This is the starting map for refinement and usually reduces the mean phase error by 1–2° below what

it would otherwise have been after the next refinement process.

3.2. Maps, dynamic density modification (DDM) and CC_s

The DDM process is based on the simple premise that, outside the region occupied by the fragment, the higher the map density the greater the probability that the position is occupied by true electron density from the structure. This leads to a density-modification procedure based on the value of σ, the standard deviation of the map density,

$$\rho' = \begin{cases} 0 & \text{if } \rho \leq 0 \\ \rho \tanh[0.2(\rho/\sigma)^{3/2}] & \text{if } \rho > 0 \\ kn\sigma & \text{if } \rho' > kn\sigma \end{cases}, \qquad (2)$$

where k is a constant given by the user (default value 3) and n is the number of the DDM iteration cycle but with a maximum value of 5. This density-modification function is best explained by its appearance as shown in Fig. 1. Negative density, a non-physical value, is made equal to zero. Over the remainder of the range the lowest density, which has the least probability of corresponding to real density, is down-weighted compared with higher density. The cutoff at higher densities reduces the bias introduced by the fragment but it is made higher as the density improves to avoid removing newly acquired structural information.

An important feature of *ACORN* is that the phase-refinement procedure is automatic and requires no user intervention. It is therefore necessary to have some check that the refinement procedure is actually producing better phases and also to know when the refinement process has terminated. To this end, *ACORN* uses a simple but amazingly efficient monitor of progress, namely CC_s, the calculation of the correlation coefficient for the small-amplitude E values

$$CC_s = \frac{\langle |E_o E_c| \rangle - \langle |E_o| \rangle \langle |E_c| \rangle}{\sigma_o \sigma_c}, \qquad (3)$$

where E_c is derived from the Fourier transform of a modified density map and the σs are standard deviations. In finding CC_s, the terms used are those with smaller E_o that were not included in the original map. The remarkable relationship between CC_s and mean phase error is shown in Fig. 2 as stages of DDM refinement progress for the structure of the ENT domain of the human EMSY protein (Chavali *et al.*, 2005; PDB code 1uz3) solved using *ACORN*. This is typical of that found for the many structures investigated so far.

When a map has been modified by DDM not only are Fourier coefficients found that enable CC_s to be calculated, but additionally the Fourier coefficients that were in the unmodified map have also been changed both in magnitude and phase. When the next map is calculated the Fourier coefficients are weighted by

$$W = \tanh\left[\frac{|E_o||F'_c|}{2(\Sigma')^{1/2}}\right], \qquad (4)$$

where $|F'_c|$ is the magnitude of the Fourier coefficient of the map scaled to the values of $|E_o|$ in shells and Σ' is the mean square of the values of $|F'_c|$.

3.3. Sayre-equation refinement (SER)

From time to time the refinement, as indicated by CC_s, seems to be progressing but slows down significantly. In such cases it is often beneficial to perform one or two cycles of SER. This modifies the phases in such a way that Sayre's equation

$$E(\mathbf{h}) = \theta(\mathbf{h}) \sum_{\mathbf{k}} E(\mathbf{k}) E(\mathbf{h} - \mathbf{k}) \qquad (5)$$

is better satisfied for both large and small normalized structure factors. In keeping with the basic philosophy of *ACORN*,

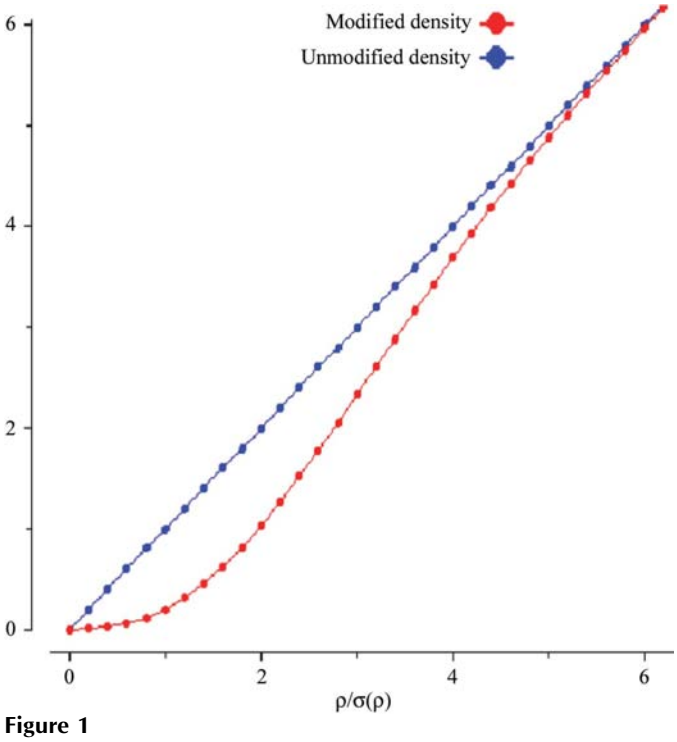

Figure 1
Density-modification function of DDM.

