Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

A flexible and efficient procedure for the solution and phase refinement of protein structures

J. Foadi,[a] M. M. Woolfson,[a] E. J. Dodson,[b] K. S. Wilson,[b]* Yao Jia-xing[b] and Zheng Chao-de[c]

[a]Physics Department, University of York, York YO10 5DD, England, [b]York Structural Biology Laboratory, Chemistry Department, University of York, York YO10 5DD, England, and [c]Institute of Physics, Beijing 100080, People's Republic of China

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

Received 8 May 2000
Accepted 27 June 2000

An *ab initio* method is described for solving protein structures for which atomic resolution (better than 1.2 Å) data are available. The problem is divided into two stages. Firstly, a substructure composed of a small percentage (∼5%) of the scattering matter of the unit cell is positioned. This is used to generate a starting set of phases that are slightly better than random. Secondly, the full structure is developed from this phase set. The substructure can be a constellation of atoms that scatter anomalously, such as metal or S atoms. Alternatively, a structural fragment such as an idealized α-helix or a motif from some distantly related protein can be orientated and sometimes positioned by an extensive molecular-replacement search, checking the correlation coefficient between observed and calculated structure factors for the highest normalized structure-factor amplitudes $|E|$. The top solutions are further ranked on the correlation coefficient for all E values. The phases generated from such fragments are improved using Patterson superposition maps and Sayre-equation refinement carried out with fast Fourier transforms. Phase refinement is completed using a novel density-modification process referred to as dynamic density modification (DDM). The method is illustrated by the solution of a number of known proteins. It has proved fast and very effective, able in these tests to solve proteins of up to 5000 atoms. The resulting electron-density maps show the major part of the structures at atomic resolution and can readily be interpreted by automated procedures.

1. Introduction

Solution of a crystal structure requires the assignment of a set of sufficiently accurate phases in order for an atomic model to be identified in the electron density and subsequently refined. For small structures, for which atomic resolution data are available, statistical procedures can develop such a phase set *ab initio* from multiple random starting sets. However, the statistical relationships between reflections become weaker as the number of atoms increases. Recent developments such as those encoded in *Shake and Bake* (Weeks & Miller, 1999; Xu *et al.*, 2000) and *SHELXD* (Sheldrick & Gould, 1995; Sheldrick, 1997) have extended direct-methods approaches to small proteins, up to around 1000 atoms to date, again provided atomic resolution data are available. These methods depend on refining phases and then picking possible atomic sites for model development. More commonly used approaches for obtaining the initial phase set in macromolecular crystallography involve the determination of experimental (MAD or MIR) phases or establishment of a molecular-replacement solution.

Here, we describe a rapid and novel approach for developing an initial phase set calculated from a small percentage (~5%) of the scattering matter of the unit cell, using a combination of Patterson superposition, Sayre's equation and density modification. As programmed at present, the method requires atomic resolution data. The technique is coded into a program suite we call *ACORN*,

tall oaks from little acorns grow (David Everett, 1769–1813).

The starting fragment can be of various types, as described in the next two sections. We have used (i) a set of experimentally determined anomalous scattering centres, (ii) a small idealized piece of secondary structure and (iii) fragments or indeed whole structures of homologous proteins. Types (ii) and (iii) are initially positioned using MR techniques either within *ACORN* or by other programs.

A flow chart of the *ACORN* program is shown in Fig. 1.

2. Experimentally positioned anomalous scatterers

The experimental approach to MAD phasing has depended on positioning metals or other anomalous scatterers such as Se and Br from the observed anomalous differences alone. Direct-methods procedures using single-wavelength anomalous differences at 3 Å have been shown to be sufficient to position many Se sites (*e.g.* Deacon *et al.*, 1999). Dauter *et al.* (1999, 2000) have recently shown that it is possible to measure sufficiently accurate data to determine the position of much lighter atoms such as S and Cl atoms from their small but significant anomalous signal. Dauter *et al.* (1999) found the positions of ten S and seven Cl atoms in lysozyme from the anomalous signal measured with 1.54 Å wavelength radiation using direct-methods programs such as *RANTAN* (Yao, 1981), *SHELX* (Sheldrick & Gould, 1995) or *Shake and Bake* (Weeks *et al.*, 1994). The incidence of S atoms in proteins is typically about 1%.

Such anomalous scatterers provide an excellent starting 'fragment' which can be oriented and positioned in the true space group and can be fed directly into a single phase-refinement run. This may prove to be highly important in extending the method to lower resolution.

3. Positioning a fragment using molecular replacement

Many protein structures have been solved by the method of molecular replacement, in which a molecule from a known structure is taken as the starting model for the unknown structure. An approximate orientation and position relative to the symmetry axes in the cell is found by matching the interatomic vectors for a range of orientations and positions with Patterson density, then scoring the potential solutions with the correlation coefficients for intensities or amplitudes. The *AMoRe* program (Navaza, 1994) is one way of doing this. The method becomes less successful as the percentage sequence identity between the model and target falls. Molecular-replacement searches have also been successful when the model fragment represents a fairly small percentage of the asymmetric unit of the protein under investigation (see review by Turkenburg & Dodson, 1996). Representative examples of this kind of work include Oh (1995), Chantalat *et al.* (1996), Bernstein & Hol (1997) and Kissinger *et al.* (1999). However, it has not been very successful with small secondary features such as idealized α-helices, where the search vectors are not distributed within a spherical volume.

Although proteins come in virtually infinite variety in terms of their amino-acid sequence, the local physical conformation of many small regions tend to be based on a limited number of motifs or secondary structural elements. Here, we investigate the extent to which this property can be exploited to derive a complete protein structure from such a motif: either a small fragment of a known structure with some sequence homology or an idealized secondary structural element such as an α-helix or small section of β-sheet. We have found that such a fragment can be orientated and sometimes positioned by exhaustively testing all possible orientations and checking the

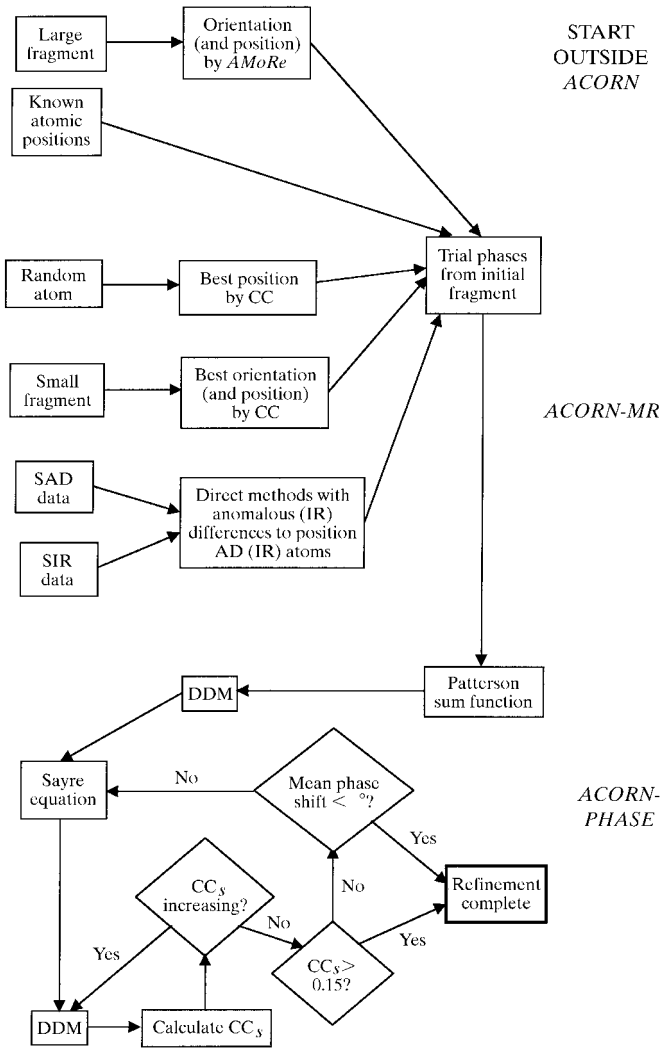

Figure 1
Flow diagram for *ACORN*. The positioning operations in parentheses are not carried out if the phase development is performed in space group *P*1.

correlation coefficient between observed and calculated structure factors for the highest normalized E values.

