Acta Crystallographica Section D
Biological Crystallography
ISSN 1399-0047

Protein phasing at non-atomic resolution by combining Patterson and *VLD* techniques

Rocco Caliandro,[a] Benedetta Carrozzini,[a] Giovanni Luca Cascarano,[a] Giuliana Comunale,[b] Carmelo Giacovazzo[a]* and Annamaria Mazzone[a]

[a]Istituto di Cristallografia, CNR, Via G. Amendola 122/O, 70126 Bari, Italy, and [b]DiCEM, Università degli Studi della Basilicata, 75100 Matera, Italy

Correspondence e-mail: carmelo.giacovazzo@ic.cnr.it

Received 11 February 2014
Accepted 5 May 2014

Phasing proteins at non-atomic resolution is still a challenge for any *ab initio* method. A variety of algorithms [Patterson deconvolution, superposition techniques, a cross-correlation function (C map), the *VLD* (*vive la difference*) approach, the FF function, a nonlinear iterative peak-clipping algorithm (*SNIP*) for defining the background of a map and the *free lunch* extrapolation method] have been combined to overcome the lack of experimental information at non-atomic resolution. The method has been applied to a large number of protein diffraction data sets with resolutions varying from atomic to 2.1 Å, with the condition that S or heavier atoms are present in the protein structure. The applications include the use of *ARP/wARP* to check the quality of the final electron-density maps in an objective way. The results show that resolution is still the maximum obstacle to protein phasing, but also suggest that the solution of protein structures at 2.1 Å resolution is a feasible, even if still an exceptional, task for the combined set of algorithms implemented in the phasing program. The approach described here is more efficient than the previously described procedures: *e.g.* the combined use of the algorithms mentioned above is frequently able to provide phases of sufficiently high quality to allow automatic model building. The method is implemented in the current version of *SIR*2014.

1. Notation

ρ, ρ_p: electron densities of the target structure and the model structure, respectively.

$\rho_q = \rho - \rho_p$: ideal difference Fourier synthesis. Summed to ρ_p it exactly provides the true electron density ρ, irrespective of the quality of ρ_p.

N: the number of atoms in the unit cell for the target structure.

p: the number of atoms in the unit cell for the model structure. Usually $p \leq N$, but it may also be that $p > N$.

f_j, $j = 1, \ldots, N$: atomic scattering factors for the target structure (thermal factor included).

$F = \sum_{j=1}^{N} f_j \exp(2\pi i \mathbf{h}\mathbf{r}_j) = |F|\exp(i\varphi)$: structure factor of the target structure.

$F_p = \sum_{j=1}^{p} f_j \exp(2\pi i \mathbf{h}\mathbf{r}'_j) = |F_p|\exp(i\varphi_p)$, where $\mathbf{r}'_j = \mathbf{r}_j + \Delta\mathbf{r}_j$: structure factor of the model structure.

$F_q = F - F_p = |F_q|\exp(i\varphi_q)$: structure factor of the ideal difference structure.

$E = A + iB = R\exp(i\varphi)$, $E_p = A_p + iB_p = R_p\exp(i\varphi_{\rm p})$, $E_q = A_q + iB_q = R_q\exp(i\varphi_q)$: normalized structure factors.

$\Sigma_N = \sum_{j=1}^{N} f_j^2$, $\Sigma_p = \sum_{j=1}^{p} f_j^2$.

R'_p, R'_q: structure factors pseudo-normalized with respect to the target structure (*i.e.* $R'_p = |F_p|/\Sigma_N^{1/2}$, $R'_q = |F_q|/\Sigma_N^{1/2}$).

$D = \langle\cos(2\pi\mathbf{h}\Delta\mathbf{r})\rangle$; the average is performed per resolution shell.

$\sigma_A = D(\Sigma_p/\Sigma_N)^{1/2}$.
$\sigma_R^2 = \langle|\mu|\rangle\Sigma_N$; $\langle|\mu|^2\rangle$ is the measurement error.
$e = 1 + \sigma_R^2$.
$I_i(x)$: modified Bessel function of order i.
$m = \langle\cos(\varphi - \varphi_p)\rangle = I_1(X)/I_0(X)$, where $X = 2\sigma_A RR_p/(e - \sigma_A^2)$.
EDM: electron-density modification.
RESID $= \sum_{\mathbf{h}} |R - R_p| / \sum_{\mathbf{h}} R$; the sum is over the measured $\mathbf{h}$ reflections.
CORR is the correlation between the final electron-density map provided by the phasing procedure and the map calculated using published data.
RES: experimental data resolution.

2. Introduction

Direct methods solved the phase problem for small molecules and made feasible the *ab initio* crystal structure solution of small proteins using data at atomic resolution. Well documented representative computer programs include *SnB* (Weeks *et al.*, 1994; Rappleye *et al.*, 2002), *SHELXD* (Sheldrick, 2008), *ACORN* (Foadi *et al.*, 2000), *SIR*2002 (Burla *et al.*, 2002) and *SIR*2004 (Burla *et al.*, 2005). Until 2006, the largest unknown protein solved *ab initio* was cytochrome c_3 (PDB entry 1gyo; Frazão *et al.*, 1999), with 2024 non-H protein atoms in the asymmetric unit, solved by *SHELXD*. In 2006, Mooers and Matthews solved the unknown structure of bacteriophage P22 lysozyme (PDB entry 2anv), with 2268 non-H protein atoms in the asymmetric unit, using *SIR*2002 (Mooers & Matthews, 2006) .

Structural complexity, however, is not the strongest obstacle to *ab initio* protein crystal structure solution: indeed, data resolution is a more severe condition because it can reduce the amount of information accessible by a diffraction experiment when the resolution is not atomic. Two approaches are available today for solving proteins at non-atomic resolution *via ab initio* methods.

(i) Patterson deconvolution techniques. Their successful use requires that heavy atoms are present in the asymmetric unit with a sufficiently large occupancy factor: success may be hindered if they show exceedingly large thermal parameters. Such techniques, integrated with active use of the extrapolated reflections (Caliandro *et al.* 2005*a*,*b*, 2007*b*; see also Yao *et al.*, 2005), allowed the solution, at non-atomic resolution (1.65 Å), of a large-size protein structure (PDB entry 1e3u; eight Au atoms and 7890 non-H atoms in the asymmetric unit) and also successful solution of PDB entry 1buu, a protein with one Ho atom and 1282 non-H atoms in the asymmetric unit and 1.92 Å data resolution (Caliandro *et al.*, 2008).

(ii) The *ARCIMBOLDO* approach (Rodríguez *et al.*, 2009, 2012). This tries to locate small molecular fragments (*e.g.* α-helices) *via* molecular-replacement techniques, which however may return a huge number of partial solutions (*i.e.* hundreds or thousands) with very similar figures of merit. All of the potential solutions are used in searching for additional new fragments: at this stage the best trials may be recognized *via* suitable figures of merit. EDM procedures, including structure-factor extrapolation beyond the experimental resolution, then improve the phases and make the electron-density map interpretable. *ARCIMBOLDO* is very demanding in terms of computational power: the calculations are distributed on a computer grid and executed in a parallel manner.

This paper deals with further development of the phasing approaches based on Patterson deconvolution techniques. It combines three methods that have previously been used in separate contexts: (i) the cross-correlation function $C(\mathbf{u})$, also denoted the C map, which was recently introduced and crystallographically characterized by Carrozzini *et al.* (2010); (ii) the *VLD* (*vive la difference*) algorithm, originally proposed by Burla, Caliandro *et al.* (2010) for *ab initio* crystal structure solution and implemented in computer programs by Burla, Giacovazzo *et al.* (2010, 2011) and Burla, Carrozzini *et al.* (2012); and (iii) the *SNIP* algorithm, originally developed by Ryan *et al.* (1988) as a method for discriminating peaks from background.

In §3 we will briefly recall the properties of such methods, in §4 we will define the phasing procedure into which they are integrated and in §5 we will describe its application to proteins, with particular interest in those with data resolutions between 1.5 and 2.1 Å.

3. The main basic tools

The phasing procedure described in §4 requires the integration of numerous algorithms, which are briefly described below. The sequence of their application is postponed to §4.

3.1. Patterson deconvolution methods and superposition techniques

From the Patterson map $P(\mathbf{u})$, the implication transformation $I_s(\mathbf{r})$ for the sth symmetry operator $\mathbf{C}_s$ is calculated as

$$I_s(\mathbf{r}) = P(\mathbf{r} - \mathbf{C}_s\mathbf{r})/n_s, \tag{1}$$

where n_s is the number of symmetry operators that give rise to the same Harker section (Harker, 1936). The symmetry minimum function

$$\mathrm{SMF}(\mathbf{r}) = \min_{s=1}^{\bar{m}}[I_s(\mathbf{r})] \tag{2}$$

is then derived. A peak search on the SMF($\mathbf{r}$) map generates a list of high-intensity peak positions ($\mathbf{r}_\mathrm{H}$) hopefully corresponding to heavy-atom positions. For each of the above peaks, the minimum superposition function

$$S(\mathbf{r}) = \min[P(\mathbf{r} - \mathbf{r}_\mathrm{H}), \mathrm{SMF}(\mathbf{r})] \tag{3}$$

is calculated to filter the SMF function and to provide a better starting model for the target structure. For further details of these techniques, see Buerger (1948, 1959), Simpson *et al.* (1965), Nordman (1966), Richardson & Jacobson (1987), Sheldrick (1992) and Pavelčík *et al.* (1992).