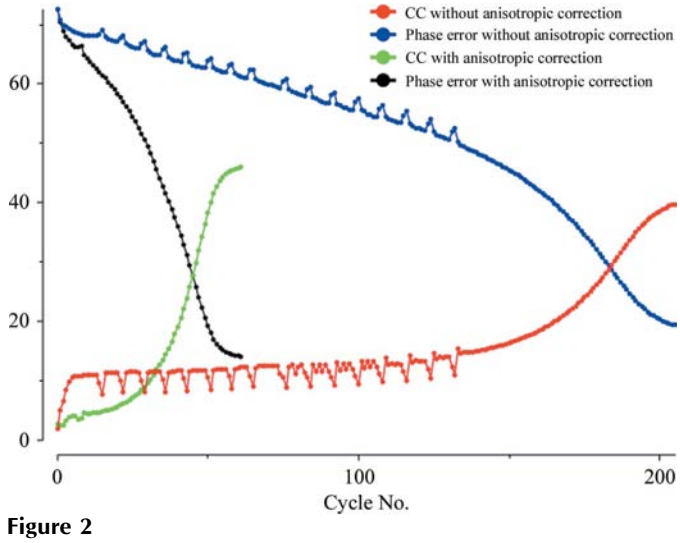

Figure 2
The phase errors (°) and CC_s (%) *versus* the cycle number of DDM for 1uz3 with and without data correction for anisotropy.

Table 1
A representative set of structures solved using *ACORN*.

N_{atoms}: the number of atoms in the asymmetric unit. Fragment: the starting fragment used to provide initial phases for *ACORN*. Located by: the method used to locate the fragment, where appropriate.

Structure	PDB code	Space group	Resolution (Å)	N_{atoms}	Fragment/located by	Reference
Deuterolysin	1eb6	$P2_1$	1.00	1665	1 Zn/*ACORN* or Patterson	McAuley *et al.* (2001)
Crustacyanin	1i4u	$P2_12_12_1$	1.15	3383	12 S/*SnB*	Gordon *et al.* (2001)
HFBII hydrophobin	1r2m	$C2$	1.00	1261	1 Mn/SAD Patterson	Hakanpaa *et al.* (2004)
CBM36	1w0n	$P2_12_12_1$	0.80	1160	1 Se/SAD Patterson	Jamal-Talabani *et al.* (2004)
EMSY protein	1uz3	$P2_1$	1.10	1921	75-atom α-helix/*ACORN*	Chavali *et al.* (2005)
Collagen model peptide	1wzb	$P2_1$	1.50	767	MR (1k6f)/*ACORN*	Kawahara *et al.* (2005)
Calcyclin-binding protein	2a26	$C222_1$	1.20	1493	Se-based phases refined by *ACORN*	Santelli *et al.* (2005)
Feruloyl esterase	1uwc	$P1$	1.08	4909	Phase extension with *ACORN*	McAuley *et al.* (2004)

which is based on the use of the FFT, the implementation of a single cycle of SER requires six separate FFTs and so is quick and easy to apply (Foadi *et al.*, 2000). The curious thing is that inserting SER has an initially detrimental effect on both CC_s and the mean phase error, but accelerates the subsequent refinement.

4. Successful applications

A number of structures have been reported recently in which the *CCP*4 version of *ACORN* has been used to obtain the solution for proteins where data to better than 1.3 Å resolution are available. Table 1 lists some of these using various types of starting phases.

We have revisited the solution of the human chromatin regulator (Chavali *et al.*, 2005; PDB code 1uz3) using the data deposited by the authors. This largely helical structure belongs to space group $P2_1$ with 204 residues (1550 atoms) in two chains. The data nominally extend to a resolution of 1.1 Å, but are somewhat anisotropic. We corrected the anisotropy using the *Phaser* utility (Storoni *et al.*, 2004) and started *ACORN* from a ten-residue idealized helical model representing 3% of the scattering matter in the cell for both the deposited and the corrected data. With the uncorrected set, *ACORN* took 206 cycles to reach convergence. In marked contrast, with the set corrected for anisotropy convergence was reached after 61 cycles, which took 114 s of CPU time on a single-processor 1.8 GHz Athlon workstation. Fig. 2 shows the progress of CC_s in the refinement and the reduction of the mean phase error for the strong E values.