In space group $P1$, the molecular-replacement problem is simpler: there are only three rotational angle parameters to determine. For all other space groups, the positioning of the fragment requires up to three further parameters to place it correctly with respect to the possible origins of the unit cell. The program described here presently functions in $P1$, *i.e.* data in high-symmetry space groups are expanded to the required hemisphere and the appropriate origin shift applied after the structure is solved. The ease of working in $P1$ is well described (Sheldrick & Gould, 1995; Burla *et al.*, 2000; Xu *et al.*, 2000). This simplifies the algorithm, but a larger fragment of the asymmetric unit must be used as search model. This limitation does not of course apply when using an experimentally positioned constellation.

We have so far restricted our attention to structures for which atomic resolution data are available. The methods have proved highly successful. The steps in the procedure we have devised are as follows.

(i) Selection of a fragment containing 1–8% of the unit-cell contents.

(ii) Orientation, and positioning if necessary and possible, of the fragment to match some part of the target structure.

(iii) Patterson superposition with the oriented and positioned residue of the fragment.

(iv) Refinement using (*a*) the Sayre equation and (*b*) density modification.

First some details of these stages will be described and then the results of applying the procedure will be discussed.

3.1. Selection of a fragment

Although the method does not require knowledge of a related known structure, if such a structure was available then that structure or some part of it would be taken as the fragment. In such an event, the usual procedures of molecular replacement might also lead to a solution. Indeed, correct solutions have been obtained with models having 30% or less sequence homology to the new crystal form. For the present procedure, in general a fragment (including symmetry-related fragments if these positions have been found) with 1–8% of the total scattering matter in the *unit cell* has been found to be adequate to start the phasing cascade. Positioning such a small fragment in the true unit cell can prove difficult, but in some test cases it has been sufficient to orient, for example, a single α-helix in the new crystal form and treat the crystal as though it belonged to the $P1$ space group with an arbitrary origin. The true origin can be easily identified once all the atom positions are determined.

3.2. Orienting and positioning the fragment

There are two procedures that have been successfully used to orient and position reasonably sized fragments. The first of these uses programs such as *AMoRe* (Navaza, 1994), which finds many reasonable matches between the interatomic vector set of the fragment and the Patterson function for the target structure, then grades the solutions using a correlation coefficient between model and observed structure factors. The problem is factorized to find first the orientation and then to follow this with a translation to find the best position relative to the symmetry axes.

A second, slower, approach is to orient and position the fragment simultaneously to give the best correlation coefficient between the structure amplitudes for the fragment and those for the target structure. Kissinger *et al.* (1999) have developed an efficient approach that finds the best orientation and translation by an evolutionary search, using the conventional correlation coefficient (CC) defined as

$$\begin{aligned} \mathrm{CC} &= \frac{\langle |E_{\rm frag}E_o| \rangle - \langle |E_{\rm frag}| \rangle \langle |E_o| \rangle}{[(\langle |E_{\rm frag}^2| \rangle - \langle |E_{\rm frag}| \rangle^2)(\langle |E_o^2| \rangle - \langle |E_o| \rangle^2)]^{1/2}} \\ &= \frac{\langle |E_{\rm frag}E_o| \rangle - \langle |E_{\rm frag}| \rangle \langle |E_o| \rangle}{\sigma_{\rm frag}\sigma_o}, \end{aligned} \tag{1}$$

where $E_{\rm frag}$ is obtained by normalizing the structure factor for the fragment $F_{\rm frag}$, E_o is the observed normalized structure factor for the target structure and the σs are the appropriate standard deviations for the quantities giving the averages. For a discussion of normalization procedures, see Blessing *et al.* (1998).

We have developed a similar procedure, calculating an initial finely sampled set of structure factors by classical techniques, performing a complete search of all orientations and using a similar CC to assess the results. We have usually found it sufficient, at least in an initial search, to use only those structure factors for which $|E_o| > \sim 1.0$, although much better discrimination of the correct orientation is obtained if all structure factors are used in determining the CCs.

These CCs are used in three different ways in the program *ACORN*. The first ranking of likely fragment orientations is made using (i) only the correlation coefficient for the higher $|E|$ values, usually somewhat greater than unity. This is referred to as $\mathrm{CC}_{|E|>X}$, where X is the cutoff value. The ranking of the highest sets from this set is made using (ii) all data and referred to as $\mathrm{CC}_{\rm all}$. In the later stages of the procedure, the convergence of the phase refinement is monitored by (iii) the correlation coefficient for the weak $|E|$ values only, referred to as CC_s.

We orient fragments using spherical polar coordinates (ψ, φ, χ). With 3° steps in these coordinates, more than 250 000 orientations must be tested and this clearly is not practicable. Instead, the fragment is first positioned in a large 'cell' with edges three times the maximum dimension of the fragment and structure factors are calculated by classical techniques to the resolution limit of the target-structure data (Lattmann & Love, 1970; Navaza, 1994). The calculated structure factors vary sufficiently slowly between neighbouring reciprocal-lattice points for the set to approximate to the continuous Fourier transform of the fragment and to allow reasonably accurate values of each $|F_{\rm frag}|$ to be found by interpolation of this 'continuous' transform. It is also possible to find the structure factors for symmetry-related fragments by adding their contributions with appropriate phase shifts.

With a modest workstation capacity it typically takes 0.05 s to interpolate 20 000 structure factors for a particular orientation and position of the fragment from the tabulated Fourier transform of the large cell. Like most people, we have found it convenient to factorize the problem into first finding the orientation of a fragment and then its position in relation to symmetry elements. In the orientation search with a single fragment, the observed data are expanded to a complete hemisphere of reciprocal space and the CC calculated for this data set, *i.e.* we have worked in *P*1 and hence avoided the translation problem.

The first pass calculates the CCs for different orientations using only reflections for which $|E_o| > E_{\rm lim}$, where $E_{\rm lim}$ is close to unity. If there are 250 000 orientations and 20 000 structure factors, then this takes under 4 h of computer time. Using this filter, the best orientations are usually in the top 100–200 of the $CC_{|E|>\rm limit}$, although not necessarily at the very top of that list. To give a margin of safety, we often take up to the top 1000 phase sets with the highest values of $CC_{|E|>\rm limit}$. Subsequently, values of $CC_{\rm all}$ are calculated for these orientations using all the observed structure factors. For 1000 orientations and, say, 100 000 observed reflections, this final stage takes less than 5 min. This two-stage procedure normally brings the best orientation very close to, if not to, the top of the $CC_{\rm all}$ list.

4. Phase refinement

4.1. Patterson superposition

Before embarking on the main refinement process, which involves density modification, we have often found it advantageous to produce a semi-sharpened Patterson sum-function map based on the fragments. The Fourier coefficients for this map are $|E_o||F_o|E_{\rm frag}$: hence, the phases are simply those for the fragment structure factors. This map becomes the starting point for density modification, described below. It is of course dominated by the peaks corresponding to the fragment, but has additional low-level density that after density modification gives a small but valuable phase improvement, often of about 2°. This is quite significant in relation to the phase improvements given by the following cycles of density modification.

Other forms of superposition function may prove to be better, although some initial trials with a minimum function have taken more CPU time without giving a better result than the sum function.