3.2. The cross-correlation function $C(\mathbf{u})$

If ρ and ρ_p are the target and a model structure, respectively, then

$$C(\mathbf{u}) = \rho(\mathbf{r}) \otimes \rho_p(\mathbf{r}) = \int_S \rho(\mathbf{r})\rho_p(\mathbf{r}+\mathbf{u})\,\mathrm{d}\mathbf{r}$$
$$= \frac{1}{V}\sum_{\mathbf{h}} |F_{\mathbf{h}}F_{p\mathbf{h}}| \exp i(\varphi_{\mathbf{h}} - \varphi_{p\mathbf{h}}) \exp(-2\pi i \mathbf{h}\cdot\mathbf{u}). \quad (4)$$

The coefficient $|F_{\mathbf{h}}F_{p\mathbf{h}}|\exp i(\varphi_{\mathbf{h}} - \varphi_{p\mathbf{h}})$ is a complex number, and therefore $C(\mathbf{u})$ is acentric (it is centric only if both $\rho_{\mathbf{r}}$ and $\rho_{p\mathbf{r}}$ are centric). Its space group is the symmorphic variant of the space group of the target structure (*e.g. Pm* against *Pc*).

$C(\mathbf{u})$ is not available during the phasing process, essentially because the $\varphi_{\mathbf{h}}$s, the target phases in which we are interested, are unknown. The approximating function

$$C'(\mathbf{u}) = \frac{1}{V}\sum_{\mathbf{h}} m_{\mathbf{h}} |F_{\mathbf{h}}F_{p\mathbf{h}}| \exp(-2\pi i \mathbf{h}\cdot\mathbf{u}), \quad (5)$$

obtained by assuming $\varphi_{\mathbf{h}} \simeq \varphi_{p\mathbf{h}}$, may then be used, particularly if the $m_{\mathbf{h}}$ values are sufficiently large. The space group of $C'(\mathbf{u})$ is centric, thus coinciding with the Patterson space group (*e.g. P*2/*m* if the space group of the target is *Pc*). The fundamental advantage of the C' map over the Patterson map is that the vectors between model atoms and the vectors between model and nonmodel atoms are present and satisfy the Laue symmetry. The amount of noise in the C' map is however reduced with respect to the Patterson map because vectors between nonmodel atoms (mostly vectors between light atoms) should be weak or absent.

The $C'(\mathbf{u})$ function has recently been combined with the implication transformations to solve medium-sized structures (*i.e.* with up to about 400 non-H atoms in the asymmetric unit), even in the case of light atoms (Caliandro, Carrozzini *et al.*, 2013). It has never been applied to proteins.

3.3. The *VLD* (*vive la difference*) approach

This is based on the properties of the Fourier transform, and originates from study of the joint probability distribution function $P(R, R_p, R_q, \varphi, \varphi_p, \varphi_q)$. *VLD* suggests the following coefficient for the difference Fourier synthesis:

$$\left[(mR - \sigma_A R_p) - R'_p(1-D)\left(\frac{e - \sigma_A^2}{1-\sigma_A^2}\right)\right]\exp(i\varphi_p), \quad (6)$$

which contains the classical Read (1986) difference term

$$(mR - \sigma_A R_p)\exp(i\varphi_p) \quad (7)$$

and the flipping term

$$-R'_p(1-D)\left(\frac{e-\sigma_A^2}{1-\sigma_A^2}\right)\exp(i\varphi_p). \quad (8)$$

VLD does not require an atomic model and works on electron-density maps: in these conditions it is frequent to assume

$$\Sigma_p = \Sigma_N \quad (9)$$

so that (6) becomes

$$\Delta E \simeq (mR - R_p)\exp(i\varphi_p). \quad (10)$$

The *VLD* algorithm may be schematized in three macro-steps:

(i) the difference Fourier synthesis (10) is calculated, conveniently modified and inverted, irrespective of the quality of the model;

(ii) the corresponding Fourier coefficients E_q are combined with the normalized structure factors of the model structure through the tangent formula

$$\tan\varphi = \frac{R_p \sin\varphi_p + w_q R_q \sin\varphi_q}{R_p\cos\varphi_p + w_q R_q \cos\varphi_q}, \quad (11)$$

where $w_q = [2(1 - \sigma_A)]^2$;

(iii) the observed Fourier synthesis, calculated *via* the phases φ defined by (11), is submitted to EDM cycles. At the end a new model structure is obtained and the program returns to (i).

VLD was originally designed for *ab initio* phasing and has been applied to a wide range of structures, from small molecules to proteins, provided that the data have atomic resolution. In a recent paper (Carrozzini *et al.*, 2013), *VLD* was successfully combined with molecular-replacement techniques and used as a powerful tool for phase extension and refinement, irrespective of the data resolution. In this paper, we will check its usefulness for extending and improving the poor phases obtained by Patterson deconvolution techniques at non-atomic resolution, in particular the set of phases produced by application of the C' map.

3.4. Nonlinear iterative peak-clipping algorithm (*SNIP*)

This was originally proposed by Burgess & Tervo (1983) and further developed by Ryan *et al.* (1988) as a method for estimating the background in a spectrum. In one dimension, given the original spectrum $y(i)$, where $i = 1, \ldots, n$ indicates the channel of the spectrum, a new value in the ith channel is calculated after p iterations as

$$y_1(i) = y(i)$$
$$y_2(i) = \min\left[y_1(i), \frac{y_1(i+2) + y_1(i-2)}{2}\right]$$
$$\vdots$$
$$y_p(i) = \min\left[y_{p-1}(i), \frac{y_{p-1}(i+p) + y_{p-1}(i-p)}{2}\right]. \quad (12)$$

An estimate of the background $b(i)$ is obtained after Nclip iterations: $b(i) = y_{\mathrm{Nclip}}(i)$. Thus, Nclip is a free parameter of the algorithm representing both the number of iterations and the width of the clipping window. The latter should be chosen taking into account the width of the peaks one wants to preserve in the spectrum (ideally, the Nclip channel should cover the half-width of the peaks). The algorithm was extended to multi-dimensional data by Morháč *et al.* (1997), Morháč & Matoušek (2008) and Morháč (2009) in such a way that it is able to recognize useless information (background, combinations of coincidences of the background with peak ridges) from useful information contained in n-fold coincidence peaks of n-dimensional maps. *SNIP* was first applied to diffraction data by Caliandro, Di Profio *et al.* (2013) for the quantitative analysis of powder data. Here, it is applied for the first time to three-dimensional crystallographic maps for improving phasing. In doing this, we have introduced boundary conditions to the algorithm [*i.e.* $y(ix, iy, iz) = y(ix + nx, iy + ny, iz + nz)$, where nx, ny, nz are the number of grid points along the three unit-cell axes, and we have set the parameter Nclip according to the data resolution. As a general rule, Nclip = int(n/Res), where Res is the data resolution in Å and $n = 3$ or 4 if Res is higher or lower than 1.6 Å, respectively. In our procedure *SNIP* is used in the Patterson map analysis to discriminate peaks from the background according to

$$\hat{P}(\mathbf{u}) = P(\mathbf{u}) - b(\mathbf{u}) \tag{13}$$

and to better estimate their positions and intensities.

3.5. FF function

The FF function is defined as

$$\mathrm{FF}(\mathbf{u}) = \frac{1}{V}\sum_{\mathbf{h}} |F_{\mathbf{h}}|^2 \exp(2i\varphi_{\mathbf{h}} - 2\pi i \mathbf{h}\mathbf{u}),$$

and provides information on the sum of the atomic positional vectors (Burla *et al.*, 2006*a*). If the space group is centric and the inversion centre is at $\mathbf{X}_0$, then FF($\mathbf{u}$) will show a very strong peak at $2\mathbf{X}_0$. In the Patterson deconvolution step the FF function may be actively used during the EDM cycles, when the residual centrosymmetry of the $S(\mathbf{r})$ map makes it more difficult to drive the phases towards the acentric solution. FF($\mathbf{u}$) is applied to reduce the centric nature of the $S(\mathbf{r})$ map and to monitor, *via* the ratio FF($2\mathbf{X}_0$)/P(0), the gradual disappearance of the pseudo-inversion centre. Superposition of the Patterson map and the FF function (both shifted by $\mathbf{r}_{\mathrm{H}}$, the positional vector of a model atom) leads to a modified map

$$\mathrm{PFF}(\mathbf{r}) = \min[P(\mathbf{r} - \mathbf{r}_{\mathrm{H}}), \mathrm{FF}(\mathbf{r} + \mathbf{r}_{\mathrm{H}})]$$

with less noise than the two original maps, since false peaks do not systematically overlap (Burla *et al.*, 2006*a*). The quality of the FF function depends on the quality of the phases, and it generally improves with the EDM cycles; however, to save computing time it is only used in the first two EDM cycles of the Patterson deconvolution step.

3.6. Structure-factor extrapolation, also called *free lunch*

A method has been proposed to reduce the drawbacks generated by the limited data resolution (Caliandro *et al.*, 2005*a*,*b*). The current set of phases and the corresponding observed moduli are used to extrapolate the moduli and phases of nonmeasured reflections both beyond and behind the experimental resolution. The method modifies the current observed electron density and Fourier-inverts the modified map: structure factors are then extrapolated up to the desired resolution. The amplitudes of the extrapolated reflections are replaced by the observed moduli when they are in the data: otherwise, extrapolated moduli and phases are actively used in all of the phasing steps to increase the rate of success of the phasing procedure.

