Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Structure determination of a small protein through a 23-dimensional molecular-replacement search

Nicholas M. Glykos and Michael Kokkinidis

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Structure determination of a small protein through a 23-dimensional molecular-replacement search

Nicholas M. Glykos[a]* and Michael Kokkinidis[a,b]

[a]IMBB, FORTH, PO Box 1527, 71110 Heraklion, Crete, Greece, and [b]Department of Biology, University of Crete, PO Box 2208, 71409 Heraklion, Crete, Greece

Correspondence e-mail: glykos@imbb.forth.gr

Received 29 October 2002
Accepted 3 February 2003

PDB Reference: A31P Rop, 1gmg, r1gmgsf.

The crystal structure of a 4-α-helical bundle protein has been determined by the application of a 23-dimensional molecular-replacement search performed using a stochastic method. The search model for the calculation was a 26-residue-long polyalanine helix amounting to less than 13% of the total number of atoms in the asymmetric unit of the target crystal structure. The crystal structure determination procedure is presented in detail, with emphasis on the molecular-replacement calculations.

1. Introduction

The recent explosive growth in computational power available to the crystallographic community has allowed the once prohibitively expensive multidimensional molecular-replacement searches to establish themselves as a standard item in the toolbox of most laboratories. The majority of the available methods (and corresponding computer programs) reduce the computational cost of these searches by (i) focusing on six-dimensional problems and (ii) employing stochastic optimization methods such as genetic algorithms (Chang & Lewis, 1997) or evolutionary programming (Kissinger *et al.*, 1999, 2001), although new (non-stochastic) optimization methods have emerged (Jamrog, 2002). We have recently described (Glykos & Kokkinidis, 2000, 2001) an alternative $6n$-dimensional molecular-replacement procedure which is based on the simultaneous determination of the rotational and translational parameters of all molecules present in the crystallographic asymmetric unit of a target structure.

Here, we present results from the successful application of this method to a 23-dimensional molecular-replacement problem that had defied all our attempts to tackle it through the application of traditional molecular-replacement methods.

2. Target crystal structure preliminaries

The target crystal structure for the calculations described in this report is the monoclinic form of the Ala31→Pro mutant of the repressor of primer (Rop) protein. Wild-type Rop is a homodimeric RNA-binding protein which, owing to its extensive structural, biochemical and thermodynamical characterization, is regarded as the paradigm of a canonical (left-handed, all-antiparallel) 4-α-helical bundle. We have shown (Glykos *et al.*, 1999) that a single point mutation of Rop at position 31 (the A31P mutant; Fig. 1) is sufficient to change the topology of the protein and to convert the bundle from the left-handed all-antiparallel form to a right-handed mixed parallel and antiparallel form. The original structure determination of this mutant was achieved through the analysis of

an orthorhombic crystal form (space group $C222_1$) containing only one monomer (half-bundle) in the crystallographic asymmetric unit, with the complete 4-α-helical bundle formed through the application of a crystallographic twofold axis as shown in Fig. 1 (Glykos & Kokkinidis, 1999).

A second (monoclinic) crystal form of this same mutant can be obtained under virtually identical crystallization conditions, with the only consistent difference being an increase in the pH of half a unit (although it is possible to obtain crystals of both forms under the same crystallization conditions). The monoclinic crystal form belongs to space group $C2$ and contains the equivalent of one complete 4-α-helical bundle per asymmetric unit and only 35% solvent. The unit-cell parameters of this crystal form show a consistent time-dependent variation, with freshly made crystals giving values of a = 94.4, b = 24.2, c = 64.6 Å, β = 130.4°, whereas a three-month-old crystal gave values of a = 91.3, b = 23.5, c = 63.2 Å, β = 131.0°. Of all the data sets available for the monoclinic form, only two were eventually used. The first is a 1.9 Å resolution data set against which the structure was refined and the second is a 3 Å data set with high multiplicity and 100% completeness which was used for the structure determination. Tables 1 and 2 show the relevant statistics for these two data sets.

As mentioned above, the target crystal structure was known (from solvent-content considerations) to contain the equivalent of one 4-α-helical bundle in the crystallographic asymmetric unit. One important question (that kept occuring throughout the course of our calculations) was whether the asymmetric unit contained one complete 4-α-helical bundle or contained two independent monomers (half-bundles) with the other two halves generated through the application of the crystallographic twofold axes. This question was important because depending on the answer we could (or could not) use as a search model a complete bundle (and not just the monomer), with a concomitant increase in the anticipated signal. Examination of the self-rotation function and of the low-resolution native Patterson function supplemented by packing considerations all provided evidence consistent with the hypothesis that the target crystal structure contained one complete 4-α-helical bundle per asymmetric unit and not two independent monomers. In the following paragraphs, we present a summary of the evidence that was obtained from these studies, not least because the final structure proved that all these indications had been incorrectly interpreted.

Figure 1
Schematic diagram of the structure of A31P Rop in the orthorhombic form. The two monomers are colour-coded and the position and orientation of the intramolecular twofold axis is also shown. Note that in this form the intramolecular symmetry axis coincides with a crystallographic twofold axis. The figure was prepared with the program *Raster3D* (Merritt & Bacon, 1997).

Table 1
Low-resolution data-set statistics.

The data were collected on a MAR Research imaging plate in four successive passes to permit a sufficiently accurate measurement of all low-resolution reflections. The X-ray source was Cu $K\alpha$ radiation produced by a Rigaku RU-300 rotating-anode generator and focused and monochromated using a double-mirror optics system. The data were processed and internally scaled with the programs *DENZO* and *SCALEPACK* (Otwinowski & Minor, 1997). Values in parentheses are for the resolution shell 3.08–3.02 Å

Unit-cell parameters	
a (Å)	92.3
b (Å)	23.8
c (Å)	63.5
β (°)	130.2
Resolution range	∞–3.0
R_{sym}	0.062 (0.131)
Completeness (%)	100 (100)
Multiplicity	8.2 (3.6)
$\langle F/\sigma(F)\rangle$	44.0 (11.7)

Table 2
High-resolution data-set statistics.

The data were collected on a MAR Research imaging plate in two successive passes to permit a sufficiently accurate measurement of low-resolution reflections. The X-ray source was Cu $K\alpha$ radiation produced by a Rigaku RU-300 rotating-anode generator and focused and monochromated using a double-mirror optics system. The data were processed and internally scaled with the programs *DENZO* and *SCALEPACK* (Otwinowski & Minor, 1997). Values in parentheses are for the resolution shell 1.97–1.90 Å.

