Published in final edited form as:

J Struct Biol. 2004 October ; 148(1): 98–109. doi:10.1016/j.jsb.2004.04.006.

Crystal structure of *Bacillus subtilis* YckF: structural and functional evolution[★]

R. Sanishvili^{a,1}, R. Wu^a, D.E. Kim^b, J.D. Watson^c, F. Collart^b, and A. Joachimiak^{a,*}

^a Structural Biology Center and Midwest Center for Structural Genomics, Biosciences Division, Argonne National Laboratory, 9700 South Cass Avenue, Building 202, Argonne, IL 60439, USA

^b Biosciences Division, Argonne National Laboratory, 9700 South Cass Avenue, Building 436D, Argonne, IL 60439, USA

^c EMBL—European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SD, UK

Abstract

The crystal structure of the YckF protein from *Bacillus subtilis* was determined with MAD phasing and refined at 1.95 Å resolution. YckF forms a tight tetramer both in crystals and in solution. Conservation of such oligomerization in other phosphate sugar isomerases indicates that the crystallographically observed tetramer is physiologically relevant. The structure of YckF was compared to with its ortholog from *Methanococcus jannaschii*, MJ1247. Both of these proteins have phosphate hexulose isomerase activity, although neither of the organisms can utilize methane or methanol as source of energy and/or carbon. Extensive sequence and structural similarities with MJ1247 and with the isomerase domain of glucosamine-6-phosphate synthase from *Escherichia coli* allowed us to group residues contributing to substrate binding or catalysis. Few notable differences among these structures suggest possible cooperativity of the four active sites of the tetramer. Phylogenetic relationships between obligatory and facultative methylotrophs along with *B. subtilis* and *E. coli* provide clues about the possible evolution of genes as they loose their physiological importance.

Keywords

Protein structure initiative; Crystal structure; MAD phasing; Oligomerization; Tetramer; Putative active site; Catalytic Glu-152; Evolutionary pathway; Gene hybridization; Diminished physiological role

1. Introduction

One of the central ideas of the protein structure initiative (PSI)² is to solve the three-dimensional structures of enough members of a protein family to be able to predict the structures and functions of the remaining members (Chance et al., 2002; Sanchez et al., 2000).

[★]The submitted manuscript has been created by the University of Chicago as Operator of Argonne National Laboratory (“Argonne”) under Contract No. W-31-109-ENG-38 with the US Department of Energy. The US Government retains for itself, and others acting on its behalf, a paid-up, nonexclusive, irrevocable worldwide license in said article to reproduce, prepare derivative works, distribute copies to the public, and perform publicly and display publicly, by or on behalf of the Government.

*Corresponding author. Fax: 1-630-252-5517. andrzej@anl.gov (A. Joachimiak).

¹Present address: GM/CA Collaborative Access Team, Biosciences Division, Argonne National Laboratory, 9700 South Cass Avenue, Building 436D, Argonne, IL 60439, USA.

As a result of incorporating this goal into the target selection process, most proteins studied by the PSI have no close sequence homologs among proteins with known structures. However, another purpose of this initiative, or perhaps its benefit, is to investigate mechanisms of molecular function and evolution by comparative studies of protein orthologs in different species along the phylogenetic path. This topic becomes particularly interesting when a protein, while structurally very similar to its counterparts in other species, serves a different physiological function. Good examples of such proteins are YckG and YckF from *Bacillus subtilis*. Their respective counterparts with 3-hexulose-6-phosphate synthase (HPS) and 6-phospho-3-hexuloisomerase (PHI) activities play a central role in methylotrophs—bacteria that utilize methane and methanol for energy and as a carbon source.

HPS and PHI are part of a ribulose monophosphate pathway (Dijkhuizen et al., 1992; Hanson and Hanson, 1996; Johnson and Quayle, 1965; Kemp and Quayle, 1967) in methylotrophs. PHI catalyzes the isomerization reaction between fructose-6-phosphate and hexulose-6-phosphate, while the synthesis of the latter from ribulose 5-phosphate and formaldehyde is catalyzed by HPS. It was thought that these reactions and the enzymes involved were absent in nonmethylotrophs. However, it was recently demonstrated (Yasueda et al., 1999) that YckG and YckF have formaldehyde-induced HPS and PHI activities, respectively, even though *B. subtilis* cannot use methane or methanol as an energy source. The function of YckG and YckF in *B. subtilis* may instead be formaldehyde detoxification (Yasueda et al., 1999).

By studying similar proteins from different organisms, in conjunction with their different physiological functions, interesting aspects of molecular evolution can be addressed. The Berkley Structural Genomics Center (<http://www.strgen.org/>) recently solved the crystal structure of MJ1247 from the archaebacterium *Methanococcus jannaschii* (Protein Data Bank entry 1jeo) (Martinez-Cruz et al., 2002), a protein homologous to YckF. Its activity was also shown to be similar to that of YckF. MJ1247 was tetrameric in solution and in crystals. Four putative active sites were identified, each built by three subunits of the tetramer. Here we report the crystal structure of the *yckF* gene product, YckF, at 1.95 Å resolution. Sequence and structural similarities with MJ1247 (~35%) and with the isomerase domain of glucosamine-6-phosphate synthase, GlmS (~24%) [Protein Data Bank entries 1moq (Teplyakov et al., 1998)] and 1jxa (Teplyakov et al., 2001), are discussed along with several potential implications.

