Acta Crystallographica Section D
Biological Crystallography
ISSN 0907-4449

Structure of a yeast hypothetical protein selected by a structural genomics approach

S. Eswaramoorthy, S. Gerchman, V. Graziano, H. Kycia, F. W. Studier and S. Swaminathan*

Biology Department, Brookhaven National Laboratory, Upton, NY 11973, USA

Correspondence e-mail: swami@bnl.gov

Received 2 June 2002
Accepted 1 October 2002

PDB References: YBL036C, MIRAS structure, 1b54; YBL036C, MAD structure, 1ct5.

Yeast hypothetical protein YBL036C (SWISS-PROT P38197), initially thought to be a member of an 11-protein family, was selected for crystal structure determination since no structural or functional information was available. The structure has been determined independently by MIR and MAD methods to 2.0 Å resolution. The MAD structure was determined largely through automated model building. The protein folds as a TIM barrel beginning with a long N-terminal helix, in contrast to the classic triose phosphate isomerase (TIM) structure, which begins with a β-strand. A cofactor, pyridoxal 5′-phosphate, is covalently bound near the C-terminal end of the barrel, the usual active site in TIM-barrel folds. A single-domain monomeric molecule, this yeast protein resembles the N-terminal domain of alanine racemase or ornithine decarboxylase, both of which are two-domain dimeric proteins. The yeast protein has been shown to have amino-acid racemase activity. Although selected as a member of a protein family having no obvious relationship to proteins of known structure, the protein fold turned out to be a well known and widely distributed fold. This points to the need for a more comprehensive base of structural information and better structure-modeling tools before the goal of structure prediction from amino-acid sequences can be realised. In this case, similarity to a known structure allowed inferences to be made about the structure and function of a widely distributed protein family.

1. Introduction

Protein sequences generated by the Human Genome Project and other associated projects are revealing many proteins with related sequences that can be grouped together as families of similar folds and/or functions. Some of these groups have functional information but not structural information, while others lack both. It is also possible for groups distant in sequence similarity to have structural or fold similarity. Grouping of proteins on the basis of sequence homology is based on the assumption that related proteins will have similar function. However, if structural information is available, it may be possible to determine the function even if the sequence similarity is distant (Murzin & Patthy, 1999). The function of a protein is better determined by the three-dimensional structure than by the sequence itself (Oliver, 1996). Amino-acid residues separated in sequence space come together to form a functional site in three-dimensional space. Often, distantly related sequences contain similar sequence motifs that form a functional site. Accordingly, the three-dimensional structures of proteins are required at the atomic level to ultimately understand the function. Also, protein domains with similar sequences share similar folds, although

occasionally proteins with low sequence similarity may also share similar folds (Orengo *et al.*, 1994). Hence, the determination of the three-dimensional structure of one representative member of a protein family will provide a model for the entire family and thus provide insight into their possible function (Moult & Melamud, 2000). The structural genomics approach thus provides a general approach to study the function of a class of proteins (Bork & Eisenberg, 2000). Such an approach has become possible because of technological advances in methods of cloning, expression, purification and structure determination on a large scale. The sequence information coupled with structural and functional information would provide an information-intensive database that could be used for various applications including medicine, agriculture *etc.* (Montelione & Anderson, 1999).

The aim of the Structural Genomics Project is to understand the structure and function of the several thousand proteins involved in the genome (Burley *et al.*, 1999). For a target organism we selected the yeast *Saccharomyces cerevisiae*, since it is a eukaryote with a known genome sequence and with many genes having human homologues, is easily manipulated biochemically and genetically for studies of protein function and is the focus of intensive study by a large research community who could benefit from almost any structure determined (Sanchez & Sali, 1998). Details of proteins selected for study in this Structural Genomics Project can be found on our home page http://www.proteome.bnl.gov. The crystal structure of a yeast hypothetical protein, YBL036C (SWISS-PROT P38197) referred to as P007 is presented here.

2. Materials and methods

2.1. Target selection

Yeast hypothetical protein YBL036C (SWISS-PROT P38197) was chosen for structure determination. At the time of selection, this protein belonged to a family of 11 proteins, UPF0001 (uncharacterized protein family) in SWISS-PROT, with a wide evolutionary range, including human. The sequence homology among the family members extended over almost the full length of each protein, but no functional information was available for any of them. None of the proteins had significant sequence similarity by *PSI-BLAST* analysis to any protein whose structure was available in the PDB. A structure for the yeast protein would provide a template for homology modeling of other proteins in the family and might provide clues to the function of this well conserved and widely distributed family. At the present time, Pfam lists 43 proteins as members of this family (Pfam accession No. PF01168).

2.2. Cloning and expression

Genomic DNA was isolated from the wild-type *S. cerevisiae* strain S288c obtained from the Yeast Genetics Stock Center at the University of California at Berkeley. Primers were ordered from Life Technologies. DNA polymerase PFU, purchased from Stratagene, was used for PCR reactions. All chemicals used for reagents and buffers were reagent grade and solutions were made up with Milli-Q water. All purification steps were carried out at 295 K.

The gene representing the coding sequence of P007 was obtained by PCR using yeast genomic DNA as a template. Primers were designed which were complementary to the 5′ and 3′ regions of the coding sequence and which included a *Nde*I or a *Bam*HI restriction site, respectively, for insertion into the T7 expression vector pET13a (Gerchman *et al.*, 1994). Sequences were verified with fluorescent sequencing techniques. The resultant clone was transformed into *Escherichia coli* strain B834(DE3), which contains the T7 RNA polymerase gene on the host chromosome under the control of the *lac* promoter. Protein expression resulted from induction with isopropyl-β-D-thiogalactopyranoside (IPTG; Studier *et al.*, 1990).