The extrapolation limit is automatically calculated by the program if the experimental resolution is less than 1.5 Å; otherwise, the limit is fixed to 1.2 Å (we do not extend the extrapolation to 1 Å resolution to avoid the contribution of the generated reflections in the phasing process dominating the contribution of the observed reflections). The selected extrapolated reflections are actively used in the subsequent electron density: their weights are calculated by extrapolating the σ_A curve up to the extended resolution.

3.7. FOM

The quality of the current phases is estimated *via* the figure of merit (see Burla *et al.*, 2013 for a related FOM),

$$\mathrm{fFOM2} = \mathrm{fFOM} \cdot \mathrm{CC(all)}_{\mathrm{current}}, \tag{14}$$

where

$$\mathrm{fFOM} = \frac{\mathrm{RAT}_{\mathrm{current}}}{\mathrm{RAT}_{\mathrm{initial}}} \frac{\mathrm{CC(all)}_{\mathrm{current}}}{\mathrm{CC(all)}_{\mathrm{initial}}} \frac{\mathrm{CC(large)}_{\mathrm{current}}}{\mathrm{CC(large)}_{\mathrm{initial}}}.$$

Let R_{calc} be the amplitudes of the normalized structure factors obtained by the inversion of a small percentage (3.5%, corresponding to the pixels with highest intensity) of the current electron-density map. Then RAT = $\mathrm{CC}_{w,R}/\langle R^2_{\mathrm{calc}}\rangle_{\mathrm{weak}}$, where the average $\langle R^2_{\mathrm{calc}}\rangle_{\mathrm{weak}}$ is calculated over 30% of the measured reflections (those with the weakest $|F_{\mathrm{obs}}|$ values). $\mathrm{CC}_{w,R}$ is the correlation coefficient between the largest R_{obs} amplitudes (about 70% of the total) and the corresponding σ_A weights.

CC is the correlation factor between R_{obs} and R_{calc}: all, large and weak indicate the overall set of normalized structure factors, the subset (70%) of the largest $|F_{\mathrm{obs}}|$ values and the subset (30%) of the weakest values.

While fFOM estimates the relative phase improvement (from the initial to the current state), fFOM2 includes an absolute estimate of the quality of the phases since it involves the current value of CC.

3.8. *RELAX* procedure

The *RELAX* procedure (Burla, Carrozzini *et al.*, 2000; Caliandro *et al.*, 2007*a*) aims at automatically placing in the correct position a model structure that is correctly oriented and misplaced. The set of phases obtained at the end of the phasing procedure is expanded in $P1$ and refined *via* EDM techniques. The origin shift to apply to the current electron-density map is identified, and the phases are recalculated and automatically returned to the correct space group, where a final EDM refinement is performed. It proved extremely useful when applied to the *VLD* approach (Burla *et al.* 2012) and to revisited direct-methods approaches (Burla *et al.*, 2013). Unexpectedly, *RELAX* also plays an important role in Patterson deconvolution methods (see the last part of §5), the success of which implies that the correct position of the heavy atoms (and therefore the correct origin in the space group) has previously been found.

4. The phasing procedure

The phasing procedure may be decomposed into three main moduli. The first concerns the Patterson deconvolution step, the second phase refinement and the third automatic model building.

4.1. Patterson deconvolution step

$P(\mathbf{u})$ is computed by using proper weights in order to more readily locate the vectors between heavy atoms. A list of peak positional vectors (*i.e.* the pivot peaks) are obtained *via* the function SMF($\mathbf{r}$) (see equation 2). Since high-intensity peaks may occasionally lie on Harker sections without corresponding to interatomic vectors between symmetry-equivalent heavy atoms, a multisolution approach is introduced. NPIVOT peaks are selected (those with the highest intensity) and for each pivot peak the following deconvolution procedure is applied.

The positional parameters and the thermal factor of the pivot peak undergo a least-squares refinement when an atom heavier than Ca is present in the structure. The C' map is then calculated, based on the pivot peak position, through (5). A new minimum superposition function is then computed not *via* the standard (3) but *via*

$$S'(\mathbf{r}) = \min[C'(\mathbf{r} - \mathbf{r}_{\rm H}), {\rm SMF}(\mathbf{r})]. \qquad (15)$$

The same algorithm is applied for space group $P1$, provided that the SMF map is replaced by the Patterson map in (3) and (15).

The $S'(\mathbf{r})$ map is subjected to EDM cycles, where filtering procedures are applied to break down the residual Patterson symmetry and the pseudo-translational symmetries generated by the deconvolution process. They have not been changed with respect to our previous implementation (Caliandro *et al.*, 2007*a*). The only modification introduced here is that now the C' map plays the role of the Patterson map in the old algorithm.

Background subtraction (BS) through the *SNIP* algorithm is activated if at least one atom heavier than Ca is included in the cell contents and the crystal symmetry is lower than tetragonal. In fact, we observed that (i) light-atom peaks could be potentially included in the estimated background and (ii) higher symmetry makes two or more different directions equivalent and therefore increases the chances of random coincidences of background in some dimensions and peaks in others.

SNIP operates on the three-dimensional maps $P(\mathbf{u})$, SMF($\mathbf{r}$), $S(\mathbf{r})$, $C'(\mathbf{r})$ and $S'(\mathbf{r})$ and, separately, on the two-dimensional sections $P(\mathbf{r} - \mathbf{C}_s\mathbf{r})$ generated throughout the Patterson module. BS substitutes the common map modification, which is usually performed by putting to zero all pixels above a given threshold, usually depending on the standard deviation of the map. While the common map modification strongly changes the shape of the peaks (indeed, their tails below the threshold are truncated), the BS modification preserves the shape of the peaks and only reduces their height. Preserving the shape of the peaks reduces the series-termination error produced by the repeated FFTs operated in the framework of EDM procedures. An example is given in Fig. 1, where the Harker section of the protein structure with PDB code 2f14 (space group $P2_1$) is shown before (Fig. 1*a*) and after (Fig. 1*b*) the application of a threshold cut. In Fig. 1(*b*) all of the pixels with intensity below $\langle\rho\rangle + 0.66\sigma_\rho$ are put to zero, where $\langle\rho\rangle$ and σ_ρ are the average and standard deviation, respectively, of the pixel intensities calculated on the whole Patterson map. It can be noticed that the threshold cut preserves the highest part of the peaks and truncates their tails: as a consequence, the relative height of the peaks is not modified and the lower part of the map is cleared up. Instead, BS filtering modifies the shape of the peaks and changes their relative height. Fig. 1(*c*) reports the same section modified by BS filtering: the highest peak, corresponding to the Hg Harker vector, is more discriminated in Fig. 1(*c*), although its absolute height is smaller than in Figs. 1(*a*) and 1(*b*).

From the above description, it is clear that the procedure attempts to solve a protein structure *via* NPIVOT phasing trials, each corresponding to a selected pivot peak. In difficult cases NPIVOT may attain a value of 100 or more.

4.2. Direct-space refinement step

The direct-space refinement module described here is an evolution of that implemented in *SIR*2011 (Burla *et al.*, 2012): it has been designed to increase the phasing efficiency and to produce higher quality electron-density maps. The new procedure uses some of the tools described in §3 and is applied to the electron-density maps provided by the Patterson deconvolution step.

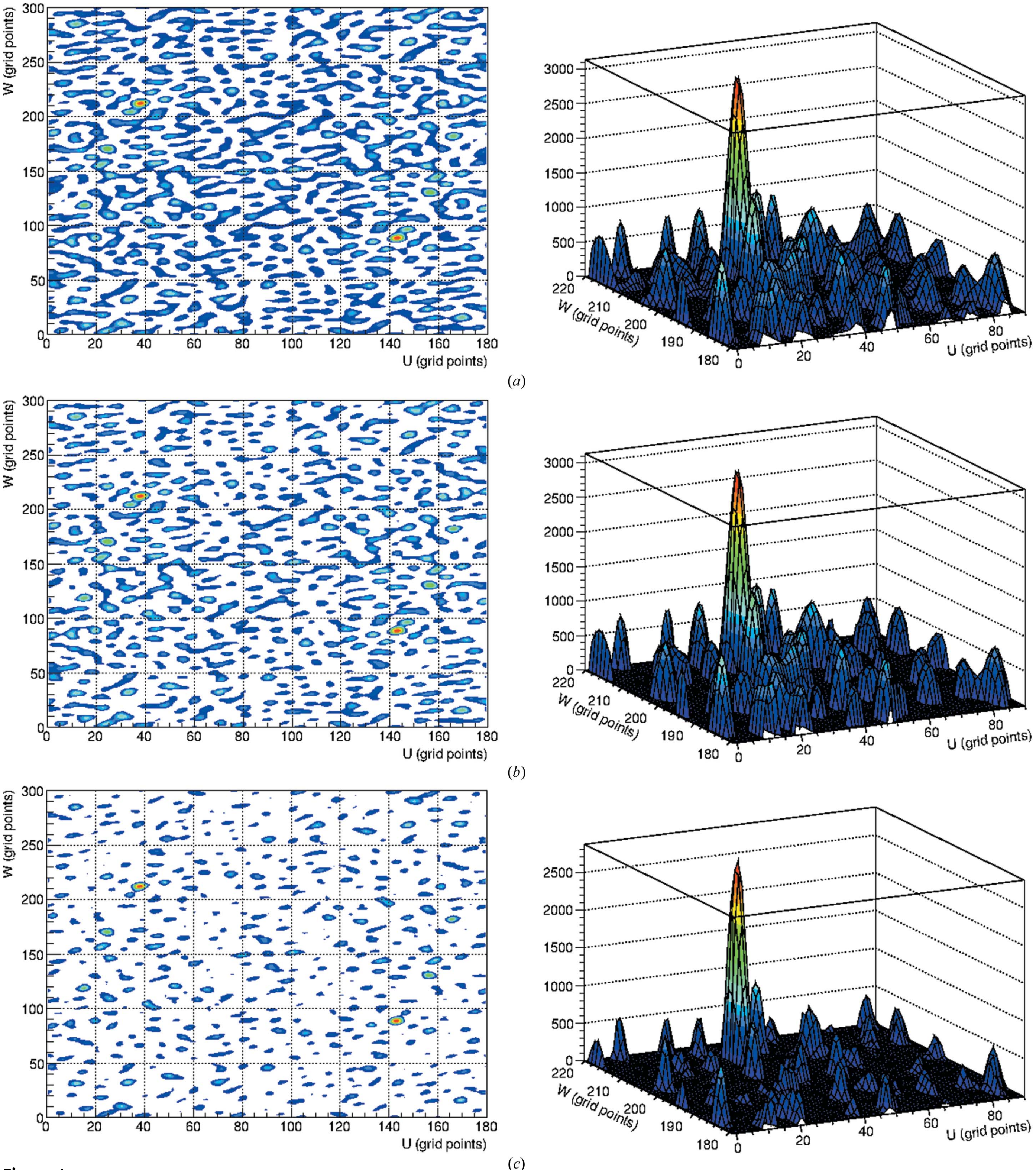
The phase-refinement process is iterative. If the structure is not solved, as indicated by fFOM2, up to 15 iterations may be performed. If the correct solution is recognized, the iterations are interrupted and the final set of phases is submitted to *ARP/wARP* (Perrakis *et al.*, 1999) for automatic building of the molecular model.