2. Experimental

2.1. Cloning and purification

The open reading frame of the *B. subtilis* YckF protein was amplified from genomic DNA with a recombinant KOD HiFi DNA polymerase (Novagen) from *Thermococcus kodakaraensis* using conditions and reagents provided by the vendor (Novagen). The gene was cloned into a pMCSG7 vector (Stols et al., 2002) using a modified ligation independent cloning protocol (Dieckman et al., 2002). This process generated an expression clone producing a fusion protein with an N-terminal His₆ tag and a recognition site (ENLYFQ ↓ S) for the tobacco etch virus (TEV) protease. The fusion protein was over-produced in *Escherichia coli* BL21(DE3)/MAGIC (Wu et al., 2000) *E. coli* strain. The magic plasmid codes for three rare-triplet tRNAs (AGG for Arg, AGA for Arg, and ATA for Ile) which were controlled by a T7 promoter and a kanamycin-resistance gene. A selenomethionine (Se-Met) derivative of the expressed protein was prepared as described earlier (Walsh et al., 1999). The protein was purified by resuspension of isopropylthiogalactoside (IPTG)-induced bacterial cells in binding buffer (500mM NaCl,

²Abbreviations used: MAD, multiple-wavelength anomalous dispersion; PSI, protein structure initiative; HPS, 3-hexulose-6-phosphate synthase; PHI, 6-phospho-3-hexuloisomerase; RuMP, ribulose monophosphate pathway; GlmS, glucosamine-6-phosphate synthase; CCD, charge coupled device; BLAST, basic local alignment search tool; SSM, secondary structure matching; SIS, sugar isomerase; COG, clusters of orthologous groups; RMSD, root mean square deviation; PGI, phosphoglucose isomerase.

5% glycerol, 20mM Hepes, pH 8.0, 10mM imidazole, and 10mM β -mercaptoethanol). The cells were lysed by adding lysozyme to 1 mg/ml in the presence of a protease inhibitor mixture (Sigma P8849) (0.25 ml/5 g cells) and sonicating for 2–3 min. After clarification by centrifugation for 30 min at 17,000 rpm (Beckman) and passage through a 0.2- μ m filter, the lysate was applied to Ni-NTA Superflow resin (Qiagen) and unbound proteins removed by washing with 10 volumes of binding buffer. The protein was eluted from the column with 250mM imidazole, and the fusion tag cleaved with recombinant His-tagged TEV protease. Target protein was purified from the His tag, undigested protein and TEV protease by application of the solution to a second Ni-NTA column. The buffer in the purified protein was exchanged with 10mM Tris-HCl, pH 7.6, 50mM NaCl on a PD-10 column (Pharmacia) and concentrated using a BioMax concentrator (Millipore). Before crystallization, any particulate matter was removed from the sample by centrifugation for 20 min at 14 000 rpm at 4°C.

2.2. Size-exclusion chromatography

The molecular weight of YckF protein in solution was estimated by size-exclusion chromatography on a Superdex-200 10/30 column (Pharmacia) calibrated by ribonuclease A (13.7 kDa), chymotrypsinogen A (25.0 kDa), ovalbumin (43.0 kDa), and bovine serum albumin (67.0 kDa) as standards. The calibration curve (not shown) of K_{av} versus log molecular weight was plotted using the equation: $K_{av} = (V_e - V_o)/(V_t - V_o)$, where V_e is elution volume of the protein, V_o is the column void volume, and V_t is the total bead volume. The obtained peak corresponded to a molecular weight of 74.8 kDa—close enough to 79.6 kDa, the weight of the tetramer calculated from amino acid sequence, suggesting that YckF is a tetramer in solution.

2.3. Crystallization

Crystals of native YckF protein were grown with standard hanging-drop vapor diffusion at ambient temperatures. They appeared overnight and grew to their maximum sizes of approximately $0.20 \times 0.15 \times 0.1 \text{ mm}^3$ in 1–2 days. In the initial trials, crystals grew in over 50 conditions of five commercial screens (Crystal Screen, Crystal Screen 2, Crystal Screen Cryo—Hampton Research; and Wizard I and Wizard II—Emerald Biostructure). The vast majority were hexagonal, with slightly varying cell dimensions. There was at least one set of conditions with tetragonal crystals and one more with most likely monoclinic (although this was not verified with X-rays). A number of crystals from several different conditions were tested. Those grown from a 10 mg/ml solution of protein with 14% PEG 8k, 80mM Na cacodylate pH 6.5, 160mM Ca acetate and 20% glycerol appeared to be of best quality, diffracting slightly better than 2.0Å resolution. Crystals were mounted on cryoloops (Hampton Research) directly from the crystallization drops and flash-frozen in liquid nitrogen. Crystals belonged to the space group P6₅22 with the cell parameters at 100K $a = b = 72.08 \text{ Å}$, $c = 245.56 \text{ Å}$, and $\alpha = \beta = 90^\circ$, $\gamma = 120^\circ$. Crystals of Se-Met-modified protein grew in the same conditions and were isomorphous to those of native protein (Table 1). The Matthews coefficient (Matthews, 1968) for these crystals was 2.3, corresponding to about 45% solvent content with two protein molecules in the asymmetric unit.

2.4. Data collection

After determining the crystal parameters and its radiation tolerance as part of the procedure for optimization of diffraction experiment, a three-wavelength MAD experiment was carried out on the 19BM beamline of the Structural Biology Center at the Advanced Photon Source. A single, $\sim 0.2 \times 0.15 \times 0.1 \text{ mm}^3$ crystal of Se-Met substituted YckF protein was used. First, the fluorescence spectrum in the 25-eV range immediately before and after the Se absorption edge was recorded from the tip of the crystal. The CHOOCH program (Evans and Pettifer, 2001) was used to calculate f' values from f'' and to plot f' and f'' vs. energy. Diffraction data were collected at the Se absorption peak, at the rising inflection point, and at remote energy (13 keV)

above the absorption edge. To achieve high redundancy, 360° of data were collected at each wavelength in 180° + 180° inverse-beam geometry. The highest resolution of diffraction was not the goal at this point. Instead, determining criteria for choosing the sample-to-detector distance, the exposure time, and the rotational width of each frame were to avoid overloaded and/or overlapped reflections. Radiation damage of the sample was monitored by *R* factor, *B* factor, mosaicity, and cell parameter changes as a function of frame number. No radiation damage was observed.

An additional data set was later collected from a single crystal of native protein using the 19ID beamline of the Structural Biology Center. To avoid overloading the low resolution reflections while also collecting high resolution shells, this data set was collected in two steps. All data were measured with custom-built 3×3 tiled CCD detectors (Westbrook, 1997) with a sensitive area of 210×210mm², a fast-duty cycle (~1.7 s) and 3000×3000 pixels across its sensitive area. The d*TREK program suite (Pflugrath, 1999) was used for beamline controls, data collection, and visualization. All data were integrated and scaled with the program package HKL2000 (Otwinowski and Minor, 1997). Some of the basic statistics of data collection and processing are given in Table 1.