Unit-cell parameters	
a (Å)	94.39
b (Å)	24.25
c (Å)	64.53
β (°)	130.40
Resolution range (Å)	36–1.9
R_{sym}	0.075 (0.291)
Completeness (%)	96.40 (91.6)
Multiplicity	8.2 (4.1)
$\langle F/\sigma(F)\rangle$	24.8 (4.1)

The dimensions of the minimal orthogonal box that can completely enclose the structure of A31P Rop (as determined in its orthorhombic form) are 49 × 28 × 25 Å^3. The very short (23.8 Å) b axis of the monoclinic form indicated that the major bundle axis (the axis parallel to the longest bundle dimension) must lie on the ac plane. Examination of the low-resolution native Patterson function (shown in Fig. 2) reinforced this interpretation through the repetitive appearance of elongated

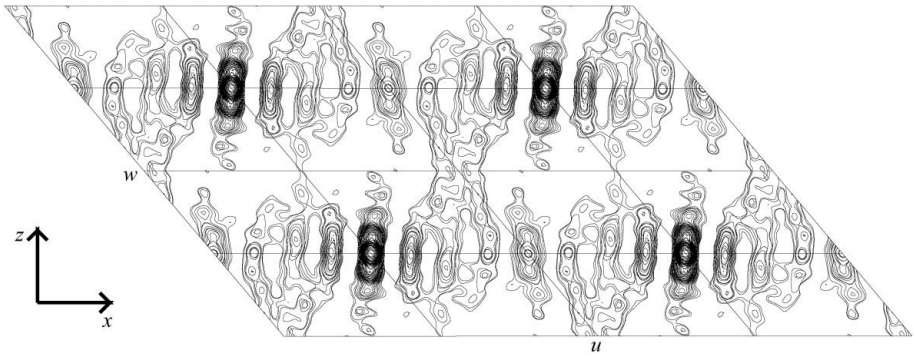

Figure 2
Low-resolution (5 Å) native Patterson function for the monoclinic form of A31P Rop. The volume shown corresponds to fractional coordinates ranging from −0.50 to 1.50 along *u* and *w* and from 0.0 to 0.072 along *v* (three sections). Grid lines have been drawn every 0.50 units. Contours are drawn every 3% of the origin peak, with the first contour at 6%. The orientation of the orthogonal frame (Brookhaven convention) used throughout the paper is also depicted (the orthogonal *y* axis coincides with the crystallographic *y* axis and is perpendicular to the plane of the paper and directed away from the viewer). The figure was prepared with the programs *NPO* and *PLTDEV* from the *CCP*4 suite of programs (Collaborative Computational Project, Number 4, 1994).

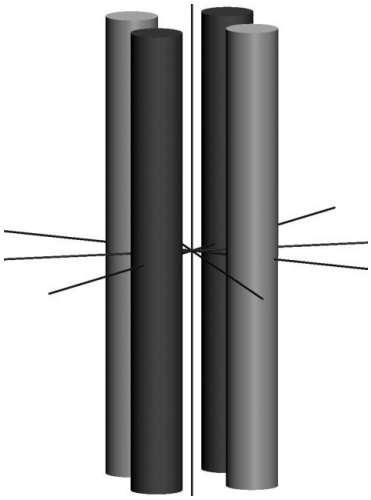

Figure 3
Schematic diagram of a low-resolution approximation to the structure of a 4-α-helical bundle, in which each cylinder corresponds to an α-helix and the connections have been omitted. The axes drawn on this graph depict the orientation of the various approximate symmetry elements that arise when this low-resolution approximation is valid.

features roughly aligned with a direction perpendicular to the *ab* plane (or, equivalently, roughly parallel to the orthogonal *z* axis in the Brookhaven orthogonalization convention). The interpretation of these elongated Patterson features in terms of the direction of individual helices was further strengthened by examination of the self-rotation function. Before presenting the results of the self-rotation function, it is worth discussing the symmetry elements expected from a structure of this kind. As shown in Fig. 3, a 4-α-helical bundle at low resolution has quite a number of approximate symmetry elements: an approximate fourfold axis parallel to the major bundle axis and a series of twofold axes located in a plane normal to the major bundle axis and spaced approximately every 45°. At high resolution (and for the case of a homodimeric bundle like Rop), most of these symmetry elements disappear, leaving behind only one twofold axis (corresponding to the intramolecular dyad axis).

Fig. 4 shows the $\kappa = 90°$ and $\kappa = 180°$ sections from self-rotation functions calculated using two different resolution ranges. The appearance of the function calculated with low-resolution data is in excellent agreement with the anticipated features: there is just one peak on the $\kappa = 90°$ section, indicating there to be just one major bundle axis, whose direction is approximately parallel to the orthogonal *z* axis (compare with the orthogonal frame depicted in Fig. 2). The $\kappa = 180°$ section shows a series of twofold axes spaced approximately every 45° and in a plane perpendicular to the approximate fourfold axis (plane *ab* in the crystallographic frame or, equivalently, *xy* in the orthogonal frame). When the resolution of the data used for the calculation was modified to include only high-resolution (4–2 Å) data, most of these features disappeared, leaving only one prominent peak on the $\kappa = 180°$ section corresponding to a non-crystallographic twofold axis almost parallel to the crystallographic *a* axis[1]. Finally, examination of a series of native Patterson functions calculated using different resolution ranges showed the absence of pseudo-origin peaks from the function, thus indicating there to be no non-crystallographic evenfold axes (on a general position) parallel to the crystallographic twofold axis (which is exactly what would be expected if the intramolecular twofold axis was indeed parallel to *a*).

All this evidence was interpreted in terms of a model containing one complete 4-α-helical bundle in the asymmetric unit with its intramolecular twofold approximately parallel to the crystallographic *a* axis and its major bundle axis aligned with the orthogonal *z* axis. The major evidence against a model containing two half-bundles was the presence of only one set of approximate fourfold and twofold axes in the low-resolution self-rotation function and of the presence of a non-crystallographic twofold axis along *a* which persisted even when using only high-resolution data[2].

[1] The peak at $\omega = 90$, $\varphi = 90°$ corresponds to the crystallographic twofold axis (parallel to *b*). The peak at $\omega = 0°$ (corresponding to an axis with orientation parallel to the orthogonal *z* axis) arises from the interaction between the crystallographic twofold (along *b*) and the non-crystallographic twofold (along *a*).

[2] Clearly, the absence of a second set of approximate fourfold and twofold axes from the low-resolution self-rotation function could be interpreted to mean that the two bundles had approximately the same orientation (*i.e.* both with their major bundle axes parallel to orthogonal *z*). The problem with this idea was that if the two half bundles had the same orientation, then there should be a non-crystallographic twofold axis parallel to *y* relating the two monomers (or its symmetry equivalent) and this should generate an outstanding peak in the native Patterson function (which was not observed).

As a last cautionary tale (illustrating just how easy it can be to obtain evidence consistent even with completely wrong hypotheses), Fig. 5 shows a permutation map[3] produced for the target A31P Rop structure long before the correct structure was known. This map was calculated by assigning phases to the nine strongest *h*0*l* reflections to 8 Å resolution (with two of these reflections serving as origin-fixing reflections) and, as is obvious from this figure, the map fully supports the (wrong) hypothesis that the asymmetric unit contains one complete 4-α-helical bundle placed at a general position and not two crystallographically independent half-bundles. Although this map was chosen for further analysis exactly because it exhibited those (erroneously anticipated) features, it is nevertheless remarkable that the same set of amplitudes when combined with different phase sets can produce electron-density maps fully consistent with an incorrect hypothesis but still perfectly acceptable on physicochemical grounds (see also discussion of the *ab initio* method in §4).

