Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

A mixture of fortunes: the curious determination of the structure of *Escherichia coli* BL21 Gab protein

Bernhard Lohkamp and Doreen Dobritzsch*

Department of Medical Biochemistry and Biophysics, Karolinska Institutet, Sweden

Correspondence e-mail: doreen.dobritzsch@ki.se

Received 1 October 2007
Accepted 10 January 2008

PDB Reference: Gab protein, 2r6s, r2r6ssf.

In protein crystallography, monodisperse protein samples of high purity are usually required in order to obtain diffraction-quality crystals. Here, crystals were reproducibly grown from a protein sample before its homogeneity had been determined. The sample was obtained after the first attempt to purify a recombinant target protein from an *Escherichia coli* cell lysate. Subsequent analysis revealed that it was a mixture of about 50 different proteins with no predominant species. Diffraction data were collected to 2.1 Å and the space group was identified as *I*422. A molecular-replacement search with models of the expected target did not give a solution, which suggested that a contaminating *E. coli* protein had been crystallized. A PDB search revealed 256 structures determined in space group *I*422, of which 14 are *E. coli* proteins and two have unit-cell parameters similar to those observed. Molecular replacement with these structures showed a clear solution for one of them, the Gab protein. The structure is presented and compared with the deposited structure, from which it shows small but significant differences. The refined model contains bicine and sulfate as bound ligands, which provide insights into possible substrate-binding sites.

1. Introduction

Proteins crystallize through recognition processes between the protein molecules. In protein crystallography, monodisperse protein of high purity is usually required to obtain crystals for diffraction experiments, since the presence of small impurities can degrade diffraction quality (Caylor *et al.*, 1999). Studies on the effect of impurities on crystallization have mostly used lysozyme as a model protein. These investigations have revealed that impurities are often correlated with an increase in the mosaicity and temperature factors as well as the formation of ill-shaped and twinned crystals (Skouri *et al.*, 1995; Lorber *et al.*, 1993). Although in some cases the disorder caused by impurities can be reduced using a simple seeding technique (Caylor *et al.*, 1999), crystallization is usually performed with samples of the most pure protein available in order to obtain crystals of high quality and to eliminate the risk of crystallizing a protein that is not the target protein. Pure protein is also more likely to be monodisperse, which increases the probability of producing crystals (Ferré-D'Amaré & Burley, 1997). To achieve both purity and monodispersity, extensive purification protocols and elaborate screening for optimal buffer conditions are sometimes necessary. Nevertheless, there are several reports in which contaminating proteins that constitute lower than 5% of the total protein content have been crystallized instead of the target protein (see, for example, Cámara-Artigas *et al.*, 2006). These contaminating proteins are often easily crystallisable, such as

doi:10.1107/S0907444908001091

Table 1
Data-collection statistics.

Values in parentheses are for the highest resolution shell. n/a, not available.

	2r6s (this work)	1jr7 (Chance *et al.*, 2002)
Space group	*I*422	*I*422
Unit-cell parameters (Å)	*a* = *b* = 120.9, *c* = 137.2	*a* = *b* = 120.5, *c* = 136.6
Wavelength (Å)	0.931	0.9878
Resolution range (Å)	45.4–2.1 (2.21–2.10)	20.0–2.0 (2.07–2.0)
No. of observed reflections	147495 (21141)	442770 (n/a)
No. of unique reflections	29931 (4316)	33222 (n/a)
Multiplicity	4.9 (4.9)	5.0 (3.0)
$R_{\rm merge}$† (%)	13.1 (63.2)	7.9 (23.9)
$\langle I/\sigma(I)\rangle$	10.5 (2.4)	13.0 (4.2)
Completeness	99.9 (100.0)	96.9 (94.1)

† $\sum_{hkl}\sum_i |I_i(hkl) - \langle I(hkl)\rangle| / \sum_{hkl}\sum_i I_i(hkl)$, where $I_i(hkl)$ is the intensity measurement for the *i*th observation of reflection *hkl* and $\langle I(hkl)\rangle$ is the average intensity for multiple measurements for this reflection.

lysozyme. On the other hand, crystallization can be used as a step in protein purification (Jakoby & William, 1971; for an example, see Arai *et al.*, 1981), which has applications in industry (Judge *et al.*, 1995). Furthermore, microcrystallinity of protein preparations can serve as an indicator that purification by other methods has progressed (see, for example, Blundell & Johnson, 1976).

In an attempt to determine the crystal structure of dihydropyrimidinase (DHP), the second enzyme in the pyrimidine catabolic pathway in mammals, we crystallized a protein from an undefined mixture of proteins. The utilized sample was derived from a first round of chromatographic steps intended for the purification of DHP from *Dictyostelium discoideum* recombinantly expressed in *Escherichia coli*. Crystals appeared between large amounts of precipitate and were obtained before the purity of the sample had been tested by SDS–PAGE. Diffraction data collected from these crystals allowed structure determination and identification of the crystallized protein with the help of a previously deposited structure. The crystallized protein was recognized as Gab protein from an *E. coli* B strain.

2. Materials and methods

2.1. Sample preparation and crystallization

The protein sample was prepared in a similar manner to that described previously (Lohkamp *et al.*, 2006; Gojkovic *et al.*, 2003). In brief, harvested *E. coli* BL21 cells were resuspended in lysis buffer *A* [50 m*M* sodium phosphate pH 7.0, 300 m*M* NaCl, 10%(*v*/*v*) glycerol, 1 m*M* DTT, 0.1 m*M* PMSF and one EDTA-free Complete Protease Inhibitor Cocktail Tablet (Roche Diagnostics) per 30 ml of buffer] and lysed by four passes through a French press. After cell debris had been removed by centrifugation, DNA was removed by the addition of streptomycin. To remove streptomycin, the supernatant was applied onto a G-25 gel-filtration column equilibrated with buffer *B* (as buffer *A* but without DTT and protease-inhibitor cocktail). Protein was eluted with buffer *B*. The total protein-containing fractions were pooled and subjected to metal-affinity chromatography by applying the sample onto a 4 ml Ni–NTA column (Chelating Sepharose Fast Flow, Amersham Biosciences). After washing the column with buffer *B* containing 50 m*M* imidazole, protein was eluted in a linear gradient of buffer *B* containing 50–250 m*M* imidazole. Peak fractions were pooled and precipitated with ammonium sulfate (70% saturation at 273 K) to remove imidazole. The pellet was dissolved in buffer *C* [100 m*M* sodium phosphate pH 7.0, 10%(*v*/*v*) glycerol] and applied onto an S-12 gel-filtration column equilibrated with buffer *C*. Elution of the protein was achieved with buffer *C*. Peak fractions were pooled. The protein solution was changed to storage buffer [100 m*M* sodium phosphate pH 7.0, 10%(*v*/*v*) glycerol, 1 n*M* $ZnCl_2$] and concentrated to 14.5 mg ml^{-1}. The sample was aliquoted and stored at 193 K until further use.

