Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Automatic α-helix identification in Patterson maps

Rocco Caliandro,[a]* Domenica Dibenedetto,[b] Giovanni Luca Cascarano,[a] Annamaria Mazzone[a] and Giovanni Nico[c]

[a]Institute of Crystallography, CNR, via Amendola 122/o, 70126 Bari, Italy, [b]Research School for Simulation Sciences, University of Aachen, Aachen, Germany, and [c]Istituto per le Applicazione del Calcolo 'Mauro Picone', CNR, via Amendola 122/o, 70126 Bari, Italy

Correspondence e-mail:
rocco.caliandro@ic.cnr.it

Received 26 July 2011
Accepted 2 November 2011

α-Helices are peculiar atomic arrangements characterizing protein structures. Their occurrence can be used within crystallographic methods as minimal *a priori* information to drive the phasing process towards solution. Recently, brute-force methods have been developed which search for all possible positions of α-helices in the crystal cell by molecular replacement and explore all of them systematically. Knowing the α-helix orientations in advance would be a great advantage for this kind of approach. For this purpose, a fully automatic procedure to find α-helix orientations within the Patterson map has been developed. The method is based on Fourier techniques specifically addressed to the identification of helical shapes and operating on Patterson maps described in spherical coordinates. It supplies a list of candidate orientations, which are then refined by using a figure of merit based on a rotation function calculated for a template polyalanine helix oriented along the current direction. The orientation search algorithm has been optimized to work at 3 Å resolution, while the candidates are refined against all measured reflections. The procedure has been applied to a large number of protein test structures, showing an overall efficiency of 77% in finding α-helix orientations, which decreases to 48% on limiting the number of candidate solutions (to 13 on average). The information obtained may be used in many aspects in the framework of molecular-replacement phasing, as well as to constrain the generation of models in computational modelling programs. The procedure will be accessible through the next release of *IL MILIONE* and could be decisive in the solution of new unknown structures.

1. Introduction

Modern proteomic and structure-based drug-discovery projects demand fast and efficient platforms for high-throughput crystallography (Blundell *et al.*, 2002). In this respect, automated crystal structure-determination pipelines have been organized which consist of a number of crystallographic software programs executed by several decision-makers (Panjikar *et al.*, 2005; Keegan & Winn, 2007; Long *et al.*, 2008). The most frequent phasing method used to achieve structure solution is molecular replacement (MR), which is typically preceded by a step in which suitable protein models are searched for in the Protein Data Bank (PDB) or generated by *ab initio* or comparative modelling programs (Rigden *et al.*, 2008; Caliandro *et al.*, 2009). Difficult cases, for which a homologous model close enough to the target structure is not available, escape from this scheme. In these cases, limited experimental resolution and/or a lack of heavy atoms in the crystal often also prevents solution by *ab initio* or SAD/MAD

doi:10.1107/S0907444911046282

phasing approaches. To handle such challenging cases, phasing procedures have been developed which make use of minimal *a priori* knowledge, consisting of limited well conserved domains, heavy atoms (when present) or ideal α-helix polyalanine fragments (Dodson & Woolfson, 2009; Bernstein *et al.*, 1977). Here, the bottleneck is constituted by the MR search, which has to cope with the difficult task of finding the correct position in the unit cell of a very small fraction of the total scattering matter. As a result, a double drawback occurs: on one hand, a long list of possible solutions is returned by the MR rotation and translation steps; on the other, they all have similar figures of merit, so they cannot be properly ranked. Recently, a procedure for *ab initio* phasing at 2 Å resolution has been proposed based on a brute-force approach (Rodríguez *et al.*, 2009). It has the MR location of model fragments such as small α-helices as a first step and phase refinement and extension by electron-density modification and model-building cycles as a second step. The drawback of this approach is that a large number of equally probable candidate solutions are obtained after the first step and all of them are submitted to the second step, which is typically much slower than the first step.

This problem triggered the idea underlying the present paper: to collect the maximum amount of *a priori* information about conserved structural fragments from crystallographic data prior to phasing. In order to pursue this aim, the Patterson map appears to be the ideal source of information, since it is directly available from diffraction intensities *via* a transform and does not need phases to be determined. On the other hand, α-helices are the ideal structural fragments to consider for proteins, since they are present in 90% of all known protein structures and have very distinctive features. The concept of recognizing molecules with pseudo-helical symmetry in Patterson maps was pursued in the very first studies of DNA (Cochran *et al.*, 1952; Franklin & Gosling, 1953, 1955), RNA (Kim & Rich, 1969; Sakurai *et al.*, 1971) and protein structures (Magdoff *et al.*, 1956) and, more recently, by Kondo *et al.* (2008). Thumiger (2008) was the first to propose the use of the Fourier transform to highlight the repetition of the Patterson function related to the pitch of α-helices. He applied it to a single protein composed of three long parallel helices (PDB entry 1s2x). A proof of principle for an automatic application of the method was made by Mazzone *et al.* (2011), which identified α-helices in Patterson maps calculated from several protein structures deposited in the PDB. The present paper deals with real crystallographic data and demonstrates the possibility of identifying α-helices in Patterson maps of large numbers of proteins.

The paper is organized as follows: the procedure is described in §2, together with the criteria used to check the results. The applications are described in §3, in which the results obtained from three tests, representing an easy, an intermediate and a difficult case, are reported in detail and the averaged efficiencies resulting from extensive tests are given. Possible applications of the information obtained by the procedure are outlined in §4.

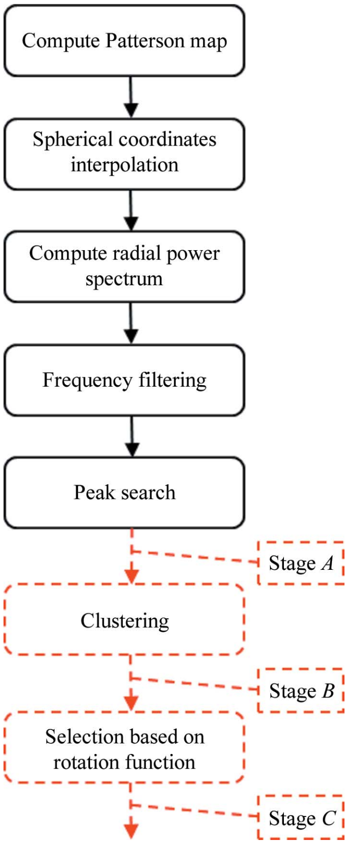
Compute Patterson map
Spherical coordinates interpolation
Compute radial power spectrum
Frequency filtering
Peak search
Stage *A*
Clustering
Stage *B*
Selection based on rotation function
Stage *C*

Figure 1
Scheme of the α-helix identification procedure. Boxes indicating operations belonging to step 1 (in unbroken black lines) and step 2 (in dashed red lines) are drawn. Three selection stages may be identified in step 2 and are denoted by the letters *A*, *B* and *C*.

2. Methods

Our approach for identifying α-helices in the Patterson map of proteins consists of two steps.

(i) Finding the α-helix candidate orientations within the Patterson map.

(ii) Selecting the best candidate orientations. Three selection stages may be identified in this step.

In addition, procedures for automatic checking of the results of the method have been developed, which include determination of the axis orientation of the helices of the target protein from its known atomic coordinates and comparison of this orientation with those found by the procedure. The algorithm is shown schematically in Fig. 1 and the single steps are described in detail below.

2.1. Finding the α-helix candidate orientations

As is well known, the peaks in the Patterson map correspond to interatomic vectors of the crystal structure that it refers to. The map is defined as the Fourier transform of the diffracted intensities,

$$P(\mathbf{u}) = \mathrm{FT}[|F(\mathbf{h})|^2] = \mathrm{FT}\left\{\sum_{i<j}^{N} f_i f_j \exp[-2\pi i \mathbf{h}(\mathbf{r}_i - \mathbf{r}_j)]\right\}, \quad (1)$$

where $P(\mathbf{u})$ is the Patterson map, $|F(\mathbf{h})|$ are the measured structure-factor amplitudes for reflection $\mathbf{h}$, f_i and $\mathbf{r}_i$ are the scattering factor and position, respectively, of the ith atom and N is the number of atoms in the crystal cell. Although this interpretation is strictly valid only at atomic data resolution, the Patterson map retains information about interatomic patterns even at lower resolution. If α-helices are considered, they have two distinctive features that can be used for their recognition: their directionality (their atoms are arranged preferentially along their axis) and periodicity (they roughly repeat themselves with a given pitch). The former suggests the use of spherical coordinates, where the radial coordinate can be used to monitor the data variability; the latter suggests the use of Fourier filtering, which, operating on the radial coordinate, can highlight the features occurring at a given frequency. The characteristics of the signal sought are as follows. Each turn of an α-helix contains 3.6 residues. If the side-chain contribution is neglected and the backbone atoms are fitted by an idealized helix, this helix has a pitch of p_α = 5.4 Å. The radial profile of an α-helix along the direction of its axis resulting from a Patterson map calculated at 3 Å resolution is shown in Fig. 2. It was obtained by considering a 15-residue α-helix extracted from the structure with PDB code 3kut (with $P1$ symmetry), rotating it around the x axis (with length a = 26.3 Å) and calculating the structure factors for reflections to 3 Å resolution. Besides the expected periodicity of the Patterson peaks, a decrease in their intensity is shown because the number of interatomic peaks of given length decreases as the length increases. In Fig. 2 the difference between a polyalanine and an all-atom helix profile may also be appreciated, indicating that the noise introduced by the side-chain variability along the helix marginally affects the signal. A Fourier transform may be applied to the Patterson radial profile along a given direction $\hat{\mathbf{n}}$, according to the formula

$$G_{\hat{\mathbf{n}}}(\nu) = \mathrm{FT}[P_{\hat{\mathbf{n}}}(r)] = \int\limits_{R_{\min}}^{R_{\max}} P_{\hat{\mathbf{n}}}(r)\exp(-2\pi i\nu r)\,\mathrm{d}r, \tag{2}$$

where $R_{\min}$ and $R_{\max}$ define the radial range considered for filtering and ν is the dual variable of r. As pointed out in Thumiger (2008), the power spectrum $|G_{\hat{\mathbf{n}}}(\nu)|^2$ is a useful quantity for the recognition of secondary-structure elements in the Patterson map, since it can highlight signals occurring at the characteristic frequency $1/p_\alpha$.

