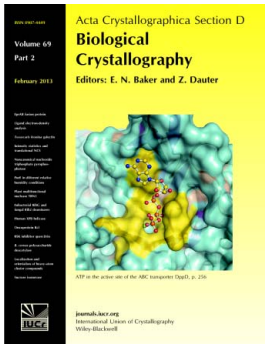
Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449
Editors: **E. N. Baker and Z. Dauter**

Structure determination through homology modelling and torsion-angle simulated annealing: application to a polysaccharide deacetylase from *Bacillus cereus*

Vasiliki E. Fadouloglou, Maria Kapanidou, Athanasia Agiomirgianaki, Sofia Arnaouteli, Vassilis Bouriotis, Nicholas M. Glykos and Michael Kokkinidis

Acta Cryst. (2013). D**69**, 276–283



Acta Crystallographica Section D: Biological Crystallography welcomes the submission of papers covering any aspect of structural biology, with a particular emphasis on the structures of biological macromolecules and the methods used to determine them. Reports on new protein structures are particularly encouraged, as are structure–function papers that could include crystallographic binding studies, or structural analysis of mutants or other modified forms of a known protein structure. The key criterion is that such papers should present new insights into biology, chemistry or structure. Papers on crystallographic methods should be oriented towards biological crystallography, and may include new approaches to any aspect of structure determination or analysis. Papers on the crystallization of biological molecules will be accepted providing that these focus on new methods or other features that are of general importance or applicability.

Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Structure determination through homology modelling and torsion-angle simulated annealing: application to a polysaccharide deacetylase from *Bacillus cereus*

Vasiliki E. Fadouloglou,[a,b] Maria Kapanidou,[a,b] Athanasia Agiomirgianaki,[b,c] Sofia Arnaouteli,[b,c] Vassilis Bouriotis,[b,c] Nicholas M. Glykos[a]*‡ and Michael Kokkinidis[b,c]*‡

[a]Department of Molecular Biology and Genetics, Democritus University of Thrace, University Campus, 68100 Alexandroupolis, Greece, [b]Institute of Molecular Biology and Biotechnology, PO Box 1527, 71110 Heraklion, Crete, Greece, and [c]Department of Biology, University of Crete, PO Box 2208, 71409 Heraklion, Crete, Greece

‡ Correspondence concerning computational crystallography should be directed to NMG. All other requests should be directed to MK.

Correspondence e-mail: glykos@mbg.duth.gr, kokkinid@imbb.forth.gr

Received 4 October 2012
Accepted 6 November 2012

PDB Reference:
polysaccharide deacetylase, 4hd5

The structure of BC0361, a polysaccharide deacetylase from *Bacillus cereus*, has been determined using an unconventional molecular-replacement procedure. Tens of putative models of the C-terminal domain of the protein were constructed using a multitude of homology-modelling algorithms, and these were tested for the presence of signal in molecular-replacement calculations. Of these, only the model calculated by the *SAM-T*08 server gave a consistent and convincing solution, but the resulting model was too inaccurate to allow phase determination to proceed to completion. The application of slow-cooling torsion-angle simulated annealing (started from a very high temperature) drastically improved this initial model to the point of allowing phasing through cycles of model building and refinement to be initiated. The structure of the protein is presented with emphasis on the presence of a C^{α}-modified proline at its active site, which was modelled as an α-hydroxy-L-proline.

1. Introduction

It may sound like a provocative proposition, but the solution of the macromolecular phase problem from first principles (also known as *ab initio* structure determination) may have become an obsolete endeavour: as discussed by Levitt (2007), the rate of discovery of new protein folds is slowing down; indeed, the number of unique folds as defined by *CATH* v.3.4 (Cuff *et al.*, 2011) for the period 2007–2010 (when this version of *CATH* was released) declined steadily, with only nine new folds being recorded in 2009 from 7383 structures deposited in the Protein Data Bank (http://www.pdb.org/pdb/statistics/contentGrowthChart.do?content=fold-cath). The implication of these trends appears to be inescapable: with a sufficiently sensitive and accurate method of homology-based modelling and a dependable set of molecular-replacement programs, the great majority of proteins can be solved directly from the native diffraction data (but, regrettably, not strictly *ab initio*). Automated pipelines based on these or similar ideas have already appeared in mainstream packages, for example the program *AMPLE* from the *CCP*4 suite of programs (Winn *et al.*, 2011) and the *phenix.mr_rosetta* procedure from the *PHENIX* suite (Adams *et al.*, 2010). The greatest difficulty with these approaches appears to be the low accuracy of present-day modelling algorithms, which remains significantly lower than that required for successful structure determination *via* molecular replacement (Giorgetti *et al.*, 2005); however, it should be noted that several groups have reported procedures that have been shown to improve the accuracy of homology modelling for crystallographic applications (Qian *et*

Table 1
Data-collection and structure-refinement statistics.

Values in parentheses are for the highest resolution shell. Ramachandran statistics were obtained from *MolProbity* (Chen *et al.*, 2010).