P07pLK 834DE3 cells were cultured in 1 l of TBYG broth (containing 10 g bacto-tryptone, 5 g NaCl, 5 g yeast extract and 0.4% glucose) in a 3 l Fernbach flask shaken at a constant temperature of 310 K. The medium was supplemented with 25 μg ml^{-1} of kanamycin to maintain the plasmid. Upon reaching an optical density of 0.6 at 600 nm, the culture was induced with IPTG to a final concentration of 0.5 m*M*. After 3 h, the culture medium was harvested by centrifugation at 7000 rev min^{-1} in a Sorvall GSA 3000 rotor for 10 min. Approximately 6 g of wet cell paste was obtained per litre of culture medium.

2.3. Purification of protein

E. coli cells (∼5 g) containing the overexpressed protein were thawed and resuspended in a homogenizer in 50 ml lysis buffer (50 m*M* Tris–HCl pH 8.0, 1 m*M* EDTA) containing 0.05 m*M* phenylmethylsulfonyl fluoride and 0.05 m*M* benzamidine hydrochloride. The cells were lysed by the addition of lysozyme (Sigma) to 1 mg ml^{-1} and 0.8%(*v*/*v*) deoxycholate to a final concentration of 0.04%. The cell suspension was incubated on ice for about 15 min or until high viscosity was attained. To the cell lysate were added 1 m*M* $MnCl_2$, 10 m*M* $MgCl_2$ and 20 μg ml^{-1} DNAse I. The mixture was incubated on ice for 15 min. Nucleic acids were precipitated with the addition of 10%(*v*/*v*) Polymin P to a final concentration of 0.5%. The mixture was centrifuged for 15 min at 12 000 rev min^{-1} in a Sorvall SS-34 rotor. The pellet was re-extracted with 25 ml of lysis buffer and centrifuged again as described above. The supernatants were combined and loaded onto a 1 × 20 cm Fractogel EMD TMAE-650 (M) column (EM Separations Technology). The target protein was eluted with a 10 column-volume linear gradient between 0 and 0.5 *M* NaCl with 25 m*M* HEPES pH 8.0. Fractions containing the protein of interest were pooled, concentrated and loaded onto two SEC columns connected in series (TSK G3000PW and G3000SW, both 21.5 mm × 60 cm) equilibrated with 25 m*M* MES pH 6.5 and 200 m*M* NaCl. Fractions containing pure protein were identified by electrophoresis on 8–25% SDS–PAGE, pooled, concentrated on a Centriprep-10 concentrator

and quantified by measuring the UV absorbance at 280 nm, with a molar extinction coefficient of 29 890 M^{-1} cm^{-1}.

2.4. Crystallization and data collection

Crystals were grown in acetate buffer pH 4.6 by the hanging-drop vapor-diffusion method. Mother liquor containing equal volumes of protein (10 mg ml^{-1}) and precipitant consisting of 0.2 *M* ammonium sulfate, 0.1 *M* sodium acetate and 30% PEG MME 2000 yielded crystals suitable for X-ray diffraction. Data from the native crystals were collected at 100 K at beamline X8C of the National Synchrotron Light Source, Brookhaven National Laboratory using a MAR Research image plate and were processed with *DENZO* (Otwinowski & Minor, 1997). $R_{\rm sym}$ is 0.064 for 17 243 reflections and the completeness is 99.5%. The crystals belong to the orthorhombic space group $P2_12_12_1$ and diffracted to better than 2.1 Å resolution. Data were collected from a number of heavy-atom derivative crystals (Au, Hg, Pt *etc.*) at NSLS beamlines X12B and X12C. For every derivative, the wavelength was adjusted to the corresponding absorption edge of the heavy atom in order to optimize the anomalous signal. Most of the derivative crystals diffracted to 2.7 Å or better.

Selenomethionine-derivatized protein was subsequently prepared and crystals were grown under similar conditions to check the feasibility of accelerated structure determination by the multiple-wavelength anomalous dispersion (MAD) procedure (Ramakrishnan & Biou, 1997). MAD data from a selenomethionine-derivative crystal were collected at the X12C beamline of the NSLS. Three data sets, at peak (0.9800 Å), edge (0.9803 Å) and remote (0.9300 Å) wavelengths, were collected with the inverse-geometry method using an automated procedure (Skinner & Sweet, 1998). Data were processed with *DENZO* and merged with *SCALEPACK* (Otwinowski & Minor, 1997). Data-collection statistics for MAD data are included in Table 1.

2.5. Structure determination and refinement

2.5.1. MIRAS method. The structure was solved by the multiple isomorphous replacement with anomalous scattering (MIRAS) method. Good heavy-atom derivative data were obtained with crystals soaked in potassium gold cyanide, mersalyl acid and potassium hexachloroplatinate(IV). The heavy-atom positions were obtained using the program *PHASES* (Furey & Swaminathan, 1997) and were refined using the isomorphous difference with the native and the anomalous difference of the derivative data. The number of heavy-atom positions obtained for each derivative and the refinement statistics are presented in Table 1. Phases were further improved by solvent flattening. The electron-density map with these phases revealed a TIM-barrel structure with a few gaps. The backbone model was built in *O* using the 'baton' option. The side chains were manually fitted into the electron density. 24 residues of the possible 257 could not be seen in the electron density and the polypeptide was discontinuous in two places.

Table 1
Crystal data and phasing statistics.

MIR data. Values in parentheses are for the outermost shell (2.2–2.12 Å).

		Derivatives		
	Native	Au	Hg	Pt
Crystal data				
Maximum resolution (Å)	2.12	2.7	2.6	2.6
Total reflections	72799	28295	20396	50390
Unique reflections	17243	8711	8493	9586
Completeness (%)	99.5 (96.4)	99.2	87.7	99.8
$R_{\rm sym}$†	0.064 (0.206)	0.060	0.048	0.108
Phasing				
$R_{\rm iso}$‡		0.115	0.223	0.057
$R_{\rm ano}$§		0.035	0.033	0.039
Resolution (Å)		3.0	3.0	3.0
No. of heavy atoms		3	2	2
Phasing power¶ (iso)		1.49	1.09	1.00
Phasing power (ano)		1.70	1.36	1.09
⟨FOM⟩††		0.514		
⟨FOM⟩ (after solvent flattening)		0.878		

MAD data. Values in parentheses are for the outermost shell (2.07–2.00 Å).