Each iteration is arranged into five blocks, and the generic block may be described *via* the following steps.

(i) The *VLD* step. In the first block of the first iteration the phases φ_p, corresponding to the heavy-atom substructure model supplied by the Patterson deconvolution step, are used to calculate the difference electron-density map (10). This, suitably modified (4% of the pixels with positive and negative density, those with the largest absolute values, are accepted unchanged, while the rest are set to zero) and inverted, provides the set of phases φ_q to be used in (11). New phases φ_p are thus obtained with weights defined by the tangent formula.

(ii) The EDM step. In accordance with §3.3, EDM cycles follow. In particular, three EDM macro-cycles, each constituted by eight or ten micro-cycles $\rho \rightarrow \varphi \rightarrow \rho$, are performed. The modification of the current map is mainly based on the inversion of small density percentages (up to 10%) and includes powering of the map (Refaat & Woolfson, 1993) and the inversion of small negative domains (Burla *et al.*, 2003). The molecular envelope (Wang, 1985; Leslie, 1987) is used as a mask (different weights are assigned to pixels falling inside or outside the envelope) to reduce the intensities of false peaks and clean the density map.

Owing to the limited information available from data at non-atomic resolution, the role of the extrapolated reflections (beyond and behind the experimental resolution) is crucial for the success of the phasing process. Structure factors (moduli and phases) of unobserved reflections are extrapolated *via* Fourier modification and inversion of the current density map



(*a*)

(*b*)

(*c*)

Figure 1

Protein with PDB code 2f14, space group $P2_1$. Top (left) and side enlarged (right) views of the Harker section of (*a*) the original Patterson map, (*b*) the map modified by using a threshold cut and (*c*) the map submitted to BS.

and are actively used in the calculation of the next electron-density function.

The phases φ_p corresponding to the model available at the end of each block are used to start the next block.

(iii) The *RELAX* step. At the end of each trial, under suitable conditions as specified below, the final phases are submitted to *RELAX* in order to recover possible positional misfitting of the electron-density map.

4.3. Stopping the phasing step and automatic model building

In any multisolution phasing approach, the procedure should be interrupted at the end of that phasing trial for which a suitable figure of merit suggests that the correct solution has been obtained. A useful FOM may be the sequence coverage obtained by an automatic building program such as *ARP/wARP*: the program should stop if a sufficiently large coverage is attained, while further phasing trials should be explored in the case of low coverage. Unfortunately, when the order of the pivot peak leading to the correct solution is low in the ranked list, the application of *ARP/wARP* to the electron-density map available at the end of each phasing trial is very time-consuming. The computing time problem is emphasized by a feature common to any direct-methods/Patterson approach: the phasing procedure randomly chooses one of the two possible structural enantiomorphs. Therefore, *ARP/wARP* must be applied a second time if satisfactory sequence coverage is not obtained for the enantiomorph randomly chosen by the phasing process.

In order to reduce the computing time without renouncing the goal of automatically providing a sound structural model, our phasing procedure (i) calculates fFOM2 at the end of each phasing trial and (ii) establishes a threshold TRFOM2 such that *ARP/wARP* is only applied to trials for which

$$\mathrm{fFOM2} > \mathrm{TRFOM2}. \qquad (16)$$

Unfortunately, such a simple approach is not very effective. Indeed, as for any statistical FOM, it may occur that (16) is obeyed but the correct solution is not attained: the program then stops at a false solution. It may also occur that high-quality phases are available but (16) is not obeyed: in this case the trial is abandoned, the solution is missed and further trials are explored in vain. (The reader should consider the fact that Patterson techniques are intrinsically different from direct methods. In the latter high-quality phases may be obtained from different random starting sets. In Patterson techniques the correct solution is hardly obtained from bad pivot peaks.)

To reduce the frequency of the above-described failures, we use the following algorithm:

(i) TRFOM2 is initially set to 3.0, a reasonable value suggested by our experimental tests;

(ii) *ARP/wARP* is applied by default to the first phasing trial, irrespective of the value of fFOM2 (it is always assumed that the first trial is the most favourable one);

(iii) if the correct solution is not attained, the following five trials are grouped into a batch;

(iv) if (16) is not verified for any of the five solutions, *ARP/wARP* is applied to the trial solution in the group with the largest value of fFOM2 (indeed, for some structures the condition fFOM2 > 3 may be not satisfied even for good solutions). A new batch is then explored.

If the coverage is larger than 0.75 the program stops, otherwise TRFOM2 is updated (it is set equal to the largest value of fFOM2 obtained in the preceding trials) and the program explores the next trials.

5. Practical applications

In order to check the real degree of efficiency of our phasing procedure, we have applied it to a wide set of proteins constituted of 126 test structures subdivided into five categories, with each category being representative of a type of potential phasing difficulty. Most of the test structures were employed by Burla *et al.* (2006*a*,*b*) and by Caliandro *et al.* (2008) to check the level of efficiency of previous Patterson-based algorithms. Some new entries have been included to make the number of structures in each test statistically meaningful.

In our experience, Patterson deconvolution methods are mostly sensitive to three parameters: the weight of the heavy atoms present in the molecule, the structural complexity and the data resolution. In order to maintain a sufficiently small number of categories, the ratio heavy-atom scattering power/total scattering power has not been explicitly taken into account in defining the categories. However, this ratio will be considered when the success or failure of cases is discussed.

In accordance with the above criteria, we subdivided the test structures into five subsets. RES, NASYM and NPROT denote the experimental data resolution, the number of non-H atoms in the asymmetric unit and the number of test structures, respectively.

Subset 1: RES $\leq$ 1.25 Å, NASYM $\leq$ 2000, at least one atomic species heavier than Ca. NPROT = 22.

Subset 2: RES $\leq$ 1.25 Å, NASYM $\leq$ 2000, no atomic species heavier than Ca. NPROT = 29.

Subset 3: RES $\leq$ 1.25 Å, 2000 < NASYM $\leq$ 6500, at least one atomic species heavier than Ca. NPROT = 24.

Subset 4: 1.25 Å < RES $\leq$ 1.5 Å, NASYM $\leq$ 2000, at least one atomic species heavier than S. NPROT = 22.

Subset 5: 1.5 Å < RES $\leq$ 2.1 Å, NASYM < 8000, at least one atomic species heavier than Fe. NPROT = 29.

The results are shown in Tables 1–5: they provide a wide and meaningful overview of the practical efficiency of a modern *ab initio* Patterson technique. In all our applications the above-described default procedure is used without any user intervention (the *ARP/wARP* model-building step is also performed automatically). As a further useful detail, we add that the number of heavy atoms is not a critical parameter for the success of the phasing procedure. The user should only provide the heavy-atom species if he has such information; indeed, the highest peaks in the SMF map are explored as possible pivot atoms for Patterson deconvolution, but the map is never interpreted in terms of atomic species. If more than

Table 1
Crystallographic data and phasing results for the test structures with RES ≤ 1.25 Å, NASYM ≤ 2000 and at least one atomic species heavier than Ca.

For each test structure, PDB code is the Protein Data Bank code, S.G. is the space group, RES is the data resolution (Å), Resid is the number of residues in the asymmetric unit, HA is the heavy-atom species and the corresponding number of atoms in the asymmetric unit, Trial is the trial at which the correct structure solution was found, fFOM2 is the value of the corresponding figure of merit, CC_P is the correlation between the electron-density map obtained at the end of the phasing procedure and the electron-density map corresponding to the published structure and COV is the sequence coverage obtained by applying *ARP*/*wARP* to the best map as selected by our phasing procedure.