4.2. Sayre-equation refinement

We have found that one or two cycles of Sayre-equation refinement before embarking on density modification can give initial phase improvement. This involves deriving new phase estimates from previous ones by the use of the Sayre equation

$$E_{\mathbf{h}} = \frac{f_{\mathbf{h}}}{g_{\mathbf{h}} V} \sum_{\mathbf{k}} E_{\mathbf{k}} E_{\mathbf{h}-\mathbf{k}}, \tag{2}$$

where $f_{\mathbf{h}}$ is the atomic scattering factor corresponding to normalized structure factors, $g_{\mathbf{h}}$ is the scattering factor for a 'squared' atom and V is the volume of the unit cell. The Sayre equation strictly applies only to equal-atom structures, for which it can be derived from the integration of the volume of overlap of two spheres that

$$f_{\mathbf{h}} = N^{1/2} \quad \text{and} \quad g_{\mathbf{h}} = \frac{4\pi s_m^3}{3N}\left(1 - \frac{3s_{\mathbf{h}}}{4s_m} + \frac{s_{\mathbf{h}}^3}{16 s_m^3}\right), \tag{3}$$

where N is the number of atoms in the unit cell, $s_{\mathbf{h}}$ is the value of $(2\sin\theta)/\lambda$ for the reflection of index $\mathbf{h}$ and s_m is the maximum value of s for the observed data, assumed complete. Although the protein fragments usually consist of O, N and C atoms, we normally apply the equation as though the atoms are equal. Simplifying the equation to

$$E_{\mathbf{h}} = C_{\mathbf{h}} \sum_{\mathbf{k}} E_{\mathbf{k}} E_{\mathbf{h}-\mathbf{k}}, \tag{4}$$

we seek to minimize the function

$$S = \sum_{\mathbf{h}} |E_{\mathbf{h}} - C_{\mathbf{h}} \sum_{\mathbf{k}} E_{\mathbf{k}} E_{\mathbf{h}-\mathbf{k}}|^2, \tag{5}$$

or

$$S = \sum_{\mathbf{h}} \left[|E_{\mathbf{h}}| \cos\varphi_{\mathbf{h}} - C_{\mathbf{h}} \sum_{\mathbf{k}} |E_{\mathbf{k}} E_{\mathbf{h}-\mathbf{k}}| \cos(\varphi_{\mathbf{k}} + \varphi_{\mathbf{h}-\mathbf{k}}) \right]^2 + \sum_{\mathbf{h}} \left[|E_{\mathbf{h}}| \sin\varphi_{\mathbf{h}} - C_{\mathbf{h}} \sum_{\mathbf{k}} |E_{\mathbf{k}} E_{\mathbf{h}-\mathbf{k}}| \sin(\varphi_{\mathbf{k}} + \varphi_{\mathbf{h}-\mathbf{k}}) \right]^2. \tag{6}$$

For a minimum, $(\partial S/\partial\varphi_{\mathbf{l}}) = 0$ for all $\mathbf{l}$. We now derive the terms in $(\partial S/\partial\varphi_{\mathbf{l}})$ from (6). First, we consider the terms arising from having $\mathbf{l} = \mathbf{h}$. These are

$$-|E_{\mathbf{l}}| \sin\varphi_{\mathbf{l}} \left[|E_{\mathbf{l}}| \cos\varphi_{\mathbf{l}} - C_{\mathbf{l}} \sum_{\mathbf{k}} |E_{\mathbf{k}} E_{\mathbf{l}-\mathbf{k}}| \cos(\varphi_{\mathbf{k}} + \varphi_{\mathbf{l}-\mathbf{k}}) \right] + |E_{\mathbf{l}}| \cos\varphi_{\mathbf{l}} \left[|E_{\mathbf{l}}| \sin\varphi_{\mathbf{l}} - C_{\mathbf{l}} \sum_{\mathbf{k}} |E_{\mathbf{k}} E_{\mathbf{l}-\mathbf{k}}| \sin(\varphi_{\mathbf{k}} + \varphi_{\mathbf{l}-\mathbf{k}}) \right]. \tag{7}$$

The first terms in the brackets disappear, so the remaining terms are

$$C_{\mathbf{l}} |E_{\mathbf{l}}| \sin\varphi_{\mathbf{l}} \sum_{\mathbf{k}} |E_{\mathbf{k}} E_{\mathbf{l}-\mathbf{k}}| \cos(\varphi_{\mathbf{k}} + \varphi_{\mathbf{l}-\mathbf{k}}) - C_{\mathbf{l}} |E_{\mathbf{l}}| \cos\varphi_{\mathbf{l}} \sum_{\mathbf{k}} |E_{\mathbf{k}} E_{\mathbf{l}-\mathbf{k}}| \sin(\varphi_{\mathbf{k}} + \varphi_{\mathbf{l}-\mathbf{k}}). \tag{8}$$

We now derive the terms in $(\partial S/\partial\varphi_{\mathbf{l}})$ from (6) when $\mathbf{l} = \mathbf{k}$ or $\mathbf{l} = \mathbf{h} - \mathbf{k}$. These are

$$2 \sum_{\mathbf{h}} \left[|E_{\mathbf{h}}| \cos\varphi_{\mathbf{h}} - C_{\mathbf{h}} \sum_{\mathbf{k}} |E_{\mathbf{k}} E_{\mathbf{h}-\mathbf{k}}| \cos(\varphi_{\mathbf{k}} + \varphi_{\mathbf{h}-\mathbf{k}}) \right] \times C_{\mathbf{h}} |E_{\mathbf{l}} E_{\mathbf{h}-\mathbf{l}}| \sin(\varphi_{\mathbf{l}} + \varphi_{\mathbf{h}-\mathbf{l}}) - 2 \sum_{\mathbf{h}} \left[|E_{\mathbf{h}}| \sin\varphi_{\mathbf{h}} - C_{\mathbf{h}} \sum_{\mathbf{k}} |E_{\mathbf{k}} E_{\mathbf{h}-\mathbf{k}}| \sin(\varphi_{\mathbf{k}} + \varphi_{\mathbf{h}-\mathbf{k}}) \right] \times C_{\mathbf{h}} |E_{\mathbf{l}} E_{\mathbf{h}-\mathbf{l}}| \cos(\varphi_{\mathbf{l}} + \varphi_{\mathbf{h}-\mathbf{l}}). \tag{9}$$

We now find the coefficient of $\sin\varphi_{\mathbf{l}}$ in (9). This is, removing a factor of $|E_{\mathbf{l}}|$,

$$2\sum_{\mathbf{h}}|E_{\mathbf{h}}E_{\mathbf{h-l}}|C_{\mathbf{h}}\cos(\varphi_{\mathbf{l-h}}+\varphi_{\mathbf{h}}) - 2\sum_{\mathbf{h}}C_{\mathbf{h}}^2|E_{\mathbf{l-h}}|\Big[\cos\varphi_{\mathbf{l-h}}\sum_{\mathbf{k}}|E_{\mathbf{k}}E_{\mathbf{h-k}}|\cos(\varphi_{\mathbf{k}}+\varphi_{\mathbf{h-k}}) - \sin\varphi_{\mathbf{l-h}}\sum_{\mathbf{k}}|E_{\mathbf{k}}E_{\mathbf{h-k}}|\sin(\varphi_{\mathbf{k}}+\varphi_{\mathbf{h-k}})\Big]. \quad (10)$$

We now write

$$G_{\mathbf{h}} = (1/V)\sum_{\mathbf{k}}E_{\mathbf{k}}E_{\mathbf{h-k}} = |G_{\mathbf{h}}|\exp(i\psi_{\mathbf{h}}) \quad (11)$$

and

$$Q_{\mathbf{h}} = (1/V)\sum_{\mathbf{k}}C_{\mathbf{k}}E_{\mathbf{k}}E_{\mathbf{h-k}}. \quad (12)$$

Then, including the term in (8) and again excluding the factor $|E_{\mathbf{l}}|$, the coefficient of $\sin\varphi_{\mathbf{l}}$ is

$$A_{\mathbf{l}} = C_{\mathbf{l}}VR_e(G_{\mathbf{l}}) + 2VR_e(Q_{\mathbf{l}}) - 2V\sum_{\mathbf{h}}C_{\mathbf{h}}^2|E_{\mathbf{l-h}}G_{\mathbf{h}}|\cos(\varphi_{\mathbf{l-h}}+\psi_{\mathbf{h}}), \quad (13)$$

where R_e designates 'real part'.