Unit-cell parameters (Å, °)	a = 36.45, b = 52.76, c = 93.64, β = 95.8
Space group	$P2_1$
Molecules per asymmetric unit	1
Resolution range (Å)	46.6–1.90 (1.95–1.90)
Total No. of reflections	88542
No. of unique reflections	21606
Completeness (%)	81.3 (59.8)
Multiplicity	3.8 (2.0)
$R_{\rm merge}$	0.16 (0.33)
$\langle I/\sigma(I)\rangle$	15.6 (4.1)
R factor	0.178 (0.211)
Free R factor	0.203 (0.252)
Mean B value (Å^2)	16.3
R.m.s.d. bond lengths (Å)	0.019
R.m.s.d. bond angles (°)	1.959
Estimated coordinate error (based on $R_{\rm free}$) (Å)	0.149
Ramachandran (%)	
Favoured	97.8
Allowed	99.7
Outliers	0.3

al., 2007; DiMaio *et al.*, 2011). The difficulties are accentuated if we consider the number of cases in which molecular replacement *per se* was successful (in the sense that the rotational/translational parameters could correctly be determined) but the model was not sufficiently accurate or complete to successfully initiate phasing. Again, several groups have reported various computational methods that aim to improve/refine the molecular-replacement solutions of even very poor initial models (Brunger *et al.*, 2012; Terwilliger *et al.*, 2012; Rodríguez *et al.*, 2009). Here, we present the results from a structure determination that during its course exemplified both the difficulty in obtaining a consistent molecular-replacement solution starting from similarity-based models and also the problems in refining this solution to the point of being able to successfully initiate phasing.

2. Preliminaries

The target protein is BC0361, a polysaccharide deacetylase from *Bacillus cereus*, which forms part of a larger research project of this group (Fadouloglou *et al.*, 2007, 2009; Kokkinidis *et al.*, 2012). It is a 360-amino-acid protein (UniProt entry Q81IM3) which is predicted to have peptidoglycan deacetylase/xylanase/chitin deacetylase activity. Its C-terminal (catalytic) domain is consistently predicted to fold as a deformed $(\beta/\alpha)_8$ TIM-like barrel with five or six strands and to carry the His-His-Asp zinc-binding motif characteristic of this protein superfamily. The N-terminal domain is highly variable in this family and no consistent predictions were available (see §3 for further details).

The protein was cloned, overexpressed and purified as follows. The gene fragment encoding the protein without the signal peptide (residues 24–360) was amplified from *B. cereus* genomic DNA by PCR using the primers GGAATTCCATATGATGAGCCAAGAACCTAAA to generate an *Nde*I restriction site and CCGCTCGAGTTACTTAATTGAAGAAGC to generate an *Xho*I restriction site at the 5′- and the 3′-termini. The PCR fragment was inserted into the pRSET A vector to generate the respective expression plasmid. Transformed *Escherichia coli* DE3 pLysS cells were grown in LB medium to an optical density (OD_{600}) of 0.6. Overexpression was then induced with 0.3 m*M* IPTG and the cultivated cells were harvested by centrifugation after 8 h incubation at 293 K. The cell pellets were suspended in buffer *A* (50 m*M* HEPES–NaOH pH 6.8) and the cells were lysed by the addition of lysozyme. The cell lysate was centrifuged (14 000*g* for 30 min at 277 K) and the filtered supernatant was loaded onto an SP Sepharose column equilibrated with buffer *A*. The fractions were eluted using a gradient from buffer *A* to buffer *B* (50 m*M* HEPES–NaOH pH 6.8, 1 *M* NaCl). Fractions containing the protein were pooled and the recombinant BC0361 was further purified by gel-filtration chromatography using a Sephacryl S-200 column equilibrated with buffer *C* (50 m*M* HEPES–NaOH pH 6.8, 200 m*M* NaCl).

The protein was crystallized using hanging-drop vapour diffusion with 25–30%(*w*/*v*) PEG 3350, 100 m*M* Tris–HCl pH 7.5–8.0. The crystals belonged to space group $P2_1$, with the unit-cell parameters shown in Table 1 and one monomer per asymmetric unit (in agreement with its oligomerization state in solution; Fadouloglou *et al.*, 2008). The crystals diffracted to at least 1.8 Å resolution on a conventional X-ray source and a relatively complete data set was collected from two crystals. The data were measured on an in-house MAR Research imaging-plate detector mounted on a Rigaku RU-3HR rotating-anode X-ray generator using Cu $K\alpha$ radiation focused and monochromated *via* a double nickel-coated mirror system. The rotation method was used throughout, with an oscillation range of 1°. Indexing, integration and scaling were performed with *MOSFLM* (Leslie & Powell, 2007) and *AIMLESS* (Evans, 2006) and resulted in data that were useful to 1.9 Å resolution (Table 1). The strong low-resolution and medium-resolution data (to 2.5 Å resolution, important for the molecular-replacement calculations) had a completeness of 96% and a multiplicity of 4.9 owing to the merging (at the scaling stage) of data from two crystals (which is also the reason for the relatively high overall $R_{\rm merge}$ value of 0.16). Merging data from two crystals was necessary owing to the low (60%) completeness of the higher resolution data set.

3. Structure determination

BC0361 is a typical twilight-zone case with respect to molecular-replacement calculations. A sequence-similarity search performed against the sequences in the PDB using *BlastP* (Altschul *et al.*, 1997) gave as the best hit a high-scoring matching segment with a length of only 123 (of 360) residues and a corresponding sequence identity of 28% (PDB entry 2j13; Oberbarnscheidt *et al.*, 2007). All significant matches corresponded to portions of the catalytic C-terminal domain. No significant similarity (at the sequence level) could be detected *via* *BlastP* for the variable N-terminal domain. Similar results were obtained from *HHpred* (Hildebrand *et al.*,

2009), which identified a segment of 143 residues in length with a sequence identity of 20% in PDB entry 2c1g (Blair *et al.*, 2005).