	Edge	Peak	Remote
Wavelength (Å)	0.9803	0.9800	0.9300
Maximum resolution (Å)	2.0	2.0	2.0
Total reflections	124616	127233	138645
Unique reflections	19755	19758	20510
Completeness (%)	94.3 (67.0)	94.5 (71.3)	97.5 (94.9)
$R_{\rm sym}$	0.064 (0.199)	0.065 (0.240)	0.067 (0.253)
Phasing statistics			
Resolution (Å)	2.0		
Se atoms	3		
Phasing power (iso)	4.51	3.41	
Phasing power (ano)	1.98	1.80	
⟨FOM⟩ (acentric/centric)	0.441/0.389		
⟨FOM⟩ (after solvent flattening)	0.926		

Refinement statistics.

	MIR	MAD
Refinement program	*CNS*	*REFMAC*
Resolution (Å)	50.0–2.1	12.5–2.0
No. of reflections	16442	19399
R value†	0.222	0.196
$R_{\rm free}$	0.277	0.245
No. of atoms		
Protein	1841	1815
Cofactor	15	15
Water molecules	162	206
R.m.s.d.		
Bond lengths (Å)	0.008	0.016
Bond angles	1.3°	0.035 Å

† $R_{\rm sym} = \sum_h \sum_i |I_i(h) - \langle I(h)\rangle| / \sum_h \sum_i |I_i(h)|$, where $I_i(h)$ is the intensity measurement for a reflection h and $\langle I(h)\rangle$ is the mean intensity for this reflection. ‡ $R_{\rm iso} = \sum |F_{PH}^2 - F_P^2| / \sum (F_{PH}^2 + F_P^2)$. § $R_{\rm ano} = \sum |F(+)^2 - F(-)^2| / \sum [F(+)^2 + F(-)^2]$; $R_{\rm cullis} = \sum \big||F_{PH} \pm F_P| - |F_{H,\rm cal}|\big| / \sum |F_{PH} \pm F_P|$ for centric reflections; $R_{\rm kraut} = \sum |F_{PH,\rm obs} - F_{H,\rm cal}| / |F_{PH,\rm obs}|$ for acentric reflections. ¶ Phasing power = $\langle F_H\rangle / E$(iso) or $\langle 2F''(\rm cal)\rangle / E$(ano). †† ⟨FOM⟩ is the mean figure of merit. † *R* value = $\sum_i \big||F_{i,\rm obs}| - k|F_{i,\rm cal}|\big| / \sum_i |F_{i,\rm obs}|$.

The structure was initially refined with *X-PLOR* (Brünger *et al.*, 1987), reserving 10% of the reflections for $R_{\rm free}$ calculation. The initial model gave an *R* value of 0.321 and an $R_{\rm free}$ of 0.384 after refinement. Clear density unaccounted for by the protein atoms was assumed to arise from a cofactor. The

cofactor was modeled as pyridoxal 5′-phosphate (PLP) since UV/VIS absorption spectrum measurements indicated the presence of PLP and also because addition of PLP greatly increased the protein yield during purification. The R value and $R_{\rm free}$ dropped to 0.297 and 0.346, respectively, upon the inclusion of PLP for refinement. A composite omit map was calculated at this stage using *CNS* and the model was further adjusted (Brünger *et al.*, 1998). The model was further refined with *CNS*. Finally, 161 water molecules were added in stages to account for the residual density in the σ_A-weighted $2F_o - F_c$ and $F_o - F_c$ maps and refined. The final R and $R_{\rm free}$ values are 0.222 and 0.277, respectively, for 16 442 reflections (working set) between 50 and 2.1 Å resolution.

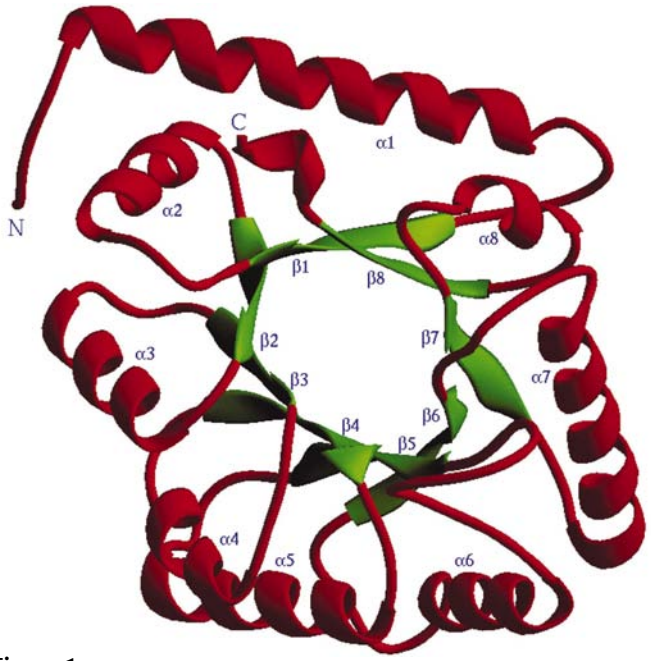

Figure 1
A *RIBBONS* (Carson, 1991) representation of the fold of the yeast hypothetical protein with the numbering of the secondary-structure motifs. α-Helices and β-strands are represented in red and green, respectively.