For crystallization, the sample stock solution was diluted with buffer (50 m*M* Tris–HCl pH 7.5, 100 m*M* NaCl, 1 m*M* DTT) to a final protein concentration of 6 mg ml^{-1}. DTT was added to a final concentration of 1 m*M*. Protein crystallization was performed in 24-well plates using the hanging-drop vapour-diffusion method at 298 K. 1.5 µl protein solution was mixed with an equal amount of precipitant solution from the 1 ml reservoir. Various commercial grid screens were deployed in the search for crystallization conditions. Protein crystals appeared after approximately one week in an ammonium sulfate grid screen (pH 4–9, 0.8–3.2 *M* ammonium sulfate; Hampton Research) in a condition with 100 m*M* bicine pH 9.0 and 1.6 *M* ammonium sulfate.

2.2. Data collection and analysis

Diffraction data were collected using synchrotron radiation with a wavelength of 0.931 Å at beamline ID14-3 of the ESRF (Grenoble, France). A MAR CCD 165 mm detector was used. For data collection, the crystal was frozen in a cold nitrogen stream at 100 K. Prior to freezing, the crystal was briefly soaked in mother liquor containing 20%(*v*/*v*) glycerol. 180° of data were collected using an oscillation angle of 1°. The higher resolution diffraction spots appeared to be diffuse. The observed intensities were indexed and integrated with *MOSFLM* (Leslie, 1992). Because there were signs of radiation damage, only 125° of data were merged and scaled with *SCALA* and structure-factor amplitudes were derived using *TRUNCATE* (Collaborative Computational Project, Number 4, 1994).

The diffraction data could be indexed and scaled in *I*-centred tetragonal, *I*- and *F*-centred orthorhombic and *C*-centred monoclinic symmetry space groups. Assessment of the data with *POINTLESS* (Evans, 2006) indicated that *I*422 with unit-cell parameters *a* = *b* = 120.9, *c* = 137.2 Å is the correct space group. The data-collection statistics for this space group are given in Table 1.

2.3. Protein identification and structure refinement

The Matthews coefficient (Matthews, 1968) was calculated for space group *I*422 using the molecular weight of 56 kDa for *D. discoideum* DHP, the original target protein. This gave acceptable values for packing with one molecule per asymmetric unit. Various available models homologous to *D. discoideum* DHP were probed in molecular-replacement searches using a number of programs, *e.g. MOLREP* (Vagin & Teplyakov, 1997) and *Phaser* (Read, 2001).

A PDB search was limited to structures of *E. coli* proteins determined from crystals of space group *I*422 and with unit-cell parameters the same as those observed here within a range of ±10 Å. The two structures identified, 5-aminolevulinic acid dehydratase (PDB code 1b4e) and Gab protein (PDB code 1jr7), were utilized as search models in molecular replacement using *MOLREP* (Vagin & Teplyakov, 1997).

Iterative refinement and model building was accomplished using *REFMAC*5 (Murshudov *et al.*, 1997) and *WinCoot* (Emsley & Cowtan, 2004; Lohkamp *et al.*, 2005). Water molecules were placed in masked $2F_o - F_c$ maps using the algorithm in *Coot* and some were added manually in difference maps. For comparison, σ_A-weighted maps for 1jr7 were obtained from the Electron Density Server at Uppsala University (Kleywegt *et al.*, 2004). Structure-similarity searches were performed using *DALI* (Holm & Sander, 1995) and *SSM* (Krissinel & Henrick, 2004). Superpositions of non-identical structural models were performed using the secondary-structure matching algorithm as implemented in *SSM*. All figures were prepared with *PyMOL* (DeLano, 2002) unless stated otherwise. Refinement statistics are given in Table 2.

Table 2
Refinement and quality of the model.

	2r6s (this work)	1jr7 (Chance *et al.*, 2002)
Refinement		
R factor† (%)	16.2	19.3
Free *R* factor‡ (%)	19.9	24.0
Free *R* reflections (%)	5.1	5.3
No. of residues built	299	306
No. of protein atoms	2547	2490
No. of solvent atoms	226	320
No. of heteroatoms (Fe, bicine, sulfate, glycerol)	165	1
Mean *B* values (Å^2)		
Wilson *B* factor	29.9	20.3
Protein	36.8	22.8
Solvent	44.6	36.0
Fe	27.0	62.4
Others (bicine, sulfate, glycerol)	61.4	—
Overall	38.8	24.3
Model quality		
Ramachandran plot		
Most favoured (%)	98.1	96.4 (295/306)
Generously allowed (%)	100.0	98.7 (302/306)
Disallowed (%)	0.0	0.6 (2/306)
R.m.s.d. bond distances (Å)	0.014	0.009
R.m.s.d. angles (°)	1.481	1.8

† $R = \sum \big||F_{\rm obs}| - |F_{\rm calc}|\big| / \sum |F_{\rm obs}|$, where $F_{\rm obs}$ and $F_{\rm calc}$ are the observed and calculated structure-factor amplitudes. ‡ As for *R*, but using only a random subset of data excluded from the refinement.

3. Results and discussion

3.1. Protein crystallization and structure solution

Protein crystals were obtained reproducibly from an ammonium sulfate grid screen. The crystals were small and appeared between large amounts of precipitate. An SDS–PAGE of the utilized protein sample revealed that it consisted of approximately 50 different proteins (see below and Fig. 1). Since large amounts of precipitated protein were present in the drops, it may be concluded that a large number of proteins were precipitated, which resulted in in-drop purification of the crystallized protein. This would result in fewer interactions with other protein species and more with identical protein species, initiating the crystallization process. Here, the fact that the crystallized protein forms higher aggregates, tetramers and possible octamers (see below), may have aided the nucleation and promoted the crystallization process, especially since the oligomer symmetry corresponds to the crystallographic symmetry.

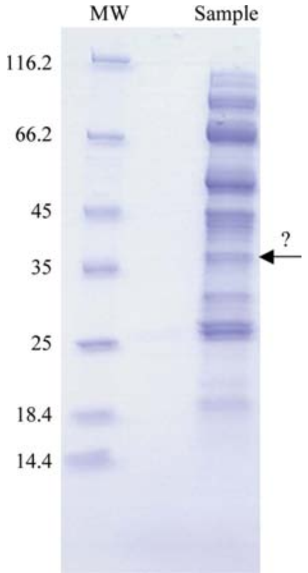

Figure 1
SDS–PAGE analysis of the sample from which the Gab protein was crystallized. Based on the molecular weight, a band corresponding to the Gab protein is marked by an arrow. The weights of the molecular-weight markers are given in kDa on the left.