Undeterred by the sequence-similarity results, we constructed a relatively large set of putative models using structures taken directly from the PDB, as well as a number of models produced by various modelling programs and servers using their default settings. These included PDB entries 1ny1, 1w17, 1w1a, 2c1g, 2cc0, 2j13, 3n2q, 3rxz, 3s6o and models produced by *MODELLER* as obtained through the ModBase database (Fiser & Sali, 2003), *SWISS-MODEL* (Arnold *et al.*, 2006), *I-TASSER* (Zhang, 2008), *CPHmodels* (Nielsen *et al.*, 2010), *M4T* (Rykunov *et al.*, 2009), *HHpred* (Hildebrand *et al.*, 2009), *EsyPred3D* (Lambert *et al.*, 2002), *Phyre*2 (Kelley & Sternberg, 2009) and *SAM-T*08 (Karplus, 2009). In the cases where the modelling programs returned more than one model, or when modelling was attempted for both domains, these were treated as independent models. This set of putative models was further augmented through the application of normal-mode analysis as implemented in *CCP*4. The resulting (approximately 60) distinct coordinate files were then tested for the presence of signal in molecular-replacement calculations performed with *Phaser* (McCoy *et al.*, 2007) and *Qs* (Glykos & Kokkinidis, 2000*a*, 2003). During the course of the investigation, we also tested searches using combinations of models for the two domains. With the exception of the one case discussed below, no convincing, persistent and consistent solution could be identified: the *Z*-scores from *Phaser* were in the range 3.0–4.0 and the [1 − Corr(*Fo*,*Fc*)] *versus* time distributions obtained from the *Qs* runs were identically uniform and unconvincing.

The model that eventually led to structure determination was produced by the *SAM-T*08 (Karplus, 2009) algorithm using the default settings of the server. *SAM-T*08 returned models for the whole protein, and these were divided and treated as independent N- and C-terminal domain models. The best scoring *SAM-T*08 model for the C-terminal domain (residues 142–360) gave a unique, consistent and persistent solution with *Phaser* (*Z*-scores of 4.7 and 5.3 for the rotation and translation functions and an LLG score of ∼30; these were twice the scores obtained from the other models) and a very clear solution with *Qs* as shown in Fig. 1.

Although the *SAM-T*08 model was sufficiently accurate to allow an unambiguous molecular-replacement solution, the statistics (following rigid-body refinement to 2.5 Å resolution) were far from hopeful: with an *R* factor of ∼0.58, a linear correlation coefficient of ∼0.22 and a mean figure of merit of ∼0.15 for all data to 2.5 Å resolution, progress was expected to be difficult. Indeed, all of our attempts to initiate phasing through model extension and refinement using various combinations of rigid-body refinement with *REFMAC* (treating individual secondary-structure elements as separate bodies) and density modification, followed by numerous runs of *Buccaneer* (Cowtan, 2012), *ARP*/*wARP* (Cohen *et al.*, 2008) and/or *SHELXE* (Sheldrick, 2010), diverged at all resolutions and parameter combinations that we tested. In retrospect, and as shown in Fig. 2(*a*), this was not very surprising: following least-squares superposition, the C^{α} r.m.s. deviation between the final (refined) structure and the homology-derived model was 2.51 Å, which was apparently too large for our calculations to be convergent (noting that this deviation concerns only half of the structure, with the other half being completely absent from the model).

The calculation that finally allowed the structure determination to proceed to completion was based on torsion-angle simulated annealing performed with *CNS* (Brünger *et al.*, 1998). To minimize the impact of possibly serious (offset sequence) errors in the homology-derived structure, this calculation was performed using a poly-alanine version of the *SAM-T*08 C-terminal domain model. In our calculations, we directly used all available data to 1.9 Å resolution and performed a grand total of 400 repetitions of torsion-angle simulated annealing using a slow-cooling protocol

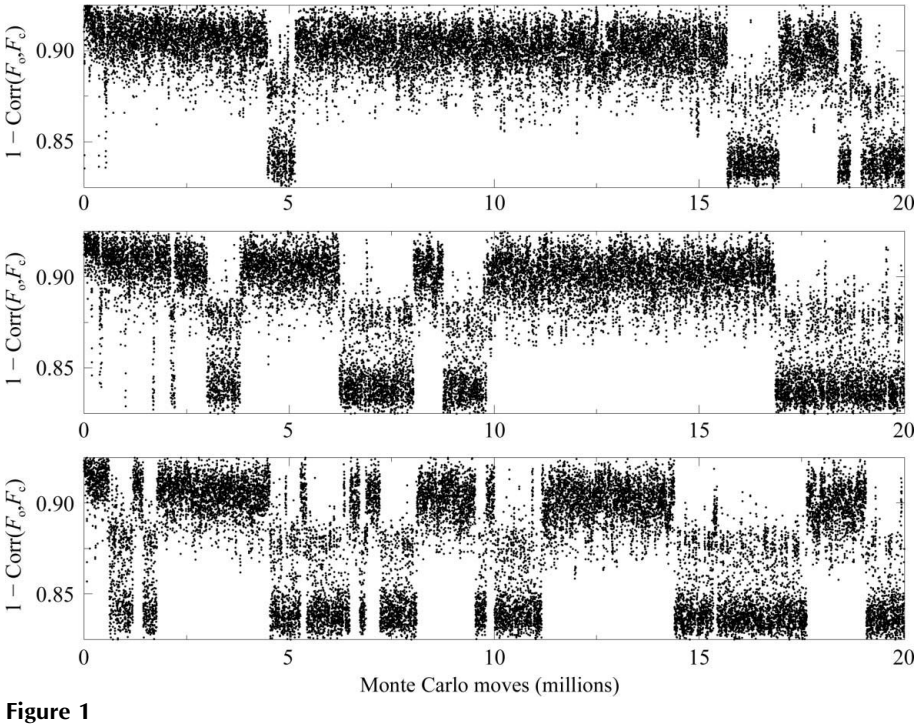


Figure 1
Molecular-replacement calculations using the *SAM-T*08 model. The three graphs show the evolution of 1.0 − Corr(F_o,F_c) *versus* Monte Carlo moves for three minimizations performed with the program *Qs* using all data between 15 and 3 Å resolution and the (default) Boltzmann annealing schedule. The sudden decreases in the target-function values (corresponding to increases in correlation) are the hallmark that a clear solution has been found. During the minimizations the correct solution is visited several times (for example, four times in the first run and 11 in the third) before the system continues exploring other configurations.

with $T_0 = 8000$ K, $\Delta T = 20$ K, a final temperature of 300 K and the maximum-likelihood target. No positional or Cartesian simulated-annealing refinement was performed. Fig. 3 shows a histogram of the $R_{\rm free}$ values obtained from these 400 repetitions. The structure with the lowest $R_{\rm free}$ value (0.496) is indicated by an arrow in Fig. 3 and is depicted in Fig. 2(*b*).