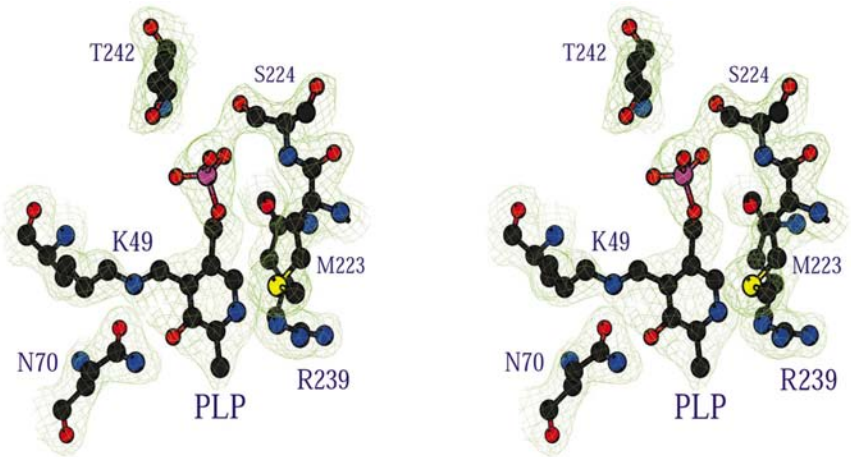

Figure 2
σ_A-weighted $2F_o - F_c$ map showing the electron density for PLP and residues in the putative active site. Contours are drawn at the 1σ level.

MAD procedure. The selenium positions were obtained with *SOLVE* (Terwilliger & Berendzen, 1999) using 3.0 Å data and the experimental phases were calculated. The phases were refined by *SHARP* (de La Fortelle & Bricogne, 1997) and further improved by solvent flattening and phase extension with *SOLOMON* (Abrahams & Leslie, 1996). The data collected at the remote wavelength (0.93 Å) were used for refinement of the model. With the experimental phases and structure amplitudes of reflections extending to 2.0 Å resolution as input, *ARP*/*wARP* (Perrakis *et al.*, 1999) completed the initial model building. The *ARP*/*wARP* model had five chains and 211 residues (of the possible 257 residues) with a connectivity index of 0.95. The side-chain building/docking of *ARP*/*wARP* placed 111 residues (53%) of the side chains correctly. Also, the electron density for the cofactor was filled with free atoms. The remaining side chains and the cofactor, pyridoxal 5′-phosphate, were built with *O*. Water molecules were added with *ARP*/*wARP* and *REFMAC* (Murshudov *et al.*, 1997) was used for refinement. The final R and $R_{\rm free}$ values are 0.196 and 0.245, respectively. 28 amino acids are missing in this model, as in the case of the MIR model.

3. Results and discussion

3.1. Structure and fold

This hypothetical protein exists as a monomer and folds as a TIM barrel (Banner & Waley, 1975), a known superfold adopted by many protein families. The structure comprises eight consecutive α/β motifs with parallel β-strands (β1–β8) forming a barrel covered by α-helices (α1–α8), as shown in Fig. 1. The electron density for the model and the cofactor is very good except for two loops and the C-terminal region (Fig. 2). The first three residues are missing because of poor electron density in that region. Five residues, Val35–Ala39, between α1 and β1, eight residues, Asn187–Lys194, between α6 and β6, and 11 residues, Ala247–Ile257, at the C-terminal are not visible in the electron-density map. Accordingly, these two loops of the protein, together with the 11 C-terminal residues, are not included in the model. The cofactor is covalently bound to Lys49 at the C-terminal end of the first strand, β1. The side-chain conformation of residues Lys100, Glu204 and Glu229 must be treated with caution as the electron density is weak for these side chains.

The TIM barrel starts with a long α-helix unlike most TIM-barrel structures, *e.g.* triose phosphate isomerase (Banner & Waley, 1975), xylanase (Natesh *et al.*, 1999) and mandelate racemase (Neidhart *et al.*, 1990) listed in TIM-DB (Pujadas & Palau, 1999), which all start with a β-strand. This type of barrel has recently been seen in a few other structures containing the cofactor pyridoxal 5′-phosphate (PLP; Kern *et al.*, 1999; Shaw *et al.*, 1997).

Table 2
Hydrogen bonds present in the core of the TIM barrel.

Residue	Residue	Distance (Å)
PLP258 O1P	Gly241 N	2.8
PLP258 O1P	Ser224 OG	2.5
PLP258 O2P	Water O	3.0
Water O	Ser224 N	2.8
PLP258 O3P	Thr242 OG	3.3
PLP258 O3	Asn70 ND2	2.7
PLP258 N1	Arg239 NH2	3.2
PLP258 N1	Arg239 NE	2.8
Arg239 NH1	Ser220 O	3.0
Arg239 NH1	Ser220 OG	3.0
Arg239 NH2	Thr182 O	3.0
Ser220 OG	Glu112 OE2	2.6
Glu112 OE1	His89 ND1	2.9
Glu112 OE1	Glu237 OE2	2.7
Glu237 OE1	Water O	2.9

3.2. Pyridoxal 5′-phosphate in TIM barrel

The PLP is covalently bound to Lys49 and is located at the C-terminal side of the barrel. Lys49, Asn70, Ser224, Arg239, Gly241 and Thr242 constitute the putative active site. It is located in the middle of the opening at the C-terminal end of the barrel, with C4A of the pyridine ring forming a covalent bond with Lys49 N^{ε} at the surface of the barrel (Fig. 3). O3 of the PLP forms a hydrogen bond with Asn70 $N^{\delta 2}$. N1 of the pyridine ring forms a hydrogen bond with N^{ε} and NH2 of Arg239. The rest of the barrel is filled with a network of hydrogen-bonding interactions between side chains. All the nitrogen centers of Arg239 are involved in hydrogen bonding. Arg239 NH1 forms a bifurcated hydrogen bond with O and O^{γ} of Ser220, while Arg239 NH2 forms a hydrogen bond with Thr182 O. Other side chains involved in hydrogen-bond network are those of Ser220, Glu112, Glu237 and His89. This hydrogen-bond network extends up to the N-terminal side of the barrel.

Residues, Ser224, Gly241 and Thr242 interact with the phosphate group of the cofactor. O1P of PLP is hydrogen bonded to Gly241 N and Ser224 O^{γ}. O2P interacts with Ala225 N through a water molecule (Wat314). O3P interacts with Thr242 $O^{\gamma 1}$. Table 2 lists the hydrogen-bonding interactions present in the core of the TIM barrel. The pyridine ring sits between hydrophobic residues Val47 and Ile91 on one side and Met223 on the other side.