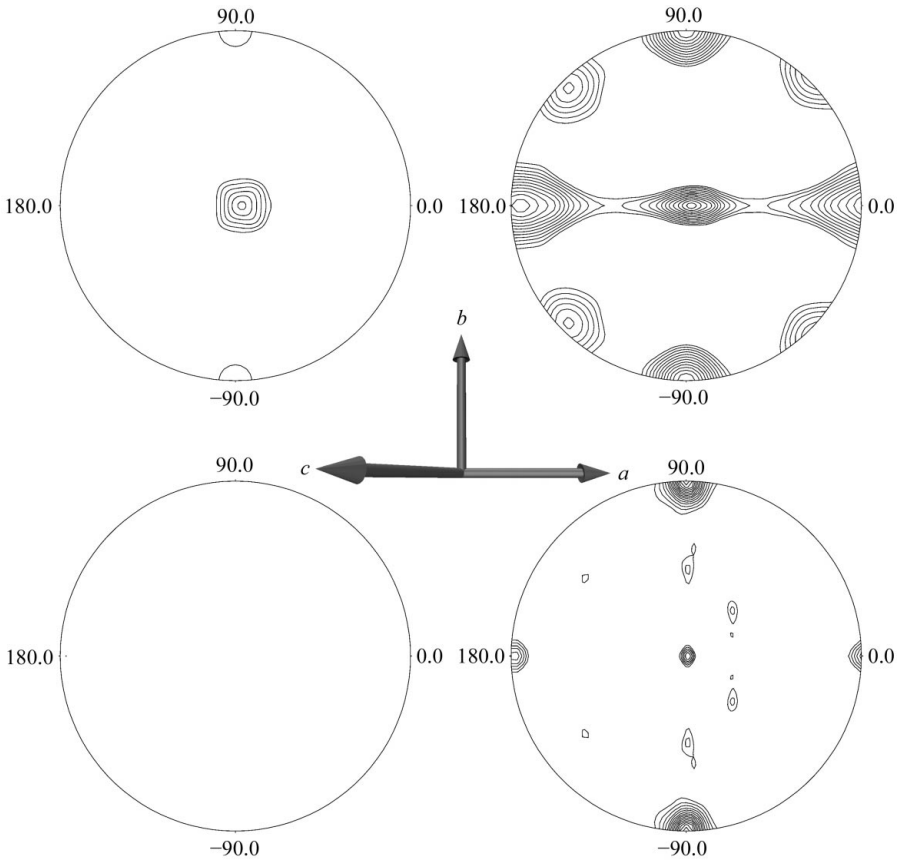

Figure 4
Self-rotation functions. The top row shows stereographic projections of the $\kappa = 90^\circ$ and $\kappa = 180^\circ$ sections for a function calculated using 15–4 Å data. The lower row shows the same sections from a function calculated using 4–2 Å data. The orthogonal frame has its *x* axis horizontal (and in the plane of the paper), its *y* axis vertical and its *z* axis perpendicular to the plane of the paper and directed towards the viewer. In all diagrams the Brookhaven orthogonalization is adopted, the relationship of which to the crystallographic frame is shown schematically between the diagrams. The self-rotation functions were calculated and drawn using the program *POLARRFN* from the *CCP*4 suite of programs (Collaborative Computational Project, Number 4, 1994).

3. Conventional molecular-replacement approaches

Given the similarity of the crystallization conditions for the two crystal forms and the availability of the structure of the same protein from the orthorhombic form, we had been expecting that the determination of the monoclinic form would be a trivial molecular-replacement exercise. This proved to be deceiving: numerous attempts using the majority of the available molecular-replacement methods and corresponding computer programs all failed to give a convincing and consistent solution. The following paragraphs summarize some of these attempts.

The programs *AMoRe* (Navaza & Saludjian, 1997; Navaza, 1994), *MOLREP* (Vagin & Teplyakov, 1997) (both from the *CCP*4 suite of programs; Collaborative Computational Project, Number 4, 1994), *CNS* (Brünger *et al.*, 1998) and *X-PLOR* (Brünger, 1992) were all used in numerous attempts to tackle the molecular-replacement problem using various models, resolution ranges, data sets and program-specific parameters [such as integration radii, Patterson correlation (PC) refinement protocols (Brünger, 1990, 1997*b*) *etc.*]. In the course of these attempts, we tried using search models ranging from a 4-α-helical bundle complete with turns and side chains down to just one 26-residue-long polyalanine helix. Of all these attempts, we shall only mention here a few of the most ambitious and systematic approaches that we have made using as platforms the aforementioned programs (and corresponding algorithms). Additionally, we will exclude from our discussion all those attempts based on using a whole bundle as a search model that were (in retrospect) doomed to fail.

[3] The basis of the permutation-syntheses method (Woolfson, 1954) is the following: given a set of observed amplitudes for a small number of centrosymmetric structure factors, Fourier syntheses are calculated for all their unique phase (sign) combinations and the resulting electron-density maps are then examined for the presence of correct or, as is usually the case, for the absence of unreasonable features, thus allowing the identification of putatively correct phase combinations and in favourable cases the extraction of useful structural information. The case of the monoclinic A31P Rop structure is particularly well suited for an application of this method because the one and only centrosymmetric zone corresponds to a projection of the crystal structure (down the [010] axis) that was known to be just one molecule thick and would thus consist of projections of 4-α-helical bundles seen edge-on.

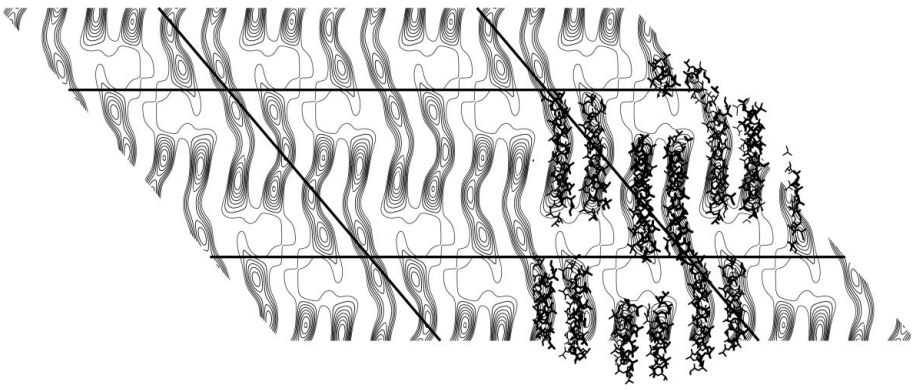

Figure 5
The contour-line plot is a low (8 Å) resolution permutation map for the monoclinic form of A31P Rop (corresponding to a projection down the [010] axis). This was calculated with the program *Pepinsky's Machine* (Glykos, 1999) after assigning phases to nine *h*0*l* reflections (two of which served as origin-fixing reflections). The atomic models (near the right-hand corner of the figure) illustrate the interpretation of this map in terms of projections of 4-α-helical bundles seen edge-on. These atomic models were built by translating (as a rigid body) a polyalanine model of the orthorhombic A31P Rop structure and generating all its symmetry equivalents within a sphere around its molecular centre.