2.5. Sequence and structure comparisons

The amino acid sequence of YckF was submitted to BLAST (Altschul et al., 1990) at National Center for Biotechnology Information (NCBI) (www.ncbi.nlm.nih.gov/BLAST/) and a conserved domain search performed. The sequences for the matched conserved domain alignments were extracted and entered into a FASTA format file along with the sequence for YckF. The multiple sequence alignment was generated from the FASTA file using ClustalW (Higgins et al., 1994). The publication quality alignments were generated from the ClustalW output using SAShtml (Milburn et al., 1998). ClustalW was also used to obtain the phylogenetic trees, which were then viewed using Treeview (Page, 1996).

Structural comparisons were made using the SSM (www.ebi.ac.uk/msd-srv/ssm) rapid graph-matching algorithm. Active-site-based structural comparisons were carried out via a template-based method. First, a three-dimensional template was created from invariant residues in the proposed active site. Then it was rapidly scanned with JESS program (Barker, 2003) against target structures, superposing the structures by the active site residues used to develop the template.

3. Results

3.1. Structure determination and refinement

MAD phasing of YckF data up to 2.7Å resolution was carried out with the CNS program suite (Brunker et al., 1998). Using the data collected at the inflection point, phases were improved and extended to 2.5Å with density modification as implemented in CNS. It is noteworthy that apart from slightly different resolutions, the electron densities before and after density modification were almost identical, pointing to the high quality of the initial phases—probably due to the high *I*/ σ (*I*) values and high redundancy (Table 1). With these phases, the initial model was built with the ARP/wARP program (Perrakis et al., 1999). Due to the resolution limit of 2.5Å, only about 50% of the main chain and very few side chains were built automatically; nevertheless, the quality of the experimental electron densities facilitated rapid building of the remainder of the model interactively with the QUANTA program (Accelrys). The model was then refined against the 1.95Å native data with one global cycle of CNS consisting of rigid body, simulated annealing, *B*-factor, and positional refinements. This step was used to rapidly improve possible poor geometry and/or fit of parts of manually built model; and to eliminate the phase bias from a free atom model used in the model building step with ARP/wARP. All

subsequent refinements were performed with REFMAC5 (Murshudov et al., 1997), as implemented in CCP4 (Collaborative Computational Project, 1994). Phasing and refinement parameters are shown in Table 2. The coordinates of YckF have been deposited into the Protein Data Bank (PDB) (Berman et al., 2000) with the PDB entry code 1M3S.

3.2. Spatial model of YckF

The final model of YckF consists of two monomers in the asymmetric unit, molecule A and B. Residues 136–140 in molecule A and 173–175 in molecule B could not be fitted unambiguously due to disorder. The rest of the residues are well refined except for Lys-2 and Glu-24 in molecule A and Lys-121, Lys-134, Thr-176, Met-177, and Phe-178 in molecule B, whose side chains are incomplete. Several side chains have more than one conformation. Of these, Asp-31, Ser-53, Met-56, Met-59, and Leu-155 in molecule A and Ser-53, Met-59, Asn-64, Asn-114, and Asp-135 in molecule B can be reliably modeled as distinct rotamers, while the disorder of Ser-17, Ser-22, Asn-64, Met-129, Asp-173, and Met-177 in molecule A and Glu-25, Glu-78, Lys-92, and Lys-169 in molecule B is more complicated. These side chains cannot be modeled unambiguously into the existing smeared electron density, most likely because of disorder at more than one χ angle. One of the three leftover residues of TEV digestion at the N terminus was modeled as Gly-0 in both molecules. There are 203 solvent molecules in the asymmetric unit.

The monomer of YckF has a modified Rossmann fold with five β strands— β_1 (residues 39–42), β_2 (65–67), β_3 (81–85), β_4 (107–112), and β_5 (125–128)—and six α helices— α_1 (2–19), α_2 (22–36), α_3 (44–61), α_4 (91–104), α_5 (117–123), and α_6 (148–170)—arranged in $\alpha_1\alpha_2\beta_1\alpha_3\beta_2\beta_3\alpha_4\beta_4\alpha_5\beta_5\alpha_6$ topology (Fig. 1A). The β strands form a parallel β sheet with a $\beta_2\beta_1\beta_3\beta_4\beta_5$ structure, which is sandwiched between helices α_2 , α_3 , and α_6 on one side and α_4 and α_5 on the other. The second longest helix, α_1 , is extended from the globule and has very little interaction with the rest of the polypeptide chain. In our crystals, YckF forms a tight tetramer, which can be characterized as a dimer of dimers. Major contributors to the formation of the dimer are interactions of the helix α_6 of one monomer with helices α_1 and α_6 of another, and of α_3 helices from the two monomers (Figs. 1B and C). The α_1 helices extend from one monomer across to the other, in a manner somewhat reminiscent of domain swapping. Extensive interactions between these helices include the formation of hydrogen bonds together with hydrophobic and electrostatic interactions. The dimer, in turn, forms a tetramer (Fig. 1C), with its replica generated by a molecular dyad. In the tetramer, each monomer interacts with all three remaining monomers. Buried surface area between monomers A and B is approximately 4400 Å², which corresponds to about 21% per monomer. Buried surface area between monomers A and C is approximately 2300 Å² (11%) and between A and D is approximately 415 Å² (2%). Thus, the strongest interactions are within the dimers A/B or C/D, while the dimers are held together mostly by the interactions of A with C and B with D. Each monomer has approximately 35% of its surface involved in the oligomerization.

3.3. Sequence and structural homology

The NCBI BLAST search (Altschul et al., 1990) indicates strong similarity with several families of sugar phosphate isomerases, either as individual proteins or as domains of larger ones. Examples include GutQ (COG0794.1), SIS (pfam01380.8), sis (LOAD_sis.7), AgaS (COG2222.1), GlmS (COG0449.1), and transcriptional regulator RpiR (COG1737.1). The amino acid sequence alignment (Fig. 2A) identifies several residues that are strictly conserved across these families of isomerases, few of which are represented by recently solved structures:

- isomerase (SIS) domain [PDB entry 1moq (Teplyakov et al., 1998)] of glucosamine-6-phosphate synthase (GlmS),
- the entire GlmS from *E. coli* (PDB entry 1jxa (Teplyakov et al., 2001)),

- GutQ [PDB entry 1je0 (Martinez-Cruz et al., 2002)],
- sis [PDB entries 1c7q and 1c7r (Chou et al., 2000)].