The summation in the last term is V times the real part of the inverse Fourier transform of the product of ρ and the function η, which is the Fourier transform of C^2G. We designate this term as $V^2R_e(X_{\mathbf{l}})$.

A similar analysis may be performed to find the coefficient of $\cos\varphi_{\mathbf{l}}$; this is

$$B_{\mathbf{l}} = C_{\mathbf{l}}VI_m(G_{\mathbf{l}}) + 2VI_m(Q_{\mathbf{l}}) - 2V^2I_m(X_{\mathbf{l}}), \quad (14)$$

where I_m designates 'imaginary part'.

Taking $(\partial S/\partial\varphi_{\mathbf{l}}) = 0$, the Sayre-equation tangent formula becomes

$$\tan\varphi_{\mathbf{l}} = \frac{B_{\mathbf{l}}}{A_{\mathbf{l}}} = \frac{C_{\mathbf{l}}VI_m(G_{\mathbf{l}}) + 2VI_m(Q_{\mathbf{l}}) - 2V^2I_m(X_{\mathbf{l}})}{C_{\mathbf{l}}VR_e(G_{\mathbf{l}}) + 2VR_e(Q_{\mathbf{l}}) - 2V^2R_e(X_{\mathbf{l}})}. \quad (15)$$

In practice, with real data and when phases are well away from their true values, the formula given in (13) can be unstable. This is because the final terms in the numerator and denominator tend to be counterbalancing but large. We have found empirically that an alternative version of the equation,

$$\tan\varphi_{\mathbf{l}} = \frac{C_{\mathbf{l}}VI_m(G_{\mathbf{l}}) + \alpha VI_m(Q_{\mathbf{l}}) - \beta V^2I_m(X_{\mathbf{l}})}{C_{\mathbf{l}}VR_e(G_{\mathbf{l}}) + \alpha VR_e(Q_{\mathbf{l}}) - \beta V^2R_e(X_{\mathbf{l}})}, \quad (16)$$

gives good results without being unstable, where the values of α and β are chosen such that

$$\alpha\sum_{\mathbf{h}}|Q_{\mathbf{h}}| = \tfrac{1}{5}\sum_{\mathbf{h}}|C_{\mathbf{h}}G_{\mathbf{h}}|, \qquad \beta V\sum_{\mathbf{h}}|X_{\mathbf{h}}| = \tfrac{2}{3}\sum_{\mathbf{h}}|C_{\mathbf{h}}G_{\mathbf{h}}|. \quad (17)$$

The whole basis of *ACORN* depends on the ability to calculate Fourier transforms efficiently and (16) lends itself to this general approach. For a single cycle of the Sayre equation, (i) calculate ρ as FT of E, (ii) calculate G as FT^{-1} of ρ^2, (iii) calculate σ as FT of CE, (iv) calculate Q as FT^{-1} of $\sigma\rho$, (v) calculate η as FT of C^2G and (vi) calculate X as FT^{-1} of $\eta\rho$. The real and imaginary parts of G, Q and X are then used in (16).

Since the application of the Sayre equation in this way does not involve the evaluation of the individual terms of the convolution summation (2), the complete data set can be used, which is important in protein applications.

4.3. Refinement by density modification

DDM (dynamic density modification) is a development from LDE (low-density elimination; Shiono & Woolfson, 1991, 1992).

For DDM, one first finds ρ_{sd}, the standard deviation of the map density.

$$\rho_{\mathrm{sd}} = (\langle\rho^2\rangle - \langle\rho\rangle^2)^{1/2} = \big\langle(\rho - \langle\rho\rangle)^2\big\rangle^{1/2}. \quad (18)$$

The density-modification process is

$$\rho' = 0, \quad \rho < 0 \qquad \rho' = \rho\tanh\left[0.2\left(\frac{\rho}{\rho_{\mathrm{sd}}}\right)^{3/2}\right], \quad \rho \geq 0. \quad (19)$$

Following this, high values of ρ' are truncated to $kn_c\rho_{\mathrm{sd}}$, where n_c is the number of the DDM refinement cycle for up to five cycles and $n_c = 5$ thereafter. The value of k is normally three but may be set to some other value.

4.4. The phase-refinement procedure

When the Patterson superposition map is calculated it is first subjected to DDM and then an inverse Fourier transform on the modified map gives new phases. These become the starting point for two cycles of Sayre-equation refinement whereby (16) is used to generate modified phases. Thereafter, all refinement uses cycles of density modification and DDM with Sayre's equation refinement carried out at the points where the CC_s stop increasing steadily.

Normally refinement is only carried out for reflections for which $|E_o| > 1.0$ or some other limit that gives a suitable number of reflections. Although the maps are calculated using only these larger normalised structure factors, once the maps have been modified by DDM their transform generates Fourier coefficients for all reflections, including those with smaller values of $|E|$. These latter coefficients are used to calculate CC_s (1). These reflections have made no contribution to the phasing and hence can be used as a cross-validation set. This turns out to be a powerful and to date infallible guide to the progress of the refinement. Typically, CC_s starts with a value of order 0.04 and reaches around 0.4 when the refinement is complete and a meaningful solution found.

When maps are calculated, the Fourier coefficients are weighted according to the Fourier coefficient of the previous modified map (PMM). Thus, the coefficient used is $W|E_o|\exp(i\varphi_c)$, where φ_c is the phase from the PMM,

$$W = \tanh\left[\frac{|E_o||F_c'|}{2(\Sigma')^{1/2}}\right]. \quad (20)$$

$|F_c'|$ is the magnitude of the Fourier coefficient of the PMM scaled to the values of $|E_o|$ in shells of reciprocal space and Σ'

Table 1
Application of *ACORN*.

(*a*) Application to known protein structures.

Protein	PDB code	Space group	Unit-cell dimensions (Å, °)	Resolution (Å)	N_{obs}	Completeness (%)	Protein MW (kDa) (No. residues)
Cytochrome c_6	1ctj	$R3$	$a = 52.11, b = 52.11, c = 81.02$	1.1	32653	98.5	9924 (89 + haem)
Cytochrome c_{553}	1b7c	$P2_12_12_1$	$a = 37.14, b = 39.42, c = 44.02$	0.97	38889	99.9	7578 (92 + haem)
Catalase		$P4_22_12$	$a = 105.79, b = 105.79, c = 105.00$	0.95	456662	99.1	56994 (502 + haem)
Lysozyme	3lzt	$P1$	$a = 26.65, b = 30.80, c = 33.63, \alpha = 88.3, \beta = 107.4, \gamma = 112.2$	0.92	46809	92.2	14795 (129)
Lysozyme	1bwi	$P4_32_12$	$a = 78.84, b = 78.84, c = 36.89$	1.00	55996	90.0	14795 (129)
RNAse AP1		$P2_1$	$a = 32.01, b = 49.73, c = 30.67, \beta = 115.83$	1.17	23853	81.5	11387 (106)
		$P2_1$	$a = 32.10, b = 52.00, c = 30.80, \beta = 115.90$	1.08	33223	88.5	11387 (106)
Penicillopepsin	1bxo	$C2$	$a = 96.98, b = 46.65, c = 65.71, \beta = 115.57$	0.90	156181	79.9	34575 (323)
Cellulase	8a3h	$P2_12_12_1$	$a = 54.45, b = 69.88, c = 77.32$	1.00	153756	96.5	34510 (303)

(*b*) Computations with *ACORN*.

%age indicates the percentage of the scattering matter of the fragment; CC st and end are the correlation coefficients of the observed and calculated E values for the weak reflections at the beginning and end of phase refinement, respectively; PE st and end are the corresponding phase errors.