The programs *CNS* and *X-PLOR* were both used in several attempts involving a search model consisting of the crystal structure of the Rop monomer (as determined from the orthorhombic form) and several different parameterizations for the resolution range and PC-refinement protocol. In the most systematic of these attempts, we used 15–3 Å data and successive PC-refinement steps (all iterated to completion) using as rigid bodies the whole monomer, individual helices and three-residue-long pieces of the whole protein. This, and a similar attempt using as a search model a single (26-residue-long) polyalanine helix, all resulted in a uniform distribution of the PC metric.

Similarly, the program *MOLREP* was used with several different models (including a single monomer and a polyalanine helix) and resolution ranges, again with no convincing results, at least as judged by our attempts to further refine the most promising solutions[4].

The program *AMoRe* was used with several different search models, resolution ranges (including but not limited to 15–4, 15–3.5, 15–3, 12–4 and 12–3 Å) and integration radii (ranging from 10 to 18 Å). From all these calculations, we will only describe here one of the most ambitious of our attempts which was performed using a 26-residue long polyalanine helix as a search model. This semi-systematic search was conducted as follows.

(i) A cross-rotation function was calculated (Navaza, 1994; Navaza & Saludjian, 1997) using data in the resolution range 8–2 Å and with an integration radius of 15 Å. The choice of a high-resolution limit of 2 Å is justified on the grounds that a polyalanine helix is expected to be a very accurate search model, thus allowing the use of all high-resolution data (in the hope of increasing the signal-to-noise ratio for the correct solutions). In retrospect, this expectation turned out to be correct: for each of the four helices of the final (refined) structure, the r.m.s. deviations between their main-chain atoms and the search model were only 0.49, 0.30, 0.54 and 0.36 Å. The low-resolution limit of 8 Å was chosen in order to exclude the very strong terms (usually between 12 and 10 Å) that arise from the packing of the helices in the bundle and not from the structure of the helices themselves.

(ii) Each and every orientation corresponding to a positive peak of the cross-rotation function (15 in total) was used to calculate a translation function (Crowther & Blow, 1967; Navaza, 1994) using data in the resolution range 8–2.5 Å. From each of these translation functions we stored the best 50 solutions, giving a total of 750 models for the first helix.

(iii) For each of the 750 models of the first helix and for each of the 15 orientations from the cross-rotation function, we calculated translation functions for a second helix (using again data in the resolution range 8–2.5 Å). By keeping only the top 50 peaks from each translation function, we reduced the number of two-helix models to 562 500 models. Of these 562 500, we only stored for further calculation the orientations and positions corresponding to 29 638 models for which the addition of the second helix simultaneously increased the linear correlation coefficient and decreased the *R* factor.

(iv) Similarly, for each of the 29 638 two-helix models, we calculated translation functions for a third helix, again only considering the top 50 peaks from each translation function, thus obtaining a total of 22.22 million three-helix models. Of these 22 million, only 273 258 models showed in a simultaneous increase of the linear correlation coefficient and a decrease of the *R* factor upon addition of the third helix.

At this stage – and before embarking on a computationally expensive four-helix search – the statistics for the best three-helix solutions were examined: the best *R* factor was 0.583 and the best linear correlation coefficient was 0.37. Such a low agreement with the observed data was taken to indicate that a correct solution had not been found. This interpretation was further strengthened by (i) the uniform distribution of both statistics and (ii) by our failure to meaningfully refine (using rigid-body simulated annealing; Glykos & Kokkinidis, 1999) several of the best (with respect to their statistics) three-helix models. For these reasons, it was finally decided to abandon the search for the fourth helix, assuming that the correct solution had been missed, possibly owing to the small number of the cross rotation function peaks used for the translation searches.

A similar (but less extensive) semi-systematic search which was performed using as a search model the A31P Rop monomer and low-resolution (12–4 Å) data also failed to give an outstanding solution.

[4] 'Promising' here refers to solutions that resulted in packing arrangements that were free from serious steric clashes.

4. Other approaches

Well over 20 heavy-atom-containing compounds have been used in a limited search for a useful derivative. As is usually the case with this protein (and its mutants), most compounds failed to show any sign of specific binding. Platinum-containing compounds did show signs of specific heavy-atom binding and were systematically examined both with respect to their size and reactivity towards protein groups and with respect to soaking time and concentration. Unfortunately, all data sets collected from crystals soaked in these compounds turned out to be non-isomorphous, even when compared with native data sets collected from crystals grown in the same hanging drop. Attempts to determine the heavy-atom structure for the most promising of these derivatives all failed to provide a consistent solution (Gazi, 2000).

One other ambitious attempt to determine the structure of the monoclinic form of A31P Rop involved an *ab initio* method aiming at providing phases for a few of the strongest reflections within the 4 Å sphere. The rationale behind this approach is that the presence of regularities in the structure of the Rop 4-α-helical bundle (both in the packing of the helices and within the helices themselves) gives rise to a few outstandingly strong reflections, most of which are located in two resolution zones: the first at around 12 Å resolution (arising from the packing of the helices) and the second at around 4 Å resolution (arising from the internal helix symmetry). An algorithm was developed in the hope that these relatively few reflections could be phased based solely on our expectations of the characteristics of a protein-like electron-density map.

In summary, the empirical algorithm we have devised is based on a simulated-annealing search of a set of phases (for the given set of reflections) such that the value of a target function calculated in real space (from the electron-density maps) is minimized. The empirical function that we sought to minimize (using a reverse Monte Carlo algorithm; McGreevy & Pusztai, 1988; Keen & McGreevy, 1990) is given by

$$\mathcal{R}(\Phi) \simeq G_{\rm special} G_{\rho_{\rm max}} \sigma_{\rm out} (1 + \mathcal{L}_1)(1 + \mathcal{L}_2) \ldots (1 + \mathcal{L}_n).$$

$\mathcal{R}$ is the function whose value we aim to minimize. Note that although what we change is the phase set Φ, the value of $\mathcal{R}$ is calculated in real space (from the electron-density map obtained through a Fourier transformation of the observed amplitudes combined with the current phase set Φ).

$G_{\rm special}$ is proportional to the number of grid points of the electron-density map that are located on a special (symmetry-wise) position and have a density above a user-defined threshold (for example, above 1 r.m.s.d. of the whole map). This enforces the requirement that protein maps should be devoid of high density on special positions.

$G_{\rho_{\rm max}}$ is proportional to the number of grid points of the electron-density map that have a density higher than a user-defined threshold (for example, 4 r.m.s.d. of the whole map). This drives the algorithm away from a U-atom-like solution (whether attainable or not).

$\sigma_{\rm out}$ is the r.m.s. deviation of the densities of the grid points located outside an envelope drawn at a user-defined threshold (say, 1 r.m.s.d. of the whole map). This enforces solvent flatness outside the (assumed) protein-occupying region of the map.