The crystal structure of YckF is remarkably similar to that of its ortholog from *M. jannaschii*, MJ1247 (1je0) (Figs. 1B and 3B), and to the isomerase domain of GlmS (1moq, 1jxa) (Fig. 3), while sequence identity is ~35 and ~24%, respectively. The root mean square deviation (RMSD) between 415 matching C α atoms of YckF and MJ1247 tetramers, excluding the first and the last α helices of all monomers, is 1.63Å. Monomers of these two proteins superimpose with the RMSD of 1.35Å for 157 atoms (molecule A of YckF). For comparison, the RMSD between the two monomers in the asymmetric unit of YckF crystals is 0.57Å for 177 C α atoms. The main differences between YckF and MJ1247 are concentrated in the segments of the molecules that are somewhat extended from the globule. The N-terminal helix α_1 (residues 2–19 in YckF nomenclature) and adjacent residues are drawn closer to the globule in MJ1247 and the orientation of α_1 helices in two proteins differs by approximately 30° (Fig. 1B). The extended stretch of polypeptide chain at the C-terminus, 168–180, contains a short α helix in MJ1247, while in YckF it contains two β turns. The poorly ordered region of 131–148 between β_5 and α_6 is longer in YckF and extended into the solvent. It is noteworthy that each of these regions is also characterized by the most variability in amino acid sequence across several phosphate hexulose isomerase (PHI) proteins (Fig. 2B). Despite the substantial differences in the orientations of helices α_1 , which play roles in the oligomerization of both YckF and MJ1247, their tetrameric architecture is virtually identical (Fig. 3B). Such similarity of crystallographic tetramers, in spite of different crystallization conditions and different crystal forms, is further indication of its stability and physiological relevance.

The functional unit of GlmS is a homodimer, each monomer of which comprises two distinct domains (Teplakov et al., 2001), glutaminase and isomerase. The isomerase domains of GlmS, forming a tight dimer, are responsible for its oligomerization, while two glutaminase domains are situated on either side of the central dimer. The monomer of the isomerase domain itself consists of two topologically identical subunits. Comparison of the YckF monomer with one of these subdomains results in an RMSD of approximately 2.24Å for 133 matching atoms. Moreover, the YckF dimer is also structurally similar to the isomerase domain of GlmS (Fig. 3), with an RMSD of 2.73Å for 274 matching C α atoms. If only the central helices (α_3 and α_6) and the β strands are compared, the RMSD is 2.22 for 243 atoms. The tetramer of YckF is also very similar to the physiological dimer of this domain, with an RMSD of 3.91Å for 463 matching atoms.

The structure of phosphoglucose isomerase (PGI) from *Bacillus stearothermophilus* (PDB entries 1c7q and 1c7r) (Chou et al., 2000), maintains the same level of sequence similarity and overall molecular architecture as YckF, MJ1247, and the isomerase domain of GlmS. However, there are significant differences as well. For example, while the functional unit of PGI is also oligomeric, it is made of heterosubunits. Therefore, despite considerable similarity of the YckF monomer with the larger of the subunits of the PGI monomer, their oligomeric makeups are quite different. There are also considerable differences in the details of the active site and of the ligand binding between MJ1247 and GlmS on one side and PGI on the other. Our discussions, therefore, will be based mainly on the structures of YckF, MJ1247, and GlmS.

4. Discussion

4.1. Amino acid conservation and active site identification

The presence of a RuMP pathway in nonmethylophs (Reizer et al., 1997), consisting of HPS and PHI activities, is intriguing because these organisms do not use methane for energy or as a carbon source. Some nonmethylophs can still utilize this pathway for detoxification purposes, as has been shown for *B. subtilis* (Yasueda et al., 1999) and *M. jannaschii* (Martinez-

Cruz et al., 2002). In *B. subtilis*, YckG and YckF enzymes have HPS and PHI activities, respectively (Yasueda et al., 1999).

Sequence analysis of the YckF protein shows strong conservation of several extended regions and isolated residues within the PHI proteins (Fig. 2). Some of the same regions show broader conservation throughout the sugar phosphate isomerase families. Structural superposition of MJ1247 from *M. jannaschii* (YckF ortholog) with GlmS, along with the fact that both had ligands bound in their crystal structures, helped identify a substrate-binding site in the former (Martinez-Cruz et al., 2002). Remarkable structural and oligomerization similarity of YckF with both the MJ1247 and GlmS isomerase domains infers the location of the active site in YckF as well. As in MJ1247, there are four deep cavities in the tetramer of YckF, containing four active sites. Residues 43–47, 70–74, 86–91, 100–130, 147–152, and 170–184, donated from three different monomers, form the active site (Fig. 4). Two of these regions from monomer A, residues 43–47 and 86–91, are loops between β_1 and α_3 , and β_3 and α_4 , respectively. They are well conserved in sugar phosphate isomerases (Fig. 2A) regardless of their substrates, indicating their importance in binding the indiscriminate phosphate group of the substrate. The phosphate group would be bound at the positive pole of helix α_3 , where residues 45–47 also provide the backbone nitrogens for more specific interactions, together with the well conserved Ser-47 side chain.

The 86–91 loop provides strategic orientation of the backbone nitrogens and of the hydroxyl groups of highly conserved Ser-88 and invariant Ser-86 and Thr-91. These three hydroxyls could also contribute to catalysis, but their strict conservation throughout the sugar-phosphate isomerase families, despite their diverse substrates, probably indicates that their primary role is in binding the phosphate group.

The Glu-90 side chain is well within the active site but it is highly variable in PHI proteins, and therefore is unlikely to have a significant catalytic role. The only other extended region that is simultaneously conserved in PHI proteins and other sugar phosphate isomerases is between Lys-100 and Pro-130 and encompasses a α_4 – β_4 – α_5 crossover. While this region does not seem to be directly involved in binding the substrate, it undoubtedly stabilizes the binding loop 86–91 and determines its conformation with numerous interactions. This crossover is also spatially conserved in both MJ1247 and GlmS. Another contribution to the active site from monomer A is the segment 147–152, which is very close to the postulated phosphate-binding site and is at the end of an extended and very flexible loop, 128–147.