PDB	S.G.	RES	Resid	HA	Trial	fFOM2	CC_P	COV
1fy2	$C2$	1.22	220	Cd 1	1R1	4.0	0.90	99
1e29	$C222_1$	1.21	138	Fe 1, Ca 3	1	4.7	0.89	99
1i76	$P2_12_12_1$	1.20	165	Zn 2, Ca 2	1	6.4	0.93	99
1cku	$P2_12_12_1$	1.20	170	Fe 4	1R2	6.5	0.95	99
1irn	$R3$	1.20	53	Zn 1	1	5.7	0.94	98
1a6g	$P2_1$	1.17	151	Fe 1	1	5.2	0.88	99
1a6n	$P2_1$	1.14	151	Fe 1	1	5.4	0.87	99
1ctj	$R3$	1.11	89	Fe 1	1	4.5	0.87	98
1jm1	$P6_1$	1.11	203	Fe 2	1	9.4	0.96	97
1iro	$R3$	1.11	53	Fe 1	1	7.3	0.93	98
1a6k	$P2_1$	1.10	151	Fe 1	1	7.6	0.92	99
1swz	$P3_221$	1.06	164	Rb 5	1	7.5	0.97	99
1nls	$I222$	1.03	238	Mn 1, Ca 1	1	8.9	0.97	99
8rxn	$P2_1$	1.03	52	Fe 1	1R2	6.6	0.94	98
1mfm	$P2_12_12_1$	1.02	153	Cd 9, Cu 1, Zn 1	1	7.2	0.95	99
1mso	$R3$	1.01	102	Zn 2	1	7.4	0.93	98
1a6m	$P2_1$	1.01	151	Fe 1	1	8.4	0.94	99
1eb6	$P2_1$	1.01	177	Zn 1	1	9.6	0.96	99
2bf9	$C2$	0.99	35	Zn 1	1	5.4	0.95	96
1c75	$P2_12_12_1$	0.97	71	Fe 1	1	10.6	0.96	98
2fdn	$P4_32_12$	0.94	55	Fe 8	1	8.5	0.98	97
1b0y	$P222$	0.94	85	Fe 4	1R2	6.5	0.95	98

Table 2
Crystallographic data and phasing results for the test structures with RES ≤ 1.25 Å, NASYM ≤ 2000 and no atomic species heavier than Ca.

For the column headings, see Table 1. For structures classified in the PDB file as antibiotics, protein type, an asterisk in the Resid column indicates that the number of non-H atoms in the asymmetric unit is given instead of the number of residues. Correspondingly, n.a. in the COV column indicates that *ARP*/*wARP* has not been applied. Dashes in the Trial, fFOM2, CC_P and COV columns indicate that the correct solution has not been found.

PDB	S.G.	RES	Resid	HA	Trial	fFOM2	CC_P	COV
9pti	$P2_12_12_1$	1.24	58	P 1, S 8	64	5.4	0.90	98
2knt	$P2_1$	1.20	58	P 1, S 6	5	4.4	0.89	98
1bx7	$P4_32_12$	1.20	51	S 11	1R2	4.5	0.93	88
1b9o	$C2$	1.15	124	Ca 1	1	5.5	0.91	99
1d4t	$P2_1$	1.14	115	S 3	—	—	—	—
1bkr	$P2_1$	1.10	108	S 4	4	7.8	0.95	99
1nkd	$C2$	1.10	59	S 4	—	—	—	—
1igd	$P2_12_12_1$	1.10	61	S 1	—	—	—	—
1a0m	$I4$	1.09	32	S 10	1	7.4	0.94	46
1sho	$P4_32_12$	1.09	207*	Cl 6	35	11.0	0.96	n.a.
1kf3	$P2_1$	1.05	125	S 13	6	7.7	0.94	99
†	$P1$	1.05	444*	Cl 12	1	7.7	0.90	n.a.
1cex	$P2_1$	1.02	197	S 5	38R2	8.3	0.94	99
1exr	$P1$	1.01	151	Ca 5	1	7.1	0.94	99
3erl	$C2$	1.01	40	S 7	1	5.8	0.93	97
1hhz	$P3_221$	0.99	354*	Cl 6	17	11.1	0.96	n.a.
352d	$P1$	0.99	1847*	Ca 9	12	5.4	0.77	n.a.
1aho	$P2_12_12_1$	0.98	64	S 8	1	7.7	0.96	98
3ltz	$P1$	0.96	132	S 10	1	8.0	0.92	99
1lzt	$P1$	0.96	129	S 10	1	8.2	0.92	99
1byz	$P1$	0.95	54	Cl 1	2	7.7	0.97	97
1ick	$P2_12_12_1$	0.95	250*	Mg 1, P 1	5	10.2	0.98	n.a.
1aa5	$P4_32_12$	0.92	200*	Cl 8	12R2	9.1	0.94	n.a.
2pvb	$P2_12_12_1$	0.91	110	Ca 2	1	10.5	0.97	99
1hhu	$P2_1$	0.89	505*	Cl 8	4R2	11.0	0.95	n.a.
1hhy	$P6_322$	0.89	208*	Cl 4	3	10.7	0.91	n.a.
1dy5	$P2_1$	0.88	251	S 27	15	8.9	0.97	99
3pyp	$P6_3$	0.86	125	S 6	1	8.8	0.96	99
1cnr	$P2_1$	0.85	48	S 6	1R2	7.4	0.91	97

† Loll *et al.* (1998).

one heavy atom is present, more than one pivot atom can lead to the same crystal solution.

For all our calculations, we employed the Intel Xeon E5-2690, 2.9 GHz, 64 bit Intel Fortran compiler.

In the following, we analyze the experimental outcome for each category: as a rule of thumb, we will consider a structure to be solved when $CC_P > 0.75$, where CC_P is the correlation between the electron-density map obtained at the end of the phasing procedure and the published map.

5.1. Subset 1

This subset of 22 structures (see Table 1) is representative of the ideal conditions for Patterson techniques (high resolution, an atomic species with atomic number larger than that of Ca, less than 2000 non-H light atoms in the asymmetric unit). All of the test structures were solved in trial 1 with a high-quality final electron-density map (the minimum value of CC_P was 0.87): evidently, high-resolution data allow easy recognition of the heavy-atom positions. Furthermore, all of the final electron-density maps were automatically interpreted by *ARP*/*wARP* (the minimum sequence coverage was 0.96 for 2bf9).

5.2. Subset 2

This subset of 29 test structures may be challenging for procedures based on Patterson techniques even if their diffraction data have atomic resolution; indeed, heavy atoms with atomic number larger than that of Ca are not allowed. The most immediate consequence is the following: the SMF may provide pivot peaks that are not useful for successful application of the superposition techniques. The crystal structure solution is then delayed to late trials or is not attained.

The average quality of the final electron-density maps provided by our phasing procedure, when successful, is high: the minimum value of CC_P was 0.77, which was obtained for 352d (a DNA structure): for the other 25 solved test structures $CC_P > 0.88$.

Only three structures remained unsolved: 1nkd, 1igd and 1d4t. All these structures have S as a heavy-atom species. For 1igd the SMF does not provide the S position (it is immersed in the map background), whereas correct S positions were found for 1nkd and 1d4t but it was impossible to recovery the full structure from the S substructure. The reason is the following: the phases defined by the S substructures are different, on average, from the corresponding target phases by more than 80°, and the phase-refinement procedure was not able to reduce such a large error.

Table 3
Crystallographic data and phasing results for the test structures with RES ≤ 1.25 Å, 2000 < NASYM ≤ 6500 and at least one atomic species heavier than Ca.

For the column headings, see Table 1. Dashes in the Trial, fFOM2, CC_P and COV columns indicate that the correct solution has not been found.

PDB	S.G.	RES	Resid	HA	Trial	fFOM2	CC_P	COV
1n8k	$P1$	1.21	749	Zn 4	14	5.2	0.87	99
1gyo	$P3_1$	1.20	212	Fe 8	1	6.4	0.94	99
1e9g	$P2_12_12_1$	1.20	567	Mn 8	3	5.5	0.94	99
1suf	$C2$	1.18	633	Ni 1, Fe 10	1	6.3	0.93	99
1heu	$P1$	1.17	750	Cd 4	1	6.6	0.89	99
1wkq	$C222_1$	1.16	313	Zn 16	3	5.7	0.95	99
1p6o	$R32$	1.15	322	Zn 2, Ca 2	1	6.8	0.93	98
1su7	$C2$	1.14	633	Ni 4, Fe 10	1	5.7	0.94	99
1het	$P1$	1.12	750	Zn 4	14	5.2	0.83	99
1moo	$P2_1$	1.11	257	Hg 1, Zn 1	1	4.6	0.94	99
1kdv	$P6_2$	1.10	372	Ca 1	—	—	—	—
1pwl	$P1$	1.09	317	Br 1	4	7.2	0.93	99
2c9v	$P2_1$	1.08	310	Zn 2, Cu 2	1	6.8	0.95	97
1w8f	$P1$	1.07	481	Ca 8	21	7.9	0.92	99
1q6z	$I222$	1.05	528	Mg 1, Ca 3	—	—	—	—
2anv	$C2$	1.04	293	Sm 3, I 6	1	7.8	0.94	99
1mnz	$I222$	1.01	389	Mg 1, Ca 1	—	—	—	—
1uzv	$P2_1$	1.00	468	Ca 8	8	7.7	0.96	99
1ea7	$P2_1$	0.96	315	Ca 6, S 1	1	8.7	0.97	94
2bw4	$P2_13$	0.91	340	Cu 2	1	9.9	0.98	99
1ix9	$P2_1$	0.91	410	Mn 4	1	8.9	0.97	99
1gwe	$P4_22_12$	0.88	498	Fe 1	1	11.0	0.90	99
1pjx	$P2_12_12_1$	0.87	314	Ca 2	2	8.4	0.98	89
1us0	$P2_1$	0.68	314	Br 1	1	19.2	0.97	99

ARP/*wARP* was applied to 18 of the 26 solved structures: 17 of them showed a sequence coverage larger than 0.88 and the other one (1a0m) showed a coverage of 0.46 with a CC_p value of 0.94. We did not apply *ARP*/*wARP* to eight test structures for the following reasons. Five of them (1hhu, 1hhz, 1hhy, 1sho and 1aa5) are classified in the PDB file as protein-type antibiotics. A sixth test structure, the only one without a PDB code, is also an antibiotic. The remaining two (1ick and 352d) are classified in the PDB files as DNA-type molecules.