$\mathcal{L}_1, \ldots, \mathcal{L}_n$ are probably the major parameters that drive the algorithm towards protein-like maps: if there are n polypeptide chains per asymmetric unit, then we expect the correct map to contain (at a given user-defined density threshold) n equally sized pieces of connected (at the given threshold) islands of density. If, for example, we know that we have four helices in the asymmetric unit (assuming that their connections have high temperature factors and will not be visible in the electron-density maps), then we would expect the correct map to contain (at a given density level) four and only four equal pieces of continuous density, each of which would contain 1/4 of the number of grid points above the given density cutoff. The parameters $\mathcal{L}_1, \ldots, \mathcal{L}_n$ encode exactly this requirement, with $\mathcal{L}_1$ being the deviation between the expected size of the isolated island of density (1/4 for the example above) and that (in the map) for the largest continuous fragment. Similarly, $\mathcal{L}_2$ measures the deviation between the expected size and that observed for the second largest fragment *etc.* Obviously, if there is just one chain per asymmetric unit, then we would expect all density above, say, 1σ to belong to one single connected piece of density. Our observation that it is the $\mathcal{L}_1, \ldots, \mathcal{L}_n$ parameters which 'drive' the algorithm towards protein-like maps is in agreement with the studies of Baker *et al.* (1993).

The empirical function described above is the result of a series of tests performed both with hypothetical all-helical structures and with real data from the orthorhombic form of A31P (whose structure was known beforehand). These tests indicated that the best that one could hope for with this algorithm was to find a sufficiently accurate solution embedded in a large set of maps, each of which exhibited the expected (protein-like) characteristics and had similar values of the target function $\mathcal{R}(\Phi)$.

To give an indication of the results obtained by this method and to illustrate one of the major problems encountered with its application, Fig. 6 shows one of the most promising electron-density maps obtained for the monoclinic form of A31P Rop. The density is clearly and convincingly organized in the form of a 4-α-helical bundle. As the polyalanine segments (which were manually fitted to this map as rigid bodies) show, almost all density features in the map can be accounted for by these four helices. However, the protein structure suggested by this map (though perfectly acceptable on both physical and chemical grounds) is also totally incorrect: there is one bundle on a general position instead of two crystallographically independent half-bundles (with the other two halves generated through the application of crystallographic twofold axes). In retrospect, it is not surprising that all our attempts to further refine this (and several other) putative solutions have been unsuccessful.

To avoid a lengthly discussion, we summarize what we believe to be the major problem with the application of this

Table 3
Final statistics for the five independent *Queen of Spades* minimizations.

The values shown correspond to those reported by the program after completion of 10 000 steps of Monte Carlo rigid-body minimization of the best solutions encountered during each of the five minimizations.

Minimization	$1.0 - \mathrm{Corr}(F_o, F_c)$	Free value
1	0.2437	0.3509
2	0.2465	0.6189
3	0.2466	0.5131
4	0.2557	0.6295
5	0.2227	0.3175

empirical *ab initio* method. The problem is an inherently present contradiction: to increase the probability that the global minimum of the target function $\mathcal{R}(\Phi)$ corresponds to the correct structure (or its enantiomorphic image), the number of reflections that enter the calculation must be maximized. This contradicts the practical requirement that to increase the probability of finding the global minimum of the target function, the number of reflections that we seek to phase should be kept to a minimum (approximately 40–60 reflections with present-day computing capabilities). With only 40–60 phased reflections, our tests indicate that it is almost certain that the phase set for which the target function is minimized will *not* correspond to the correct phase set and only by chance will such a phase set be produced[5]. Additional problems arise from the target function *per se*, the most important being that the 'connectivity'-describing parameters $[(1 + \mathcal{L}_2) \ \ldots \ (1 + \mathcal{L}_n)]$ completely ignore the shape of the connected density (and thus encode no information about the chemical expectations we have from a medium-resolution protein electron-density map).

5. Stochastic molecular replacement

The target crystal structure was essentially determined through the application of a 23-dimensional molecular-replacement search performed with the program *Queen of Spades* (Glykos & Kokkinidis, 2001). Although the best solution from this method had one of the helices placed with the wrong polarity (and mistranslated by one helical turn in a direction parallel to the helical axis), it was nevertheless sufficiently close to the correct solution to allow structure determination to proceed to completion. Before discussing in more detail the application of this method and the results obtained from it, we feel that we should state that these results have not been obtained automatically during our first attempt with the program. On the contrary, several months of CPU time had been expended on various unsuccessful attempts before the successful run (presented below) appeared[6].

For the successful run, we used the strongest 70% of all reflections in the resoution range 15–3.5 Å with $F/\sigma(F) > 2.0$ (999 reflections in total). 10% of these reflections were reserved for statistical cross-validation (Brünger, 1997*a*). The target function for the minimization was $[1.0 - \mathrm{Corr}(F_o, F_c)]$, where $\mathrm{Corr}(F_o, F_c)$ is the linear correlation coefficient between the observed and calculated structure-factor ampli-

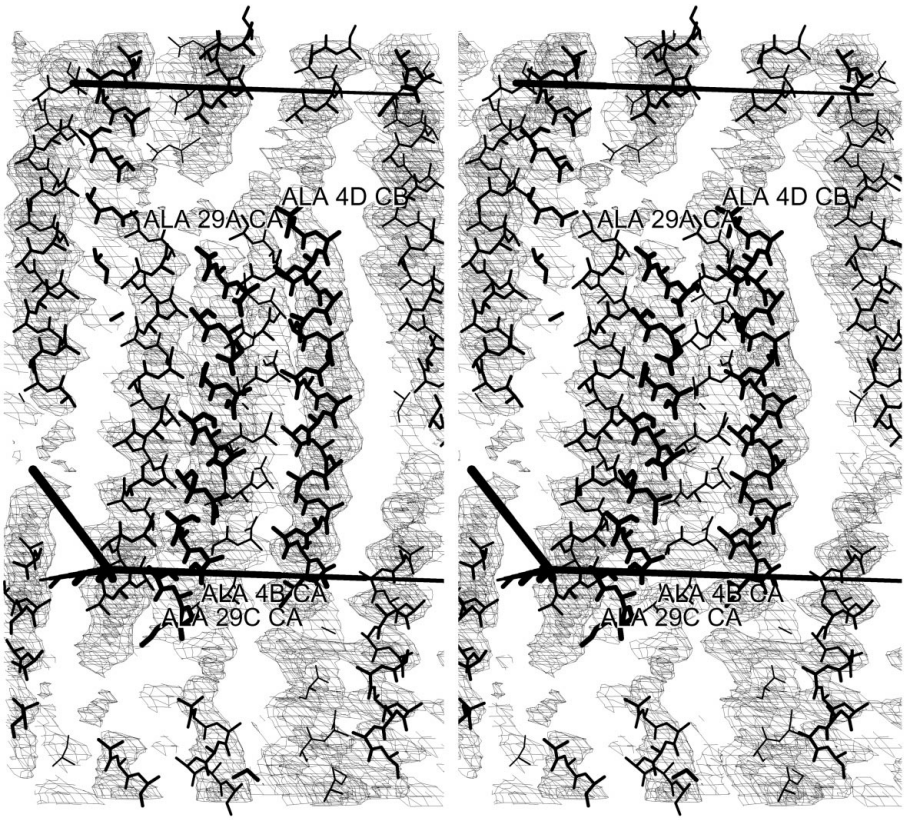

Figure 6
Stereodiagram depicting one of the most convincing electron-density maps obtained for the monoclinic form of A31P *via* the *ab initio* method described in the text. This map was produced by phasing the 26 strongest reflections to 4.4 Å resolution (corresponding to 3.5% of all reflections to that resolution). The polyalanine models that are shown superimposed on the map are the result of manually fitting (as rigid bodies) four copies of a 26-residue-long α-helix into the density and generating all their symmetry equivalents. The orientation and position of the unit cell axes as well as the N- and C-terminal amino acids of the helices are also indicated. The figure was prepared with the program *Xfit* from the *XtalView* suite of programs (McRee, 1992).