In view of these findings, a procedure to identify α-helix directions in the Patterson maps of proteins has been implemented consisting of the following steps.

(i) Normalized structure factors $E(\mathbf{h})$ are calculated according to the Wilson procedure (Wilson, 1942) and reflections with resolution lower than 3 Å are selected for further analysis.

(ii) The Patterson map $P(\mathbf{u})$ is calculated with coefficients $|E(\mathbf{h})|^2$ by using the discrete Fourier transform algorithm implemented in the *FFTW* routines (Frigo & Johnson, 2005). By default, this is calculated in the crystallographic (Cartesian) coordinate system. Let us denote $\mathbf{U}_\mathrm{C} = (u_x^c, u_y^c, u_z^c)$, the vector $\mathbf{u}$ in matrix notation.

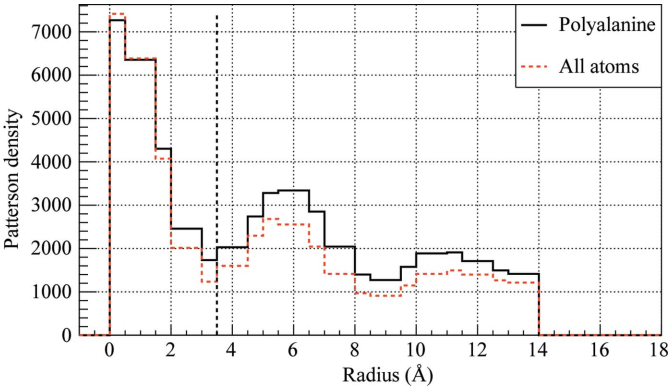

Figure 2
Radial profile of the Patterson map of a polyalanine (unbroken line) and all-atoms (dashed line) α-helix of 15 residues (SPESLRSKVDEAVAV) in the direction coinciding with its axis. The dashed line indicates the minimum radius used for the Fourier analysis.

(iii) A grid is defined by using spherical coordinates (r, θ, φ) with zenith direction z and azimuth axis x. The variable $t = \cos\theta$ is introduced in place of the inclination angle θ to ensure a uniform sampling along the z axis. The ranges used are

$$r \in (R_{\min}, R_{\max}), \quad t \in (-1, 1), \quad \varphi \in (-\pi, \pi), \tag{3}$$

which correspond to a semi-sphere including the independent part of the Patterson map (restrictions introduced by the space-group symmetry are accounted for in a later stage of the procedure). The value of $R_{\min}$ is chosen to be 3.5 Å to avoid contributions from the Patterson origin peak (see the dashed line in Fig. 2), while $R_{\max}$ is related to the lengths a, b, c of the unit-cell axes by the equation

$$R_{\max} = \max(a, b, c)/2, \tag{4}$$

since the information from the Patterson map is all contained in half the cell. The spacing is chosen as 0.5 Å for the radial axis and 0.05 and 3° for the variables t and φ, respectively.

(iv) The Patterson map is converted into spherical coordinates according to the equation

$$P(\mathbf{U}_\mathrm{S}) = \mathrm{Interp}\{P[\mathbf{A}^{-1}g^{-1}(\mathbf{U}_\mathrm{S})]\}, \tag{5}$$

where $\mathbf{U}_\mathrm{S} = (u_r, u_t, u_\varphi)$ is the vector $\mathbf{u}$ expressed in spherical coordinates, g is a function which allows conversion from Cartesian to spherical coordinates in an orthonormal frame, $\mathbf{U}_\mathrm{S} = g(\mathbf{U}_\mathrm{CO})$,[1] $\mathbf{A}$ is a matrix that orthonormalizes the crystallographic coordinates, $\mathbf{U}_\mathrm{CO} = \mathbf{A}\mathbf{U}_\mathrm{C}$, and Interp indicates the operation of linear interpolation in a grid of 19 points centred on the point to be interpolated. During this operation, the density of the points falling outside the half-cell is set to zero in order to avoid sampling of the centrosymmetric part of the Patterson map. This is the reason for the truncation at 14 Å of the profile shown in Fig. 2.

[1] For our choice of variables, $g(u_x, u_y, u_z) = [(u_x^2, u_y^2, u_z^2)^{1/2}, u_z/(u_x^2, u_y^2, u_z^2)^{1/2}, \tan^{-1}(u_y/u_x)]$ and $g^{-1}(u_r, u_t, u_\varphi) = [u_r(1 - u_t^2)^{1/2}\cos(u_\varphi), u_r(1 - u_t^2)^{1/2} \times \sin(u_\varphi), u_r u_t]$, where $\mathbf{U}_\mathrm{CO} = (u_x, u_y, u_z)$.

(v) A power spectrum $|G_{\hat{\mathbf{n}}}(\nu)|^2 = \mathrm{FT}[P(u_r, u_t, u_\varphi)]$ is calculated for each direction $\hat{\mathbf{n}}$ defined by the pair (t, φ) within the ranges of (3).

(vi) The grid number k corresponding to the digital frequency $\overline{\nu} = k\{1/[n_{\mathrm{pad}}(R_{\max} - R_{\min})]\}$ closest to the characteristic frequency $1/p_\alpha$ of the helix is given by

$$k = \mathrm{Int}\left[\frac{n_{\mathrm{pad}}(R_{\max} - R_{\min})}{P_\alpha}\right], \tag{6}$$

where n_{pad} is a multiplicative factor introduced by applying the technique of zero-padding to the Fourier transform, with the aim of increasing the frequency sampling. It is set to a value of 4.

(vii) The power-spectrum map is modified by considering only the values in a slice centred on $\overline{\nu}$ defined by three sections at constant frequency and having intensities $I > \langle I \rangle + \sigma(I)$, where the mean $\langle I \rangle$ and the standard deviation $\sigma(I)$ values are calculated within the selected slice. This map will be denoted $|G_{\hat{\mathbf{n}}}(\overline{\nu})|^2$.

(viii) A peak search is performed on $|G_{\hat{\mathbf{n}}}(\overline{\nu})|^2$ within the angular intervals of (3). In the presence of crystallographic symmetry they are restricted as reported in Appendix *A*. The peak search is made by a 19-point interpolation within the selected sections.

As an example, Fig. 3 shows the power spectrum calculated from the radial profile of Fig. 2 for increasing values of the parameter n_{pad}. The peaks at $\overline{\nu}$ corresponding to the helix signal are highlighted by arrows. They are well separated from the peak at the origin for padding factors greater than 1. For $n_{\mathrm{pad}} = 4$ the peaks at $\overline{\nu}$ of the all-atoms (violet) and polyalanine (black) α-helix profiles coincide.

2.2. Selecting the candidate orientations

The peaks obtained at the end of the procedure described above represent the candidate α-helix orientations available at selection stage *A*. They are checked and improved using the methods described in this section.

As a first step, the list of peaks undergoes a hierarchical clustering performed using the group-average method, which has the effect of grouping similar orientations, thus reducing their numbers and improving their estimates. The distance matrix is calculated by referring to the intersections of the helix axes with a sphere of radius 1 Å, while the threshold that defines the clusters is chosen according to the method described in §2.3.2. The original list of peaks is replaced by a list of representative orientations calculated by averaging the orientations belonging to the same cluster. The height of the peaks, as calculated by the peak-search procedure, is used to weight the average and to assign a height to the representative orientations, which is chosen as the maximum among the heights of the in-cluster peaks. Both the distance matrix and the representative orientations are calculated by considering the symmetry operations of the Laue group to which the protein space group belongs. The representative orientations are sorted according to their heights and represent the set of solutions at selection stage *B*.