3.3. Comparison with alanine racemase and ornithine decarboxylase

A search for families of structurally similar proteins (*FSSP*; Holm & Sander, 1996) brings up alanine racemase (ARC) and ornithine decarboxylase (ODC) structures with *Z* scores of 18.6 and 16.2 (high *Z* scores indicate good structural similarity); the r.m.s.d.s are 2.7 and 2.8 Å, respectively, although these proteins did not show up in the family of proteins with similar sequence during the initial target-selection procedure (Kern *et al.*, 1999; Shaw *et al.*, 1997). When compared with the classical TIM barrel of triose phosphate isomerase, the *Z* score is 9.5 and the r.m.s.d. is 3.8 Å. P007, ARC and ODC all contain an N-terminal α-helix unlike the classic TIM-barrel fold, although the length of the α-helix is longer in P007. The major difference between the yeast hypothetical protein and the other two structures is that the yeast hypothetical protein is a single-domain structure while the other two are two-domain structures with a TIM-barrel domain and a β-sheet domain. Also, ARC and ODC exist as dimers, unlike P007. The agreement between P007 and ARC was poor when the C^{α} atoms of the entire TIM barrel were used in the least-squares fit. The best matches show good agreement for four strands (β1, β2, β7 and β8) and part of two more strands (β3 and β6). The rest of the structure did not match, although the topology is the same. Fig. 4 shows the superposition of P007 and ARC. Brute-force alignment using *LSQMAN* (Kleywegt & Jones,

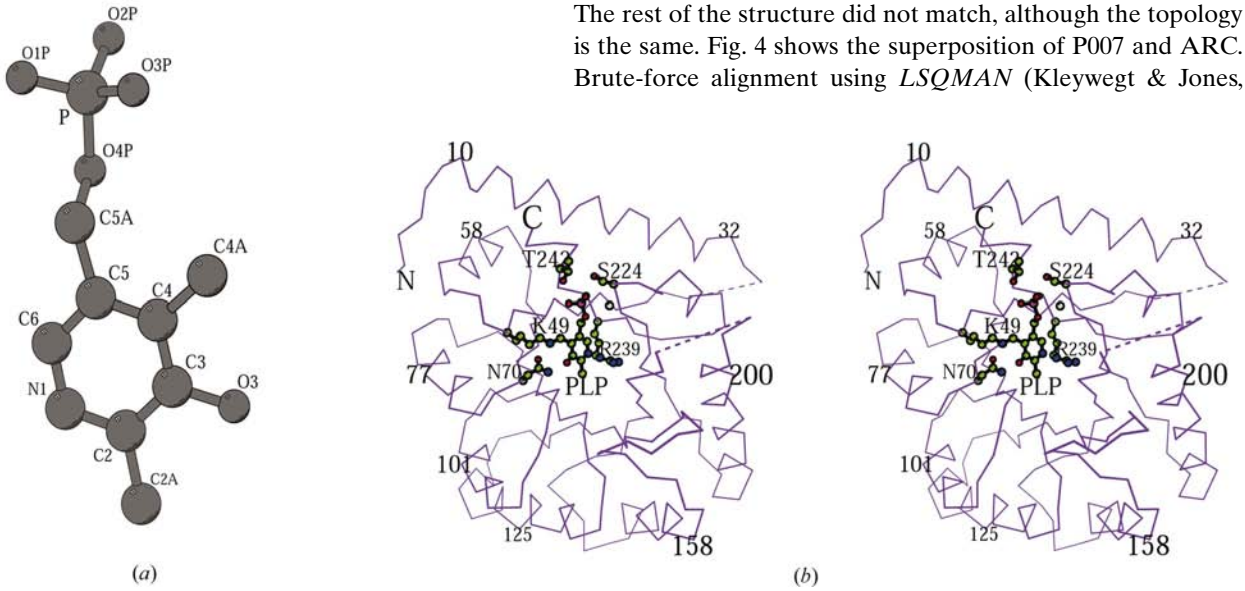

Figure 3
(*a*) A schematic diagram with the numbering scheme of PLP. (*b*) Stereoview of the active site of the yeast protein P007. The cofactor and residues involved in the active site are shown as a ball-and-stick model along with the C^{α} trace of the protein. The cofactor PLP is covalently bound to Lys49 and makes a hydrogen bond with Arg239. The phosphate group interacts with Ser224 and Thr242. The O3 of the pyridine ring makes a hydrogen bond to Asp70.

Figure 4
Superposition of P007 and the TIM-barrel domain of ARC shows similarity near the active site. The helices and the strands at the other side deviate considerably. The brute-force alignment of *LSQMAN* matched 154 atoms with an r.m.s.d. of 1.72 Å.

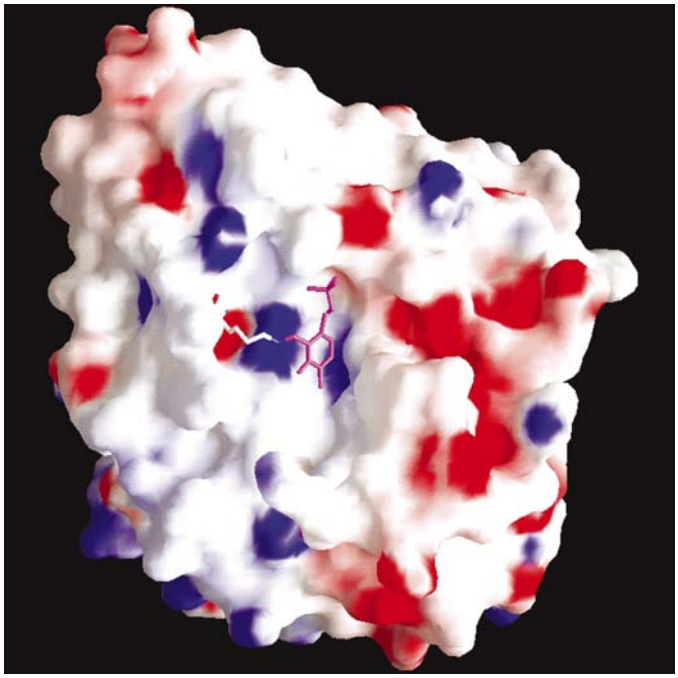

Figure 5
Surface potential (Nicholls *et al.*, 1991) and active-site cavity of P007, with the cofactor placed in the cavity. The negative and positive potentials are represented in red and blue, respectively. The stick model in pink is PLP and that in white is lysine, with N^{ε}, which makes the Schiff base, in blue.