[5] It is exactly for this reason that our computer program writes out the currently available 'best' phase set every a few thousand Monte Carlo moves.

[6] The successful run could be identified as a promising solution even before any further analysis was performed on it. This was owing to the presence of a distinct and relatively deep minimum in the graph showing the evolution of the values of the target function and of its corresponding free set *versus* time (as shown in Fig. 7). The presence of such a sudden drop of both indicators is to our experience a dependable indicator that a correct or partially correct solution has been encountered in the course of the simulation.

tudes. A Boltzmann annealing schedule was used with the temperature T at step k given by $T = T_o/\log k$, where T_o is the starting temperature for the minimization [set to 0.070 (arbitrary units) for the successful minimization]. We performed five independent minimizations each lasting 50 million Monte Carlo moves and taking approximately 36 h of CPU time on a below-average (by present-day standards) computing machine[7]. The search model (four identical and independent copies) was a 26-residue-long polyalanine helix (extracted from PDB code 1rpo, residues 4–29) and consisting of 129 atoms.

Table 3 shows the statistics for each of the five minimizations and Fig. 7 shows the evolution of the average values of the target function *versus* Monte Carlo moves for the fifth minimization. The local minimum (shown in magnification in the lower panel of this figure) is 'not so deep as a well, nor so wide as a church-door; but 't is enough, 't will serve', as indicated in Fig. 8, which shows a view of the packing of the search models down the [010] axis. The search models are clearly organized as two independent 4-α-helical bundles centred on the crystallographic twofold axes. The bundles centred on the twofold axes at (1/2, 1/2) and equivalent (by crystallographic symmetry) are well formed and are identical (in terms of fold and topology) to the orthorhombic A31P structure. The bundles at (0, 0) and equivalent appeared to have one helix mistranslated by one helical turn (in a direction parallel to the helical axis) and with the wrong polarity.

This solution immediately explained our failure to interpret correctly all the evidence presented in §2: the two independent monomers – although not related by a non-crystallographic twofold axis – are symmetrically juxtaposed about the orthogonal z axis. The result is that the self-vectors of the two monomers (when translated to the origin of the Patterson function) will appear as if related by a non-crystallographic twofold axis along x, although the true (intramolecular) twofolds coincide with the crystallographic twofold axes. In other words, what we have been trying to interpret in terms of the intramolecular symmetry was the result of the 'special' packing of the two independent monomers.

6. Structure completion and refinement

To keep the structure-determination procedure as free from subjective judgements as possible, we proceeded as follows.

(i) For each of the four polyalanine helices (from the *Queen of Spades* solution), we generated a second helix with the same position and orientation but with the inverse polarity. The result of taking all unique combinations of these eight helices was a set of 16 four-helix models with all possible combinations of helix polarities.

(ii) Each and every one of these 16 four-helix models was subjected to rigid-body simulated-annealing refinement at successively higher resolution (up to 3 Å resolution) as previously described (Glykos & Kokkinidis, 1999). The best solution from this procedure (both in terms of the R factor and of the $R_{\rm free}$ value) corresponded to a solution in which the polarity of one of the helices was the inverse of that suggested by the *Queen of Spades* solution (as expected, see previous section).

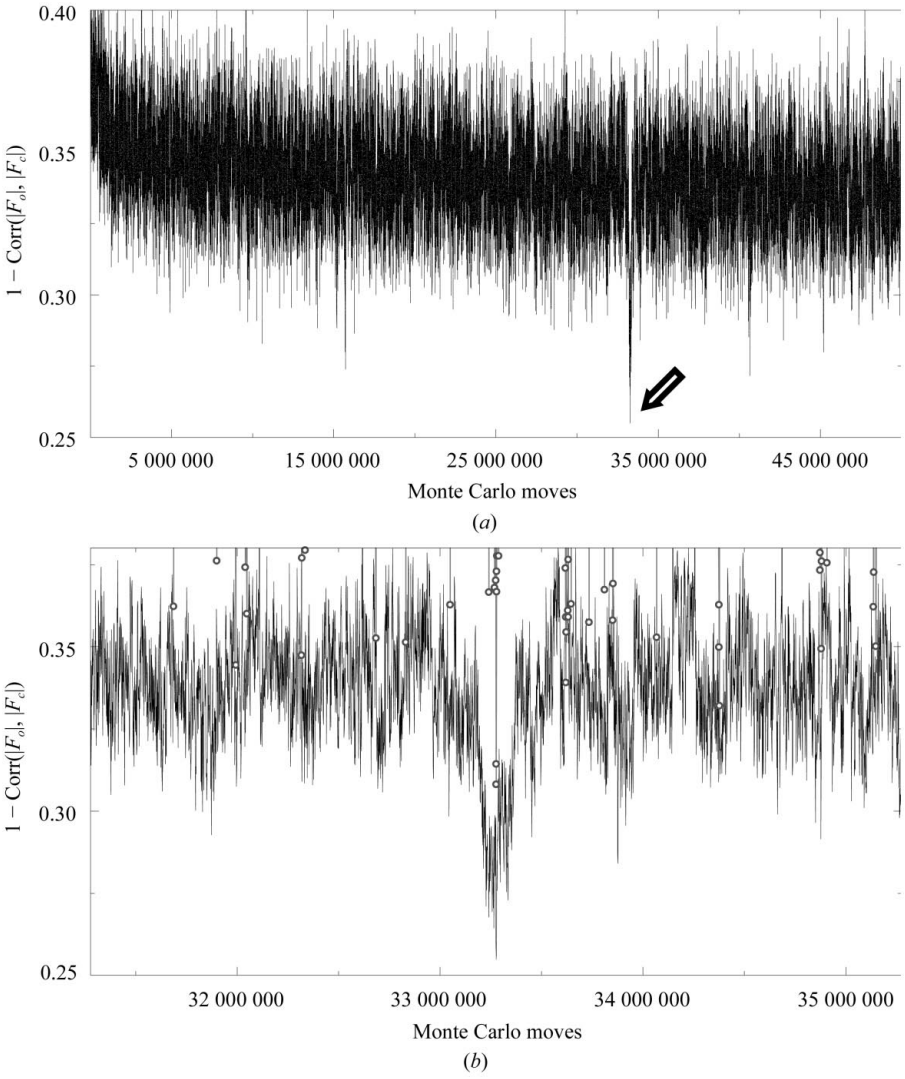

Figure 7
Evolution of the average values of the target function 1.0 − Corr(F_o, F_c) *versus* Monte Carlo moves for the successful *Queen of Spades* minimization. (*a*) the complete run; (*b*) a magnification of the local minimum located at around 33 million moves (indicated by an arrow in *a*). The open circles in (*b*) correspond to the free (cross-validated) values of the target function.