Glu-152, which is invariant in PHI proteins but is not conserved in the PSI group (Fig. 2), is the most feasible nucleophile in the active site, and therefore has been postulated (Martinez-Cruz et al., 2002) to be the catalytic residue. Another clue for the involvement of Glu-152 in catalysis is its conformational rigidity as it is “squeezed” between α_3 helices from monomers A and B. Other contributors to such stabilization are hydrogen bonds and the electrostatic interactions of its carboxyl with the guanidinium groups of Arg-46 (from monomer A) and Arg-57 (from monomer B) which, in turn, are stabilized by further hydrogen bonds with the neighboring residues His-60, Ser-149, and Asp-160 (Fig. 4).

It is noteworthy that His-60 is also invariant in PHI proteins. Furthermore, Ser-149 and Glu-160 change only to Thr and Asp, respectively, conserving the hydroxyl groups. Arg-57 and His-60 are part of the well preserved C-terminal end of α_3 , which is donated by monomer B and forms a bottom of the active site (Fig. 4). Another contribution of the monomer B is an irregular polypeptide chain from Lys-170 at the C-terminal end of the molecule. Together with the loop 128–147, it may play a role in providing conformational flexibility, which may be needed for isomerization of the substrate to take place. It should be also noted that this region is somewhat removed from the binding site, and that His-181 is the only residue that can directly, or

indirectly (by stabilizing His-60 with stacking interactions) be involved in catalysis. Yet this region has a very well preserved segment at the C-terminus, which, curiously enough, interacts more with the neighboring monomer than with the active site. It is tempting to suggest that the C-terminus of the molecule is involved in some sort of cooperativity between the active sites. Indeed, it could transmit the conformational changes in the active site to the α_4 - β_4 - α_5 segment of monomer D (Fig. 4), which intimately sets the conformation of binding loop 86–91 of the active site formed between monomers D, C, and A. The side chain of His-181 and its equivalent His-176 in MJ1247 occupy different conformations, possibly because of the bound ligand in the latter structure. This may argue against the role of His-181 in stabilization of Arg-57 and for its involvement in substrate-binding and/or catalytic activity. Residues 70–74 from monomer D are also at the bottom of the active site, contributing mostly to stabilization of the oligomer with strong hydrophobic interactions with monomers A and B.

The biggest structural differences between YckF and MJ1247, concentrated in three areas— α_1 , regions 131–148, and 168–180 (in YckF numbering)—may be related to substrate binding. Indeed, the last two loops are in more “closed” conformation (Fig. 4) in MJ1247, which has a citrate bound in the active site. Also, the 131–148 loop interacts with the α_1 helix, possibly causing its rotation by about 30° from the orientation of a substrate-free tetramer. If this is the case, such reorganization would be transferred to another active site via α_1 - α_1 interactions (Figs. 1B and 4)—another strong contribution to the cooperativity of the active sites. Coincidentally, orientations of α_1 helices are much closer in MJ1247 and the GlmS isomerase domain, both of which have ligands bound in the active site. Ligand binding induced conformational changes, including movement of α helix, was reported for a phosphoglucose isomerase (Arsenieva and Jeffery, 2002; Jeffery et al., 2001; Lee et al., 2001).

4.2. Possible evolutionary path of some PHI proteins

Components of the RuMP pathway, HPS and PHI activities, have been observed in some nonmethylophs. The role of the RuMP pathway in these organisms was not immediately clear because nonmethylophs do not use methane as a carbon and/or energy source. However, a new role for this pathway has been discovered in some organisms. In *B. subtilis*, for example, it was shown (Yasueda et al., 1999) that HPS and PHI enzymes (YckG and YckF, respectively) were activated only in the presence of formaldehyde. Several experiments have suggested that this pathway is involved in formaldehyde detoxification (Yasueda et al., 1999).

In *E. coli*, the protein with sequence similarity to HPS had been identified (Reizer et al., 1997) but no enzymatic activity was detected (Yasueda et al., 1999). At the same time, no sequence homologs of PHI are known in *E. coli*. However, YckF, a PHI from *B. subtilis*, displays remarkable structural similarity with the isomerase domain of glucosamine-6-phosphate synthase (GlmS) from *E. coli*. This enzyme consists of two distinct domains, isomerase and glutaminase. The isomerase domain, in turn, is made of two nearly identical subdomains, which have been suggested to be a result of gene duplication (Teplyakov et al., 1998). Monomer or YckF has weak sequence similarity (~24%) but almost identical structure with each of these subdomains. Moreover, the supramolecular organization of the functional units of the YckF and GlmS isomerase domains (Fig. 3) and their active sites (Fig. 4) is well preserved. It is noteworthy that the sugar-phosphate isomerase and the glutaminase domains of GlmS are both structurally and functionally independent and maintain catalytic activity when separated by controlled chymotrypsin proteolysis or expressed separately (Denisot et al., 1991; Leriche et al., 1996).

When following a declining role of PHI and HPS proteins from methylophs to *B. subtilis* to *E. coli*, it is tempting to suggest that as its physiological importance diminished, YckF was combined with another protein to form GlmS in *E. coli*. This theory would be consistent with

the fact that *E. coli* contains a “dysfunctional” HPS sequence, which has either mutated enough to render it inactive or has acquired a different function (Yasueda et al., 1999). Disappearance of the PHI half of the RuMP pathway via its incorporation into a larger, multifunctional protein could be a result of, or a trigger for, extensive changes in the HPS half.

A similar situation exists in the molybdopterin pathway of various organisms. In *E. coli*, the MoaB protein does not have a critical physiological role and is replaced by, or is redundant with, another protein, MogA, to which it has extensive sequence and structural similarity (Sanishvili et al., submitted). In higher organisms, MoaB appears to be a part of larger, multifunctional protein—Cnx1 in plants, cinnamoin in insects, and gephyrin in mammals. Each of these proteins is also involved in molybdopterin cofactor synthesis. MoaB of *E. coli* forms a tight hexamer, which is also maintained in larger and multidomain proteins (e.g., gephyrin). This trend of incorporating proteins into larger, multifunctional and multi-domain proteins is perhaps a tool for drastic changes during evolution. In that case, hybridization of genes, rather than autopolyploidy, would be a more powerful mechanism for creating new physiological processes, as discussed by Spring (2003).