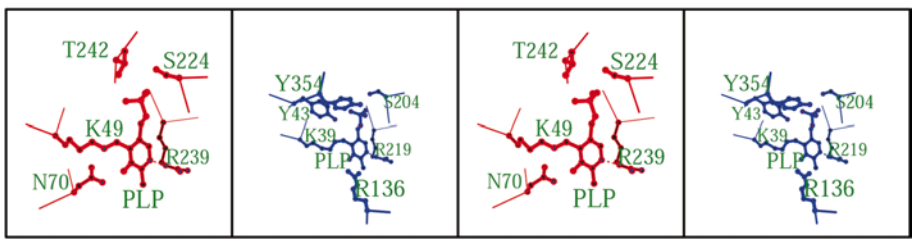

Figure 6
Comparison of the active sites of the P007 and ARC structures. Stereoview of the active sites placed side-by-side (P007 is in red and ARC is in blue). The residues Lys49, Arg239 and Ser224 of P007 are the same in ARC. Asn70 of P007 is replaced by an arginine in ARC. Additional residues Tyr43 and Tyr354 of ARC are not present in P007. Tyr265 from the second monomer of ARC is not shown in the figure.

1997) gives an r.m.s.d. of 1.72 Å for 154 C^{α} atoms. It is interesting to note that in spite of this structural similarity, the structure of P007 could not be determined by the molecular-replacement method by us using ARC as a search model.

P007 is the only single-domain structure known to contain a PLP molecule in the TIM-barrel fold. The electrostatic potential surface (Fig. 5) shows the depth of the cavity and the location of the cofactor in P007. The mode of binding of PLP in P007 is similar to that in ARC and ODC. A lysine from the C-terminal end of the first β-strand of the TIM barrel forms a Schiff base with the PLP in all three structures. The residues involved in the PLP site of P007 and ODC are different except for the Schiff base. Specifically, N1 of PLP interacts with an arginine in P007 and ARC, but with an aspartic acid in ODC. The putative active site and the conserved residues of P007 resemble ARC. The hydrogen-bonding interactions, N1 of the pyridine ring to N^{ε} of an arginine and O1P of the phosphate group to O^{γ} of a serine, are the same, in addition to the Schiff base (Fig. 6). Apart from these three conserved interactions, there are some similarities and dissimilarities. In ARC, Tyr43, Tyr354 and Arg136 interact with PLP, although Tyr354 belongs to the second domain and not to the TIM-barrel domain. The interactions of the tyrosines are through their hydroxyl groups. Spatially, Thr242 of P007 is between the positions of Tyr43 and Tyr354 and its hydroxyl group interacts with the PLP. The most important difference is the interaction of Tyr265 (not shown in the figure) of the second monomer in ARC. This interaction is absent in P007 since P007 exists as a monomer. It has been reported (Watanabe *et al.*, 1999) that Lys39 abstracts hydrogen from D-alanine to convert it into L-alanine while Tyr265 does the same thing for L-alanine. In the absence of an interaction similar to Tyr265, it is not clear how the racemization from L- to D-alanine can take place in P007. In ARC, His166 forms a bridge between Tyr265 and Arg219, and helps in making Tyr265 negatively charged (Sun & Toney, 1999), while in P007 this His166 is absent. These differences suggest that if P007 is an alanine racemase, the conversion from D to L could be explained in a similar way to that by ARC, but not the conversion from L to D. Since three of the PLP-site interactions are the same for P007 and ARC, it was suspected that P007 may be an amino-acid racemase and accordingly it was tested for D-alanine racemase activity. Though it exhibited D- to L-alanine racemase activity, further tests are required to verify L- to D-alanine activity and to rule out other racemase activities. Based on the absence of the second domain, which presumably determines the specificity, Godzik (personal

communication) speculates that P007 may be a non-specific racemase.

The structure determination of this protein helped to model 45 other proteins in this family with the TIM-barrel fold by sequence similarity using *Modeler* (Sanchez & Sali, 1998). Since the models are based on sequence similarity, neither ARC nor ODC is in the list of proteins modeled. A careful comparison of the PLP site might give some clue to the activity of these proteins. This structure has also been used as a template in *ProDom* to model a family of 42 uncharacterized proteins (including P007) that might bind to PLP. Members of the family are found in prokaryotic and eukaryotic proteins, including proline synthetase-associated protein in eubacteria and human. The sequence alignment of these proteins with

Figure 7
Sequence alignment and secondary-structural alignment of proteins modeled with P007 as template. The SWISS-PROT ID is given for each protein. Identical residues across the family are colored red. When residues are conserved in majority of the proteins or when the changes in amino-acid type are conservative, yellow is used. Residues interacting with PLP in P007 are either identical or highly conserved across the members of the family.