A further selection of the candidate solutions is achieved by means of a figure of merit developed in analogy to the rotation function used in MR programs. A polyalanine helix is used as a model consisting of N_{sample} atoms. It is placed in the same orientation as the candidate solution to be checked by multiplying its atomic coordinates in the orthonormal Cartesian frame $\mathbf{r}_j^{\mathrm{ort}}$, $j = (1, N_{\mathrm{sample}})$ by the rotation matrix $\mathbf{M}(z, \varphi)$, depending on the spherical angles of the given candidate orientation. The rotation-dependent part of the structure factors is then calculated as

$$\sum_{s=1}^{m} |\gamma(\overline{\mathbf{h}}\mathbf{R}_s)|^2 = \sum_{s=1}^{m} \left| \sum_{j=1}^{N_{\mathrm{sample}}} f_j \exp(i2\pi\overline{\mathbf{h}}\mathbf{R}_s\mathbf{A}\mathbf{M}\mathbf{r}_j^{\mathrm{ort}}) \right|^2, \tag{7}$$

where $\mathbf{R}_s$, $s = (1, m)$ are the rotation matrices of the symmetry space group and $\mathbf{A}$ is the same orthonormalization matrix as in (5). The correlation (CORR) between the observed amplitudes $|F(\mathbf{h})|^2$ and the sum $\sum_{s=1}^{m} |\gamma(\overline{\mathbf{h}}\mathbf{R}_s)|^2$ is expected to be larger for the correct rotation $\mathbf{M}$ (DeLano & Brünger, 1995; Caliandro *et al.*, 2006); therefore, it is used as a criterion to select the candidate orientations. All the measured reflections are used in this calculation. In contrast to MR, in our case the rotation matrix $\mathbf{M}$ is determined by only two angles, since the rotation about the axis of the helix is not specified. However, the presence of side chains in the target structure breaks the approximated rotation invariance of a polyalanine helix. To account for this, three values of CORR are calculated by rotating the sample helix by 120° around its axis oriented according to $\mathbf{M}$. The maximum of these three values is considered as the figure of merit associated with the candidate orientation. The solutions are ordered according to this figure of merit and selected using the criterion

$$\frac{\mathrm{CORR} - \mathrm{CORR}_{\min}}{\mathrm{CORR}_{\max} - \mathrm{CORR}_{\min}} > 0.6. \tag{8}$$

They represent the final candidate solutions which are available at selection stage *C*.

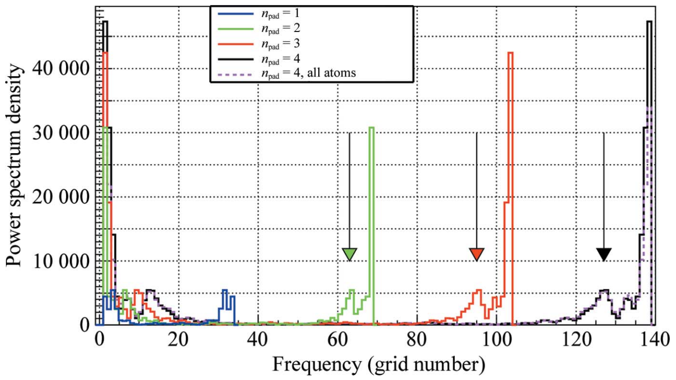

Figure 3
Radial profile of the power spectrum $|G_{\hat{\mathbf{n}}}(\nu)|^2$ of a polyalanine α-helix in the direction coinciding with its axis obtained for different values of the zero-padding parameter n_{pad}. The all-atoms α-helix profile (dashed line) is only reported for $n_{\mathrm{pad}} = 4$.

2.3. Checking the results

In order to validate the procedure, the list of candidate helix orientations needs to be checked against the true α-helix directions in the case of structures already deposited in the PDB. With this aim, an automatic checking protocol has been developed which (i) determines α-helix orientations from the coordinates of their atoms and (ii) compares these orientations with those obtained using our procedure. The protocol will be described in the following two subsections.

2.3.1. Calculus of the orientation of the helix axis. The residues belonging to α-helices are selected within the PDB file: they are all considered as alanine residues. The coordinates of the corresponding atoms are used to calculate the orientation of the helix axis by using a combination of two methods. The first method involves the following steps.

(i) Calculation of the centre of the helix $(x_1^{\rm o}, x_2^{\rm o}, x_3^{\rm o})$ and of its tensor of inertia $I_{ij} = \sum(x_i - x_i^{\rm o})(x_j - x_j^{\rm o})$, where the sum is over all atoms belonging to the helix.

(ii) Diagonalization of the tensor through calculation of its eigenvalues.

(iii) The eigenvector corresponding to the higher eigenvalue then represents the direction of the axis of the helix.

The steps of the second method are as follows.

(i) Calculation of the centre of the atoms of the first three residues of the helix.

(ii) Calculation of the centre of the atoms of the last three residues of the helix.

(iii) The direction cosines of the line passing through the two points then identify the orientation of the axis of the helix.

The directions from the two methods are combined by vector sum, since for helices with more than ten residues the two methods give results that are in agreement within a few degrees. In the case of symmetry the directions are referred to the asymmetric unit of the rotation group, as described in Appendix *A*. These estimations have been checked by the program *HELFIT* (Enkhbayar *et al.*, 2008), which fits three-dimensional data points with a continuous helix by the total least-squares method, finding an agreement within 5° in both θ and φ with our results.

2.3.2. Comparison among helix orientations. To compare two helix orientations, we used the distance between their intersection points with a sphere of radius 1 Å determined in an orthonormal frame (dist). In the case of symmetry, dist is the minimum distance among those calculated between a given direction and all the symmetry-equivalent directions to the other. The threshold value to be used for dist should be related to the error in the determination of the helix axis orientation (err). This can be estimated by considering the scheme in Fig. 4(*a*). The distance D between the intersection points on the unit sphere of the two extreme orientations for an axis passing through the helix is related to the length L and radius R of the helix by the equation

$$D = \frac{2R}{(L^2 + R^2)^{1/2}}. \tag{9}$$

For α-helices, $R \simeq 2.4$ Å and $L \simeq 1.4N_{\rm res}$, where $N_{\rm res}$ is the number of residues in the helix, so that D depends on the number of residues in the helix. A reasonable estimate for err is to consider a cone of angle θ around the true axis. This corresponds to a value of the distance in the unit sphere of $D/2$, the trend of which as a function of $N_{\rm res}$ is reported in Fig. 4(*b*). A dynamic threshold of the distance between two orientations has hence been used throughout the procedure,

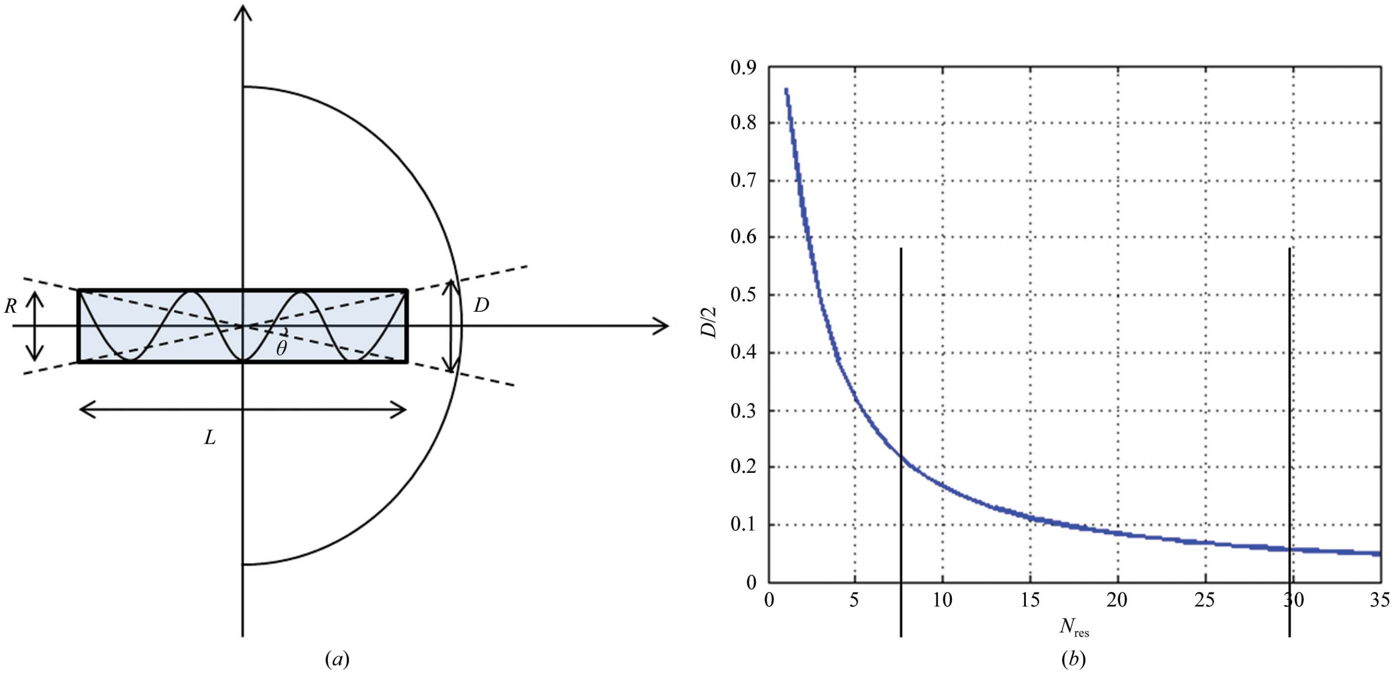

Figure 4
Scheme showing the computation of the error in the determination of the axis of an α-helix (*a*) and the curve of the estimated error as a function of the number of residues in the helix (*b*). The horizontal lines indicate the lower and upper limits of applicability of the dynamic threshold in our procedure.

the value of which depends on the length of the helix to be searched according to (9).