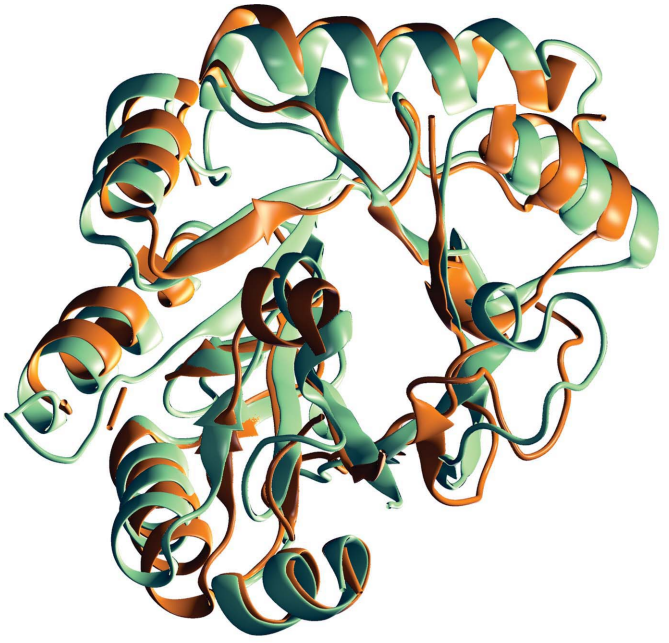


Figure 2
Torsion-angle simulated annealing. The two images compare the final structure of the C-terminal domain (coloured green) with (*a*) the structure obtained from homology modelling using the *SAM-T*08 server (coloured orange) and (*b*) the structure obtained after torsion-angle simulated annealing (same colour scheme). The corresponding C^{α} r.m.s. deviations are 2.51 and 1.69 Å, respectively.

Admittedly, torsion-angle simulated annealing performed beautifully: almost all of the secondary-structure elements of the domain converged to their refined positions, with a concomitant reduction in the C^{α} r.m.s. deviation (with respect to the refined structure of the domain) from 2.51 to 1.69 Å. Given that this is a stochastic method, it can be hypothesized that an even more accurate polyalanine model could be produced through either a larger number of repetitions or a slower annealing protocol. We have actually tested this idea (*a posteriori*) by repeating the calculation twice: in the first round we produced 2800 structures instead of 400 structures (the minimizations required the equivalent of 2 min per structure on a quad-core processor clocked at 2.4 GHz). In the second round we again produced 400 structures but this time with a ΔT of only 5 K. Although, as with any stochastic method, the results can only be interpreted probabilistically, it is probably fair to say that the extra computation involved may have been worth the wait: the test with 2800 repetitions resulted in a structure with a C^{α} r.m.s.d. (*versus* the refined domain structure) of only 1.53 Å, whereas the test with $\Delta T = 5$ K gave a structure with an r.m.s.d. of 1.59 Å.

Having obtained a reliable model for the C-terminal domain using torsion-angle simulated annealing, structure determination proceeded to completion smoothly: a large number of *Buccaneer* cycles at 1.9 Å resolution and with a very small weight for the X-ray term in *REFMAC* led to an almost complete model for the protein with an R value of 0.275, a free R value of 0.297 and a very clear $2mF_{\rm o} - DF_{\rm c}$ map. Cycles of model building with *Coot* (Emsley *et al.*, 2010) and refinement with *REFMAC* (Murshudov *et al.*, 2011) resulted in a final structure with an R value of 0.178, a free R value of 0.203 and excellent geometry (Table 1). Fig. 4 shows a large volume from the final $2mF_{\rm o} - DF_{\rm c}$ map illustrating the quality of phase determination and a detailed view of the active site of the enzyme.

A natural question that arises at this point is whether other recently described methods such as morphing (Terwilliger

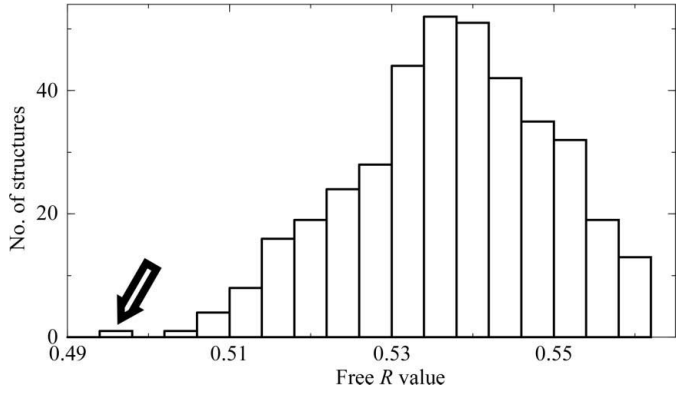


Figure 3
Distribution of the free R values from simulated annealing. The histogram depicts the distribution of the final $R_{\rm free}$ values obtained from 400 repetitions of the slow-cooling annealing protocol described in the text. The arrow points to the (single) structure which converged to an R value of 0.494 and a free R value of 0.496 for all data to 1.9 Å resolution (this structure is shown in Fig. 2*b*).