5. Summary

The crystal structure of the YckF protein—6-phospho- 3-hexulose isomerase from *B. subtilis* was solved with MAD phasing and refined to 1.95Å resolution. YckF displays extensive sequence and structural similarities with other phosphate hexulose isomerases and with phosphate sugar isomerases in general.

Conservation extends to the oligomerization of these proteins as well. Such similarity allowed us to identify a putative active site in YckF. There, the phosphate-binding site is constructed by strictly conserved and characteristic loops between α helices and β strands, employing several hydroxyl groups from serine and threonine residues and a number of backbone nitrogens. From several potentially catalytic residues, Glu-152 is identified as the most probable. The role of some of the other atomic interactions was also addressed and the possible cooperativity of the four active sites in the tetramer was explored.

Sequence and structure conservation patterns, along with the change of the physiological role of PHI proteins in several organisms, may suggest an evolutionary pathway of PHI proteins.

Acknowledgments

We wish to thank all members of the Structural Biology Center at Argonne National Laboratory and of the Ontario Centre for Structural Proteomics for their help in conducting experiments; R. Laskowski for helpful discussions regarding the manuscript; and Youngchang Kim for assistance in size-exclusion chromatography. This work was supported by the National Institutes of Health (Grant GM 62414) and by the US Department of Energy, Office of Biological and Environmental Research, under contract (W-31-109- Eng-38).

References

- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J Mol Biol* 1990;215:403–410. [PubMed: 2231712]
- Arsenieva D, Jeffery CJ. Conformational changes in phosphoglucose isomerase induced by ligand binding. *J Mol Biol* 2002;323:77–84. [PubMed: 12368100]
- Barker JATJM. An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis. *Bioinformatics* 2003;19:1644–1649. [PubMed: 12967960]
- Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The protein data bank. *Nucleic Acids Res* 2000;28:235–242. [PubMed: 10592235]
- Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Kuszewski J, Nilges M, Pannu N, Read RJ, Rice LM, Simonson T, Warren GL. Crystallography & NMR System: a new

- suite for macromolecular structure determination. *Acta Crystallogr D* 1998;54:905–921. [PubMed: 9757107]
- Chance MR, Bresnick AR, Burley SK, Jiang JS, Lima CD, Sali A, Almo SC, Bonanno JB, Buglino JA, Boulton S, Chen H, Eswar N, He G, Huang R, Ilyin V, McMahan L, Pieper U, Ray S, Vidal M, Wang LK. Structural genomics: a pipeline for providing structures for the biologist. *Protein Sci* 2002;11:723–738. [PubMed: 11910018]
- Chou CC, Sun YJ, Meng M, Hsiao CD. The crystal structure of phosphoglucose isomerase/autocrine motility factor/neuroleukin complexed with its carbohydrate phosphate inhibitors suggests its substrate/receptor recognition. *J Biol Chem* 2000;275:23154–23160. [PubMed: 10770936]
- Collaborative Computational Project. The CCP4 suite: programs for protein crystallography. *Acta Crystallogr D* 1994;50:760–763. [PubMed: 15299374]
- Denisot MA, Le Goffic F, Badet B. Glucosamine 6- phosphate synthase from *Escherichia coli* yields two proteins upon limited proteolysis: identification of the glutamine amidohydrolase and 2R ketose/aldehyde isomerase-bearing domains based on their biochemical properties. *Arch Biochem Biophys* 1991;288:225–230. [PubMed: 1898018]
- Dieckman L, Gu M, Stols L, Donnelly MI, Collart FR. High throughput methods for gene cloning and expression. *Protein Express Purif* 2002;25:8–15.
- Dijkhuizen, L.; Levering, PR.; de Vries, GE. The physiology and biochemistry of aerobic methanol-utilizing Gram-negative and Gram-positive bacteria. In: Murrell, JC.; Dalton, H., editors. *Methane and methanol utilizers*. Plenum Press; New York: 1992.
- Evans G, Pettifer RF. CHOOCH: a program for deriving anomalous scattering factors from X-ray fluorescence spectra. *J Appl Crystallogr* 2001;34:82–86.
- Hanson RS, Hanson TE. Methanotrophic bacteria. *Microbiol Rev* 1996;60:439–471. [PubMed: 8801441]
- Higgins D, Gibson TJ, Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* 1994;22:4673–4680. [PubMed: 7984417]
- Jeffery CJ, Hardre R, Salmon L. Crystal structure of rabbit phosphoglucose isomerase complexed with 5-phospho-D-arabinonate identifies the role of Glu357 in catalysis. *Biochemistry* 2001;40:1560–1566. [PubMed: 11327814]
- Johnson PA, Quayle JR. Microbial growth on C1 compounds. Synthesis of cell constituents by methane- and methanol-grown *Pseudomonas methanica*. *Biochem J* 1965;95:859–867. [PubMed: 14342526]
- Kemp MB, Quayle JR. Microbial growth on C1 compounds. Uptake of (14C)formaldehyde and (14C) formate by methane-grown *Pseudomonas methanica* and determination of the hexose labelling pattern after brief incubation with (14C)methanol. *Biochem J* 1967;102:94–102. [PubMed: 6030306]
- Lee JH, Chang KZ, Patel V, Jeffery CJ. Crystal structure of rabbit phosphoglucose isomerase complexed with its substrate D-fructose 6-phosphate. *Biochemistry* 2001;40:7799–7805. [PubMed: 11425306]
- Leriche C, Badet-Denisot MA, Badet B. Characterization of a phosphoglucose isomerase-like activity associated with the carboxy-terminal domain of *Escherichia coli* glucoseamine-6-phosphate synthase. *J Am Chem Soc* 1996;118:1797–1798.
- Martinez-Cruz LA, Dreyer MK, Boisvert HY, Martinez-Chantar ML, Kim R, Kim SH. Crystal structure of MJ1247 protein from *M. Jannaschii* at 2.0 Å resolution infers a molecular function of 3-hexulose-6-phosphate isomerase. *Structure* 2002;10:195–204. [PubMed: 11839305]
- Mathews BW. Solvent content of protein crystals. *J Mol Biol* 1968;33:491–497. [PubMed: 5700707]
- Milburn D, Laskowski R, Thornton J. Sequences annotated by structure: a tool to facilitate the use of structural information in sequence analysis. *Protein Eng* 1998;11:855–859. [PubMed: 9862203]
- Murshudov GN, Vagin AA, Dodson EJ. Refinement of macromolecular structures by the maximum likelihood method. *Acta Crystallogr D* 1997;53:240–255. [PubMed: 15299926]
- Otwinowski Z, Minor W. Processing of X-ray diffraction data collected in oscillation mode. *Methods Enzymol* 1997;276:307–326.
- Page RDM. TREEVIEW: an application to display phylogenetic trees on personal computers. *Comput Appl Biosci* 1996;12:357–358. [PubMed: 8902363]
- Perrakis A, Morris R, Lamzin VS. Automated protein model building combined with iterative structure refinement. *Nat Struct Biol* 1999;6:458–463. [PubMed: 10331874]