This example is typical of the way that *ACORN* operates to solve structures. Refinement times by DDM can vary from less than a minute to hours, but even with longer refinement times positive progress is indicated by a steady increase in the value of CC_s.

5. Extending to lower resolution

For protein structures with data to better than 1.3 Å resolution the original version of *ACORN* provides a convenient and effective means of refining an initial set of phases by density modification. The phases may come from a fragment, which could be a single metal atom, a motif such as an α-helix or even a complete homologous molecule, and *ACORN* offers several ways of finding and using such information. Fragment positions derived from either anomalous scattering or isomorphous replacement data can be particularly useful and phases from MIR or MAD data can also be used to provide an initial map for refinement.

The requirement for atomic resolution data limited the number of structures to which *ACORN* could be applied and in 2002 we derived an approach to applying *ACORN* to lower resolution data, first reported at 'The Dodson Symposium', a meeting to recognize the retirement of Eleanor and Guy Dodson held in York in September 2004, and subsequently published by Yao *et al.* (2005). The basis of the method is straightforward: it is artificially to extend the data to 1 Å and to include these data in the refinement process. It was also found beneficial to artificially fill in any gaps in the data out to the observed limit arising from effects such as saturated reflections at low resolution or substantial blind regions.

On the face of it, this seems an extraordinary step to take and one might ask the question of how unobserved data can provide information. In retrospect, it is clear that effectively setting all unobserved E values to 0.0 is a very poor Bayesian estimate. The critical question is whether or not reasonable phase estimates can be made for these added reflections and whether these estimates can be refined. It is well known that reflections with correct phases and random magnitudes will give a map showing a distorted view of the structure, thus illustrating the information content of phases independent of the magnitudes. However, given that we can find some phase estimates for the extended data, the question then arises of what magnitudes to give them. Four different ways of estimating the magnitudes were examined, from which it was concluded that the best estimate to take was $|E| = 1$ for all reflections; that is, to give each reflection the average magnitude (Yao *et al.*, 2005). In practice, the overall contribution of these estimated terms is down-weighted using the formula below.

The DDM refinement process is carried out as usual, but for the extended reflections, instead of the weight (4), a weight is used that takes into account the inevitable errors in their magnitudes. This is

$$W_{\rm ext}(\mathbf{h}) = \tanh(0.5X_{\rm ext}), \quad (6)$$

where

$$X_{\rm ext} = \left(\frac{n_{\rm obs}}{n_{\rm obs}+n_{\rm ext}}\right)^{1/2} |E_{\rm c}(\mathbf{h})|,$$

$n_{\rm obs}$ is the observed number of reflections, $n_{\rm ext}$ is the number of extended reflections and $|E_{\rm c}(\mathbf{h})|$ from the fragment or map is scaled in shells in reciprocal space to make $\langle|E_{\rm c}|^2\rangle = 1$.

A test of the extension idea was made with the structure penicillopepsin (PDB code 1bxo; Ding *et al.*, 1998), which belongs to space group *C*2 with 2977 atoms in the asymmetric unit. Although the observed data extended to 0.9 Å, they were terminated at 1.5 Å for the purpose of the test. Using all the data and a trial 400-atom fragment, the final mean phase error given by *ACORN* was 13.1°. In truncating the data to 1.5 Å it was necessary to include reflections with $|E| > 0.8$ in the map calculation. Artificially extending the data to 1.0 Å and applying *ACORN* as described above reduced the mean phase error from an initial value of 59.5° to 42.7° after 35 cycles of DDM. If the original *ACORN* without artificial extension was applied to 1.5 Å data for 1bxo, the final mean error was little different from the initial value of 59.5°.