Unlike in Table 1, the solution trial is no longer the first one: in these cases the largest peaks in the SMF do not correspond to heavy-atom positions. In one case (9pti) it was necessary to explore 64 trials to find the correct solution.

5.3. Subset 3

This subset of 24 test structures aims at establishing whether complex (up to 750 residues in the asymmetric unit) protein structures may be solved when diffraction data at atomic resolution are available and heavy atoms equal to or heavier than Ca are present in the unit cell.

21 test structures were solved by default, and the procedure provides high-quality final electron-density maps (the minimum value of CC_P was 0.83, which was obtained for 1het). The maps were automatically interpreted by *ARP*/*wARP* with high coverage values. The trial order of the correct solution is frequently smaller or equal to 3: only in a few cases is it larger.

Three test structures (1mmz, 1q6z and 1kdv) remained unsolved: in all three cases the heavy atoms have the minimum allowed weight (that of Ca) and the SMF does not provide peaks related to the heavy-atom positions. However, it is worthwhile noting that the presence in the unit cell of atoms that are not as heavy, such as Ca or Mn or Fe for example, is able to lead to the solution of even complex protein structures, provided that their data have atomic resolution.

Table 4
Crystallographic data and phasing results for the test structures with 1.25 Å < RES ≤ 1.5 Å, NASYM ≤ 2000 and at least one atomic species heavier than S.

For the column headings, see Table 1. An asterisk in the Resid column indicates that the structures are of DNA/RNA type: in these cases the number of non-H atoms in the asymmetric unit is given instead of the number of residues and *ARP*/*wARP* is not applied (n.a. in the COV column). Dashes in the Trial, fFOM2, CC_P and COV indicate that the correct solution has not been found.

PDB	S.G.	RES	Resid	HA	Trial	fFOM2	CC_P	COV
1l0z	$P2_12_12_1$	1.50	241	Xe 1, Br 32	45	4.3	0.85	99
1l1g	$P2_12_12_1$	1.50	240	Xe 1, Br 1	13	4.6	0.87	99
1w1d	$P21$	1.50	145	Au 1	1	3.4	0.80	99
2i7o	$I222$	1.50	129	Re 1, Cu 1	1	5.9	0.92	99
1jes	$P2_1$	1.49	486*	Cu 2	2	1.8	0.42	n.a.
1eao	$C2$	1.45	248	Br 4	—	—	—	—
1u0x	$C2$	1.45	184	Xe 2, Fe 1	1	—	—	—
1plc	$P2_12_12_1$	1.43	100	Cu 1	24	4.1	0.86	98
1dxc	$P6$	1.41	154	Fe 1	1	5.4	0.94	99
1dxd	$P6$	1.40	154	Fe 1	1	—	—	—
1awd	$I432$	1.40	94	Fe 2	1	6.8	0.96	98
3ebx	$P2_12_12_1$	1.40	62	S 8	1	5.0	0.87	98
1j6s*P*	$P42_12$	1.40	648*	Br 4, Ba 8	2	4.3	0.91	n.a.
1bx8	$P4_32_12$	1.38	55	S 11	1R2	6.1	0.92	86
2bpu	$P4_32_12$	1.35	130	Ho 3	1	5.7	0.93	99
1fs3	$P3_221$	1.35	124	S 12	69	5.5	0.94	99
1m1f	$P21$	1.33	212	S 8	—	—	—	—
193l	$P4_32_12$	1.33	130	Cl 1, S 10	—	—	—	—
1rwa	$P2_1$	1.31	757	Hg 3	1	5.5	0.93	97
1aac	$P2_1$	1.30	106	Cu 1	2	5.7	0.93	98
362d	$P2_12_12_1$	1.30	293*	Co 2	13	4.9	0.92	n.a.
1ikj	$C2$	1.28	184	Fe 1	1	—	—	—
1m77	$P4_32_12$	1.28	231*	Co 1	1	5.6	0.93	n.a.

5.4. Subset 4

The 22 test structures have data at quasi-atomic resolution (≤1.5 Å), a structural complexity of up to 2000 non-H atoms in the asymmetric unit and at least one atomic species heavier than S. The test aims at verifying whether it is possible to solve protein structures at quasi-atomic resolution even when the heavy atoms are not particularly heavy. 16 structures were solved with good final electron-density maps (CC_P greater than 0.80): 12 of them were well interpreted by *ARP*/*wARP* with COV > 0.85. *ARP*/*wARP* was not applied to four of them (362d, 1jes, 1m77 and 1j6s) because they were DNA/RNA-type molecules.

It is worthwhile mentioning the correct solutions of 3ebx and 1fs3, which have S as the heavy atom and data resolution equal to 1.40 and 1.35 Å, respectively. The solution of 3ebx was easy (it was attained in the first trial) because of the relatively small number of residues (62) in the asymmetric unit. The solution of 1fs3 was much more difficult (69 trials were explored to find the correct solution) because of the relatively larger number of residues (124) in the asymmetric unit.

Six structures remained unsolved (193l, 1m1f, 1dxd, 1u0x, 1ikj and 1eao). For 193l and 1m1f the failure may be ascribed to the low atomic numbers of the heavy atoms (16 and 17) and to the comparatively large number of residues (130 and 212) in the asymmetric unit: accordingly, no useful peak was found in the SMF map. No useful peaks in SMF were also found for 1eao: the probable reason is the partial site occupancy (~0.90) and the relatively higher vibrational parameter of Br atoms (greater than 20 Å^2, compared with an average $B_{\rm iso}$ of 13 Å^2 for the full structure). Approximate heavy-atom positions were found for the remaining three unsolved structures (1dxd, 1u0x and 1ikj) in spite of their reduced site-occupancy factors: the low occupancy, however, reduces the scattering power of the heavy-atom substructures. Consequently, their phases, on average, differ from the corresponding target phases by more than 80°, and the phase-refinement procedure was not able to reduce such a large error.

5.5. Subset 5

The applications in Table 5 aim at checking how feasible the *ab initio* automatic solution of protein structures is at non-atomic resolution. Of the 29 test structures with RES > 1.50 and heavy atoms heavier than Ca, 23 were automatically solved: they include one of the three structures (3ajw) with a data resolution equal to or worse than 2 Å. 20 of the above 23 structures are fully interpreted by *ARP*/*wARP* with a coverage larger than 0.92; for one of them (1ytt) COV = 0.80. For two of the above 23 structures (1n0y and 1naq) *ARP*/*wARP* does not provide a solution in spite of the high value of CC$_{\rm P}$ (0.92 and 0.79, respectively). 1iha, a RNA-type molecule, was not submitted to *ARP*/*wARP*.

The six unsolved structures (1crm, 2f14, 1arm, 1r0h, 1z1y and 1h87) may be partitioned into two subsets.

(i) A subset (2f14 and 1arm) in which the heavy-atom substructure has been perfectly defined but the location of the light atoms remains less accurate. For both such structures CC$_{\rm P}$ = 0.63: our phase-refinement algorithms were unable to improve the phase quality sufficiently to allow successful application of *ARP*/*wARP*.

(ii) A subset in which the heavy-atom substructure is correctly defined but the distribution of the light atoms in the unit cell is unreliable (1crm, 1r0h, 1z1y and 1h87). Such structures are characterized by CC$_{\rm P}$ values smaller than 0.45, *i.e.* 0.43, 0.31, 0.31 and 0.31, respectively. The reader should consider that a strong contribution to CC$_{\rm P}$ arises from the correct location of the heavy atoms. If the experimental CC$_{\rm P}$ values between our electron densities and the published electron densities are deprived of the heavy-atom contribution they decrease to 0.33, 0.23, 0.19 and 0.19, respectively. The heavy-atom phases are therefore quite far away from the full structure phase values, and the phase-refinement step was unable to improve them.

A few practical considerations are now necessary to enlighten the role of fFOM2 and of *RELAX*, and also to predict the CPU time necessary for solving the structures.

Table 5
Crystallographic data and phasing results for the test structures with RES > 1.5 Å, NASYM ≤ 8000 and at least one atomic species heavier than Ca.

For the column headings, see Table 1. An asterisk in the Resid column indicates that the structure is of DNA/RNA type: in these cases the number of non-H atoms in the asymmetric unit is given instead of the number of residues and *ARP*/*wARP* is not applied (n.a. in the COV column). A dash in the COV column indicates that the correct solution has not been found.