The two horizontal lines in Fig. 4(*b*) indicate the limits of applicability of the dynamic threshold. At the lower limit, helices with less than eight residues have diverging err values. On the other hand, more than 3.6 × 2 = 7.2 residues are necessary to complete two turns of the helix, which give rise to a complete peak in its Patterson radial profile. Therefore, the procedure cannot be applied to helices with less than eight residues. At the upper limit, err is dominated by the angular step chosen to sample the grid in spherical coordinates. By considering the steps adopted for the variables (t, φ), the dist of two orientations passing through two contiguous grid points can be calculated as a function of the spherical coordinates. From Fig. 5, it can be seen that it does not depend on φ and that it is nearly constant in t to a value of 0.062, apart from directions parallel to the z axis, where it diverges. Half of this distance, which corresponds to α-helices of 30 residues according to (9), is a reasonable estimate of the minimum sampling error. Therefore, the dynamic threshold is not applied to α-helices with more than 30 residues.

In summary, two directions are considered to be close if the following criterion is satisfied:

$$\mathrm{dist} < \begin{cases} \dfrac{2.4}{[(1.4N_{\rm res})^2 + 2.4^2]^{1/2}} & \text{if} \quad N_{\rm res} \in (8, 30) \\ 0.031 & \text{if} \quad N_{\rm res} > 30 \end{cases} . \qquad (10)$$

These arguments do not apply in case of bent helices, for which higher errors are expected. On the other hand, their identification by the algorithm here proposed is hampered by the fact that the repetition of their Patterson peaks does not occur along a straight line, so frequency filtering in the radial direction is not effective.

In the framework of the clustering procedure, a threshold value of 0.2 was used to define a cluster, which corresponds to the higher allowed value of dist.

3. Results and discussion

3.1. Test structures

The procedure has been extensively tested on known crystal structures by using diffraction data deposited in the PDB. 74 proteins were selected from the database from those containing at least one α-helix with more than seven residues, which represents the limit of applicability of the procedure. The chosen structures have a wide coverage of crystallographic symmetry, data resolution (ranging from 0.8 to 3.0 Å) and total number of residues in the asymmetric unit (ranging from 100 to 2500). A wide structural variability is also present, with proteins belonging to CATH (Orengo *et al.*, 1997) classes mainly α and α–β included in the test sample. Two further structures, 3gwh and 1gmg, were added to the test set as they have been used to validate other programs. The main crystallographic properties of the 76 test structures are reported in Table S1.[2]

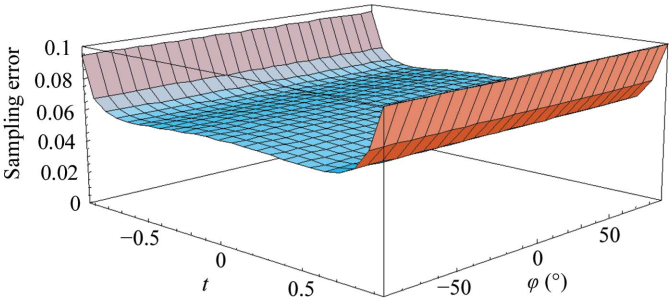

Figure 5
Distance between the intersections of two orientations passing through two consecutive grid points with a sphere of radius 1 Å as a function of the spherical coordinates (t, φ).

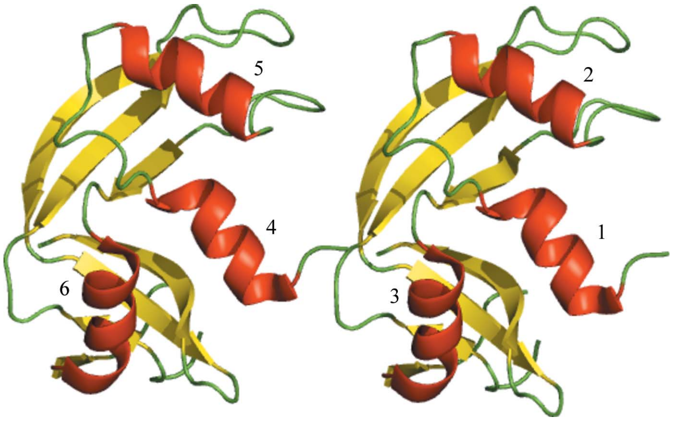

Figure 6
Crystal structure of PDB entry 1dy5. α-Helices containing more than seven residues are numbered.

The procedure is organized so that the number of residues of the longest helices in the structure ($N_{\rm resMax}$), which is usually obtainable from a bioinformatics analysis of the protein sequence, can be supplied as input. It is used to redefine the parameter $R_{\max}$ (= $1.4N_{\rm resMax}$) and to set the parameter $N_{\rm sample}$ (= $5N_{\rm resMax}$).

The performance of the procedure was monitored using two efficiencies,

$$\mathrm{eff}_{\rm helix} = \frac{\text{number of helices found}}{\text{number of helices in the structure}},$$
$$\mathrm{eff}_{\rm sol} = \frac{\text{number of true solutions}}{\text{number of solutions}}, \qquad (11)$$

where the helix is 'found' if at least one solution of the procedure is close to it. In the same way, a solution is tagged as 'true' if at least one helix of the structure has its direction close to it. The criterion for closeness is specific for each helix and is given by (10). The percentage of false positives in the sample of selected solutions is given by $1 - \mathrm{eff}_{\rm sol}$. The information on the number of α-helices with more than seven residues and their length was read from the PDB file of each test structure and used to implement criterion (10) and to calculate the efficiencies (11).

[2] Supplementary material has been deposited in the IUCr electronic archive (Reference: RR5004). Services for accessing this material are described at the back of the journal.

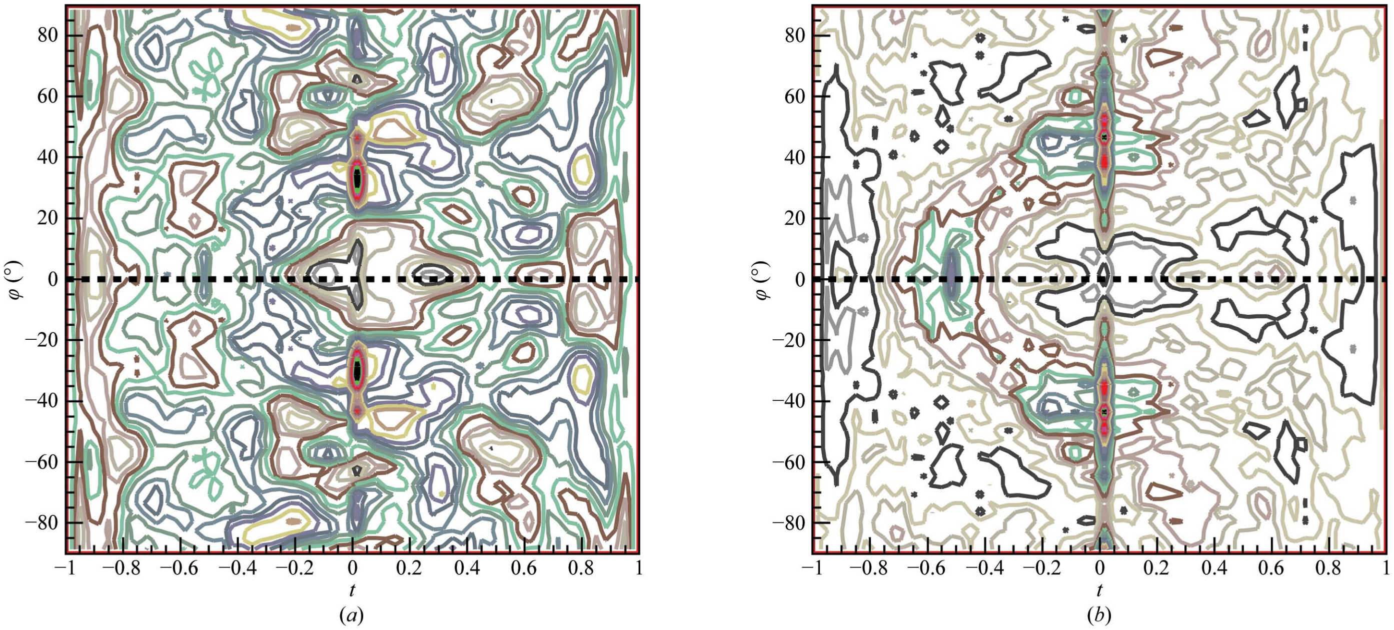

Figure 7
Projection in the plane (t, φ) of the Patterson function (*a*) and the power spectrum $|G_{\hat{\mathbf{n}}}(\nu)|^2$ (*b*) for 1dy5. The horizontal dashed line indicates the restriction introduced by the space-group symmetry.