No molecular-replacement solution could be found using search models similar to DHP from *D. discoideum*. This led us to investigate the crystal content in order to identify the protein. A number of crystals were dissolved and analysed by SDS–PAGE. However, owing to the small size of the crystals and the presence of precipitate in the crystallization drops, multiple faint bands were observed. These were analysed by MALDI–TOF and determined with low scores to be GTP cyclohydrolase I, triosephosphate isomerase and β-galactosidase. Molecular replacement with the deposited structures for these proteins did not yield a solution. Analysis of the crystallized protein sample by SDS–PAGE showed more than 50 proteins with no predominant species and hence was not useful in the identification of the crystallized protein (Fig. 1). These observations suggested that one of the *E. coli* proteins present in the sample had been crystallized. Therefore, the search for deposited structures in the PDB could be restricted.

Using the space group *I*422, the search revealed 256 deposited structures, of which 14 are *E. coli* proteins. Two of these structures have similar unit-cell parameters to the observed crystals: 5-aminolevulinic acid dehydratase (PDB code 1b4e; Erskine *et al.*, 1999) and Gab protein (PDB code 1jr7; Chance *et al.*, 2002). No solution was obtained for 5-aminolevulinic acid dehydratase. However, a clear solution was found for the Gab protein with an *R* factor of 35.1% for one molecule in the asymmetric unit. It is interesting to note that both structures have the same space group and unit-cell parameters despite being crystallized (using the same precipitant ammonium sulfate) at the greatly differing pH values of 5.6 and 9.0, respectively.

The Gab protein (CsiD) is encoded by the *csiD* gene (sometimes annotated as *ygaT*) and was identified as a nonhaem Fe^{II}-dependent oxygenase (Chance *et al.*, 2002). It is the first gene in the GABA operon of *E. coli*, *csiD-ygaF-gabDTPC*. The operon encodes proteins involved in GABA catabolism and uptake. GABA aminotransferase (*gabT*) and an NADP-dependent succinic semialdehyde dehydrogenase (*gabD*) catalyse the conversion from GABA to succinate. *GabP* and *gabC* encode a GABA-specific permease and a repressor, respectively. The functions of the proteins encoded by *ygaF* and *csiD* are unknown, although sequence alignments suggest that YgaF is an FAD-dependent oxidoreductase.

The initial model and electron-density map obtained from molecular replacement with 1jr7 were already of good quality, as expected for very similar or identical proteins. The Fe^{II} centre of the Gab protein was clearly visible in the electron density, as were water molecules and the ligand molecules bicine and sulfate.

The purification protocol used yielded high-purity *D. discoideum* DHP on other occasions and resulted in the successful determination of the DHP structure. However, it seems that the expression levels here were too low to yield pure DHP and instead a mixture of proteins was obtained. The number of proteins retained after the imidazole wash is

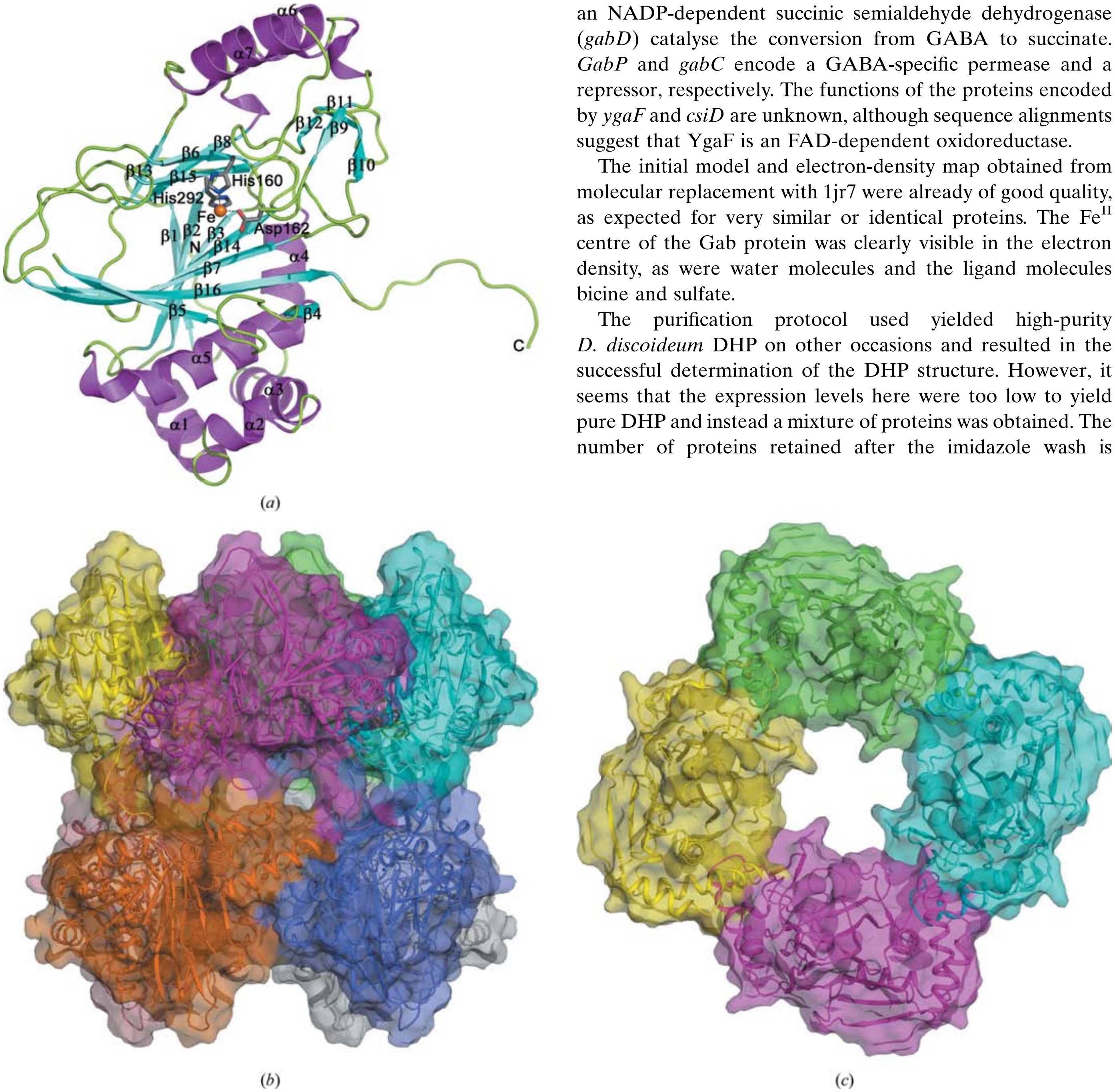

Figure 2
Gab protein monomer and potential octamer. (*a*) Gab protein monomer in ribbon representation. Fe^{II} is shown as a sphere and coordinating residues are shown in stick representation. Secondary-structure elements, iron-coordinating residues and termini are labelled. (*b*) Gab protein octamer in ribbon representation with surface. All chains are coloured differently. The top and bottom tetramers represent the smaller oligomeric units. (*c*) A top view of (*b*) with only one tetramer shown.