6. An analysis of the problems of low resolution

Data to less than atomic resolution present a challenge to *ab initio* methods of protein structure solution in different ways. It can be shown theoretically (Yao *et al.*, 2002) that for a structure with N equal non-H atoms in the asymmetric unit, M independent observed reflections and with true phases the ratio of the electron density at an atomic centre, $\rho_{\rm c}$, to the standard deviation of the electron density, σ, is given by

$$\frac{\rho_{\rm c}}{\sigma} = \left(\frac{M}{N}\right)^{1/2}. \quad (7)$$

For an average protein at 1 Å resolution $M/N \simeq 120$, so an atomic peak has density about 11σ, which makes it very distinctive. Even at a resolution of 1.7 Å, with perfect phases, a peak has density $\sim 5\sigma$, which would still show up rather well. However, with a mean phase error $\Delta\varphi$, assumed to be independent of the value of $|E|$, we have

$$\frac{\rho_{\rm c}}{\sigma} = \left(\frac{M}{N}\right)^{1/2} \cos\Delta\varphi. \quad (8)$$

With typical initial mean phase errors of 75°, this makes the ratios of peak height to standard deviation 2.8 for 1 Å resolution and 1.3 for 1.7 Å resolution. The former value gives a reasonable probability that atomic positions occur within the higher regions of map density. The latter value is well within the normal range of random fluctuation, so recognizing peaks becomes almost impossible. This shows the need for better starting phases for obtaining solutions with low-resolution data.

Another problem is that the expectation density at atomic centres disproportionately favours heavy atoms, so that S atoms appear well and many C atoms do not appear at all. This is especially true for C^{α} atoms that are connected to three others. In the resolution range 1.5–1.7 Å the first diffraction ripple owing to series termination occurs at about a bond distance. Hence, the density at a C^{α} position is heavily depressed by its neighbours. These problems are partially dealt with by artificially extending data, but we have still found it advantageous to introduce procedures specifically to enhance weak density.

7. Density enhancement

The DDM map coefficients are initially $W(\mathbf{h})E_{\rm o}(\mathbf{h})$, but as the electron density develops to show more of the structure, other kinds of coefficient become useful. We have introduced two modifications of DDM, which we now designate as DDM0, with coefficients as follows.

For DDM1,

$$W(\mathbf{h})[2|E_{\rm o}(\mathbf{h})| - |E_{\rm c}(\mathbf{h})|]\exp[i\varphi_{\rm c}(\mathbf{h})], \quad (9)$$

where for extended reflections $|E_{\rm o}(\mathbf{h})| = |E_{\rm ext}(\mathbf{h})| = 1$.

For DDM2,

$$[2m(\mathbf{h})|E_{\rm o}(\mathbf{h})| - \sigma_{\rm A}|E_{\rm c}(\mathbf{h})|]\exp[i\varphi_{\rm c}(\mathbf{h})]. \quad (10)$$

The quantity $\sigma_{\rm A}$ was first described by Srinivasan (1966) and developed for use with proteins by Read (1986). The figure of merit m depends on the value of $\sigma_{\rm A}$ and also on the agreement between $E_{\rm o}(\mathbf{h})$ and $E_{\rm c}(\mathbf{h})$. The value of $\sigma_{\rm A}$ reflects the mean phase error which we estimate from the value of $CC_{\rm s}$. When the phase error is large then $\sigma_{\rm A}$ is small and DDM2 gives virtually the same result as DDM0.

Normal cycles of refinement are carried out using DDM0, DDM1 or DDM2 (as specified by the user with DDM0 as the default) until $CC_{\rm s}$ ceases to increase. One cycle of density enhancement is then introduced that consists of the following steps.

(i) Calculate a map with DDM1 (default) or DDM2 as specified by the user.

(ii) Negative density is set to zero and then the square root is taken of the remaining density to enhance the weaker density.

(iii) For a map with coefficient magnitudes $W(\mathbf{h})|F_{\rm o}(\mathbf{h})|$ the density at each point is replaced by the average density within a 2 Å sphere centred on the point. This is perfomred by a convolution technique and enables a protein envelope to be found (Wang, 1981, 1985). All the density in solvent regions is set to zero.

(iv) Weak density is enhanced by multiplying the density at each grid point by the average density within a 2 Å sphere centred at the point.

(v) All density greater than $n\sigma$ is made equal to $n\sigma$ (default value $n = 3$).

Once these steps have been carried out, further refinement using DDM0, DDM1 or DDM2 is continued.

8. Examples of running new *ACORN*

8.1. *Campylobacter jejuni* dUTPase (PDB code 1w2y)

The space group of this structure is $P2_12_12_1$, with 430 residues (3851 non-H atoms) in the asymmetric unit. There were 52 878 data to 1.65 Å resolution and the data were artificially extended to 1 Å. A low-homology model, *Trypanosoma cruzi* dUTPase (PDB code 1ogk), was used to solve the structure by molecular replacement with 4 Å data (Moroz *et al.*, 2004).