3.2. The 1dy5 test case

The protein ribonuclease A (PDB entry 1dy5; Esposito *et al.*, 2000) has crystallographic data to 0.9 Å resolution in space group $P2_1$. Its crystal structure is shown in Fig. 6, with numbers indicating helices with more than seven residues. It is constituted of two monomers of 124 residues each in the asymmetric unit related by a quasi-perfect pseudo-translation symmetry. This particular type of noncrystallographic symmetry doubles the contribution from intramolecular vectors, so that the contribution from the second monomer does not constitute noise for the first one. The projection onto the (t, φ) plane of its Patterson function transformed in spherical coordinates and of the corresponding power spectrum $|G_{\hat{\mathbf{n}}}(\nu)|^2$ are shown

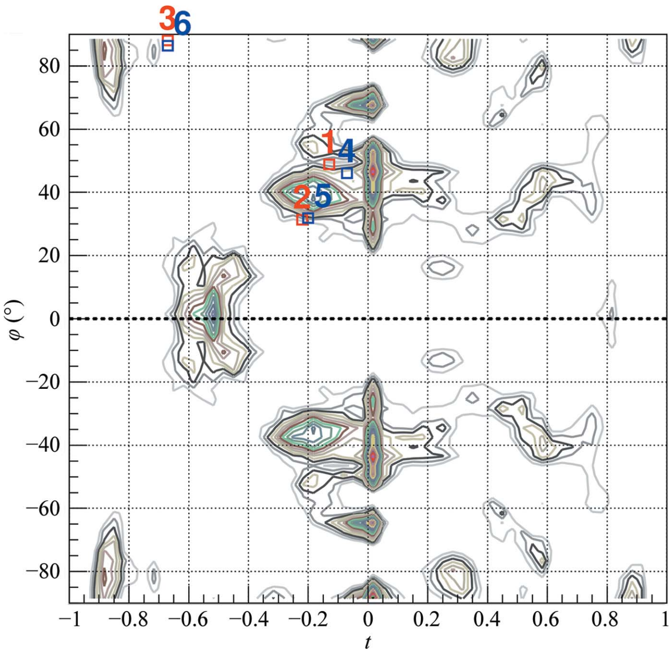

Figure 8
Section of the power spectrum $|G_{\hat{\mathbf{n}}}(\overline{\nu})|^2$ in the plane (t, φ) for 1dy5 obtained by filtering $|G_{\hat{\mathbf{n}}}(\nu)|^2$ at the characteristic frequency ν for α-helix identification. The boxes are centred on the α-helix directions calculated from the published coordinates for chain *A* (red) and chain *B* (blue). The lengths of their sides are equal to the steps used to define the grid. The dashed line indicates the restriction introduced by the space-group symmetry.

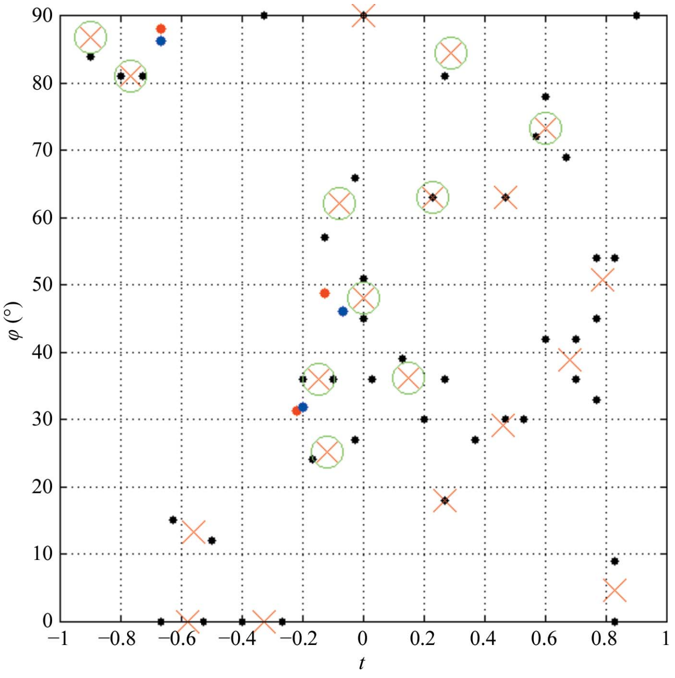

Figure 9
Representation of the true helix directions (coloured dots) and candidate solutions found at stages *A* (dots), *B* (crosses) and *C* (circles) of the procedure for the 1dy5 structure.

in Figs. 7(*a*) and 7(*b*), respectively. The monoclinic crystallographic symmetry introduces a symmetry axis, which is shown as a dashed line. The section $|G_{\hat{\mathbf{n}}}(\nu)|^2$, obtained as described in §2.1 (point v), is shown in Fig. 8, together with the α-helix directions of the published structure, shown by a box of side equal to the grid step. Because of the pseudo-translational symmetry, they occur in doublets, shown in different colours. The candidate directions found by the peak search applied to the asymmetric unit of section $|G_{\hat{\mathbf{n}}}(\nu)|^2$ are reported in Fig. 9 (black dots), together with the true α-helix directions (coloured dots) and the candidate solutions obtained at stages *B* and *C*, which are shown as crosses and circles, respectively. Further details of the number of solutions obtained at the different selection stages, the values of the efficiencies defined by (11) and the minimum distance between the candidate solutions and the true directions are given in Table 1. This resulted in all of the α-helix directions of the protein being found in the first four solutions.

Table 1
Results of the α-helix identification procedure applied to protein 1dy5.

The stages of the procedure are as indicated in Fig. 1. For each helix, the minimum distance from the direction calculated by the published coordinates (dist) and the order number (ord) are reported for the candidate solutions that satisfy the criteria for its identification (10). In the case of more solutions, the closest solution is reported in bold. In the first column, the number of residues and (t, φ) position of each helix is shown.

	dist (Å)/ord		
	Stage *A*	Stage *B*	Stage *C*
Helix 1: 11 residues (−0.13, 48.8)	0.14/**5**, 6, 12	0.13/4	0.13/3
Helix 2: 10 residues (−0.22, 31.3)	0.08/**2**, 4, 13	0.11/**2**, 9	0.11/**1**, 2
Helix 3: eight residues (−0.67, 88.0)	0.12/26	0.16/15	0.16/4
Helix 4: 11 residues (−0.07, 46.0)	0.07/5, **6**, 10	0.08/4	0.08/3
Helix 5: ten residues (−0.20, 31.8)	0.07/**2**, 4, 13	0.09/**2**, 9	0.09/**1**, 2
Helix 6: eight residues (−0.67, 86.4)	0.11/25, **26**	0.15/15	0.15/4
No. of solutions	43	20	14
$\text{eff}_{\text{helix}}$ (%)	100	100	100
eff_{sol} (%)	19	20	29

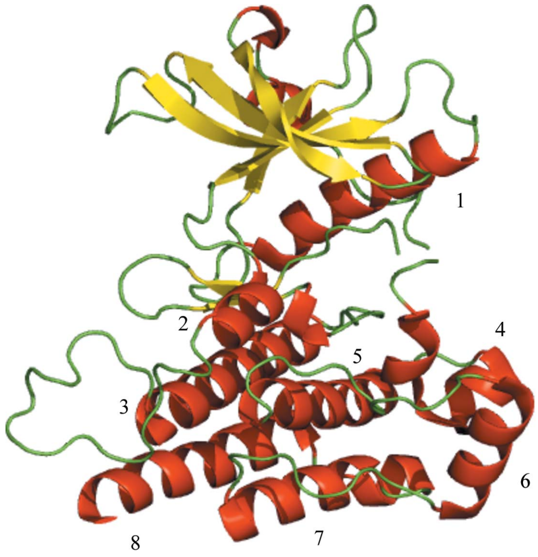

Figure 10
Crystal structure of PDB entry 2qu5. α-Helices containing more than seven residues are numbered.

3.3. The 2qu5 test case

The kinase domain (PDB entry 2qu5; Potashman *et al.*, 2007) represents a more challenging case for α-helix identification. Its structure, reported in Fig. 10, is constituted of two domains, one classified as α–β and the other as mainly α. Its space group, data resolution and number of residues in the asymmetric unit are $P2_12_12$, 2.95 Å and 314, respectively. The