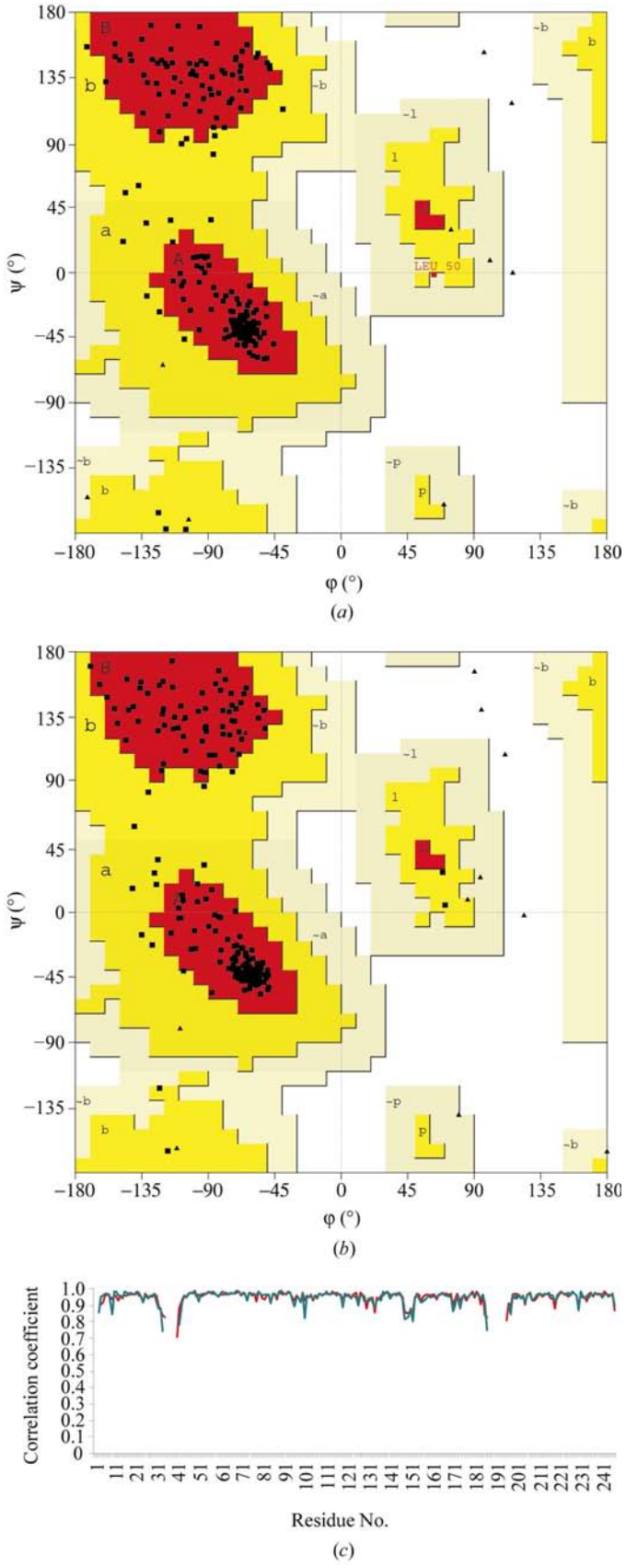

Figure 8
Ramachandran plots for (*a*) the MIR model and (*b*) the MAD model. (*c*) Real-space correlation coefficients. MIR, red; MAD, green.

Table 3
Comparison of phase sets.

Unweighted average phase differences between the experimental, solvent-flattened and final model for each method are presented. For MIRAS the experimental and solvent-flattened phases were from *PHASES*; for the others they are from *SHARP*. Except for the MIRAS experimental model, the resolution range is the same as used in the refinement.

Method	$\langle\Delta\varphi\rangle$ of experimental and final model (°)	$\langle\Delta\varphi\rangle$ of solvent-flattened and final model (°)
MIRAS	67	59.1
SIRAS	83.3	83.1
MAD	54.5	27.8
SAD	64.2	32.6

P007, together with the secondary-structural elements of P007, is given in Fig. 7. The sequence corresponding to the first α-helix does not align well with other proteins. The five residues Lys49, Asn70, Ser224, Arg239 and Gly241 involved in PLP binding in P007 are either identical or highly conserved in all of them. Also, residues corresponding to Val47, Ile91 and Met223 that sit on either side of the pyridine ring in P007 are almost conserved in the whole family. The conservation of these residues suggests that the orientation of PLP might be the same in all of them.

According to PROSITE, the consensus sequence of this family is [FW]-H-[FM]-[IV]-G-*x*-[LIV]-Q-*x*-[NKR]-K-x_3-[LIV] and extends from residues 88 to 102 of P007. Interestingly, this includes neither the conserved residue Lys49 which is covalently bound to PLP nor the other residues interacting with the PLP molecule. This is true for all the proteins shown in Fig. 7. It may be that the consensus sequence is required for substrate recognition in these proteins. The residues in the consensus sequence lie in helix α3, strand β4 and the loop connecting them. In ARC and ODC, the respective consensus sequence contains the conserved residue lysine attached to PLP. In ARC, the consensus sequence follows a left-handed helix with Lys39 at the beginning of the helix (Kleywegt, 1999). In P007, Lys49 is at the C-terminus of the β1 strand and there is no left-handed helix. However, the next residue Leu50 has positive φ–ψ values and lies just outside the generously allowed region in the Ramachandran plot, which could be a consequence of the constraint on Lys49 attached to PLP. In ODC there is also no left-handed helix for the consensus-sequence region.

3.4. Comparison of MIRAS and MAD

The crystal structure was first determined by the MIRAS method before the selenomethionine protein became available. A number of derivatives were used and the model was built manually by a conventional method using *O*. The whole process from data collection to refinement took about a month. This was largely because of the synchrotron time that was readily available to us at the time. Subsequently, selenomethionine protein became available and the structure was redetermined by the MAD method (Hendrickson & Ogata, 1997). This structure determination was performed in a high-throughput manner without any reference to the MIRAS

model. Since the model was built by *ARP*/*wARP*, the structure determination was completed in a week.

Since one of the major concerns in structural genomics projects is data-collection time and the efficient use of synchrotron beam time, a few further structure-determination trials were carried out using the available data. The SIRAS (single isomorphous replacement with anomalous scattering) method on the gold derivative was tried, as the phasing power for this derivative was high. Though the solvent-flattened phases gave a good electron-density map, the model could not be built easily as the map had lot of discontinuities. The structure was also determined by the SAD (single-wavelength anomalous dispersion) method using the data collected at the peak wavelength (data not presented). This was again performed in a high-throughput manner. This was perfomed to test whether minimum synchrotron beam time could be used for structure determination. The method worked very well and gives us confidence that SAD data are sufficient to solve the structure, as has been shown by many others. In order to estimate which experimental method gives the best initial set of phases, the experimental and solvent-flattened phases were compared with the calculated phases from the final model. In the case of the SIRAS method the final model was taken as the MIRAS model, whereas for MAD and SAD the corresponding refined models were considered. The results are presented in Table 3. The large average phase difference in the case of SIRAS is probably the reason why it was difficult to trace the polypeptide chain.