[7] All calculations described in this report were performed on a personal computer equipped with a 800 MHz Intel Pentium III processor, 512 Mbytes of random-access memory and a proper operating system (GNU/Linux, Red Hat distribution, version 6.2).

(iii) The best solution from the previous step was further refined, again with rigid-body simulated annealing, this time including all data in the resolution range 15–1.9 Å and using (in the final steps of the refinement) rigid bodies that consisted of only two alanine residues (per rigid body). This procedure converged to a four-helix polyalanine model, giving an R value of 0.451 and an $R_{\rm free}$ value of 0.472 for all data to 1.9 Å.

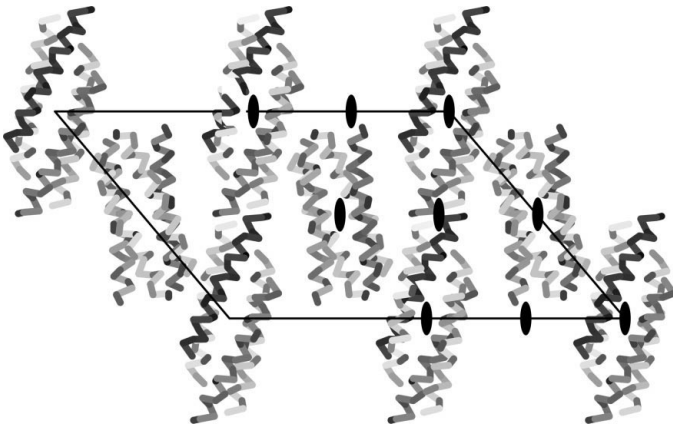

Figure 8
Packing diagram of the monoclinic A31P Rop structure corresponding to the solution obtained from the fifth *Queen of Spades* minimization. The unit-cell axes and several of the crystallographic symmetry elements of the [010] projection are indicated.

A σ_A-weighted map (Read, 1997) of the form $(2mF_o - DF_{\rm polyAla})\exp(i\varphi_{\rm polyAla})$ (where $F_{\rm polyAla}$ and $\varphi_{\rm polyAla}$ are the amplitudes and phases calculated from the four polyalanine helices) was readily interpretable in terms of the protein sequence and inter-helix connectivity and was further improved through the application of the program *ARP* (Lamzin & Wilson, 1997) as implemented in the *wARP* procedure (Perrakis *et al.*, 1997). Side chains and the connecting turns were built using the program *Xfit* from the *XtalView* suite of programs (McRee, 1992). The structure was further refined with the program *X-PLOR* (Brünger, 1992) using torsion-angle dynamics, simulated-annealing and positional (conjugate-gradient) refinement methods, interspersed with rounds of water-molecule addition, bulk-solvent correction, anisotropic scaling (between the observed and calculated structure-factor amplitudes), application of a two-line weighting scheme (Smith, 1997) and individual isotropic temperature-factor refinement. The final structure has an R factor of 0.187, an $R_{\rm free}$ value of 0.232 (for all data in the resolution 36–1.9 Å) and excellent geometry, with all residues in the core region of the Ramachandran plot and an overall G factor of 0.52 (Laskowski *et al.*, 1993). Fig. 9 shows part of the final electron-density map with the refined structure superimposed and Fig. 10 shows a comparison between the schematic diagrams of the structure of A31P Rop in the orthorhombic and monoclinic forms. As it is obvious from this figure, even the relative positions and orientations of the helices belonging to the same monomer are not well conserved, partly explaining the problems encountered while trying to determine the structure using conventional molecular-replacement approaches.

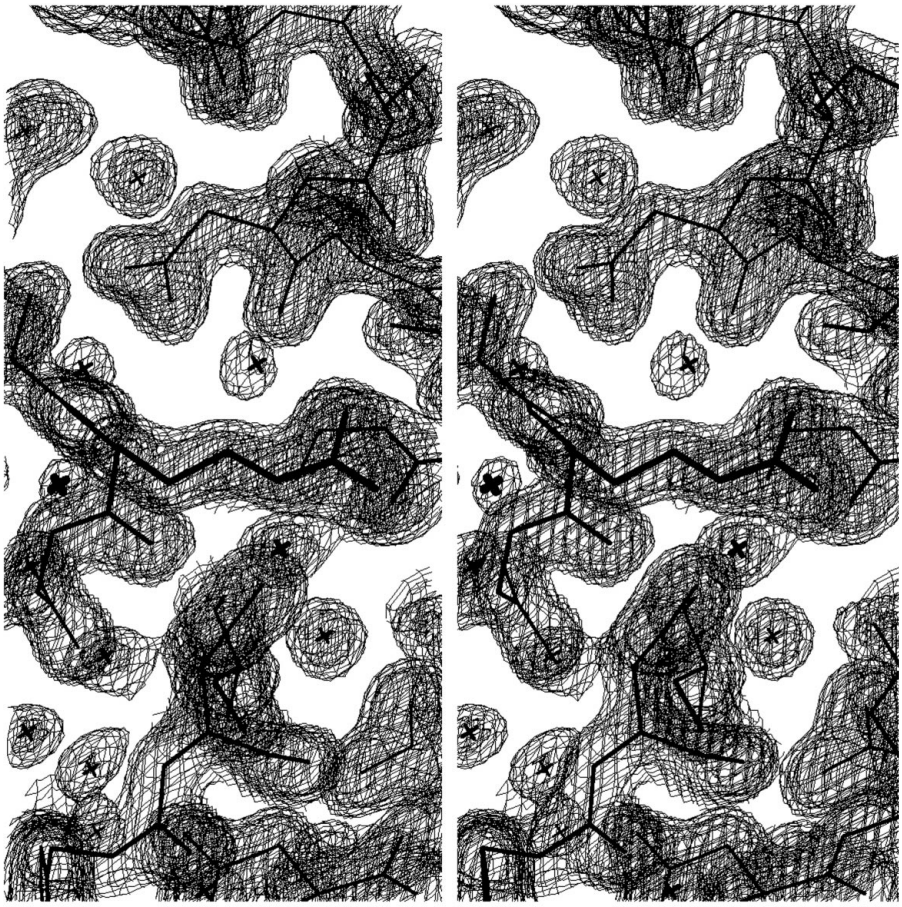

Figure 9
Stereodiagram of a portion from the final 1.9 Å resolution electron-density map for the monoclinic form of A31P Rop. Electron-density isosurfaces are drawn at 1.0 and 1.5σ above the mean density (of the whole map) and they correspond to a σ_A-weighted map of the form $(2mF_o - DF_c)\exp(i\varphi_c)$, where F_c and φ_c are the amplitudes and phases calculated from the final model (also shown superimposed). The figure was prepared with the program *Xfit* from the *XtalView* suite of programs.