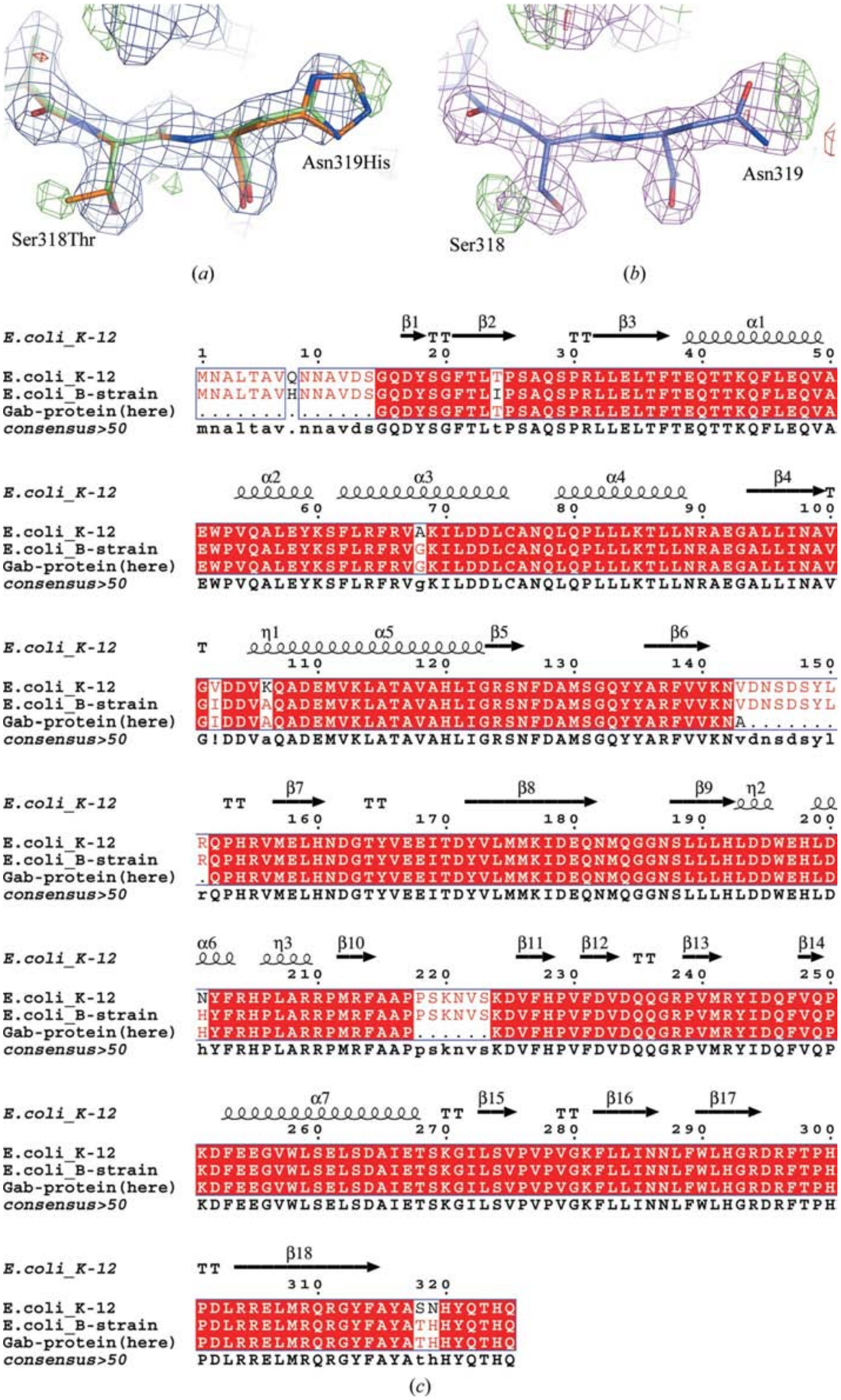
Asn319His
Ser318Thr
(a)
Asn319
Ser318
(b)
E.coli_K-12
β1 TT β2 TT β3 α1
1 10 20 30 40 50
E.coli_K-12 MNALTAVQNNAVDSGQDYSGFTLTPSAQSPRLLELTFTEQTTKQFLEQVA
E.coli_B-strain MNALTAVHNNAVDSGQDYSGFTLIPSAQSPRLLELTFTEQTTKQFLEQVA
Gab-protein(here) ....GQDYSGFTLTPSAQSPRLLELTFTEQTTKQFLEQVA
consensus>50 mnaltav.nnavdsGQDYSGFTLtPSAQSPRLLELTFTEQTTKQFLEQVA
α2 α3 α4 β4 T
60 70 80 90 100
E.coli_K-12 EWPVQALEYKSFLRFRVAKILDDLCANQLQPLLLKTLLNRAEGALLINAV
E.coli_B-strain EWPVQALEYKSFLRFRVGKILDDLCANQLQPLLLKTLLNRAEGALLINAV
Gab-protein(here) EWPVQALEYKSFLRFRVGKILDDLCANQLQPLLLKTLLNRAEGALLINAV
consensus>50 EWPVQALEYKSFLRFRVgKILDDLCANQLQPLLLKTLLNRAEGALLINAV
T η1 α5 β5 β6
110 120 130 140 150
E.coli_K-12 GVDDVKQADEMVKLATAVAHLIGRSNFDAMSGQYYARFVVKNVDNSDSYL
E.coli_B-strain GIDDVAQADEMVKLATAVAHLIGRSNFDAMSGQYYARFVVKNVDNSDSYL
Gab-protein(here) GIDDVAQADEMVKLATAVAHLIGRSNFDAMSGQYYARFVVKNA.......
consensus>50 G!DDVaQADEMVKLATAVAHLIGRSNFDAMSGQYYARFVVKNvdnsdsyl
TT β7 TT β8 β9 η2
160 170 180 190 200
E.coli_K-12 RQPHRVMELHNDGTYVEEITDYVLMMKIDEQNMQGGNSLLLHLDDWEHLD
E.coli_B-strain RQPHRVMELHNDGTYVEEITDYVLMMKIDEQNMQGGNSLLLHLDDWEHLD
Gab-protein(here) .QPHRVMELHNDGTYVEEITDYVLMMKIDEQNMQGGNSLLLHLDDWEHLD
consensus>50 rQPHRVMELHNDGTYVEEITDYVLMMKIDEQNMQGGNSLLLHLDDWEHLD
α6 η3 β10 β11 β12 TT β13 β14
210 220 230 240 250
E.coli_K-12 NYFRHPLARRPMRFAAPPSKNVSKDVFHPVFDVDQQGRPVMRYIDQFVQP
E.coli_B-strain HYFRHPLARRPMRFAAPPSKNVSKDVFHPVFDVDQQGRPVMRYIDQFVQP
Gab-protein(here) HYFRHPLARRPMRFAAP......KDVFHPVFDVDQQGRPVMRYIDQFVQP
consensus>50 hYFRHPLARRPMRFAAPpsknvsKDVFHPVFDVDQQGRPVMRYIDQFVQP
α7 TT β15 TT β16 β17
260 270 280 290 300
E.coli_K-12 KDFEEGVWLSELSDAIETSKGILSVPVPVGKFLLINNLFWLHGRDRFTPH
E.coli_B-strain KDFEEGVWLSELSDAIETSKGILSVPVPVGKFLLINNLFWLHGRDRFTPH
Gab-protein(here) KDFEEGVWLSELSDAIETSKGILSVPVPVGKFLLINNLFWLHGRDRFTPH
consensus>50 KDFEEGVWLSELSDAIETSKGILSVPVPVGKFLLINNLFWLHGRDRFTPH
TT β18
310 320
E.coli_K-12 PDLRRELMRQRGYFAYASNHYQTHQ
E.coli_B-strain PDLRRELMRQRGYFAYATHHYQTHQ
Gab-protein(here) PDLRRELMRQRGYFAYATHHYQTHQ
consensus>50 PDLRRELMRQRGYFAYAtHHYQTHQ
(c)