et al., 2012), deformable elastic network (DEN) refinement (Brunger *et al.*, 2012) or the application of jelly-body restraints with *REFMAC* could have successfully substituted for torsion-angle simulated annealing in this structure determination (noting that DEN refinement actively uses torsion-angle simulated annealing in its refinement procedure). To address the issue, we used the corresponding modules and programs from the *PHENIX* and *CCP*4 distributions using the same polyalanine starting model and the same 1.9 Å resolution data that were used in simulated annealing (see §6 for data availability). In the case of jelly-body restraints with *REFMAC*, we tested three different values of the corresponding weight term (0.01, 0.02 and 0.04) and we performed a total of 100 cycles for each run. The C^{α} r.m.s. deviation of the resulting models from the deposited domain structure were 2.18, 2.14 and 2.21 Å, respectively. In the case of morphing, we used the *phenix.morph_model* module with its default settings. The resulting model (after six cycles) had a map–model correlation after refinement of 0.403 and a C^{α} r.m.s. deviation from the deposited domain structure of 2.02 Å. Finally, the application of DEN refinement using the *phenix.den_refine* module gave a very accurate model with a C^{α} r.m.s. deviation of only 1.77 Å. With the proviso that during these tests we did not attempt to perform complete structure re-determinations and that these were one-off runs using the default settings of the programs, our tentative conclusions are that (i) application of jelly-body restraints from within *REFMAC* may not have been very successful in the case examined, (ii) morphing significantly improved the starting model but not as much as did simulated annealing (which reached r.m.s.d. values as low as 1.53 Å; see above) and (iii) DEN refinement performed significantly better than the other two methods; with an r.m.s.d. of 1.77 Å, its final model was only slightly worse than the best model produced by straight torsion-angle simulated annealing.

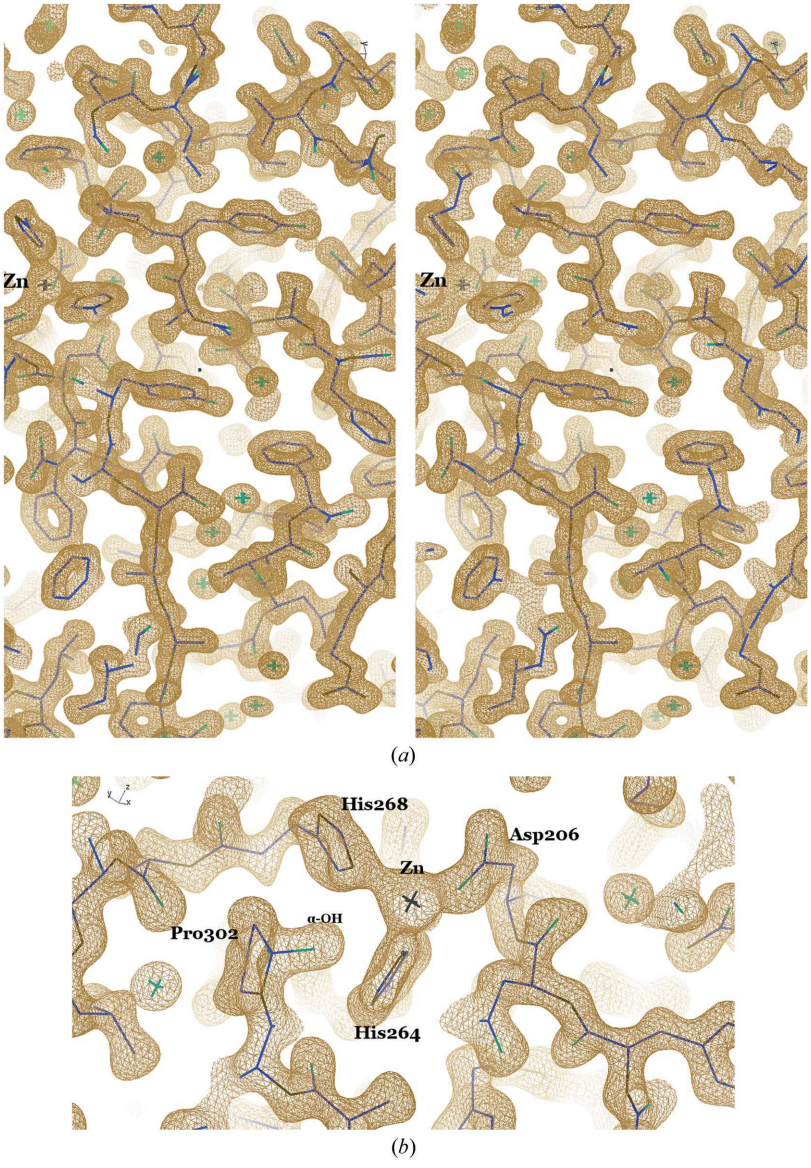


Figure 4
Maximum-entropy estimate of the final $2mF_o - DF_c$ electron-density map. (*a*) shows a wall-eyed stereo diagram of a relatively large volume from the final 1.9 Å resolution $2mF_o - DF_c$ map illustrating the quality of phase determination. The map shown is the maximum-entropy estimate (of the corresponding σ_A-derived coefficients) as produced by the program *GraphEnt* (Glykos & Kokkinidis, 2000*b*). The final model is shown superimposed using a liquorice representation. The location of the active site of the enzyme has been marked by labelling the Zn atom. (*b*) shows a detailed view of the active site with the coordinating His-His-Asp triplet clearly visible. Note the presence and closeness to the active site of the OH group of the nearby α-hydroxy-L-proline 302.