- Pflugrath JW. The finer things in Xray diffraction data collection. *Acta Crystallogr D* 1999;55:1718–1725. [PubMed: 10531521]
- Reizer J, Reizer A, Saier MH Jr. Is the ribulose monophosphate pathway widely distributed in bacteria? *Microbiology* 1997;143 (Pt 8):2519–2520. [PubMed: 9274005]
- Sanchez R, Pieper U, Melo F, Eswar N, Marti-Renom MA, Madhusudhan MS, Mirkovic N, Sali A. Protein structure modeling for structural genomics. *Nat Struct Biol* 2000;7 (Suppl):986–990. [PubMed: 11104007]
- Sanishvili R, Beasley S, Skarina T, Glesne D, Joachimiak A, Edwards A, Savchenko A. The crystal structure of *Escherichia coli* MoaB, a member of the molybdenum cofactor biosynthesis operon. 2004submitted
- Spring J. Major transitions in evolution by genome fusions: from prokaryotes to eukaryotes, metazoans, bilaterians and vertebrates. *J Struct Funct Genomics* 2003;3:19–25. [PubMed: 12836681]
- Stols L, Gu M, Dieckman L, Raffin R, Collart FR, Donnelley MI. A new vector for high-throughput, ligation-independent cloning encoding a tobacco etch virus protease cleavage site. *Protein Express Purif* 2002;25:1–7.
- Teplyakov A, Obmolova G, Badet-Denisot MA, Badet B, Polikarpov I. Involvement of the C terminus in intramolecular nitrogen channeling in glucosamine 6-phosphate synthase: evidence from a 1.6Å structure of the isomerase domain. *Structure* 1998;6:1047–1055. [PubMed: 9739095]
- Teplyakov A, Obmolova G, Badet B, Badet-Denisot MA. Channeling of ammonia in glucosamine 6-phosphate synthase. *J Mol Biol* 2001;313:1093. [PubMed: 11700065]
- Walsh MA, Dementieva I, Evans G, Sanishvili R, Joachimiak A. Taking MAD to the extreme: ultrafast protein structure determination. *Acta Crystallogr D* 1999;55:1168–1173. [PubMed: 10329779]
- Westbrook EMN, Naday I. Charge-coupled device-based area detectors. *Methods Enzymol* 1997;276:244–268. [PubMed: 9048377]
- Wu N, Christendat D, Dharamsi A, Pai EF. Purification, crystallization and preliminary X-ray study of orotidine 5'-monophosphate decarboxylase. *Acta Crystallogr D* 2000;56:912–914. [PubMed: 10930842]
- Yasueda H, Kawahara Y, Sugimoto SI. *Bacillus subtilis* yckG and yckF encode two key enzymes of the ribulose monophosphate pathway used by methylotrophs, and yckH is required for their expression. *J Bacteriol* 1999;181:7154–7160. [PubMed: 10572115]

Fig. 1.

Crystal structure of YckF and its oligomerization. (A) On a cartoon of the YckF monomer, the α helices (blue) and β strands (yellow) are labeled along with the termini. Labels on α helices are placed closer to their N-termini. Chain break (dashed line) is caused by the residues missing from the model. (B) Superposition of the YckF and MJ1247 dimers. Blue and orange correspond to YckF monomers, whereas the whole dimer of MJ1247 is shown in gray. The helices α_1 , α_3 , and α_6 , forming the dimer, are labeled. The two dimers are superimposed very well except for three regions (see arrows). The angle between the α_1 helices of YckF and MJ1247 is approximately 30° . The region between β_5 and α_6 (residues 129–147) is three residues longer in YckF and is significantly more extended from the globule (arrow 2) than in MJ1247. In one monomer of our YckF structure, this region has short α helical segment, but the corresponding region in another monomer is disordered. The region between α_6 and the C-terminus is also more disordered in YckF and acquires a more irregular conformation, whereas in MJ1247 this region has a short helix. As described in the text, these differences may be

related to the fact that of these two structures, only MJ1247 has a ligand bound in the active site. (C) Monomers of the YckF tetramer are labeled and color-coded except for β sheets, which are all green. The helices are also labeled to accentuate their involvement in oligomerization. Major contributors to the tetramer formation are helices α_4 from all four monomers and several loop regions. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

Fig. 2. Amino acid sequence similarities. (A) Sequence alignment of YckF and several sugar phosphate isomerase domains. The YckF (gi 28373593) sequence is aligned with consensus sequences of families (in bold), shown in the table below

COG0794.1, GutQ	pfam01380.9, SIS	COG1737.1, RpiR,	LOAD sis.7, sis,	COG2222.1, AgaS,	COG0449.1, GlmS
gi 20150408 (1jeo)	gi 17942686 (1jxa)	gi 15832681	gi 10835796 (1c7r)	gi 16080314	gi 20138373
gi 15926249)	gi 28373593 (1m3s)	gi 15672536	gi 20150408 (1jeo)	gi 15896738	gi 1169919
gi 7388526	gi 13432146	gi 15926019	gi 7429975	gi 19703962	gi 21759128
gi 1175696	gi 1169919	gi 15640234	gi 11499385	gi 20094354	gi 19703963
gi 15679542	gi 1176842	gi 16123113	gi 15679542	gi 15966707	gi 15897315
gi 14521911	gi 1708000	gi 16123196	gi 7388382	gi 16767784	gi 16767783
gi 18312666	gi 2507341	gi 15803086	gi 1175696	gi 15643576	gi 17230956
gi 14591680	gi 1710667	gi 16130486	gi 7388526	gi 13473143	gi 16799113

Homology with the YckF sequence is shown as percent after the name of each domain. The total number of amino acids in the original sequence and the number included in alignment are shown in the end of each sequence except for YckF, which was fully included. The PDB entries of some structures are shown in parentheses next to their gi numbers. (B) Sequence alignment of proteins with PHI activity from several organisms. Numbering and secondary structure assignment in both (A and B) follow the sequence of YckF. Color coding corresponds to the degree of conservation from none (no color) to gray to orange to red (invariant). Homology with the YckF sequence is shown as percent after the name of each organism. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

Fig. 3.