Protein	Resolution (Å)	Protein MW (kDa) (No. residues)	Fragment MW (kDa) (No. residues)	%age	Solution	CPU	CC st/end	PE st/end
Cytochrome c_6	1.1	9924 (89 + haem)	26 (1 Fe)	2	Y	5 min	0.03/0.23	56/27
Cytochrome c_{553}	0.97	7578 (92 + haem)	26 (1 Fe)	3	Y	59 s	0.05/0.40	56/15
			16 (1 S)	1	Y	106 s	0.01/0.40	72/15
			1 random atom, 50000 trials	1	Y	50 min	0.02/0.40	63/15
Catalase	0.95	56994 (502 + haem)	144 (9 S)	2	Y	8.6 h	0.01/0.49	74/13
Lysozyme $P1$	0.92	14795 (129)	48 (3 S)	1	Y	53 s	0.02/0.43	71/14
			330 (10 res. ideal α-helix)	5	Y	15.3 h	0.01/0.43	76/14
Lysozyme $P4_32_12$	1.0	14795 (129)	16 (1 S)	5	Y	38 min	0.01/0.40	78/18
RNAse AP1	1.17	11387 (106)	80 (5 S)	3.5	Y	235 s	0.06/0.18	71/33
			561 (17 res. ideal α-helix in $P1$)	5	N			
			561 (17 res. related α-helix in $P1$)	5	Y	9.7 h	0.07/0.20	71/33
	1.08	11387 (106)	80 (5 S)	3.5	Y	150 s	0.01/0.34	73/19
			561 (17 ideal α-helix in $P1$)	5	Y	23 h	0.02/0.34	77/19
Penicillopepsin	0.9	34575 (323)	32 (2 S)	0.5	N			
			5600 (57 res.) homologous in $P1$	8	Y	10.8 h	0.02/0.49	77/22
			Same, preceded by *AMoRe*	8	Y	37.7 min	0.01/0.49	76/15
			Whole homologous protein positioned by *AMoRe*	90	Y	28.6 min	0.03/0.47	70/17

is the mean square of the values of $|F'_c|$. In a successful refinement the values of W tend steadily to increase.

5. Examples

The application of the method to a set of known protein structures listed in Table 1(*a*) with data to 1.2 Å or better is now described. The program is divided into two main parts, one used to position the starting fragment (referred to as *ACORN-MR*) and the other to complete the phase refinement (referred to as *ACORN-PHASE*). The first three examples are metalloproteins. We have assumed that the positions of any metals and S atoms in the structure can be defined from the anomalous scattering terms and these fragments are investigated as single starting points for DDM. This has not been confirmed for all the structures, as in some cases the anomalous measurements had not been retained during processing, but recent results indicate this should be generally possible (Dauter *et al.*, 1999). Metalloproteins are always easier to solve using direct methods as the early cycles locate the heavy atoms that then serve as a powerful core from which the rest of the structure is developed.

The other examples do not contain a metal centre and attempts to solve the subsequent structures either from the S atoms alone or from unoriented fragments, particularly a α-helix, are described for most of them. In addition, rapid phase refinement of a complete MR model was tested for one of the larger proteins. All computations were performed on a

Silicon Graphics O2 workstation and the CPU times reported for the applications of *ACORN* in Table 1(*b*) relate to this. The importance of the data quality is shown by the RNase AP1 example and this point is discussed further below.

All computations using a starting model found from the anomalous differences were carried out using the real space-group symmetry. Those using randomly oriented fragments were performed in space group *P*1.

The CC_s, the correlation coefficient for the small $|E|$ values, acts as a sensitive indicator of the progress of the refinement. Fig. 2 shows the successive values of CC_s and of the mean phase error as the refinement progressed for penicillopepsin (discussed below). Of course, when solving an unknown structure one would not be able to follow the progress of the refinement by the value of the mean phase error. However, this result is quite representative and there is no difficulty in knowing when to cease refinement or indeed in knowing whether or not the refinement has been successful.

5.1. Metalloproteins

5.1.1. Cytochromes. The first test was with cytochrome c_6, previously solved with 1.1 Å data using *SHELX* and *E*-map peak enhancement (Frazão *et al.*, 1995). Input of the known Fe-atom position established from a sharpened difference Patterson synthesis generated a phase set with a mean phase error of 27° within 5 min CPU time and a map which resembled that computed from the refined model.

Similar computations with the recently determined structure of cytochrome c_{553} (Benini *et al.*, 2000) starting from the Fe atom or even from a single S atom gave even better results within 1–2 min, which is much faster than available through model construction with approaches such as *ARP*/*REFMAC*. Indeed, 50 000 trials starting from a single atom in a random position gave a solution of comparable quality which was clearly indicated by the correlation coefficient; this required 50 min CPU time.

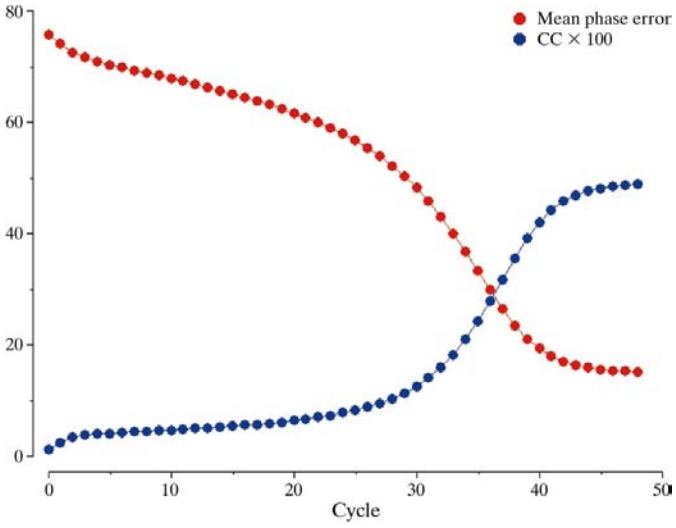

Figure 2
The changes in the correlation coefficient CC_s and in the mean phase error for penicillopepsin (PDB code 1bxo). The solution is evident from the asymptotic behaviour of the correlation coefficient alone.

5.1.2. Catalase. *Micrococcus lysodeikticus* catalase is a large protein in a high-symmetry space group whose structure has been known for some years (Murshudov *et al.*, 1992) and for which excellent diffraction data have more recently been recorded (Murshudov *et al.*, unpublished data). The iron and sulfur positions are easily found from direct methods or Patterson searches using the anomalous differences. These positions were used as a starting fragment for *ACORN-PHASE* (Table 1*b*); the electron density for the haem group computed with the resulting phase set is shown in Fig. 3. This indicates that the method is not limited to small proteins provided the experimental data are sufficient.

5.2. Proteins without a metal centre

5.2.1. Lysozymes. The method was applied to both the triclinic (Walsh *et al.*, 1998) and the tetragonal forms (Evans *et al.*, 2000) of hen egg-white lysozyme. For the *P*1 structure there were 58 376 measured reflections to 0.92 Å resolution, but we restricted the data to 1.0 Å, giving 46 809 reflections. Phase refinement was first attempted starting from the sulfur positions, which can easily be established for this protein from the anomalous measurements (Dauter *et al.*, 1999). An oriented fragment of the ten S atoms was sufficient to lead to a phase set with a CC of about 0.43 and a mean phase error of 14° and required only 53 s CPU time. Indeed, the method proved sufficiently powerful to develop an equally good phase set starting from as few as three of the S atoms (Table 1*b*). Similar results were found for tetragonal lysozyme.

The next computations centred on the positioning of a fragment as starting model. A ten-residue α-helix with C^β atoms (50 atoms in all) corresponding to 3.4% of the cell

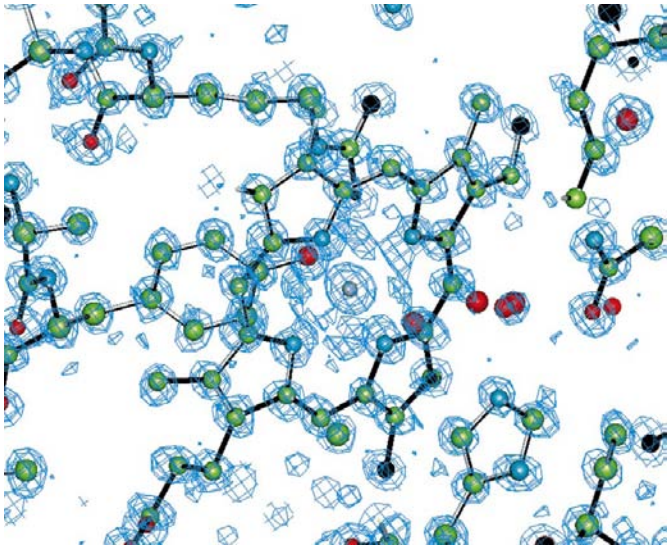

Figure 3
The electron density for the *E* map for the haem group in *M. lysodeikticus* catalase at 1.0 Å resolution, with phases computed from *ACORN*. The phase error from those computed from the refined model is 13°. The coordinates of the refined model are superimposed. The positions of all haem atoms are extremely clearly defined.