4. Structure description

Fig. 5 shows schematic diagrams of the BC0361 structure. The structure is in full agreement with the characteristics expected for a member of this protein family. The catalytic C-terminal domain is a TIM-like barrel with seven well formed β-strands and an open and accessible active-site cleft (clearly seen in Fig. 5*b* near the very top). The metal ion in the active site (modelled as zinc) is coordinated by His264, His268 and Asp206 (see Fig. 4*b*). Significant density strongly bound to the Zn atom was modelled as an acetate ion. The variable N-terminal domain is a two-layered (4 + 3) β-sandwich. This domain is not well ordered in the crystal structure and no density could be detected for its first 44 residues.

Although a meticulous description of the structure will not be undertaken (especially since the specific substrate of BC0361 is as yet unknown), there is one feature that will be discussed in some length, not least because to our knowledge this is the first report of a protein residue modified at its C^{α} atom.

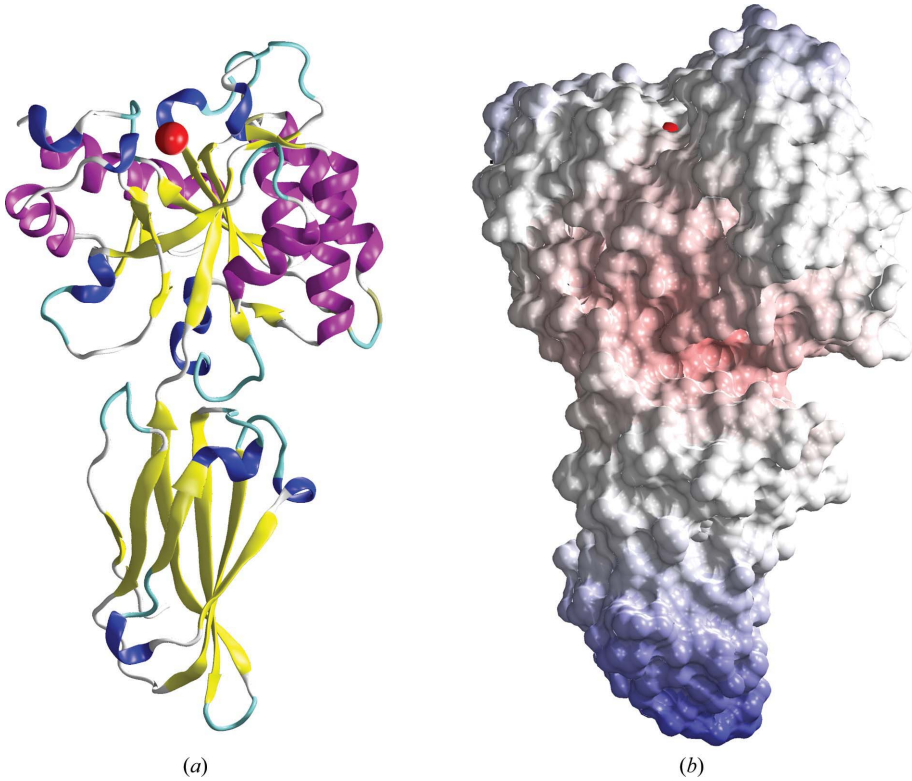
The residue in question is Pro302 and Fig. 6 shows views of the $mF_o - DF_c$ difference map at an early stage of the refinement. The difference map leaves little doubt: there is significant density (reaching up to 10.3σ above the mean density of

the map) attached to the C^{α} atom and so close to it that it is probably a direct C^{α} modification of the residue (also observed clearly in Fig. 4*b*). To tackle the question about the chemical identity of this modification we proceeded as follows.

Models were prepared of α-methyl-L-proline, α-hydroxy-L-proline and α-chloro-L-proline. All three models had acetylated N-termini and amidated C-termini to emulate the protein environment. The equilibrium geometry of the modified residues was determined with *GAMESS* (Schmidt *et al.*, 1993) at the B3LYP/6-31G(d,p) level of theory. Upon convergence, the $C^{\alpha}-OH$ distance was 1.42 Å, the $C^{\alpha}-CH_3$ distance was 1.54 Å and the $C^{\alpha}-Cl$ distance was 1.93 Å, which are in good agreement with the default bond lengths (in crystallographic dictionaries) of 1.43, 1.52 and 1.79 Å, respectively. The possibility of α-chloro-L-proline was not examined any further owing to the very long bond distance. The remaining two possibilities were (separately) incorporated into the protein model and refined with *REFMAC* to convergence. The α-hydroxy-L-proline model not only gave slightly better statistics, but the difference map was devoid of any features even at the 1.0σ level. In contrast, α-methyl-L-proline showed clear difference density at the 4σ level consistent with the presence of an erroneously long distance for the bond. These indications, together with the chemical environment of the proline (the modification group is 3.5 Å from the Zn atom in the active site; see Fig. 4) and the fact that this proline appears to be evolutionarily conserved in related proteins, allowed us to confidently model and deposit the modified proline as α-hydroxy-L-proline. One final indication concerning the putative functional importance of this modification came from a personal communication with Dr Andrew Lovering, who kindly shared his observation that this same proline modification is also present in a polysaccharide deacetylase from a different organism currently under study in his group. The question as to whether this modification is actually necessary for catalysis (and thus is the result of a previously unknown type of post-translational modification) or is a by-product of the catalytic activity of the enzyme remains to be resolved [but we should note that the *a priori* probability that we have discovered a new hitherto unknown type of post-translational modification present in this (and only this) very specific class of enzymes is very low].