Comparison of crystal structures of YckF, MJ1247 and GlmS. (A) Cartoon representation of GlmS architecture. Physiological unit of GlmS is a dimer (blue and green chains), each monomer of which comprises isomerase (hexagons) and glutaminase (rectangles) domains. Each chain of the isomerase domain contributes two, almost identical subunits (solid and checkerboard hexagons). The whole isomerase domain consists of four nearly identical subunits. YckF tetramer is formed by four identical molecules of the protein. Their positions are marked YckF A, YckF B, YckF C and YckF D on their corresponding subunits of GlmS. (B) Superposition was carried out for YckF and MJ1247 tetramers as a whole and for the dimer of GlmS isomerase domain, but for simplicity only the halves of the MJ1247 (beige) and of the GlmS (orange) oligomers are shown. The YckF tetramer is presented in blue with green β sheets. YckF and MJ1247 tetramers superimpose very well except for the regions that also differ in the monomers. The architecture of GlmS isomerase domain is very close to that of YckF and MJ1247 as well, but details of helix–helix interactions and the locations of some of the loops are somewhat different. The helices involved in the substrate binding (α_3 , α_6 , α_4 , α_5), and the β sheets, align very well in all structures. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

Fig. 4.

Stereo diagram of the putative active site of YckF and its comparison with active sites of MJ1247 and GlmS. Four monomers of YckF are color coded: (A) beige, (B) pink, (C) grey, and (D) blue. Residues in the phosphate-binding loops 43–47 and 86–91 are shown in full, while Arg-46, Arg-57, His-60, Ser-149, Glu-152, Asp-160, and His-181 are represented only by their side chains. The α helices and β strands are labeled. Some of the hydrogen bonds, discussed in the text, are shown as dashed cyan lines. The phosphate binding site of GlmS with bound glucosamine-6-phosphate is shown in yellow. Citrate molecule, bound in MJ1247 structure, occupies same site (not shown) as the substrate in GlmS. The phosphate binding site in all three proteins has virtually identical structure. Also the helices, forming the active site, superimpose very well (as seen from Fig. 3B). Few of the loops have different conformations. Two of these from MJ1247 are shown in green along with the side chain of His181. The loops in GlmS differ even more and they enclose the active site better (not shown). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this paper.)

Table 1

Basic data collection and processing statistics for YckF crystals

Number of amino acid residues/methionines 185/9 Number of molecules in the asymmetric unit 2 Crystal lattice				
Native	$P6_322$ $a = b = 72.08$, $c = 245.56$ Å			
Se-Met	$P6_322$ $a = b = 72.48$, $c = 244.87$ Å			
	Crystal 1 (Se-Met)	$\alpha = \beta = 90^\circ$		$\gamma = 120^\circ$
		Crystal 2 (native)		$\gamma = 120^\circ$
	Inflection point	Peak	Remote	Low resolution
				High resolution
Energy (keV)	12.65835 ^d	13.000		13.100
Wavelength (Å)	0.97947	0.95373		0.94644
Resolution (Å)	50–2.5	50–2.6		4–1.95
No. of observations	559 946	46 1026	501 104	262 698
Unique ^b reflections	24913	22082	22 153	25 054
Completeness (%) ^c	99.9 (100)	99.9 (100)	99.9 (100)	99.9 (100)
$I/\sigma(I)$ ^c	50.8 (8.3)	45.7 (7.7)	48.1 (6.3)	31.6 (5.3)
R_{sym}	0.068 (0.513)	0.074 (0.483)	0.082 (0.654)	0.072 (0.468)
R_{merge}	between low- and high-resolution passes on native crystals			
				0.035

^aEnergy deviation from usual values of Se absorption edge reflect calibration of the monochromator at the time of the experiment.

^bBijvoet pairs for scaling the MAD data sets were kept separately.

^cIn the last resolution shell.

Table 2

Phasing and refinement statistics for YckF structure

Phasing Energies	Resolution (Å) ^a	Number of reflections ^a	Phasing power ^a	FOM ^a
Inflection point	40–2.7 (2.81–2.7)	19 239 (2148)	5.57 (3.89)	0.46 (0.40)
Peak	40–2.7 (2.81–2.7)	19 115 (2120)	5.42 (3.89)	0.51 (0.41)
Remote	40–2.7 (2.81–2.7)	19 036 (2100)	3.98 (2.59)	0.52 (0.40)
Overall	40–2.7 (2.81–2.7)	19 497 (2191)		0.85 (0.70)
FOM after phase extension with density modification in 50–2.5 Å shell			0.94 (0.88)	
Refinement				
Resolution (Å)	62–1.95	2.0–1. 95		
Number of reflections	26 023	1745		
<i>R</i> factor %	18.9	20.1		
<i>R</i> free %	22.4	26.9		
Correlation	95.6			
Correlation free	94.3			
All nonhydrogen atoms	3012			
Solvent atoms	203			
Mean <i>B</i> -factor	18.4			
Deviations from ideal	Refined	Target		
Covalent bonds	0.024	0.021		
Bond angles	1.83	1.967		
Planarity	0.008	0.02		
Chiral centers	0.11	0.20		
Torsion angles 1	5.85	5.0		
Torsion angles 2	40.37	26.47		
Torsion angles 3	19.25	15.0		
VDW contacts	0.244	0.20		

^aIn the highest-resolution shell.