7. Discussion

We have shown that a 23-dimensional molecular-replacement search performed with a highly incomplete search model allowed us to determine a structure that has resisted all our attempts to solve it using the majority of the methods available today. Although this serves as a testimony to the power (albeit at a high computational cost) of these methods, it does leave an open

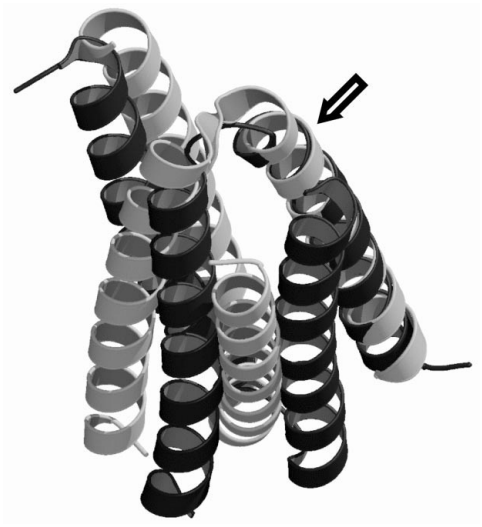

Figure 10
Superposition of the structures of the A31P Rop mutant in its orthorhombic (dark gray) and monoclinic (light gray) forms. The two structures have been superimposed on one of their helices (the pair of helices indicated by an arrow) using the program *LSQKAB* from the *CCP*4 suite of programs. The figure was prepared with the programs *MOLSCRIPT* (Kraulis, 1991) and *RASTER3D* (Merritt & Bacon, 1997).

question, namely the improbability that such a successful search would have been observed. A crude calculation can illustrate the point: if we assume that on average there are only 40 distinct configurations for each of the 23 dimensions of the search, then the search space would comprise no less than 7×10^{36} distinct configurations. With 50 million Monte Carlo moves per run, simulated annealing can only sample an infinitesimally small portion of the parameter space (the metaphorical equivalent of one femtosecond of the age of universe). Although it is possible that pure serendipity allowed us to solve the structure, we suspect that the true reason lies with the reduced effective dimensionality of the search space for the specific problem. What we mean by 'reduced effective dimensionality' is that several parameters or combinations of parameters are rapidly determined by the program owing to the presence of regularities in the target crystal structure. For example, the (020) reflection is the strongest reflection of the data set, having an intensity approximately 100 times higher than the average intensity of the reflections in the data set. This single very strong reflection (arising from the placement of the helices on two equidistant planes parallel to the *ac* plane) can severely reduce the probable locations *and* orientations of all four helices[8]. The combined effect of this (and other) regularities present in the target structure is to drastically reduce the accessible (by the algorithm) volume of the parameter space, leading to a concomitant increase of the probability to locate the global minimum of the target function.

[8] The strong (020) reflection not only determines the spacing between the helices, but also enforces such an orientation for them, so as to make all helical axes parallel to the *ac* plane.

8. Data and program availability

The structure of the monoclinic form of A31P Rop and the 1.9 Å data set (against which the structure was refined) are both available from the Protein Data Bank (Bernstein *et al.*, 1977; entries 1gmg, r1gmgsf).

The 3 Å data set used for the structure determination, together with the polyalanine search model and all files from the successful *Queen of Spades* run, are immediately available for download from http://origin.imbb.forth.gr/~glykos/Monoclinic_A31P_data.tar.gz.

The program *Queen of Spades* is free open-source software available from http://origin.imbb.forth.gr/~glykos/.

References

Baker, D., Krukowski, A. & Agard, D. (1993). *Acta Cryst.* D**49**, 186–192.
Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F. Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). *J. Mol. Biol.* **112**, 535–542.
Brünger, A. T. (1990). *Acta Cryst.* A**46**, 46–57.
Brünger, A. T. (1992). *X-PLOR Version 3.1. A System for X-ray Crystallography and NMR.* Connecticut, USA: Yale University Press.
Brünger, A. T. (1997*a*). *Methods Enzymol.* **277**, 366–396.
Brünger, A. T. (1997*b*). *Methods Enzymol.* **276**, 558–580.
Brünger, A. T., Adams, P. D., Clore, G. M., DeLano, W. L., Gros, P., Grosse-Kunstleve, R. W., Jiang, J.-S., Kuszewski, J., Nilges, M., Pannu, N. S., Read, R. J., Rice, L. M., Simonson, T. & Warren, G. L. (1998). *Acta Cryst.* D**54**, 905–921.
Chang, G. & Lewis, M. (1997). *Acta Cryst.* D**53**, 279–289.
Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.
Crowther, R. & Blow, D. (1967). *Acta Cryst.* **23**, 544–548.
Gazi, A. (2000). Masters Thesis, University of Crete.
Glykos, N. M. (1999). *J. Appl. Cryst.* **32**, 821–823.
Glykos, N. M., Cesareni, G. & Kokkinidis, M. (1999). *Structure*, **7**, 597–603.
Glykos, N. M. & Kokkinidis, M. (1999). *Acta Cryst.* D**55**, 1301–1308.
Glykos, N. M. & Kokkinidis, M. (2000). *Acta Cryst.* D**56**, 169–174.
Glykos, N. M. & Kokkinidis, M. (2001). *Acta Cryst.* D**57**, 1462–1473.
Jamrog, D. C. (2002). PhD thesis, Rice University, Houston, Texas, USA..
Keen, D. A. & McGreevy, R. L. (1990). *Nature (London)*, **344**, 423–425.
Kissinger, C. R., Gehlhaar, D. K. & Fogel, D. B. (1999). *Acta Cryst.* D**55**, 484–491.
Kissinger, C. R., Gehlhaar, D. K., Smith, B. A. & Bouzida, D. (2001). *Acta Cryst.* D**57**, 1474–1479.
Kraulis, P. J. (1991). *J. Appl. Cryst.* **24**, 946–950.
Lamzin, V. S. & Wilson, K. S. (1997). *Methods Enzymol.* **277**, 269–305.
Laskowski, R. A., MacArthur, M. W., Moss, D. S. & Thornton, J. M. (1993). *J. Appl. Cryst.* **26**, 283–291.
McGreevy, R. L. & Pusztai, L. (1988). *Mol. Simul.* **1**, 359–367.
McRee, D. E. (1992). *J. Mol. Graph.* **10**, 44–46.
Merritt, E. A. & Bacon, D. J. (1997). *Methods Enzymol.* **277**, 505–524.
Navaza, J. (1994). *Acta Cryst.* A**50**, 157–163.
Navaza, J. & Saludjian, P. (1997). *Methods Enzymol.* **276**, 581–593.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Perrakis, A., Sixma, T., Wilson, K. & Lamzin, V. (1997). *Acta Cryst.* D**53**, 448–455.
Read, R. J. (1997). *Methods Enzymol.* **277**, 110–128.
Smith, G. D. (1997). *Acta Cryst.* D**53**, 41–48.
Vagin, A. & Teplyakov, A. (1997). *J. Appl. Cryst.* **30**, 1022–1025.
Woolfson, M. M. (1954). *Acta Cryst.* **7**, 65–67.