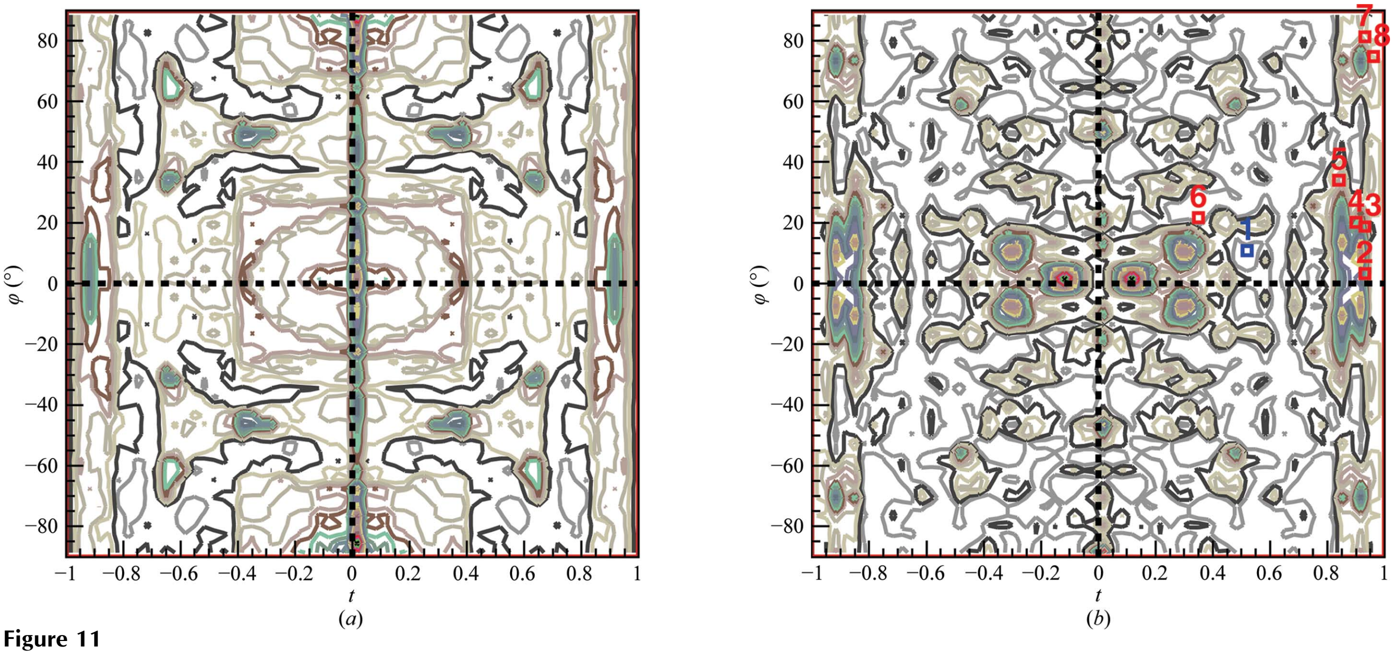

Figure 11
Projection of the power spectrum $|G_{\hat{\mathbf{n}}}(\nu)|^2$ (*a*) and section $|G_{\hat{\mathbf{n}}}(\overline{\nu})|^2$ (*b*) in the plane (t, φ) for 2qu5. The boxes are centred on the α-helix directions calculated from the published coordinates for the α–β (blue) and mainly α (red) domains. The lengths of their sides are equal to the steps used to define the grid. Dashed lines indicate the restriction introduced by the space-group symmetry.

Table 2
Results of the α-helix identification procedure applied to 2qu5.

The stages of the procedure are as indicated in Fig. 1. For each helix, the minimum distance from the direction calculated by the published coordinates (dist) and the order number (ord) are reported for the candidate solutions that satisfy the criteria for its identification (10). In the case of more solutions, the closest solution is reported in bold. In the first column, the number of residue and (t, φ) position of each helices is shown.

	dist (Å)/ord		
	Stage *A*	Stage *B*	Stage *C*
Helix 1: 17 residues (0.52, 10.8)	0.09/53	—	—
Helix 2: eight residues (0.93, 3.3)	0.04/**2**, 3, 6, 7, 9, 19, 37, 41	0.09/**2**, 3	0.09/2, **3**
Helix 3: 22 residues (0.93, 18.8)	0.07/**2**, 3, 37	0.06/2	0.06/2
Helix 4: 8 residues (0.90, 20.1)	0.04/2, **3**, 5, 6, 7, 9, 37, 41	0.08/2, **3**	0.08/**2**, 3
Helix 5: 17 residues (0.84, 34.0)	0.08/5	—	—
Helix 6: 11 residues (0.35, 21.7)	0.11/14, 16, **22**	0.12/8, **9**	0.12/**6**, 11
Helix 7: 12 residues (0.93, 81.3)	0.04/8, **10**	0.05/5	0.05/8
Helix 8: 19 residues (0.96, 74.8)	0.07/8	—	—
No. of solutions	53	19	13
$\text{eff}_{\text{helix}}$ (%)	100	63	63
eff_{sol} (%)	28	26	39

$|G_{\hat{\mathbf{n}}}(\nu)|^2$ projection and the $|G_{\hat{\mathbf{n}}}(\overline{\nu})|^2$ section, shown in Fig. 11, show completely different features. The orthorhombic symmetry manifests itself as two axes, which are shown as dashed lines. The results of the procedure are summarized in Table 2. It can be seen that helices 1, 5 and 8 are not identified within the solutions retained at step *B* and *C* despite being composed of 17, 17 and 19 residues, respectively. It is worth noting that helices 1 and 6 are very flexible (their averaged thermal factors are above 70 Å^2, with those of the remaining helices being around 40 Å^2) and that helix 5 is slightly bent.

3.4. The 1b9o test case

The efficiency of identification was found to strongly depend on the peculiar tertiary-structure arrangement of the protein. As an example, a more difficult structure for the procedure was found to be α-lactalbumin (PDB entry 1b9o; Harata *et al.*, 1999), shown in Fig. 12. It belongs to space group $P2_12_12_1$ with 1.15 Å resolution and a 71% completeness for reflections lower than 4 Å. The signal from the two helices with more than seven residues is hidden by the structural fragments composed of residues 5–21 and 101–108, which are shown as blue and black sticks, respectively, in Fig. 12. They include pieces of α-helices and 3_{10}-helices, interspersed with turns, which run along different directions. As a check, the procedure was applied to amplitudes calculated from published coordinates for different selections of atoms. The results obtained at stage *C* and reported in Table 3 prove that the structural fragment 5–21 interferes with the signal from helix 1, while the structural fragment 101–108 interferes with the signal from helix 2.

Table 3
Results of the α-helix identification procedure applied to several data samples related to the protein 1b9o.

For calculated data, the protein fragments used are reported. For each helix, the minimum distance from the direction calculated by the published coordinates (dist) and the order number (ord) are reported for the candidate solutions obtained at stage *C* which satisfy the criteria for its identification (10).

	dist (Å)/ord	
Data sample	Helix 1	Helix 2
Experimental data	—	—
Calculated data: all of the protein	—	—
Calculated data: helix 1 + helix 2 + residues 5–21 + residues 101–108	—	—
Calculated data: helix 1 + helix 2 + residues 5–21	—	0.05/2
Calculated data: helix 1 + helix 2 + residues 101–108	0.06/4	—
Calculated data: helix 1 + helix 2	0.05/1	0.05/6

3.5. The 3gwh test case

The phosphotransferase system regulation domain II (PDB entry 3gwh) is a five-helix bundle of 111 amino acids which forms a dimer with a substantial deviation from twofold symmetry. It was solved using the program *ARCIMBOLDO* (Rodríguez *et al.*, 2009), which adopts the brute-force approach described in §1. The MR search, which was accomplished by the program *Phaser* (McCoy *et al.*, 2007), produced 49 solutions (rotation + translation), with no relevant ranking for a single α-helical polyalanine fragment of 14 residues, which were then combined to search for multi-fragment solutions. The structure was solved by refining one of the 1473 solutions obtained for the three-fragment search. Our procedure gives 9/10 correct α-helix orientations at step *A*, with 188 candidate solutions, seven of which are still present at step *B*, with 31 candidate solutions. Ranking them by CORR leads to five correct orientations with 16 selected solutions (step *C*). All of the correct orientations are found in the top four solutions, considering that the three helices have very similar orientations (they differ at most by dist = 0.10). Although a direct comparison with the *ARCIMBOLDO* result is not possible, since only the problem of finding the correct orientation of a single α-helix is addressed by our procedure, we stress the fact that a list of well ranked candidate rotation solutions has been

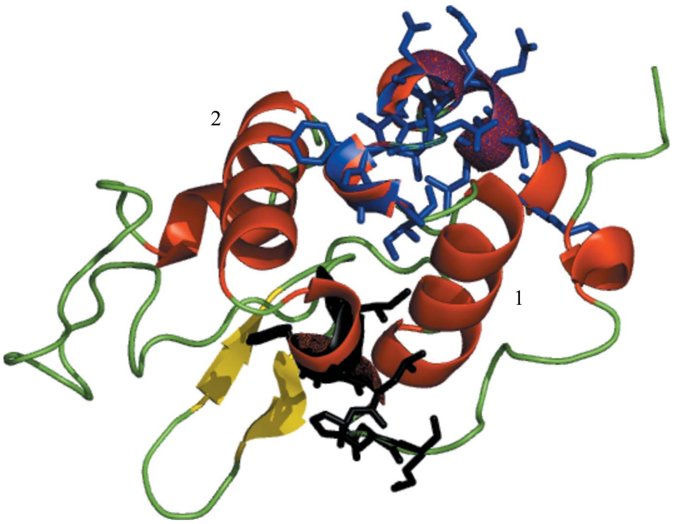

Figure 12
Crystal structure of protein 1b9o. α-Helices containing more than seven residues are numbered. Regions representing noise for helix 1 (blue) and helix 2 (black) are highlighted by sticks.

supplied which could be actively used in the framework of the *ARCIMBOLDO* approach.

To better assess the contribution of our Fourier filtering algorithm to improving the rotational search, we performed the following test: all the orientations, sampled by the same (t, φ) grid defined in point (iii) of §2.1, underwent the selection criterion based on CORR. As a result, 3721 candidate solutions were produced and 1782 were selected containing nine correct orientations. This result should be compared with that obtained by the standard procedure in step *A*. It should be noted that in the top 50 solutions only the common orientation of the three above-mentioned helices is represented and one must go to the 1350th solution to find the nine orientations represented. If a clustering procedure similar to that described in §2.2 is applied to the selected solutions 38 solutions were obtained, containing five correct orientations found in the first 11 solutions. Their mean dist value (0.06) is comparable with that obtained by the standard procedure at step *C*, but the number of solutions and their ranking is worse. The efficiencies calculated for this test are reported in Table S1 under entry $3gwh_{test}$.