5. Discussion

We have presented what we consider to be an interesting structure determination. Indeed, the original *BLAST* results (with a sequence identity of 28% for only 123 of 360 residues) would probably allow this determination to qualify as an adventurous molecular-replacement application. This brings us back to the question of how close we are to the point of only needing molecular replacement as the sole method of phase determination. Machine-learning methods (for example hidden Markov models) have greatly increased the sensitivity of similarity detection and are constantly pushing the twilight zone for confident modelling to lower sequence identities. However, this does not change the atomic accuracy requirements for structure determination using crystallographic methods. Indeed, almost all of the homology-modelling algorithms we tested gave models of the C-terminal domain that were correct in their general characteristics and (more probably than not) some of these models may have been correctly placed by the molecular-replacement programs. However, structural semblance is definitely not sufficient to successfully initiate phasing or even to allow the molecular-replacement problem to be solved confidently. What would be needed at this stage is something similar to a generalized Patterson-correlation refinement as originally proposed and implemented by Brünger



(*a*) (*b*)

Figure 5
The BC0361 structure. (*a*) shows a schematic (cartoon) representation of the structure coloured according to secondary-structure assignments by *DSSP*. The two domains are clearly visible, with the C-terminal domain at the top and the N-terminal domain below. The red sphere corresponds to the Zn atom and marks the position of the active centre. (*b*) depicts a surface representation of the protein in the same view as in (*a*). The substrate-binding cleft together with the accessible active site can be seen near the top (with the red sphere again corresponding to the Zn atom in the active site). This image was prepared with *VMD* (Humphrey *et al.*, 1996).

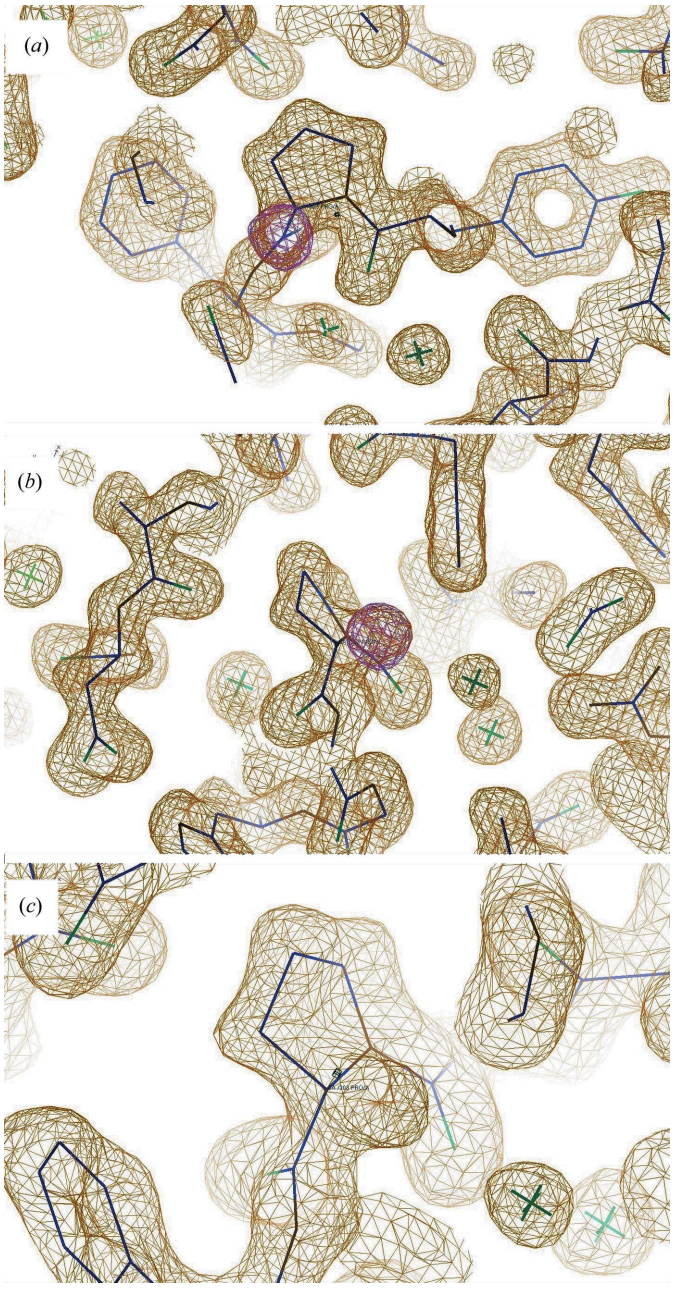


Figure 6
Pro302 and its modification. The three panels show views of the $2mF_o - DF_c$ (green) and $mF_o - DF_c$ (magenta) maps around Pro302. The difference map (only shown in *a* and *b*) is contoured at 5σ above the mean density of the map.

several years ago (Brünger, 1990) but without its limitations for high-symmetry cases. If the molecular-replacement problem can be solved, the next problem is how to refine the model (and thus the phases) while avoiding model bias. We showed how, when started at very high temperatures, torsion-angle simulated annealing can significantly improve the accuracy of the model to the point of initiating phasing. To conclude, we expect that the sheer amount of sequence-structure information present in the PDB, together with the availability of highly sensitive modelling algorithms and the development of properly tuned automated crystallographic pipelines, will significantly push the boundaries of what is currently considered to be solvable by molecular-replacement methods.

6. Data availability

The crystallographic data and the final model are available from the PDB (entry 4hd5). The *SAM-T*08-derived poly-alanine model before torsion-angle simulated annealing together with the corresponding data (in the form of an MTZ file) are available at http://utopia.duth.gr/~glykos/public/bc0361_MR.tar.

We would like to thank John S. Garavelli, Miriam Hirshberg and Andrew Lovering for sharing insight and unpublished data concerning the modified proline.