Figure 3
Alignment of Gab protein sequences and initial electron-density maps for varying residues. (*a*) The initial electron-density map obtained after molecular replacement for Gab protein residues Thr318 and His319 together with the initial model (green C^{α} atoms) and the final model (gold C^{α} atoms). The $2F_o - F_c$ map is contoured at 1.5σ and coloured blue. The $F_o - F_c$ map is contoured at ±3σ and coloured green and red, respectively. (*b*) σ_A maps from the EDS for 1jr7 and deposited structure 1jr7 in the same orientation as in (*a*). Maps are contoured and coloured as in (*a*), except that the $2F_o - F_c$ map and model are coloured purple with light blue C^{α} atoms. (*c*) Sequence alignment of Gab protein from *E. coli* K-12 strain, B strain and as observed here. The secondary-structure elements of the Gab protein are shown on top of the alignment. The figure was produced with *ESPript* (Gouet *et al.*, 1999).

surprisingly high. This may also be caused by the limited amounts of tagged proteins present in the lysate, leading to an increased load of the Ni resin with unspecific binders owing to insufficient competition.

3.2. Structure and sequence of Gab protein

The monomer of Gab protein reveals a β-strand core folded into a distorted jelly-roll motif (Fig. 2*a*). The central β-sheet consists of seven mixed strands (β1, β2, β3, β14, β7, β16 and β5). The jelly-roll motif is completed by strands β13, β8, β15 and β6. The core β-sheet is flanked by helices α4 and α5, which together with helices α1, α2 and α3 form a helical subdomain. Helices α6 and α7 associate with the smaller four-stranded β-sheet of the jelly roll and form another subdomain which includes strands β9–12. The iron ion is bound at the top of the smaller β-sheet by residues His160, Asp162 and His292. Higher assemblies of the Gab protein monomer can be formed by crystallographic symmetry. Analyses of the possible assemblies with the *PISA* server (Krissinel & Henrick, 2005) revealed that an octamer as well as a tetramer are likely to be stable. A dimer buries 1060 Å^2 (7% of the total surface area) of the monomer surface area, a tetramer 1600 Å^2 (11%) and an octamer 3200 Å^2 (21%). A tetramer was described in the previously reported structure of Gab protein and identified as the predominant species in gel-filtration experiments (Chance *et al.*, 2002). However, the additional buried surface area in the octamer is of the same magnitude as for the tetramer, which makes the octamer a likely oligomeric state. The interface between two tetramers to form the octamer is mainly formed by interactions of the C-terminal tail (Tyr316–Gln235), resulting in 24 hydrogen bonds as well as four salt bridges. Additionally, residues Gln235–236 of a loop interact with a symmetry partner (His205–Pro207), forming a small interface. The tetramer–tetramer interface is characterized by the most likely flexible C-terminal tail, a limited number of hydrophobic interactions and a patched appearance. This makes the octamer doubtful as being the smallest physiological oligomeric state of Gab protein. It seems more likely that the tetramer observed in the gel-filtration experiment may be in equilibrium with an octamer and the crystallization conditions favour this higher oligomeric state.

Table 3
Structural similarities of Gab protein to other proteins.

Parameters were calculated with *SSM*.

Protein	PDB code	No. of aligned residues	Core r.m.s.d. (Å)	Sequence identity (%)	Oligomer
Asparagine oxygenase	2og7	253	2.7	15	Monomer
Clavaminate synthase	1drt	239	2.6	15	Monomer
Taurine/α-ketoglutarate dioxygenase	1gqw	212	2.4	14	Dimer
Putative Fe^{II}/2-oxoglutarate-dependent enzyme	1y0z	205	2.7	17	Dimer
Carbapenem synthase	1nx4	194	2.6	12	Hexamer
Asparaginyl hydroxylase	1mze	143	3.7	10	Dimer
Prolyl hydroxylase	2g19	129	2.7	6	Trimer
JmjC-domain-containing histone demethylase	2yu1	129	2.9	9	Monomer
Oxidative DNA/RNA-repair enzyme AlkB	2fd8	108	2.8	12	Monomer

During initial model building, positive difference density was observed near the side chains of residues Val102, Asn201, Ser318 and Asn319 and negative difference density near Ala68 (Fig. 3*a*). This indicates that the sequence of the Gab protein crystallized here is different to the published sequence which was used as a search model. Based on the difference density and the sequence of Gab protein from an *E. coli* expression strain (the B strain BL21) the residues in the model were mutated to Ile102, His201, Thr318, His319 and Gly68, respectively. Furthermore, Lys106 was mutated to Ala as no density was observed for the side chain and because this mutation occurs in the B-strain protein. The observed difference in sequence is not surprising since the published structure was from *E. coli* strain K-12, whereas we crystallized the Gab protein from a B strain. A sequence alignment of the observed structure is in good agreement with a published sequence of Gab protein from *E. coli* B strain and highlights the differences from *E. coli* K-12 (Fig. 3*c*).