The MIR and MAD structures were considered as independent models and compared by a least-squares fit. The r.m.s. deviation between these two models is 0.266 Å (when comparing the C^{α} positions of 227 common residues). Except for lysine residues and Glu24, the side-chain conformations agreed well in the two structures. Most of the water molecules bound to the protein were well conserved. However, there was less agreement in the second-layer water molecules. Also, probably owing to the map quality, the MIR model had a lesser number of water molecules. In both cases the models are of excellent quality as judged by the Ramachandran plots (about 90% of residues in the most favored regions) and the real-space correlation coefficients between the model and the corresponding σ_A-weighted $2F_o - F_c$ electron-density maps (Fig. 8).

4. Conclusions

Proteins selected on the basis of very low sequence similarity to known structures are expected to reveal new folds in many cases. However, in the case of P007 the molecule folds in the well known and well distributed TIM-barrel fold, with the difference that P007 starts with a long N-terminal α-helix while the conventional TIM-barrel structures start with a β-strand. It is not surprising that the target selection missed molecules such as alanine racemase and ornithine decarboxylase with a similar fold and also containing PLP as cofactor, since the sequence similarity was very low. Comparison of the PLP-binding sites revealed that these sites of P007 and ARC are very similar, with only minor differences, and suggests that P007 might have a limited racemase activity. The structure determination of this protein helped in modeling several related proteins with the TIM-barrel fold by sequence similarity.

Research was supported by the National Institutes of Health (GM62529) under Prime Contract No. DEAC02-98CH10886 with the Brookhaven National Laboratory.

References

Abrahams, J. P. & Leslie, A. G. W. (1996). *Acta Cryst.* D**52**, 30–42.
Banner, D. W. & Waley, S. G. (1975). *Nature (London)*, **255**, 609–614.
Bork, P. & Eisenberg, D. (2000). *Curr. Opin. Struct. Biol.* **10**, 341–342.
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.
Brünger, A., Kuriyan, J. & Karplus, M. (1987). *Science*, **235**, 458–460.
Burley, S. K., Almo, S. C., Bonanno, J. B., Capel, M., Chance, M. R., Gaasterland, T., Lin, D., Sali, A., Studier, F. W. & Swaminathan, S. (1999). *Nature Genet.* **23**, 151–157.
Carson, M. (1991). *J. Appl. Cryst.* **24**, 958–961.
Furey, W. & Swaminathan, S. (1997). *Methods Enzymol.* **276**, 590–620.
Gerchman, S. I., Graziano, V. & Ramakrishnan, V. (1994). *Protein Expr. Purif.* **5**, 242–251.
Hendrickson, W. A. & Ogata, C. M. (1997). *Methods Enzymol.* **276**, 494–523.
Holm, L. & Sander, C. (1996). *Science*, **273**, 595–602.
Kern, A. D., Oliveira, M. A., Coffino, P. & Hackert, M. L. (1999). *Structure*, **7**, 567–581.
Kleywegt, G. J. (1999). *J. Mol. Biol.* **285**, 1887–1897.
Kleywegt, G. J. & Jones, T. A. (1997). *Methods Enzymol.* **277**, 525–545.
La Fortelle, E. de & Bricogne, G. (1997). *Methods Enzymol.* **276**, 472–493.
Montelione, G. T. & Anderson, S. (1999). *Nature Struct. Biol.* **6**, 11–12.
Moult, J. & Melamud, E. (2000). *Curr. Opin. Struct. Biol.* **10**, 384–389.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Murzin, A. G. & Patthy, L. (1999). *Curr. Opin. Struct. Biol.* **9**, 359–362.
Natesh, R., Bhanumoorthy, P., Vithayathil, P. J., Sekar, K., Ramakumar, S. & Viswamitra, M. A. (1999). *J. Mol. Biol.* **288**, 999–1012.
Neidhart, D. J., Kenyon, G. L., Gerlt, J. A. & Petsko, G. A. (1990). *Nature (London)*, **347**, 692–694.
Nicholls, A., Sharp, K. & Honig, B. (1991). *Proteins*, **11**, 281–296.
Oliver, S. G. (1996). *Nature (London)*, **379**, 597–600.
Orengo, C. A., Jones, D. T. & Thornton, J. M. (1994). *Nature (London)*, **372**, 631–634.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Perrakis, A., Morris, R. & Lamzin, V. S. (1999). *Nature Struct. Biol.* **6**, 458–463.
Ramakrishnan, V. & Biou, V. (1997). *Methods Enzymol.* **276**, 538–557.
Pujadas, G. & Palau, J. (1999). *Biologica*, **54**, 231–254.
Sanchez, R. & Sali, A. (1998). *Proc. Natl Acad. Sci. USA*, **95**, 13597–13602.
Shaw, J. P., Petsko, G. A. & Ringe, D. (1997). *Biochemistry*, **36**, 1329–1342.
Skinner, J. M. & Sweet, R. M. (1998). *Acta Cryst.* D**54**, 718–725.
Studier, F. W., Rosenberg, A. H., Dunn, J. J. & Dubendorff, J. W. (1990). *Methods Enzymol.* **185**, 61–89.
Sun, S. & Toney, M. D. (1999). *Biochemistry*, **38**, 4058–4065.
Terwilliger, T. C. & Berendzen, J. (1999). *Acta Cryst.* D**55**, 849–861.
Watanabe, A., Yoshimura, T., Mikami, B. & Esaki, N. (1999). *J. Biochem.* **126**, 781–786.