In the investigation of the new *ACORN* three different approaches were made. In the first, to test the new procedures, a starting fragment was used consisting of 400 accurately placed atoms. The initial phase error of 62.2° fell to 46.4° after 35 cycles of DDM0. Further cycles of density enhancement (ENH) and DDM1 and DDM2 cycles gave a final mean phase error of 35.7° with $CC_s = 0.137$. The benefits of using the new *ACORN* procedures are clear, in particular by comparing the results illustrated in Fig. 3, which show maps based on the original *ACORN* procedure and on the current *ACORN* using all its new features. The enhanced map shows most individual atoms quite clearly.

Using 400 perfectly placed atoms as a starting point is clearly unrealistic, so now we describe some more representative approaches to solving this structure. Using the low-homology model with which the structure had originally been solved as a starting fragment, the original mean phase error was 78.3° for 26 859 reflections with $|E_o| > 0.8$ with $CC_s = 0.0221$. After 841 cycles of *ACORN* refinement using DDM1, DDM2 and ENH the phase error was reduced to 43.4° with $CC_s = 0.1068$. The final map correlation, 0.66, was a considerable improvement on that from the original MR model.

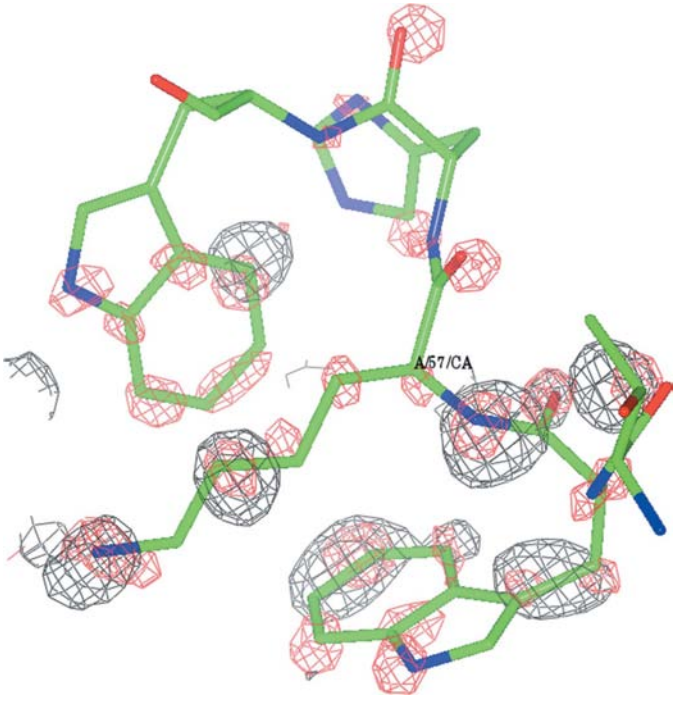

Figure 3
E-maps for 1w2y for the original (blue) and modified (red) *ACORN* procedures.

In a final approach, the original MR model was subjected to rigid-body refinement before *ACORN* was applied. After 27 cycles of refinement with DDM0 the mean phase error was 36.7° and further refinement, including ENH, gave a mean phase error of 33.4° after a total of 140 cycles. The final map, with a correlation coefficient 0.77, was easily interpretable.

8.2. Exo-mannosidase (PDB code 1uuq)

The space group of this structure is $P2_12_12_1$, with 440 residues (4129 non-H atoms) in the asymmetric unit. There were 74 907 data to 1.50 Å resolution and the data were artificially extended to 1 Å. The structure was originally solved by Dias *et al.* (2004).

Measurements at the Se peak wavelength were available to 1.6 Å and by use of *SHELXD* and *SHELXE* the Se positions and initial experimental phases were determined to 2.5 Å (16 365 reflections with mean phase error 50.0°), 2.1 Å (27 562 reflections with mean phase error 42.1°) and 1.6 Å (61 777 reflections with mean phase error 27.6°). There were 176 738 extended reflections. The phase error for the strong reflections with $|E| > 0.8$ (38 372 reflections to 1.5 Å) reduced to 22.9° from 50.0° (2.5 Å) after 54 cycles of DDM0 and DDM1, 20.2° from 42.1° (2.1 Å) after 22 cycles of DDM0 and DDM1 and 17.7° from 27.6° (1.6 Å) after 13 cycles of DDM0 and DDM1. The progress of CC_s and the mean phase error is shown in Fig. 4 for starting from 2.5 Å initial experimental phases. The resultant map was of excellent quality and showed the power of *ACORN* as an adjunct to SAD data.