Additional difference density was observed near the Fe^{II} centre of the Gab protein. This density was attributed to a bicine buffer molecule and to sulfate from the precipitant ammonium sulfate. Further sulfate ions were found in later stages of the refinement, as well as glycerol molecules from the cryosolution. Refinement of the model, including water molecules and other heteroatoms, yielded a final *R* factor of 16.2% and a free *R* factor of 19.9%. All residues are in the allowed region of the Ramachandran plot, with 98.1% in the most favourable region. The final model consists of 299 amino-acid residues of the 325 of the Gab protein from *E. coli*. The 14 N-terminal residues as well as two loops (143–148 and 218–223) could not be built owing to poor electron density in these regions.

3.3. Comparison with published structure

The structure of Gab protein determined here superimposes very well with the previously published structure (1jr7). The r.m.s.d. between all matching C^{α} atoms is 0.20 Å; however, deviations in equivalent C^{α} atoms of up to 1.0 Å and of other atoms of up to 6.5 Å are observed, in particular for amino acids adjacent to gaps in the models. The missing residues are not resolved in the electron density, indicating some flexibility in this part of the protein (residues 143–148; 144–148 in 1jr7). The final *R* and free *R* factors obtained for the Gab protein described here, 16.2% and 19.9%, respectively, are significantly lower than those for 1jr7 (19.3% and 24.0%, respectively). It is unlikely that these differences can solely be attributed to recent advances in the program *REFMAC*5, which was used to refine both structures, and in the programs used to integrate, merge and scale the diffraction data. Furthermore, since the resolution of 1jr7 is slightly higher at 2.0 Å and the data statistics, namely $R_{\rm merge}$ and $I/\sigma(I)$, are better for 1jr7, better refinement statistics would be expected for this model. These observations prompted a further investigation into the differences between the two structures. The coordinates of 1jr7, the maps obtained from the EDS and deposited structure factors were used for comparison. Firstly, the model presented here contains 299 residues, whereas 1jr7 consists of 306 residues of the 325 amino acids in the Gab protein from *E. coli*. Both models lack the 14 N-terminal residues; however, in 1jr7 the missing loop around residue 220 and parts of the loop around residue 145 are present. Inspection of the available electron-density map for 1jr7 shows that there is ill-defined density in these modelled regions. Additionally, two residues in one of the loops, Pro218 and Lys220, are outliers in the Ramachandran plot as shown by Ramachandran analysis using *MOLPROBITY* (Lovell *et al.*, 2003). Secondly, validation tools identified some wrong side-chain rotamers, which were confirmed by difference density peaks as well as by geometry analysis. Thirdly, near the side chains of the mutated residues in the structure presented here (*e.g.* Ala68, Asn201) 1jr7 shows difference density peaks similar to those observed here (Fig. 3*b*). This indicates that the sequence of the protein presented in the structure 1jr7 is indeed closer to that of the Gab protein from *E. coli* B strain rather than K-12 strain. Since the protein from which the structure 1jr7 resulted was cloned and purified using a His tag, it seems likely that the genomic DNA template used was not from *E. coli* K-12 strain but a B strain. Correction of rotamers, mutation of residues according to the B-strain sequence and omission of ill-defined loops from the 1jr7 model yielded *R* factors that were similar to those described here and no difference density was observed near the mutations.

3.4. Comparison with other Fe-dioxygenases, active site and potential ligand binding

The fold of the Gab protein resembles that of other Fe^{II} α-ketoglutarate-dependent oxygenases and identifies the protein as a member of the clavaminate synthase-like superfamily as classified in SCOP (Murzin *et al.*, 1995). Other structurally characterized families in this group are penicillin synthase-like (*e.g.* isopenicillin *N*-synthase; PDB code 1ips; Roach *et al.*, 1995), clavaminate synthase (PDB code 1drt; Zhang *et al.*, 2000), YhcH-like (PDB code 1jop), type II proline 3-hydroxylase (proline oxidase; PDB code 1e5s; Clifton *et al.*, 2001), TauD/TfdA-like (*e.g.* *E. coli* taurine/α-ketoglutarate dioxygenase; PDB code 1gqw; Elkins *et al.*, 2002), γ-butyrobetaine hydroxylase (*e.g.* carbapenem synthase, CarC; PDB code 1nx4; Clifton *et al.*, 2003) and hypoxia-inducible factor HIF inhibitor (FIH1; PDB code 1mze; Dann *et al.*, 2002). Structure-similarity searches in various databases reveals that Gab protein is most similar to asparagine oxygenase (PDB code 2og7; Strieker *et al.*, 2007) and clavaminate synthase (PDB code 1drt). Most of the secondary-structure elements are common in these structures and superimpose well. The exceptions are the termini, especially the C-terminus, and the loop regions that were not resolved in Gab, which fold to close the active site in the other structures (Fig. 4). Similarities of Gab protein to other oxygenases are observed but are often restricted to the central sheet, with only partially overlapping, smaller or very different subdomains. For comparison, the most similar oxygenases are listed in Table 3 together with alignment parameters. All of these structures show sequence identities of below 20%. However, similarities are revealed in *PSI-BLAST* (Altschul *et al.*, 1998) searches, particularly for γ-butyrobetaine hydroxylases, asparagine oxygenase and other dioxygenases such as taurine dioxygenase. Although Gab is not the only oxygenase in this family that shows a higher oligomeric assembly, none of the other proteins seem to exist as a tetramer or octamer.

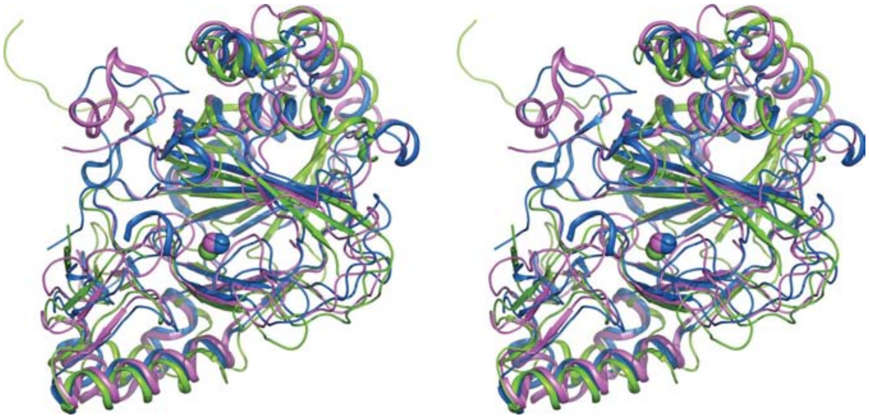

Figure 4
Stereoview of Gab protein superimposed with the structurally most similar oxygenases. Gab protein (ribbon representation) is shown in green, clavaminate synthase (PDB code 1drt) in blue and asparagine oxygenase (PDB code 2og7) in pink. Although the superposition is based on secondary structure, the iron centres superimpose well. Differences are mainly observed at the C-termini and the two loops that were not resolved in the Gab protein and that close the active site in the other oxygenases.