Table 2
Correlation coefficients.

(*a*) The ranking order and values of $CC_{|E|>11}$ for the ten highest values of CC_{all} for the *P*1 lysozyme structure. Eight of the ten starting sets led to a correct solution.

Rank on CC_{all}	CC_{all}	Rank on $CC_{|E|>1.1}$	$CC_{|E|>1.1}$	CC_s	Phase error (°)
1	0.04078	779	0.02897	0.42441	14.8
2	0.04007	343	0.03292	0.02401	85.5
3	0.03802	207	0.03481	0.01257	85.2
4	0.03744	787	0.02967	0.42795	14.6
5	0.03723	99	0.03699	0.42047	14.5
6	0.03715	8	0.04308	0.41956	14.6
7	0.03708	167	0.03396	0.42088	14.8
8	0.03620	9	0.04205	0.41718	14.9
9	0.03619	15	0.04068	0.42451	14.6
10	0.03607	24	0.03884	0.41674	14.9

(*b*) The ranking order and values of $CC_{|E|>1.1}$ for the ten highest values of CC_{all} for the 1.17 Å RNase AP1 structure. Only two starting sets led to a solution.

Rank on CC_{all}	CC_{all}	Rank on $CC_{|E|>1.1}$	$CC_{|E|>1.1}$	CC_s	Phase error (°)
1	0.05278	41	0.09109	0.19725	33.2
2	0.05197	39	0.09119	0.19767	33.2
3	0.05129	9	0.09572	0.07873	67.1
4	0.05040	149	0.08624	0.07492	77.6
5	0.04970	253	0.08416	0.07766	73.5
6	0.04725	122	0.08798	0.08848	72.2
7	0.04562	685	0.07939	0.08406	81.5
8	0.04334	365	0.08255	0.07693	85.2
9	0.04260	84	0.08855	0.08407	85.5
10	0.04249	232	0.08468	0.07343	84.4

contents was chosen. Its best orientation gave a CC_{all} of 0.041. Following two cycles of Sayre-equation refinement, *ACORN-PHASE* was run with 16 761 reflections with $|E| > 1.0$. After 46 cycles, the value of CC_s had risen from 0.0079 to 0.4244. The initial mean phase error for 16 761 reflections with $|E| > 1$ was 77.2° and at the end of the refinement it was 14.8° without weights and 13.3° with the weight *W* given in (20). The quality of the resulting electron-density map is evident from the electron density shown in Fig. 4. In fact, fragment orientations corresponding to eight of the top ten best values of CC_{all} led to the correct phase set, as shown in Table 2(*a*).

Triclinic lysozyme turned out to be a particularly easy structure to solve. In multi-trial experiments, solutions were found from unoriented random fragments and even from a two-atom fragment. It might be said that it resisted all our efforts not to solve it!

5.2.2. RNase AP1. RNase AP1 has been solved in space group $P2_1$ (Bezborodova *et al.*, 1988) and the asymmetric unit contains one molecule with 96 amino acids (808 non-H atoms including five S atoms) together with 83 ordered water molecules. Two data sets were available, one nominally to a resolution of 1.17 Å and the other to 1.08 Å. The $|E|$ distribution showed that there were some anomalies in the 1.17 Å set (see Fig. 5).

Solution of the structure starting from a fragment composed of the five S atoms was again straightforward for either data set, although it proceeded more rapidly and to a lower final phase error for the 1.08 Å set. A single run of *ACORN-PHASE* (with CPU time 150 s) produced an excellent phase set, with a CC of 0.34 and a phase error of 19°. Not surprisingly, the map was of outstanding quality.

To solve this structure in space group *P*1, using an unoriented fragment as starting model, proved more challenging than for lysozyme. RNase AP1 is known to contain a 17-residue α-helix which is slightly distorted, making up 4.8% of the cell contents. The search was initiated using two alternative fragments both consisting of a 17-residue α-helix main chain plus the attached C^β atoms (85 atoms in all). In the first model an idealized α-helix was used, while in the second it was taken from a homologous RNase in the PDB (code 9rnt; Martinez-Oyanedel *et al.*, 1991) with an 80% identical sequence of amino acids.

Details are given first for the search with the distorted helix from the PDB with the 1.17 Å resolution data set. Initially, structure factors were calculated for 12 944 observed reflections with $|E| > 1.1$ in a hemisphere of reciprocal space and for 276 200 different orientations. For orientations corresponding to the highest 1000 values of $CC_{|E|>1.1}$, structure factors for all 46 842 reflections in the hemisphere were calculated to give values for CC_{all}. Fragment orientations corresponding to the top two values of $CC_{|E|>1.1}$ led to solutions of the structure; the top 10 values of CC_{all} are given in Table 2(*b*) together with the ranking order derived from $CC_{|E|>1.1}$. For the top solution, the initial mean phase error for the 12 944 large $|E|$s was 71.1°. After 152 cycles of *ACORN-PHASE* refinement, the CC_s had increased from 0.069 to 0.197 with a final unweighted mean phase error of 33°. The quality of the resultant map is excellent.

Using an idealized 17-residue helix, we were unable to solve this structure using the 1.17 Å data. However, with the better

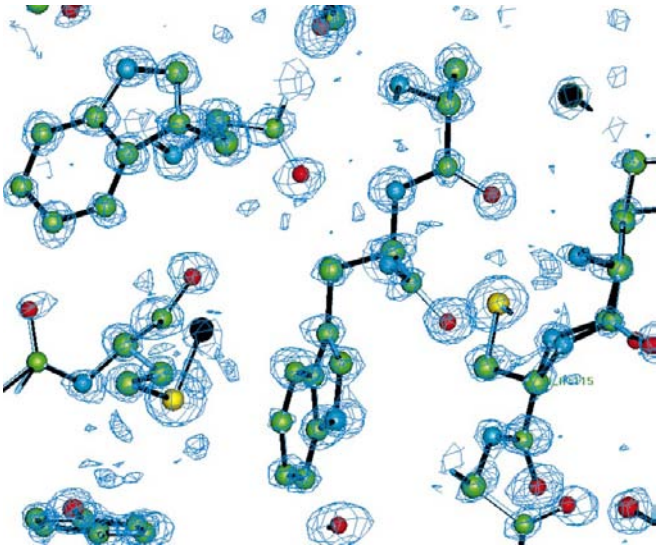

Figure 4
The electron density for the *E* map for triclinic lysozyme at 1.0 Å, with a mean phase error of 14°, again with the refined model. The positions of the two tryptophans are especially evident. The relative peak heights of the C, O and S atoms is as expected.

quality 1.08 Å data both the idealized helix (requiring no knowledge of the structure) and the distorted helix produced clear solutions with a substantially lower phase error than for the 1.17 Å data, 19° as opposed to 33° (Table 1*b*).

5.2.3. Penicillopepsin. Penicillopepsin (PDB code 1bxo; Ding *et al.*, 1998) crystallizes in space group *C*2 with 2977 non-H atoms (including 83 heterogen and 528 solvent atoms) in the asymmetric unit, *i.e.* 5954 in the unit cell. This structure was chosen with three tests in mind. Firstly, we tried unsuccessfully to develop the complete model from the S atoms. Secondly, we tested whether a whole structure can be developed from a small fragment positioned by the molecular-replacement procedure. Thirdly, we evaluated the use of *ACORN-PHASE* for rapid phase refinement from a fully oriented and positioned MR model.