In the second test with this structure, the 16 Se atoms were used as a starting fragment, giving a mean phase error of 74.1° for the strong reflections with $|E| > 0.8$. The refinement consisted of cycles of DDM0 with insertions of single cycles of SER whenever the refinement slowed. The refinement was extremely slow, but after 2712 cycles CC_s was 0.17 and the mean phase error was 30°. Another 44 cycles with DDM1 gave a final CC_s of 0.25 and a mean phase error 21.0°. An *F*-map

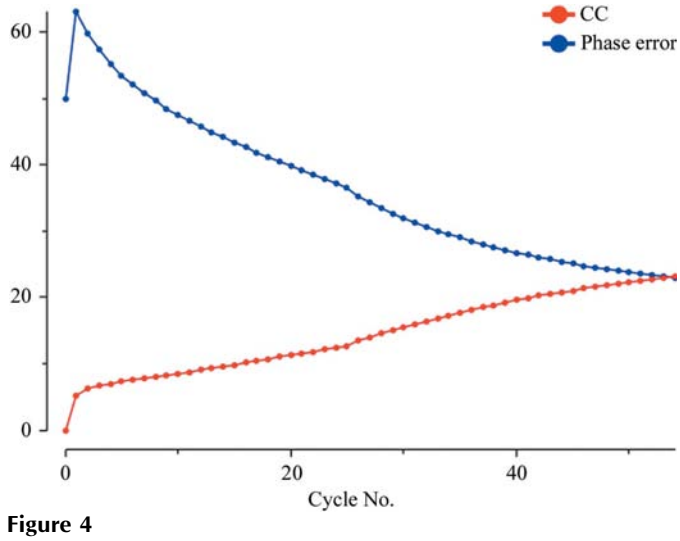

Figure 4
The phase errors (°) and CC_s (%) *versus* the cycle number for 1uuq starting from 2.5 Å initial experimental phases.

Table 2
Application of the new *ACORN*, with artificial data extension and DDM enhancement, to a set of structures from the Joint Centre for Structural Genomics archive (http://www.jcsg.org/datasets-info.shtml).

PE: phase error (°) compared with those calculated from the deposited model, followed by the number of observations in parentheses. Initial PE: phases taken from the archived data; it is not always clear exactly how these phases were derived. Run 1: *ACORN* starting from all available experimental phases. Run 2: starting phases restricted to 2.5 Å resolution. N_{cyc}: number of cycles. For 1vpj* the initial phases were derived from only nine of the 16 Se sites. The phasing was repeated to 1.74 Å using 15 sites and the *ACORN* runs repeated.

PDB code	N_{atoms}	Resolution (Å)	Initial PE	PE *ACORN* Run 1	N_{cyc}	Initial PE to 2.5 Å	PE *ACORN* Run 2	N_{cyc}
1vly	3020	1.30	59.1 (74527)	28.6 (75749)	14	23.9 (11644)	29.9 (75749)	25
1vp8	1704	1.53	46.1 (31456)	30.4 (41005)	14	35.5 (9756)	32.0 (41004)	22
1vmg	826	1.46	63.2 (24916)	25.2 (25182)	13	36.3 (5602)	25.9 (25103)	18
1vqs	5720	1.50	40.2 (24539)	37.2 (110183)	31	40.2 (24278)	37.3 (110181)	31
1vpm	3951	1.66	57.1 (60458)	36.7 (62700)	7	37.1 (18490)	41.0 (62700)	35
1vpj*	2915	1.69	75.2 (33765)	77.3 (42117)	76	74.0 (14136)	84.5 (42117)	63
1vpj	2915	1.74	65.0 (41359)	48.7 (41403)	94	58.2 (14265)	59.2 (41401)	118
1vpb	3864	1.75	57.7 (47678)	44.7 (47842)	6	42.1 (21215)	48.9 (47841)	16
1vmf	3782	1.80	55.7 (42167)	40.4 (42325)	7	36.9 (16336)	72.6 (42330)	75

with observed data and an *E*-map including extended reflections to 1.0 Å are shown in Fig. 5. The whole procedure took 9 h on a 1.1 GHz laptop, but was quite automatic and required no user intervention.

8.3. Structures from the JCSG

The Joint Center for Structural Genomics (JCSG, principal investigator Professor Ian Wilson) is one of the Structural Genomics initiatives funded by the National Institutes of Health. Many of the structures have not yet been published, but coordinates and the data against which these were refined are available from the PDB. However, for 60 of these structures, Dr Ashley Deacon has archived all the experimental data together with the derived SAD or MAD phases (http://www.jcsg.org/datasets-info.shtml) and made them available for test purposes to program developers. These have proved especially useful in evaluating the new *ACORN*.