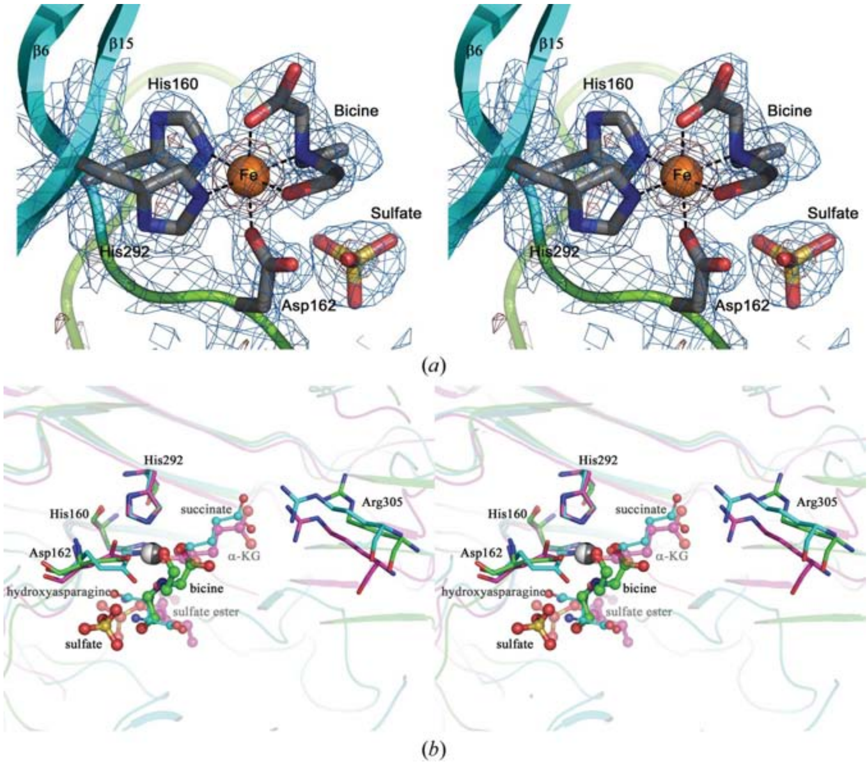

Figure 5
Stereoview of the iron coordination in Gab protein and comparison of ligand binding. (*a*) The octahedrally coordinated iron-binding site of Gab protein is shown with ligating protein residues and molecules in stick representation. A composite OMIT map is shown in blue contoured at 1.5σ and in brown contoured at 5σ. (*b*) Active site of Gab protein (green cartoon representation and C atoms) superimposed with asparagine oxygenase (cyan cartoon and C atoms) and alkylsulfatase AtsK (magenta colouring). Protein residues ligating the iron centre and the α-ketoglutarate substrate (α-KG) are shown in stick representation and labelled for the Gab protein. Ligand molecules are shown in ball-and-stick representation and labelled (black, Gab protein; dark grey, asparagine oxygenase; light grey, AtsK).

The structure of the Gab protein shows a binding site for an ion, which was identified as Fe^{II} based on homology to other Fe^{II}-dependent oxygenases. The iron is coordinated by the protein *via* a 2-His-1-carboxylate facial triad with His160 $N^{\varepsilon2}$, His292 $N^{\varepsilon2}$ and Asp162 $O^{\delta1}$ (Fig. 5*a*). Residues His160 and Asp162 are part of the characteristic and conserved His-*X*-Asp/Glu . . . His motif found in α-ketoglutarate-dependent and related oxygenases (Hegg & Que, 1997). The sequence conservation of the motif is reflected in the structural conservation of His160 from strand β6 and His292 from the neighbouring strand β15 together with Asp162 from the characteristic loop following $\beta6$ (Figs. 2*a* and 5*a*). The iron ion is nearly perfectly octahedrally coordinated, with the three remaining coordination sites occupied by bicine atoms (Fig. 5*a*). All distances of ligand atoms to the iron centre are around 2.1 Å as expected for this coordination. The iron and coordinating residues from Gab protein superimpose well with those of related oxygenases (*e.g.* those listed in Table 3). Slight deviations in the iron coordination are observed for oxygenases in which the residue equivalent to Asp162 is replaced by a Glu (*e.g.* 1y0z, 1drt, 2og7). A superposition of Gab protein with ligand complexes of related oxygenases shows that the bicine O atoms usually superimpose well with the carboxyl group at C1 and the oxo group of C2 of the α-ketoglutarate coordinating the iron. This would leave the coordination site occupied by the bicine N atom free and pointing towards the surface of a binding groove to be attacked by molecular oxygen. In these oxygenases, the carboxyl group at C5 of the α-ketoglutarate is usually bound by the guanidium group of a highly conserved Arg (here Arg305). However, in the absence of α-ketoglutarate Arg305 adopts a different conformation and forms hydrogen bonds to main-chain carbonyls as well as the side chain of Glu180 in our model of the Gab protein. In related oxygenase structures, the substrate usually coordinates the Fe^{II} where the bicine N atom binds in this case. The binding site for the substrate is found in the active-site pocket or cleft opposite to the α-ketoglutarate-binding site. Such a cleft is also present in the Gab protein. Interestingly, a sulfate ion is found here which superimposes reasonably well (sulfur distance 2.1 Å) with the sulfate of a sulfate ester ligand in alkylsulfatase AtsK (PDB code 1oik; Fig. 5*b*). However, since the substrate of Gab protein is not known it would be presumptuous to conclude that this may be a physiological sulfate- or phosphate-binding site.

4. Conclusions

It is a broadly accepted dogma that obtaining protein samples of high purity is an important aspect in successful protein crystallization. Nevertheless, small molecules, *e.g.* sugars, are readily crystallized from low-purity solution by selecting favourable conditions from assessment of their phase diagrams. Additionally, there are several reports in which attempts to crystallize protein complexes resulted in the crystallization of only one protein. Crystallizations of samples of a target protein with minor impurities have occasionally resulted in crystallization of the impurity (Cámara-Artigas *et al.*, 2006). Here, we describe for the first time the crystallization and structure determination of a protein from an uncharacterized mixture of about 50 different proteins. The crystallized protein was identified by structure solution to be the Gab protein from *E. coli*. The structure contained the ligands bicine and sulfate from the crystallization solution, clearly revealing the α-ketoglutarate-binding site as well as a putative binding site for a second substrate.