As penicillopepsin contains only two S atoms in 323 residues and one of these is disordered, it is perhaps not surprising that *ACORN-PHASE* failed to expand a fragment made up of these two atoms to a satisfactory phase set.

For the first fragment search, 57 residues (398 atoms) were taken from a related structure with 63% sequence identity (PDB code 1er8; Pearl & Blundell, 1984). The structure was positioned by *ACORN-MR* in space group *P*1 so that the fragment used amounted to 6.7% of the cell contents. After refinement by 68 cycles of *ACORN-PHASE* the value of CC_s for smaller structure factors had risen from 0.010 to 0.491. The initial and final mean phase errors for 50 000 reflections with $|E| > 1.28$ were 76.9° and 21.8°, respectively. Fig. 2 shows the successive values of CC_s and of the mean phase error. Fig. 6 shows a representative region of the resulting *E* map. The CPU time required was 46.6 h, most of which was used for the *ACORN-MR* search. We subsequently ran *ACORN-MR* using only the 2.0 Å data and produced a comparable solution in only 10.8 h CPU time.

The second test used the complete structure of the same homologous protein with 41% sequence identity as an initial model, which was orientated and positioned in the $P2_1$ cell using *AMoRe*. *ACORN-PHASE* then rapidly increased the correlation coefficient and reduced the phase error without any further reference to the model, thus avoiding the problem of phase bias which can often hamper refinement from an MR model. This required only 28 min CPU time in total. The test was repeated using only the 57-residue fragment, again positioned in the $P2_1$ cell using *AMoRe*. Once again *ACORN-PHASE* generated an excellent phase set taking only a few minutes longer.

5.2.4. Cellulase. Atomic resolution data nominally to 1.0 Å resolution are available for an inhibitor complex of the *Bacillus agaradhaerens* cellulase Cel5A solved in this laboratory (PDB code 8a3h; Varrot *et al.*, 1999). We expected to be able to solve it from the cluster of seven S atoms (two disordered) using *ACORN-PHASE*, but this proved not to be possible. The reason is probably associated with the data rather than the method. Fig. 6 shows a representation of the diffraction data for the *hk*0 zone. It is clear that the data are highly anisotropic, extending to a much lower resolution limit along *a** (around 1.2 Å) compared with *b**. This in turn will distort the current normalization procedure, which assumes an isotropic fall-off with resolution. In addition, the data were slightly incomplete, with some regions of reciprocal space completely missing (Fig. 7). To test the latter effect, we endeavoured to solve the structure from the sulfur constellation using calculated amplitudes. When the complete data to 1.0 Å were used, this led directly to a structure solution. In contrast, when those reflections absent from the measured data were excluded, no solution was obtained.

6. The effect of data quality

ACORN uses normalized structure factors, generated at present using *K*-curves, where the same correction factor is applied to all reflections within the same resolution range, *i.e.*

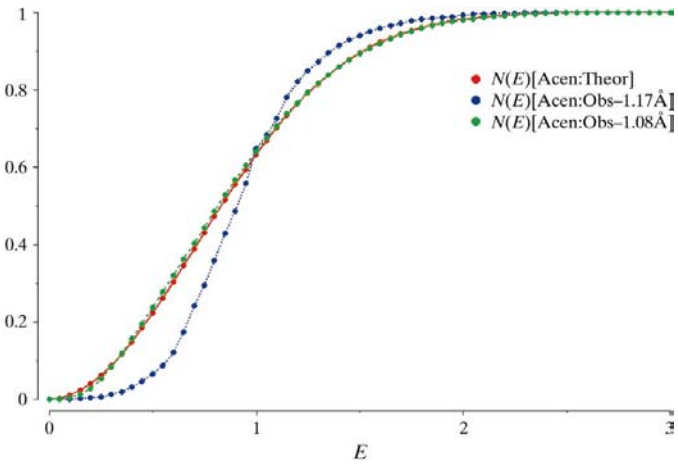

Figure 5
The *E* distribution for the two RNase AP1 data sets is shown. The 1.08 Å resolution data closely follow the expected distribution, whilst the lower quality 1.17 Å data deviate in a manner consistent with overestimation of the weak reflection amplitudes.

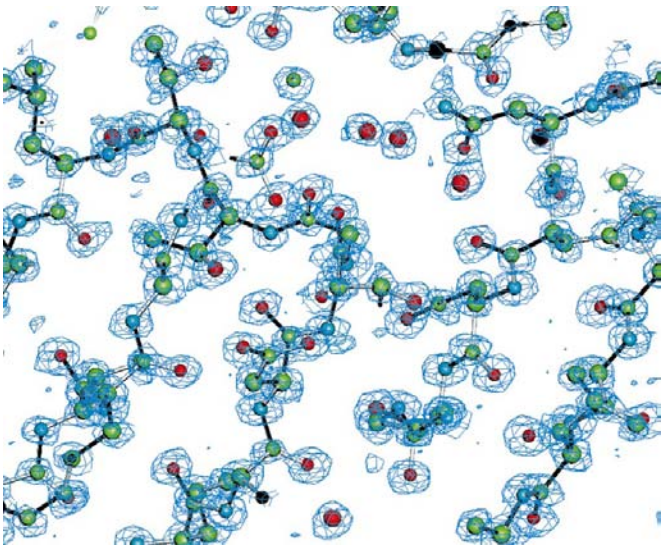

Figure 6
The electron density for the *E* map for penicillopepsin (PDB code 1bxo). The correlation of the water structure from the refined model with the density is particularly striking.

the correction is isotropic. However, many (indeed most) protein crystals diffract anisotropically so that the effective resolution is different along different axes. In general, if the normalization statistics are unusual, the solution may prove more difficult, *e.g.* for RNase AP1, where better data made the structure solution more straightforward. In addition, if the data have missing patches the structure solution may be hampered, as in the cellulase example. The use of more sophisticated data-processing and normalization procedures such as those developed by Blessing *et al.* (1998, 1999) would probably improve the performance of the algorithm. Further studies are needed to establish more robust normalization procedures and to investigate the effects of missing data. We should also reaffirm that with modern data-collection resources there is no reason not to record complete data.

7. Conclusions

This work and related work by others indicates that when atomic (better than 1.2 Å) resolution data are available then proteins with up to 500 or so amino acids are amenable to *ab initio* solution. The method is applicable even when the protein has no heavy atoms, which we take as atoms with atomic number higher than that of sulfur. In a sense, we are using the fragment as a kind of heavy atom, although it is one that we must orient.

We make the following conclusions from our study.

(i) Splitting the *ab initio* problem into two parts is logical. The density modification, *ACORN-PHASE*, can start from a single fragment established experimentally from anomalous scattering or, at the other extreme, from a fragment whose orientation has been established either by the *ACORN-MR* multisolution approach or from programs such as *AMoRe*.

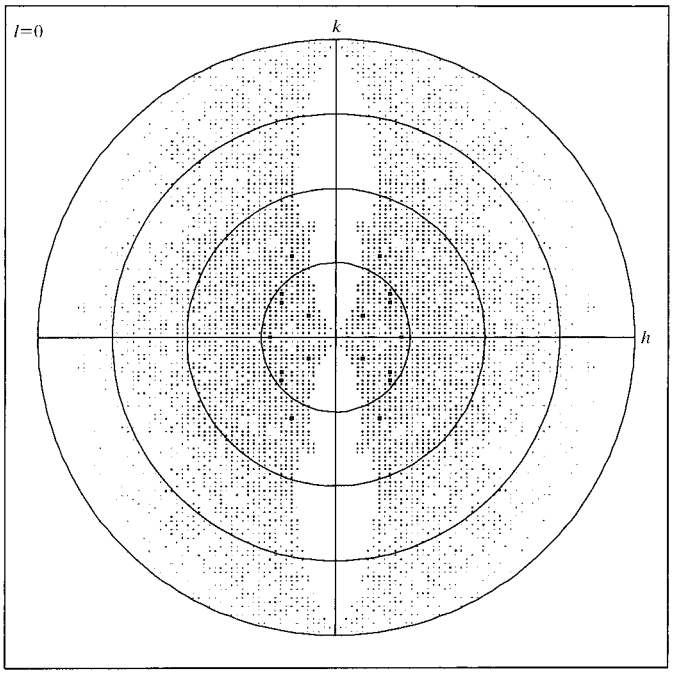

Figure 7
A representative of the diffraction pattern from cellulase Cel5A (PDB code 8a3h). The data are clearly anisotropic and the region around the a^* axis is missing from the measured set.