References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.
Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). *Nucleic Acids Res.* **25**, 3389–3402.
Arnold, K., Bordoli, L., Kopp, J. & Schwede, T. (2006). *Bioinformatics*, **22**, 195–201.
Blair, D. E., Schüttelkopf, A. W., MacRae, J. I. & Van Aalten, D. M. F. (2005). *Proc. Natl Acad. Sci. USA*, **102**, 15429–15434.
Brünger, A. T. (1990). *Acta Cryst.* A**46**, 46–57.
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.
Brunger, A. T., Das, D., Deacon, A. M., Grant, J., Terwilliger, T. C., Read, R. J., Adams, P. D., Levitt, M. & Schröder, G. F. (2012). *Acta Cryst.* D**68**, 391–403.
Chen, V. B., Arendall, W. B., Headd, J. J., Keedy, D. A., Immormino, R. M., Kapral, G. J., Murray, L. W., Richardson, J. S. & Richardson, D. C. (2010). *Acta Cryst.* D**66**, 12–21.
Cohen, S. X., Ben Jelloul, M., Long, F., Vagin, A., Knipscheer, P., Lebbink, J., Sixma, T. K., Lamzin, V. S., Murshudov, G. N. & Perrakis, A. (2008). *Acta Cryst.* D**64**, 49–60.
Cowtan, K. (2012). *Acta Cryst.* D**68**, 328–335.
Cuff, A. L., Sillitoe, I., Lewis, T., Clegg, A. B., Rentzsch, R., Furnham, N., Pellegrini-Calace, M., Jones, D., Thornton, J. & Orengo, C. A. (2011). *Nucleic Acids Res.* **39**, D420–D426.
DiMaio, F., Terwilliger, T. C., Read, R. J., Wlodawer, A., Oberdorfer, G., Wagner, U., Valkov, E., Alon, A., Fass, D., Axelrod, H. L., Das, D., Vorobiev, S. M., Iwaï, H., Pokkuluri, P. R. & Baker, D. (2011). *Nature (London)*, **473**, 540–543.
Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.
Evans, P. (2006). *Acta Cryst.* D**62**, 72–82.
Fadouloglou, V. E., Deli, A., Glykos, N. M., Psylinakis, E., Bouriotis, V. & Kokkinidis, M. (2007). *FEBS J.* **274**, 3044–3054.
Fadouloglou, V. E., Kokkinidis, M. & Glykos, N. M. (2008). *Anal. Biochem.* **373**, 404–406.
Fadouloglou, V. E., Stavrakoudis, S., Bouriotis, V., Kokkinidis, M. & Glykos, N. M. (2009). *J. Chem. Theory Comput.* **5**, 3299–3311.
Fiser, A. & Sali, A. (2003). *Methods Enzymol.* **374**, 461–491.
Giorgetti, A., Raimondo, D., Miele, A. E. & Tramontano, A. (2005). *Bioinformatics*, **21**, ii72–ii76.
Glykos, N. M. & Kokkinidis, M. (2000*a*). *Acta Cryst.* D**56**, 169–174.
Glykos, N. M. & Kokkinidis, M. (2000*b*). *J. Appl. Cryst.* **33**, 982–985.
Glykos, N. M. & Kokkinidis, M. (2003). *Acta Cryst.* D**59**, 709–718.
Hildebrand, A., Remmert, M., Biegert, A. & Söding, J. (2009). *Proteins*, **77**, Suppl. 9, 128–132.

Humphrey, W., Dalke, A. & Schulten, K. (1996). *J. Mol. Graph.* **14**, 33–38.
Karplus, K. (2009). *Nucleic Acids Res.* **37**, W492–W497.
Kelley, L. A. & Sternberg, M. J. (2009). *Nature Protoc.* **4**, 363–371.
Kokkinidis, M., Glykos, N. M. & Fadouloglou, V. E. (2012). *Adv. Protein Chem. Struct. Biol.* **87**, 181–218.
Lambert, C., Léonard, N., De Bolle, X. & Depiereux, E. (2002). *Bioinformatics*, **18**, 1250–1256.
Leslie, A. G. W. & Powell, H. R. (2007). *Evolving Methods for Macromolecular Crystallography*, edited by R. J. Read & J. L. Sussman, pp. 41–51. Dordrecht: Springer.
Levitt, M. (2007). *Proc. Natl Acad. Sci. USA*, **104**, 3183–3188.
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.
Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.
Nielsen, M., Lundegaard, C., Lund, O. & Petersen, T. N. (2010). *Nucleic Acids Res.* **38**, W576–W581.
Oberbarnscheidt, L., Taylor, E. J., Davies, G. J. & Gloster, T. M. (2007). *Proteins*, **66**, 250–252.
Qian, B., Raman, S., Das, R., Bradley, P., McCoy, A. J., Read, R. J. & Baker, D. (2007). *Nature (London)*, **450**, 259–264.
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.
Rykunov, D., Steinberger, E., Madrid-Aliste, C. J. & Fiser, A. (2009). *J. Struct. Funct. Genomics*, **10**, 95–99.
Schmidt, M. W., Baldridge, K. K., Boatz, J. A., Elbert, S. T., Gordon, M. S., Jensen, J. H., Koseki, S., Matsunaga, N., Nguyen, K. A., Su, S., Windus, T. L., Dupuis, M. & Montgomery, J. A. (1993). *J. Comput. Chem.* **14**, 1347–1363.
Sheldrick, G. M. (2010). *Acta Cryst.* D**66**, 479–485.
Terwilliger, T. C., Read, R. J., Adams, P. D., Brunger, A. T., Afonine, P. V., Grosse-Kunstleve, R. W. & Hung, L.-W. (2012). *Acta Cryst.* D**68**, 861–870.
Winn, M. D. *et al.* (2011). *Acta Cryst.* D**67**, 235–242.
Zhang, Y. (2008). *BMC Bioinformatics*, **9**, 40.