3.6. The 1gmg test case

The A31P mutant of the repressor of primer protein (PDB entry 1gmg) forms a helix–turn–helix motif that homodimerizes to form a four-helix bundle with two copies of the motif in the asymmetric unit. The structure was originally solved by the program *Queen of Spades* (Glykos & Kokkinidis, 2000) using a 26-residue polyalanine helix as the model and an intensive 23-dimensional Monte Carlo search (Glykos & Kokkinidis, 2003). Subsequently, the structure was solved by the program *Phaser* (McCoy *et al.*, 2007), which produced many potential solutions and found the correct solution after placing all four copies of the helical model. Our procedure finds only two correct orientations at stage *A*, with only one of them reaching stage *C*, which is the first of eight candidate solutions.

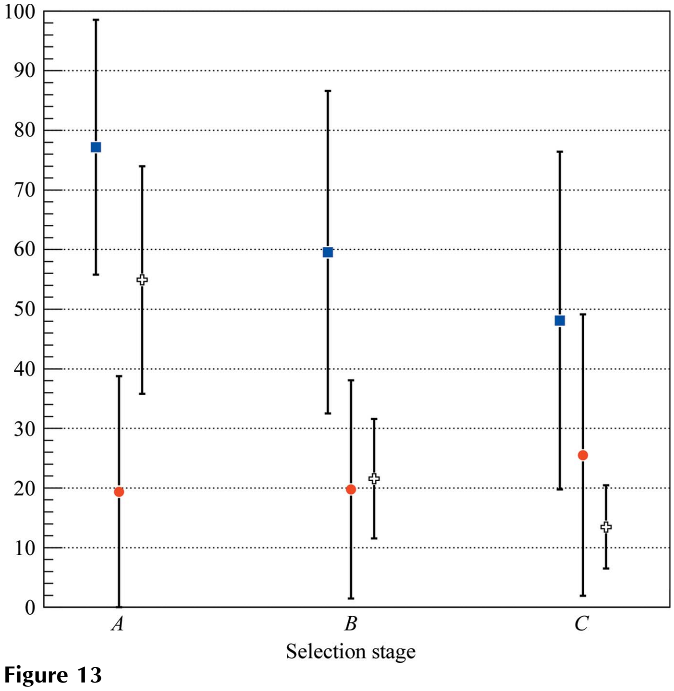

Figure 13
Efficiencies $\text{eff}_{\text{helix}}$ (squares) and eff_{sol} (circles) and number of solutions (crosses) averaged over all the test structures measured at different stages of the procedure. Error bars indicate the root-mean-square deviations of the corresponding distributions.

We can envisage the following difficulty for the application of our procedure to this test case: the four helices are roughly parallel to the *c* axis, where the sampling in spherical coordinates is less efficient (even if the *t* variable is used). On the other hand, the very short *b* axis (about half of *a* and *c*) indicates that the major bundle axis must lie on the *ac* plane. We then re-ran our procedure by referring the spherical coordinates to the *y* axis instead of the *z* axis (see Fig. 1), so that the helices are found in the equatorial plane. In this case at stage *C* all the helix directions were found in the top seven solutions, with ten selected solutions. The corresponding efficiencies are reported in Table S1 under entry $1gmg_{test}$.

3.7. Overall results

The results of the application of the procedure to all of the test structures are summarized in Table S1. The large number of test structures allows a statistical analysis of the results. The efficiency values averaged over all of the test structures and measured at each selection stage of the procedure are reported in Fig. 13. At stage *A* the highest value of $\text{eff}_{\text{helix}}$ (77%) and the lowest value of eff_{sol} (19%) are attained. At stage *B* $\text{eff}_{\text{helix}}$ decreases to 60%, with eff_{sol} remaining nearly constant. As compensation, the average number of solutions is greatly reduced by clustering the solutions, decreasing from 55 to 22. The final selection lowers $\text{eff}_{\text{helix}}$ to 48%, while eff_{sol} increases to 26%. The average number of solutions obtained at stage *C* (13) indicates that it represents a good compromise between the demands of maximizing both the efficiencies and reducing the number of solutions. On average, three of these solutions ($13\times\text{eff}_{\text{sol}}$) are close to a true α-helix orientation; the remaining ten are false positives. It is worthwhile considering that for specific applications it may be advisable to improve the ranking of the solutions rather than reducing their number. In this case, selection stage *A* would be the best choice, because on one hand it corresponds to the higher efficiency in finding α-helix orientations, while on the other it supplies solutions naturally ordered according to the height of the peaks in the $|G_{\hat{\mathbf{n}}}(\overline{\nu})|^2$ section.

The dependence of the two efficiencies on some of the relevant variables is analyzed in Fig. 14. The following can be observed.

(i) $\text{eff}_{\text{helix}}$ decreases smoothly with data resolution (Fig. 14*a*) and number of symmetry operations (Fig. 14*b*) at selection stage *A* (empty markers).

(ii) At selection stage *C* (filled markers) a different behaviour occurs, since $\text{eff}_{\text{helix}}$ increases with resolution for resolutions higher than 1.2 Å and symmetry up to tetragonal. This trend is already observed to occur at selection stage *B* (data not shown).

(iii) eff$_{\rm sol}$ increases with both data resolution and number of symmetry operations independently of the selection stage. This is a consequence of the lower number of candidate solutions found at lower resolution and higher symmetry.

No relevant dependence of the performance of the procedure has been observed on the size of the protein structure and on the number of helices to be found. The execution time of the whole procedure is very limited. Even for the larger structures results were obtained within a few minutes using a 2.6 GHz processor.

4. Outlook

A method of identifying α-helices and estimating their orientations from X-ray diffraction data of protein crystals has been described. The method is based on analysis of the Patterson function interpolated on a grid in spherical coordinates. This function was Fourier-transformed along the radial direction and the resulting power spectrum was filtered in the neighbourhood of the spatial frequency corresponding to the characteristic pitch of α-helices. A peak-search analysis of this power-spectrum section returns the possible α-helix orientations of an unknown protein structure.

The availability of such information at the very first stage of the phasing procedure is of great advantage in the framework of the MR approach. It may be used for the following.

(i) To constrain the generation of or the search for homologous models to be used by MR. In the first case the information should be implemented in programs for *ab initio* (Shortle *et al.*, 1998; Zhang, 2008) or comparative modelling (Sali & Blundell, 1993; Arnold *et al.*, 2006) and in the second case it should be used as an additional criterion in data mining from the PDB. This application is possible when more than one α-helix is present in the structure, in which case the constraint consists of their mutual orientation.

(ii) To improve the location of the model in the unit cell. This procedure could be embedded in the rotation step of MR programs and the search for the best orientation of the model structure could be improved by enhancing the α-helix signal.

(iii) To restrict the rotational MR search. The model structure could be pre-oriented by aligning the α-helices with the directions found by the procedure (this can also be performed for models containing a single α-helix). The rotational space explored during the subsequent MR run could hence be restricted. This could be particularly useful for six-dimensional MR programs (Chang & Lewis, 1997; Kissinger *et al.*, 1999; Sheriff *et al.*, 1999; Glykos & Kokkinidis, 2000; Jamrog *et al.*, 2003) which, although more powerful than standard three-dimensional plus three-dimensional searches, require larger computation time.

(iv) To reduce the number of solutions obtained by MR when a minimal model consisting of a single α-helix is used. This strategy is chosen when an homologous model is not available, but it has a disadvantage in the large number of solutions that are obtained (Rodríguez *et al.*, 2009; Dodson & Woolfson, 2009).

Depending on the specific usage, different selection steps could be most suitable: step *A* would be useful for improving the MR search and step *C* for constraining model generation or for restricting the rotational search.

The algorithm has an intrinsic efficiency of 77% in finding α-helix orientations, despite the presence of a large number of false positives within the sample of candidate solutions. The efficiency remains 48% after a severe reduction of the number of candidate solutions (13 on average). These figures are affected by data resolution (a decrease in resolution has been