(ii) Density modification appears to give much faster and smoother structure development than peak-picking approaches such as those used in *Shake and Bake* or *SHELXD*. Like these, *ACORN-PHASE* uses algorithms in both reciprocal and real space, but the density modification allows slow growth of all the electron density over the whole asymmetric unit throughout the procedure.

(iii) The correlation coefficient for the weak $|E|$ values, CC_s, is an excellent and simple discriminator for a correct solution in *ACORN-PHASE*.

(iv) The quality of data that can be recorded nowadays is stunning. In-house sources with cryogenic cooling allow the recording of multiple measurements at 1.54 Å wavelength allowing the ready positioning of anomalous scatterers such as sulfur. Synchrotron sources with efficient detectors mean that increasing numbers of protein crystals diffract to atomic resolution. The anomalous scatterers provide an excellent oriented fragment for phase refinement with *ACORN-PHASE*.

A feature of *ACORN-MR* is that no fitting of models to maps or information about the target structure (except, in some applications, recognition of a related structure in the PDB) was required to obtain the excellent quality maps. Since there is no guarantee that an arbitrary fragment taken from an unrelated structure can be well matched to a part of the target structure, we feel that the full power of the method will be realised when a multi-solution approach is used. Future developments of *ACORN-MR* will use a library of α-helices of different lengths and β-sheet backbones. At present, a solution from an unoriented fragment can take many hours on a Silicon Graphics O2 workstation. The inevitable improvements of computers within the next few years should enable trials with many potential fragments to be made much more rapidly.

We believe the *ACORN-PHASE* approach can naturally be extended to function at lower resolution. Density modification clearly improves phases at all resolution ranges; with more sophisticated algorithms, it should be possible to remove the requirement for atomic resolution data. The density distribution changes subtly but predictably with resolution. In addition, Sayre's equation has been demonstrated to improve phases even with lower resolution data. This is supported by recent experience in the solution of pseudoazurin using 1.55 Å resolution data (Mukherjee *et al.*, 2000).

Again let us say that this kind of method emphasizes the desirability of collecting the most complete and highest resolution data possible. In times past when data collection was an arduous task, there was a tendency to solve a structure from lower resolution data, collected by some means or other within a reasonable time, and then later to extend the resolution for refinement. This makes no sense when data can be acquired quickly and accurately, as it removes an opportunity for a quick and efficient solution of the structure. It is certainly worthwhile using a few hours of computer time in an attempt

to find a solution automatically before embarking on alternative procedures.

It is our intention to automate fully the procedures described here and other related procedures that may be developed in the computer package *ACORN* provided with default parameters for automatic solution attempts and with provision for user intervention when the default procedures fail.

We have recently solved *ab initio* the structure of an unknown zinc-containing proteinase with a MW of 20 kDa using this program (McAuley *et al.*, in preparation). In this structure, the model could be automatically built from the peaks in the *E* map using the *ARP*/*wARP* autotracing option (Perrakis *et al.*, 1999), as could clearly have been performed for the examples quoted above.

We gratefully acknowledge the support of BBSRC (contract No. 87/B08494) for the support of YJ and of the University of York, the Royal Society and the Institute of Physics, Beijing which has enabled us to collaborate on this project. EJD, KSW and YJ thank the BBSRC for infrastructure support through the Centre grant to the YSBL.

References

Benini, S., González, A., Rypniewski, W. R., Wilson, K. S., Van Beeumen, J. J. & Ciurli, S. (2000). Submitted.
Bernstein, B. E. & Hol, W. G. J. (1997). *Acta Cryst.* D**53**, 756–764.
Bezborodova, S. I., Ermekbaeva, I. A., Shlyapnikov, S. V., Polyakov, K. M. & Bezborodov, A. M. (1988). *Biokhimiya*, **53**, 965–973.
Blessing, R. H., Guo, D. Y. & Langs, D. A. (1998). *Direct Methods for Solving Macromolecular Structures*, NATO ASI Series Volume, Series C: *Mathematical and Physical Sciences*, Vol. 507, edited by S. Fortier, pp. 47–71. Dordrecht: Kluwer.
Blessing, R. H. & Smith, G. D. (1999). *J. Appl. Cryst.* **32**, 664–670.
Burla, M. C., Carrozzini, B., Cascarano, G. L, Giacovazzo, C. & Polidori, G. (2000). *J. Appl. Cryst.* **33**, 307–311.
Chantalat, L., Wood, S. D., Rizkallah, P. & Reynolds, C. D. (1996). *Acta Cryst.* D**52**, 1146–1152.
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.
Deacon, A. M. & Ealick, S. E. (1999). *Struct. Fold. Des.* **7**, 161–166.
Ding, J., Fraser, M. E., Meyer, J. H. & Bartlett, P. A. (1998). *J. Am. Chem. Soc.* **120**, 4610–4621.
Evans, G., Joachimiak, A., Westbrook, E. M. & Walsh, M. A. (2000). In preparation.
Frazão, C., Soares, C. M., Carrondo, M. A., Pohl, E., Dauter, Z., Wilson, K. S., Hervas, M., Navarro, J. A., De la Rosa, M. A. & Sheldrick, G. M. (1995). *Structure*, **3**, 1159–1169.
Kissinger, C. R., Gehlhaar, D. K. & Fogel, D. B. (1999). *Acta Cryst.* D**55**, 484–491.
Lattmann, E. E. & Love, W. E. (1970). *Acta Cryst.* B**26**, 1854–1857.
Martinez-Oyanedel, J., Choe, H. W., Heinemann, U. & Saenger, W. (1991). *J. Mol. Biol.* **222**, 335–352.
Mukherjee, M., Maiti, S. & Woolfson, M. M. (2000). *Acta Cryst.* D**56**, 1132–1136.
Murshudov, G. N., Melik-Adamyan, W. R., Grebenko, A. I., Barynin, V. V., Vagin, A. A., Vainshtein, B. K., Dauter, Z. & Wilson, K. S. (1992). *FEBS Lett.* **302**, 127–131.
Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.
Oh, B.-H. (1995). *Acta Cryst.* D**51**, 140–144.
Pearl, L. & Blundell, T. (1984). *FEBS Lett.* **174**, 96–101.
Perrakis, A., Morris, R. J. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.
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.
Shiono, M. & Woolfson, M. M. (1991). *Acta Cryst.* A**47**, 526–533.
Shiono, M. & Woolfson, M. M. (1992). *Acta Cryst.* A**48**, 451–456.
Turkenburg, J. P. & Dodson, E. J. (1996). *Curr. Opin. Struct. Biol.* **6**, 604–610.
Varrot, A., Schulein, M., Pipelier, M., Vasella, A. & Davies, G. J. (1999). *J. Am. Chem. Soc.* **121**, 2621–2622.
Walsh, M. A., Schneider, T. R., Sieker, L. C., Dauter, Z., Lamzin, V. S. & Wilson, K. S. (1998). *Acta Cryst.* D**54**, 522–546.
Weeks, C. M., De Titta, G. T., Hauptman, H. A., Thuman, P. & Miller, R. (1994). *Acta Cryst.* A**50**, 210–220.
Weeks, C. M. & Miller, R. (1999). *Acta Cryst.* D**55**, 492–500.
Xu, H., Hauptman, H. A., Weeks. C. M. & Miller, R. (2000). *Acta Cryst.* D**56**, 238–240.
Yao, J.-X. (1981). *Acta Cryst.* A**37**, 642–644.