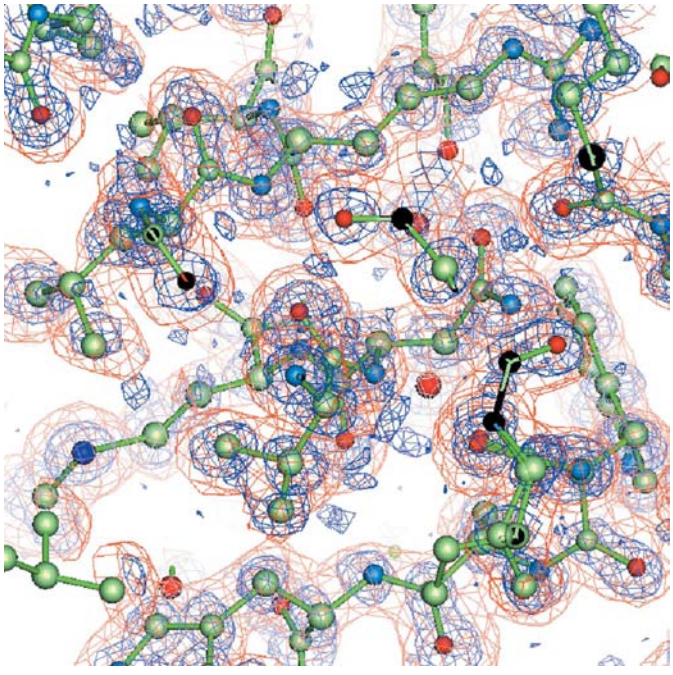

Figure 5
F-map (red) and *E*-map (blue) from modified *ACORN* for 1uuq.

We analysed the performance of the new *ACORN* in refining the experimental phases from this archive. 15 sets had data to 1.8 Å or better and we applied *ACORN* to eight of these (Table 2). For each structure we first corrected the data for anisotropy using the *Phaser* utility, as described above for 1uz3. For each structure, *ACORN* was first run using all the experimental phases provided. The resolution for these was frequently lower than that of the measured amplitudes. *ACORN* was used to extend and refine phases for all the amplitudes using the approach described in §5, *i.e.* with data extension to 1.0 Å and the enhanced density modification. In seven out the eight cases, *ACORN* was extremely successful, as can be seen from columns 4 and 5 of Table 2. In addition, the procedure converged very rapidly, taking less than 5 min even for the larger structures on a single-processor 1.8 GHz Athlon workstation.

The *ACORN* runs were repeated after truncating the resolution of the starting phases to 2.5 Å. The results are shown in columns 7 and 8 of Table 2. For six of the eight, the resulting phases, obtained with about twice as many cycles, were almost equally as good as those from the previous calculations. This simulates experiments where the SAD or MAD data are only measured to a lower resolution, while the set of observations, either for the native or SeMet protein, are pushed to the highest possible spacing. To us, this suggests that this is indeed an efficient way to obtain good phases, allowing the user to concentrate on getting highly accurate and redundant data to a limited resolution. In addition, the phase extension by *ACORN*, as stated above, makes few assumptions about the nature of the model or solvent, unlike procedures such as *ARP*/*wARP*, and might be expected to achieve a less biased map.

9. *ACORN* now and future developments

ACORN is not a substitute for the traditional physical methods of solving protein structures involving anomalous scattering or isomorphous replacement. However, the

example of 1uuq shows that it can be a useful and powerful adjunct to a physical method, regarding it rather as a refinement tool. An important aspect of *ACORN* in this regard is its lack of bias in the refinement process.

Where physical data are not available then *ACORN* can be used as a method of *ab initio* solution, starting with a fragment that can be anything from a single atom to a complete molecular-replacement molecule. It is in this area that developments are being sought to find better methods of generating starting fragments of a size and accuracy that can give suitable starting points for solutions with lower resolution data.

Another issue that is being explored is that of being able to apply *ACORN* down to resolutions of 2 Å, a limit that would bring a a high proportion of protein data sets into its orbit. As we have seen, there are powerful theoretical limitations on what can be performed with low-resolution data, but we have also seen that an unexpected development, such as data extension, can overcome such limitations to some extent.

We are grateful for the support of BBSRC for this project through grant 87/B15857. We also express our thanks to CCP4 for funding.

References

Chavali, G. B., Ekblad, C. M., Basu, B. P., Brissett, N. C., Veprintsev, D., Hughes-Davies, L., Kouzarides, T., Itzhaki, L. S. & Doherty, A. J. (2005). *J. Mol. Biol.* **350**, 964–973.