PDB	S.G.	RES	Resid	HA	Trial	fFOM2	CC$_{\rm P}$	COV
3ajw	$P6_522$	2.10	135	Hg 1	1	2.8	0.74	99
1crm	$P2_12_12_1$	2.02	260	Hg 4	1	1.9	0.43	0
1z1y	$P2_1$	2.00	316	Yb 9	2	2.5	0.31	—
1buu	$P2_13$	1.93	149	Ho 1	1	5.7	0.93	99
1yfd	$P2_12_12_1$	1.90	694	Hg 13, Fe 4	1	3.5	0.75	99
1jpr	$P2_12_12_1$	1.89	695	Hg 14, Mn 4	1	3.7	0.79	99
1naq	$P2_12_12_1$	1.81	636	Hg 18	1	2.5	0.79	0
1arm	$P2_1$	1.80	312	Hg 4, Cu 1	1	2.5	0.63	0
1pm2	$P2_12_12_1$	1.80	692	Hg 14, Mn 4	1	3.8	0.80	99
1ytt	$P2_12_12_1$	1.80	227	Yb 4	1	4.4	0.78	82
1r0h	$R3$	1.78	53	Co 1	4	2.4	0.31	—
1n0y	$C2$	1.75	168	Pb 14	1	6.0	0.92	59
1h87	$P4_32_12$	1.72	129	Gd 2	1	2.6	0.31	—
2f14	$P2_1$	1.71	258	Hg 1	1	2.5	0.63	0
2yzw	$P2_12_12_1$	1.70	283	Gd 2	1	3.9	0.77	99
1g0e	$P2_1$	1.70	259	Hg 1	1	3.6	0.83	99
1ccr	$P6_1$	1.69	112	Fe 1	1	4.4	0.83	95
1e3u	$P2_1$	1.65	978	Au 8	1	4.4	0.88	99
2p09	$P3_221$	1.65	70	Zn 1	2	6.1	0.96	98
1a70	$P2_12_12_1$	1.62	97	Fe 2	27	4.4	0.87	98
1jqc	$P2_12_12_1$	1.61	694	Hg 13, Mn 4	1	4.4	0.86	99
1r0f	$R3$	1.61	54	Ga 1	1	4.9	0.85	98
1r0g	$R3$	1.61	54	Hg 1	1	5.1	0.89	98
1iha	$C2$	1.60	414*	Rh 2, Br 2	1	4.2	0.88	n.a.
2fj9	$P6_522$	1.56	86	Pb 1, Zn 1	1	3.0	0.81	98
1paz	$P6_5$	1.56	121	Cu 1	1	5.2	0.91	99
1ix2	$I23$	1.55	205	Se 10	12	5.1	0.90	93
2fp1	$P2_1$	1.53	329	Pb 2	1	4.2	0.86	97
1r0i	$R3$	1.53	53	Cd 1	3	5.1	0.85	98

Tables 1–5 suggest that fFOM2 depends on the average phase error (as expected for any useful figure of merit) and on the data resolution. Indeed, the average fFOM2 value is larger for the correct solutions quoted in Tables 1–3 than for those quoted in Tables 4–5; 4.0 is the minimum fFOM2 value for Tables 1–3, while 3.4 and 2.5 are the minima for Tables 4 and 5, respectively.

Tables 1–5 also confirm that simply fixing a threshold for recognizing the correct solution would often lead to incorrect choices. For example, let us fix the threshold TRFOM2 to 3.0 for the structures in Table 5. Four good trial solutions would then be lost for 2f14, 1arm, 1naq and 3ajw; conversely, if we fix TRFOM2 at 2.5 then the program may stop at a trial not corresponding to the correct solution. For example, in the case of 1a70 (see Table 5) the program would stop at trial 5 instead of the correct trial 27, with fFOM2 = 2.8 and CC$_{\rm P}$ = 0.01.

In conclusion, for TRFOM2 values that are too high the program will not recognize the correct solutions, whereas for values that are too small the program stops at false solutions.

The above considerations have strong consequences not only for the ratio of successes to failures but also for the total CPU time T necessary to obtain the correct molecular model. Let us subdivide T into two parts: $t_{\rm S}$ denotes the time necessary to obtain the correct solution and $t_{\rm A}$ is the CPU time globally used for the *ARP*/*wARP* applications. $t_{\rm S}$ is expected

Table 6
For each of Tables 1–5 we show the average number of residues (⟨Resid⟩), the average CPU time (min) needed to solve the structures (⟨$t_{\rm S}$⟩) and the average CPU time (min) employed by *ARP*/*wARP* to build a satisfactory model (⟨$t_{\rm A}$⟩).

For each table the data refer to solved structures for which *ARP*/*wARP* obtained satisfactory coverage (COV > 75%)

Table	⟨Resid⟩	⟨$t_{\rm S}$⟩	⟨$t_{\rm A}$⟩
1	129	2	48
2	107	19	120
3	440	22	314
4	150	113	176
5	299	48	80

Table 7
For some structures solved *via* the *RELAX* procedure we give the PDB code, the number of the table to which the structure belongs (Table No.), the average phase error before (⟨$|\Delta\varphi_1|$⟩) and after (⟨$|\Delta\varphi_2|$⟩) the application of *RELAX* and the origin shift automatically applied.

PDB code	Table No.	⟨$|\Delta\varphi_1|$⟩ → ⟨$|\Delta\varphi_2|$⟩	Origin shift
1fy2	1	43° → 33°	0.0, 0.0, 0.0
1a6n	1	37° → 31°	0.0, 0.0, 0.0
2knt	2	36° → 31°	0.0, 0.0, 0.0
1boy	1	85° → 22°	0.50, 0.53, 0.0
1cku	1	86° → 22°	0.46, 0.50, 0.0
1hhu	2	89° → 14°	0.01, 0.0, 0.25
1cnr	2	89° → 23°	0.43, 0.0, 0.34
1cex	2	88° → 27°	0.47, 0.0, 0.02
1bx7	2	88° → 25°	0.50, 0.50, 0.48
1su7	3	88° → 29°	0.51, 0.0, 0.0
1suf	3	89° → 33°	0.48, 0.0, 0.49

to depend on RES, on the structural complexity and on the heavy-atom species and occupancy; *a posteriori* it will also depend on the order of the trial corresponding to the correct solution and on the quality of the electron-density map (an intermediate-quality map may require iterated applications of *ARP*/*wARP* to attain sufficiently large coverage). The average experimental values of $t_{\rm S}$ and of $t_{\rm A}$ (⟨$t_{\rm S}$⟩ and ⟨$t_{\rm A}$⟩) calculated for each of Tables 1–5 are reported in Table 6.

For the structures in Table 1 ⟨$t_{\rm S}$⟩ = 2 min: the very short CPU time is mainly owing to the fact that the correct solution is obtained in the first trial. The CPU spread is also small: it varies from a minimum of 0.5 min for 2bf9 to a maximum of 6 min for 1fy2. ⟨$t_{\rm A}$⟩ is also relatively small owing to the high quality of the electron-density maps submitted to *ARP*/*wARP* (⟨$t_{\rm A}$⟩ = 48 min). In this case, as in all of the cases described below, each application of *ARP*/*wARP* includes two runs: one per enantiomorph. However, a new version of *ARP*/*wARP* will probably be soon available which will check the correct enantiomorph at the beginning of the AMB step, thus approximately dividing the $t_{\rm A}$ times reported here by two (Victor Lamzin & Tim Wiegels, personal communication).

⟨$t_{\rm S}$⟩ and ⟨$t_{\rm A}$⟩ are larger for Tables 2–5, mostly because of the larger numbers of trials: obviously, an important role is also played by the size of the structure. The spread of the number of trials is mainly responsible for the $t_{\rm S}$ spread. For example, in Tables 4 and 5 the minimum $t_{\rm S}$ values correspond to 3ebx and to 1paz, respectively (1 and 2 min), both of which were solved in trial 1; the maximum $t_{\rm S}$ values correspond to 1fs3 and 1ix2, respectively (662 and 283 min), where the former was solved in trial 69 and the latter in trial 12.

Owing to the algorithm deciding how many times *ARP*/*wARP* is applied, the number of trials will severely influence $t_{\rm A}$. For example, *ARP*/*wARP* is applied 15 × 2 times to 1fs3 to allow the automatic 0.99 sequence coverage at trial 69.

We now discuss the role of *RELAX* in phasing methods based on Patterson techniques.

In a multisolution direct-methods procedure applied to a crystal structure with heavy atoms, the probability of correctly locating the heavy atoms does not depend on the trial order: any random starting set has the same probability of finding the heavy-atom positions and therefore of solving the structure. In a Patterson-based phasing technique, the highest probability corresponds to the trial exploiting the correct peak in the SMF map. If the correct SMF peak has been employed, *RELAX* may have a more limited use. Two cases may be considered.

Case 1. The heavy atoms have been correctly positioned but the subsequent phase extension and refinement steps are not able to recover satisfactory phase values for the full structure. The *RELAX* procedure, *via* phase expansion in $P1$ and subsequent EDM refinement, may lead to a better set of phases which, when returned to the correct space group, give rise to a higher quality electron-density map. In this case *RELAX* does not shift the origin; it only improves the phases *via* phase refinement in $P1$.

Case 2. The pivot peaks in the SMF map used in the superposition techniques (see equation 3) do not correspond or only approximately correspond to heavy-atom positions. The subsequent phase refinement is not able to modify the situation: the task is sometimes accomplished by the *RELAX* procedure, which shifts the map by a non-allowed origin translation (Hauptman & Karle, 1956; Giacovazzo, 1974).