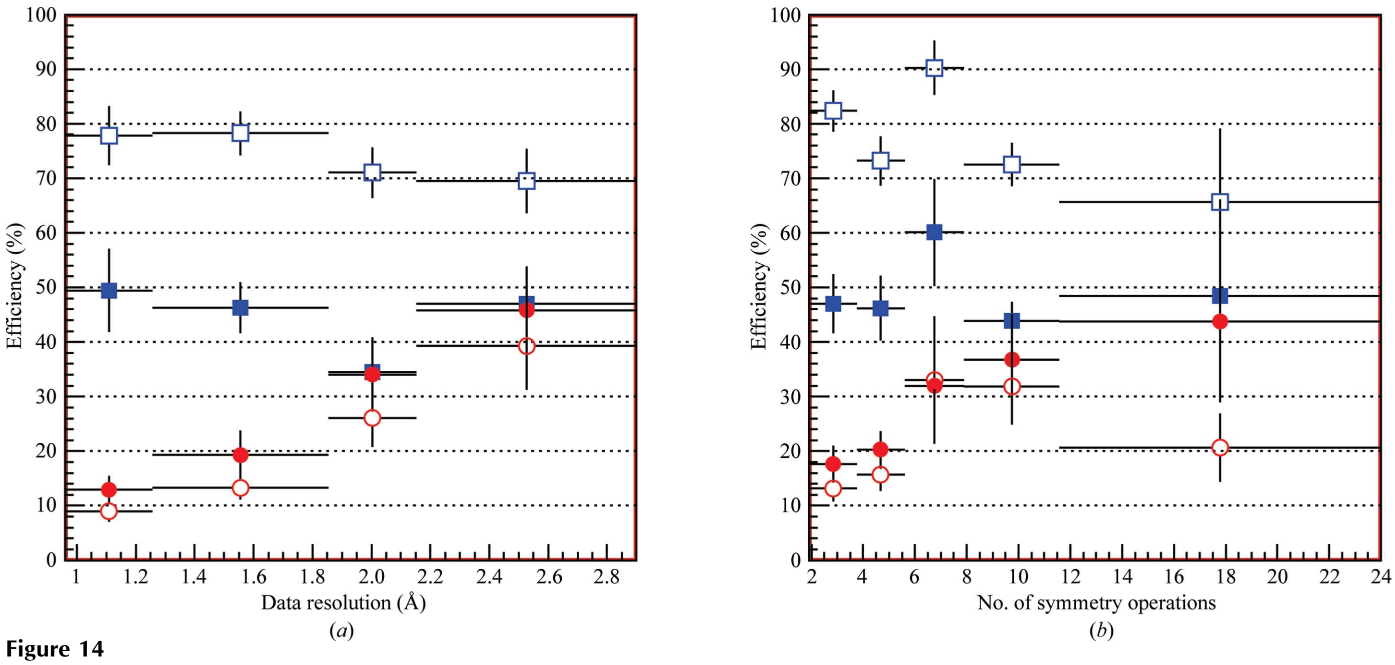

Figure 14
Efficiencies eff$_{\rm helix}$ (squares) and eff$_{\rm sol}$ (circles) measured at selection stages *A* (open markers) and *C* (full markers) averaged over all of the test structures as a function of data resolution (*a*) and number of symmetry operations (*b*). Error bars indicate the root-mean-square deviations of the corresponding efficiency values, while the size of the bins is inversely proportional to their content.

Table 4
Restrictions on the (t, φ) parameters introduced by the Patterson symmetry.

Laue class	Range of t	Range of φ
$\bar{1}$	$(-1, 1)$	$(-\pi/2, \pi/2)$
$2m$	$(-1, 1)$	$(0, \pi/2)$
mmm	$(0, 1)$	$(0, \pi/2)$
$4m$	$(0, 1)$	$(0, \pi/2)$
$4mmm$	$(0, 1)$	$(0, \pi/4)$
$\bar{3}$	$(-1, 1)$	$(0, \pi/3)$
$3m$	$(-1, 1)$	$(-\pi/2, \pi/2)$
$6m$	$(0, 1)$	$(0, \pi/3)$
$6mmm$	$(0, 1)$	$(-\pi/6, \pi/3)$
$m\bar{3}$	$(0, 1)$	$(0, \pi/2)$
$m\bar{3}m$	$(0, 1)$	$(0, \pi/4)$

detected within the range considered) and by crystallographic symmetry (the efficiency increases for higher symmetry). The performance of the procedure is nearly independent of the size of the structures and of the number of α-helices to be found. β-Sheets were not found to systematically affect the efficiency of the method, while the presence of specific structural arrangements may interfere with the signal from the α-helices in some cases, inhibiting their detection. The accuracy in the determination of the α-helix orientation depends on its length: only helices consisting of more than seven residues and that are not bent may be found with reliable precision. The automatic procedure will be included in the next release of the software package *IL MILIONE* (Burla *et al.*, 2007) devoted to protein crystal structure solution.

APPENDIX *A*
Symmetry restrictions in spherical variables

The symmetry operations of the protein space group restrict the parameter space that needs to be explored in searching for α-helix orientations. Strictly speaking, one should refer to the asymmetric unit of the Laue class to which the protein space group belongs. In *International Tables for Crystallography* the asymmetric units are given in Cartesian coordinates (Hahn & Looijenga-Vos, 2006) and we cannot use them by converting the (t, φ) variables to Cartesian coordinates, since the range of the (t, φ) variables used in our procedure does not coincide with the range of the Cartesian coordinates in the unit cell. Therefore, we found the limits of the asymmetric unit in the space defined by the parameters (t, φ) by visual inspection of the Patterson projections onto this plane. As an example, the effect of $2/m$ symmetry can be seen in Fig. 8, while that of mmm symmetry can be seen in Fig. 11. These rules depend on the symmetry operations of the Patterson space group, as they are defined in the crystallographic reference frame, and on the choice of the reference system used to define our parameters. The restrictions found are summarized in Table 4. In the framework of the procedure, they are used to restrict the peak search within the $|G_{\hat{\mathbf{n}}}(\nu)|^2$ sections.

References

Arnold, K., Bordoli, L., Kopp, J. & Schwede, T. (2006). *Bioinformatics*, **22**, 195–201.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.

Blundell, T. L., Jhoti, H. & Abell, C. (2002). *Nature Rev. Drug Discov.* **1**, 45–54.

Burla, M. C., Caliandro, R., Camalli, M., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C., Polidori, G., Siliqi, D. & Spagna, R. (2007). *J. Appl. Cryst.* **40**, 609–613.

Caliandro, R., Carrozzini, B., Cascarano, G. L., De Caro, L., Giacovazzo, C., Mazzone, A. M. & Siliqi, D. (2006). *J. Appl. Cryst.* **39**, 185–193.

Caliandro, R., Carrozzini, B., Cascarano, G. L., Giacovazzo, C., Mazzone, A. M. & Siliqi, D. (2009). *Acta Cryst.* D**65**, 477–484.

Chang, G. & Lewis, M. (1997). *Acta Cryst.* D**53**, 279–289.

Cochran, W., Crick, F. H. & Vand, V. (1952). *Acta Cryst.* **5**, 581–586.

DeLano, W. L. & Brünger, A. T. (1995). *Acta Cryst.* D**51**, 740–748.

Dodson, E. J. & Woolfson, M. M. (2009). *Acta Cryst.* D**65**, 881–891.

Enkhbayar, P., Damdinsuren, S., Osaki, M. & Matsushima, N. (2008). *Comput. Biol. Chem*, **32**, 307–310.

Esposito, L., Vitagliano, L., Sica, F., Sorrentino, G., Zagari, A. & Mazzarella, L. (2000). *J. Mol. Biol.* **297**, 713–732.

Franklin, R. E. & Gosling, R. G. (1953). *Acta Cryst.* **6**, 678–685.

Franklin, R. E. & Gosling, R. G. (1955). *Acta Cryst.* **8**, 151–156.

Frigo, M. & Johnson, S. G. (2005). *Proc. IEEE*, **93**, 216–231.

Glykos, N. M. & Kokkinidis, M. (2000). *Acta Cryst.* D**56**, 169–174.

Glykos, N. M. & Kokkinidis, M. (2003). *Acta Cryst.* D**59**, 709–718.

Hahn, Th. & Looijenga-Vos, A. (2006). *International Tables for Crystallography*, Vol. *A*, 1st online ed., edited by Th. Hahn, pp. 22–23. Chester: International Union of Crystallography.

Harata, K., Abe, Y. & Muraki, M. (1999). *J. Mol. Biol.* **287**, 347–358.

Jamrog, D. C., Zhang, Y. & Phillips, G. N. (2003). *Acta Cryst.* D**59**, 304–314.

Keegan, R. M. & Winn, M. D. (2007). *Acta Cryst.* D**63**, 447–457.

Kim, S.-H. & Rich, A. (1969). *Science*, **166**, 1621–1624.

Kissinger, C. R., Gehlhaar, D. K. & Fogel, D. B. (1999). *Acta Cryst.* D**55**, 484–491.

Kondo, J., Urzhumtseva, L. & Urzhumtsev, A. (2008). *Acta Cryst.* D**64**, 1078–1091.

Long, F., Vagin, A. A., Young, P. & Murshudov, G. N. (2008). *Acta Cryst.* D**64**, 125–132.

Magdoff, B. S., Crick, F. H. C. & Luzzati, V. (1956). *Acta Cryst.* **9**, 156–162.

Mazzone, A., Dibenedetto, D., Nico, G., Cascarano, G. L. & Caliandro, R. (2011). *Proceedings of the 2011 IEEE International Workshop on Medical Measurements and Applications (MeMeA)*, pp. 437–441.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.

Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B. & Thornton, J. M. (1997). *Structure*, **5**, 1093–1108.

Panjikar, S., Parthasarathy, V., Lamzin, V. S., Weiss, M. S. & Tucker, P. A. (2005). *Acta Cryst.* D**61**, 449–457.

Potashman, M. H. *et al.* (2007). *J. Med. Chem.* **50**, 4351–4373.

Rigden, D. J., Keegan, R. M. & Winn, M. D. (2008). *Acta Cryst.* D**64**, 1288–1291.

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.

Sakurai, T., Rao, S. T., Rubin, J. & Sundaralingam, M. (1971). *Science*, **172**, 1234–1237.

Sali, A. & Blundell, T. L. (1993). *J. Mol. Biol.* **234**, 779–815.

Sheriff, S., Klei, H. E. & Davis, M. E. (1999). *J. Appl. Cryst.* **32**, 98–101.

Shortle, D., Simons, K. T. & Baker, D. (1998). *Proc. Natl Acad. Sci. USA*, **95**, 11158–11162.

Thumiger, A. (2008). *PhD Thesis in Molecular Sciences*. University of Padua. http://paduaresearch.cab.unipd.it/979/1/Thumiger_thesis.pdf.

Wilson, A. J. C. (1942). *Nature (London)*, **150**, 150–152.

Zhang, Y. (2008). *BMC Bioinformatics*, **9**, 40.