Dauter, Z., Dauter, M., de La Fortelle, E., Bricogne, G. & Sheldrick, G. M. (1999). *J. Mol. Biol.* **289**, 83–92.

Dauter, Z., Dauter, M. & Rajashankar, K. R. (2000). *Acta Cryst.* D**56**, 232–237.

Dias, F. M. V., Vincent, F., Pell, G., Prates, J. A. M., Centeno, M. S. J., Tailford, L. E., Ferreira, L. M. A., Fontes, C. M. G. A., Davies, G. J. & Gilbert, H. J. (2004). *J. Biol. Chem.* **279**, 25517–25526.

Ding, J., Frasere, M. E., Meyer, J. H. & Bartlett, P. A. (1998). *J. Am. Chem. Soc.* **120**, 4610–4621.

Foadi, J., Woolfson, M. M., Dodson, E. J., Wilson, K. S., Yao, J.-X. & Zheng, C.-D. (2000). *Acta Cryst.* D**56**, 1137–1147.

Gordon, E. J., Leonard, G. A., McSweeney, S. & Zagalsky, P. F. (2001). *Acta Cryst.* D**57**, 1230–1237.

Hakanpaa, J., Paananen, A., Askolin, S., Nakari-Setala, T., Penttila, M. & Rouvinen, J. (2004). *J. Biol. Chem.* **279**, 534–539.

Jamal-Talabani, S., Boraston, A. B., Turkenburg, J. P., Tarbouriech, N., Ducros, V. M. & Davies, G. J. (2004). *Structure*, **12**, 1177–1187.

Kawahara, K., Nishi, Y., Nakamura, S., Uchiyama, S., Nishiuchi, Y., Nakazawa, T., Ohkubo, T. & Kobayashi, Y. (2005). *Biochemistry*, **44**, 15812–15822.

McAuley, K. E., Svendson, A., Patkar, S. A. & Wilson, K. S. (2004). *Acta Cryst.* D**60**, 878–887.

McAuley, K. E., Yao, J.-X., Dodson, E. J., Lehmbeck, J., Ostergaard, P. R. & Wilson, K. S. (2001). *Acta Cryst.* D**57**, 1571–1578.

Moroz, O. V., Harkiolaki, M., Galperin, M. Y., Vagin, A. A., González-Pacanowska, D. & Wilson, K. S. (2004). *J. Mol. Biol.* **342**, 1583–1597.

Mukherjee, M., Ghosh, S. & Woolfson, M. M. (1999). *Acta Cryst.* D**55**, 168–172.

Mukherjee, M., Maiti, S., Ghosh, S. & Woolfson, M. M. (2001). *Acta Cryst.* D**57**, 1276–1280.

Mukherjee, M., Maiti, S. & Woolfson, M. M. (2000). *Acta Cryst.* D**56**, 1132–1136.

Mukherjee, M. & Woolfson, M. M. (1995). *Acta Cryst.* D**51**, 626–628.

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

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

Santelli, E., Leone, M., Li, C., Fukushima, T., Preece, N. E., Olsen, A. J., Ely, K. R., Reed, J. C., Pellecchia, M., Lidington, R. C. & Matsuzawa, S. (2005). *J. Biol. Chem.* **280**, 34278–34287.

Sheldrick, G. M. (1997). *Proceedings of the CCP4 Study Weekend. Recent Advances in Phasing*, edited by K. S. Wilson, G. Davies, A. Ashton & S. Bailey, pp. 147–158. Warrington: Daresbury Laboratory.

Sheldrick, G. M. & Gould, R. O. (1995). *Acta Cryst.* B**51**, 423–431.

Srinivasan, R. (1966). *Acta Cryst.* **20**, 143–144.

Storoni, L. C., McCoy, A. J. & Read, R. J. (2004). *Acta Cryst.* D**60**, 432–438.

Wang, B.-C. (1981). *Acta Cryst.* A**37**, C11.

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

Weeks, C. M. & Miller, R (1999). *Acta Cryst.* D**55**, 492–500.

Woolfson, M. M. & Yao, J.-X. (1990). *Acta Cryst.* A**46**, 409–413.

Xu, H., Hauptman, H. A., Weeks, C. M. & Miller, R. (2000). *Acta Cryst.* D**56**, 238–240.

Yao, J.-X., Woolfson, M. M., Wilson, K. S. & Dodson, E. J. (2002). *Z. Kristallogr.* **217**, 636–643.

Yao, J.-X., Woolfson, M. M., Wilson, K. S. & Dodson, E. J. (2005). *Acta Cryst.* D**61**, 1465–1475.