In our procedure *RELAX* is applied at the end of the phase refinement, when fFOM2 < 4.5 (it is supposed that model phases are correct when fFOM2 > 4.5). In Table 7 some examples are shown: in the columns specifying the trial orders, the additional script R1 or R2 indicate that the solution has been obtained by using *RELAX* according to case 1 or to case 2. 1fy2, 1a6n and 2knt belong to case 1: they have average phase errors of 43, 37 and 36°, respectively, before application of *RELAX* and corresponding average phase errors of 33, 31 and 31° after *RELAX*. The rest of the structures in Table 7 belong to case 2. Sometime a large origin shift is necessary to bring the heavy atoms (and of course, the full electron-density map) into the correct position (see 1hhu and 1cnr). This case may occur when the pivot peak used for the superposition techniques does not coincide with the position of the heaviest atom: the *RELAX* origin shift then restates the correct electron density. Sometime the origin shift differs from an allowed origin by a distance of the order of one or a few angstroms: in this case the heavy-atom positions are slightly misplaced from the correct positions.

6. Conclusions

The combined use of *VLD* and Patterson-based algorithms was able to solve most of the 126 test structures *ab initio*. With respect to our previous results, the package here described shows a greater efficiency: in particular, it is able to refine phases to a quality level that permits automatic model building. That is mostly owing to the combined use of Patterson deconvolution techniques, the *C* map and the *VLD* approach.

The most favourable situation occurs when sufficiently heavy atoms, atomic resolution data and limited structural complexity coexist. In this case useful pivot peaks are easily recognized and high-quality electron-density maps are immediately obtained. Crystal structure solution becomes more difficult when the data resolution is not atomic and/or when the heavy atoms are not very heavy: in this case the order of the trial solution is usually higher, the method is more time-consuming and sometimes the structure is not solved.

A post-mortem analysis of the failures suggests that they are mostly owing to data resolution rather than to structural complexity and to too small a value of the heavy-atom scattering power/total scattering power ratio. It may also be useful to reiterate that this ratio also depends on the thermal factor of the heavy atom (with respect to the rest of the structure) and on the heavy-atom occupancy factor.

The applications clearly show that crystal structure solution may be automatically attained even for structures with data resolution close to 2 Å. This requires figures of merit and/or ancillary algorithms that are able to recognize the correct solution and to stop the program when it has been found. The phasing procedure is still not very robust when the data resolution is close to 2 Å and will hopefully be improved in the near future, but now opens new perspectives for *ab initio* crystal structure solution of proteins. Indeed, it is a good result that structures at this resolution are solved at all.

The last remark concerns the overall CPU time necessary for *ab initio* automatic solution and model building (even if it is not comparable with the time necessary to obtain good crystals and experimental diffraction data). At least for the classes of structures tested in this paper, the crystal structure solution is not very time-demanding: indeed, the model-building time is by far the most time-consuming section. Progress in this area would allow large advances for any *ab initio* technique.

The kind advice of Victor Lamzin and Tim Wiegels for the automatic use of *ARP/wARP* is gratefully acknowledged.

References

Buerger, M. J. (1948). *Phys. Rev.* **73**, 927–928.
Buerger, M. J. (1959). *Vector Space and its Application in Crystal Structure Investigation*. New York: Wiley.
Burgess, D. D. & Tervo, R. J. (1983). *Nucl. Instrum. Methods Phys. Res.* **214**, 431–434.
Burla, M. C., Caliandro, R., Camalli, M., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C., Polidori, G. & Spagna, R. (2005). *J. Appl. Cryst.* **38**, 381–388.
Burla, M. C., Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C., Polidori, G. & Siliqi, D. (2006*a*). *J. Appl. Cryst.* **39**, 527–535.
Burla, M. C., Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C., Polidori, G. & Siliqi, D. (2006*b*). *J. Appl. Cryst.* **39**, 728–734.
Burla, M. C., Caliandro, R., Giacovazzo, C. & Polidori, G. (2010). *Acta Cryst.* A**66**, 347–361.
Burla, M. C., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Polidori, G. (2003). *Acta Cryst.* A**59**, 245–249.
Burla, M. C., Carrozzini, B., Cascarano, G. L., Giacovazzo, C. & Polidori, G. (2000). *J. Appl. Cryst.* **33**, 307–311.
Burla, M. C., Carrozzini, B., Cascarano, G. L., Giacovazzo, C. & Polidori, G. (2002). *Z. Kristallogr.* **217**, 629–635.
Burla, M. C., Carrozzini, B., Cascarano, G. L., Giacovazzo, C. & Polidori, G. (2012). *J. Appl. Cryst.* **45**, 1287–1294.
Burla, M. C., Giacovazzo, C. & Polidori, G. (2010). *J. Appl. Cryst.* **43**, 825–836.
Burla, M. C., Giacovazzo, C. & Polidori, G. (2011). *J. Appl. Cryst.* **44**, 193–199.
Burla, M. C., Giacovazzo, C. & Polidori, G. (2013). *J. Appl. Cryst.* **46**, 1592–1602.
Caliandro, R., Carrozzini, B., Cascarano, G. L., Comunale, G. & Giacovazzo, C. (2013). *Acta Cryst.* A**69**, 98–107.
Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C., Mazzone, A. & Siliqi, D. (2008). *J. Appl. Cryst.* **41**, 548–553.
Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Siliqi, D. (2005*a*). *Acta Cryst.* D**61**, 556–565.
Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Siliqi, D. (2005*b*). *Acta Cryst.* D**61**, 1080–1087.
Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Siliqi, D. (2007*a*). *J. Appl. Cryst.* **40**, 883–890.
Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C. & Siliqi, D. (2007*b*). *J. Appl. Cryst.* **40**, 931–937.
Caliandro, R., Di Profio, G. & Nicolotti, O. (2013). *J. Pharm. Biomed. Anal.* **78–79**, 269–279.
Carrozzini, B., Cascarano, G. L., Comunale, G., Giacovazzo, C. & Mazzone, A. (2013). *Acta Cryst.* D**69**, 1038–1044.
Carrozzini, B., Cascarano, G. L. & Giacovazzo, C. (2010). *J. Appl. Cryst.* **43**, 221–226.
Foadi, J., Woolfson, M. M., Dodson, E. J., Wilson, K. S., Jia-xing, Y. & Chao-de, Z. (2000). *Acta Cryst.* D**56**, 1137–1147.
Frazão, C., Sieker, L., Sheldrick, G. M., Lamzin, V., LeGall, J. & Carrondo, M. A. (1999). *J. Biol. Inorg. Chem.* **4**, 162–165.
Giacovazzo, C. (1974). *Acta Cryst.* A**30**, 390–395.
Harker, D. (1936). *J. Chem. Phys.* **4**, 381.
Hauptman, H. & Karle, J. (1956). *Acta Cryst.* **9**, 45–55.
Leslie, A. G. W. (1987). *Acta Cryst.* A**43**, 134–136.
Loll, P. J., Miller, R., Weeks, C. M. & Axelsen, P. H. (1998). *Chem. Biol.* **5**, 293–298.
Mooers, B. H. M. & Matthews, B. W. (2006). *Acta Cryst.* D**62**, 165–176.
Morháč, M. (2009). *Nucl. Instrum. Methods Phys. Res. A*, **600**, 478–487.
Morháč, M., Kliman, J., Matoušek, V., Veselský, M. & Turzo, I. (1997). *Nucl. Instrum. Methods Phys. Res. A*, **401**, 113–132.
Morháč, M. & Matoušek, V. (2008). *Appl. Spectrosc.* **62**, 91–106.
Nordman, C. E. (1966). *Trans. Am. Crystallog. Assoc.* **2**, 29–38.
Pavelčík, F., Kuchta, L. & Sivý, J. (1992). *Acta Cryst.* A**48**, 791–796.
Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.
Rappleye, J., Innus, M., Weeks, C. M. & Miller, R. (2002). *J. Appl. Cryst.* **35**, 374–376.
Read, R. J. (1986). *Acta Cryst.* A**42**, 140–149.
Refaat, L. S. & Woolfson, M. M. (1993). *Acta Cryst.* D**49**, 367–371.
Richardson, J. W. & Jacobson, R. A. (1987). *Patterson and Pattersons*, edited by J. P. Glusker, B. K. Patterson & M. Rossi, pp. 310–317. Oxford University Press.

Rodríguez, D. D., Grosse, C., Himmel, S., González, C., de Ilarduya, I. M., Becker, S., Sheldrick, G. M. & Usón, I. (2009). *Nature Methods*, **6**, 651–653.

Rodríguez, D., Sammito, M., Meindl, K., de Ilarduya, I. M., Potratz, M., Sheldrick, G. M. & Usón, I. (2012). *Acta Cryst.* D**68**, 336–343.

Ryan, C. G., Clayton, E., Griffin, W. L., Sie, S. H. & Cousens, D. R. (1988). *Nucl. Instrum. Methods Phys. Res. B*, **34**, 396–402.

Sheldrick, G. M. (1992). *Crystallographic Computing 5*, edited by D. Moras, A. D. Podjarny & J.-C. Thierry, pp. 145–157. Oxford University Press.

Sheldrick, G. M. (2008). *Acta Cryst.* A**64**, 112–122.

Simpson, P. G., Dobrott, R. D. & Lipscomb, W. N. (1965). *Acta Cryst.* **18**, 169–179.

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

Weeks, C. M., DeTitta, G. T., Hauptman, H. A., Thuman, P. & Miller, R. (1994). *Acta Cryst.* A**50**, 210–220.

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