Comparison of the structure presented here with the previously deposited structure of Gab protein from *E. coli* revealed several small differences. These differences are mainly the result of imperfectly built side-chain conformers and main-chain atoms and the assignment of an incorrect primary sequence. These results emphasize not only the importance of structure validation but also the validation of the crystallized target, *e.g.* by sequencing or mass analysis.

We acknowledge the use of beamline ID14-3 at the ESRF (Grenoble, France). We would like to thank Birgit Andersen, Lund University, Sweden for technical assistance and Masatoshi Nukui, Karolinska Institutet, Sweden for help with the MALDI–TOF analysis. We would like to thank the reviewers and editors for helpful discussions during the preparation of the final manuscript.

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. (1998). *Nucleic Acids Res.* **25**, 3389–3402.

Arai, K., Yasuda, S. & Kornberg, A. (1981). *J. Biol. Chem.* **256**, 5247–5252.

Blundell, T. L. & Johnson, L. N. (1976). *Protein Crystallography*, pp. 59–82. New York: Academic Press.

Cámara-Artigas, A., Hirasawa, M., Knaff, D. B., Wang, M. & Allen, J. P. (2006). *Acta Cryst.* F**62**, 1087–1092.

Caylor, C. L., Dobrianov, I., Lemay, S. G., Kimmer, C., Kriminski, S., Finkelstein, K. D., Zipfel, W., Webb, W. W., Thomas, B. R., Chernov, A. A. & Thorne, R. E. (1999). *Proteins*, **36**, 270–281.

Chance, M. R. *et al.* (2002). *Protein Sci.* **11**, 723–738.

Clifton, I. J., Doan, L. X., Sleeman, M. C., Topf, M., Suzuki, H., Wilmouth, R. C. & Schofield, C. J. (2003). *J. Biol. Chem.* **278**, 20843–20850.

Clifton, I. J., Hsueh, L. C., Baldwin, J. E., Harlos, K. & Schofield, C. J. (2001). *Eur. J. Biochem.* **268**, 6625–6636.

Collaborative Computational Project, Number 4 (1994). *Acta Cryst.* D**50**, 760–763.

Dann, C. E. III, Bruick, R. K. & Deisenhofer, J. (2002). *Proc. Natl Acad. Sci. USA*, **99**, 15351–15356.

DeLano, W. L. (2002). *The PyMOL Molecular Graphics System.* http://www.pymol.org.

Elkins, J. M., Ryle, M. J., Clifton, I. J., Dunning Hotopp, J. C., Lloyd, J. S., Burzlaff, N. I., Baldwin, J. E., Hausinger, R. P. & Roach, P. L. (2002). *Biochemistry*, **41**, 5185–5192.

Emsley, P. & Cowtan, K. (2004). *Acta Cryst.* D**60**, 2126–2132.

Erskine, P. T., Norton, E., Cooper, J. B., Lambert, R., Coker, A., Lewis, G., Spencer, P., Sarwar, M., Wood, S. P., Warren, M. J. & Shoolingin-Jordan, P. M. (1999). *Biochemistry*, **38**, 4266–4276.

Evans, P. (2006). *Acta Cryst.* D**62**, 72–82.

Ferré-D'Amaré, A. & Burley, S. K. (1997). *Methods Enzymol.* **276**, 157–166.

Gojkovic, Z., Rislund, L., Andersen, B., Sandrini, M. P. B., Cook, P. F., Schnackerz, K. D. & Piškur, J. (2003). *Nucleic Acids Res.* **31**, 1683–1692.

Gouet, P., Courcelle, E., Stuart, D. I. & Metoz, F. (1999). *Bioinformatics*, **15**, 305–308.

Hegg, E. L. & Que, L. J. (1997). *Eur. J. Biochem.* **250**, 625–629.

Holm, L. & Sander, C. (1995). *Trends Biochem. Sci.* **20**, 478–480.

Jakoby, W. B. & William, B. J. (1971). *Methods Enzymol.* **22**, 248–252.

Judge, R. A., Johns, M. R. & White, E. T. (1995). *Biotechnol. Bioeng.* **48**, 316–323.

Kleywegt, G. J., Harris, M. R., Zou, J., Taylor, T. C., Wählby, A. & Jones, T. A. (2004). *Acta Cryst.* D**60**, 2240–2249.

Krissinel, E. & Henrick, K. (2004). *Acta Cryst.* D**60**, 2256–2268.

Krissinel, E. & Henrick, K. (2005). *CompLife 2005*, edited by M. R. Berthold, R. Glen, K. Diederichs, O. Kohlbacher & I. Fischer, pp. 163–174. Berlin, Heidelberg: Springer-Verlag.

Leslie, A. G. W. (1992). *Jnt CCP4/ESF–EAMCB Newsl. Protein Crystallogr.* **26**.

Lohkamp, B., Andersen, B., Piškur, J. & Dobritzsch, D. (2006). *Acta Cryst.* F**62**, 36–38.

Lohkamp, B., Emsley, P. & Cowtan, K. (2005). *CCP4 Newsl.* **42**, 7.

Lorber, B., Skouri, M., Munch, J.-P. & Giegé, R. (1993). *J. Cryst. Growth*, **128**, 1203–1211.

Lovell, S. C., Davis, I. W., Arendall, W. B. III, de Bakker, P. I. W., Word, J. M., Prisant, M. G., Richardson, J. S. & Richardson, D. C. (2003). *Proteins*, **50**, 437–450.

Matthews, B. W. (1968). *J. Mol. Biol.* **33**, 491–497.

Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.

Murzin, A. G., Brenner, S. E., Hubbard, T. & Chothia, C. (1995). *J. Mol. Biol.* **247**, 536–540.

Read, R. J. (2001). *Acta Cryst.* D**57**, 1373–1382.

Roach, P. L., Clifton, I. J., Fülöp, V., Harlos, K., Barton, G. J., Hajdu, J., Andersson, I., Schofield, C. J. & Baldwin, J. E. (1995). *Nature (London)*, **375**, 700–704.

Skouri, M., Lorber, B., Giegé, R., Munch, J.-P. & Candau, J. S. (1995). *J. Cryst. Growth*, **152**, 209–220.

Strieker, M., Kopp, F., Mahlert, C., Essen, L. O. & Marahiel, M. A. (2007). *ACS Chem. Biol.* **2**, 187–196.

Vagin, A. & Teplyakov, A. (1997). *J. Appl. Cryst.* **30**, 1022–1025.

Zhang, Z., Ren, J., Stammers, D. K., Baldwin, J. E., Harlos, K. & Schofield, C. J. (2000). *Nature Struct. Biol.* **7**, 127–133.