Acta Crystallographica Section D
**Biological Crystallography**
ISSN 0907-4449

# Structure of the inositol-1-phosphate cytidylyltransferase from *Thermotoga maritima*

**Oleg V. Kurnasov,[a] Hung-Jie Daniel Luk,[b] Mary F. Roberts[b] and Boguslaw Stec[a]***

[a]Sanford–Burnham Institute for Medical Research, La Jolla, CA 92037, USA, and [b]Merkert Chemistry Center, Boston College, Chestnut Hill, MA 02467, USA

Correspondence e-mail: bog.stec.2010@gmail.com

Received 22 April 2013
Accepted 2 June 2013

**PDB Reference:** inositol-1-phosphate cytidylyltransferase, 4jd0

The unique steps in the synthesis of an unusual osmolyte in hyperthermophiles, di-*myo*-inositol-1,1′-phosphate (DIP), involve the production of CDP-inositol and its condensation with an inositol-1-phosphate molecule to form phosphorylated DIP. While many organisms fuse both activities into a single enzyme, the two are separate in *Thermotoga maritima*. The crystal structure of the *T. maritima* inositol-1-phosphate cytidylyltransferase, which as a soluble protein may transiently associate with its membrane-embedded partner phospho-DIP synthase (P-DIPS), has now been obtained. The structure shows a conserved motif of sugar nucleotide transferases (COG1213) with a structurally reinforced C-terminal Cys bonded to the core of the protein. A bound arsenosugar identifies the location of the active site for inositol 1-phosphate. Based on homologous structures from several species and the identification of the crucial conserved aspartate residue, a catalytic mechanism for this enzyme is proposed as well as a mode for its association with P-DIPS. This structure imposes constraints on the mode of association, communication and temperature activation of two separate enzymes in *T. maritima*. For the first time, a working model for the membrane-bound P-DIPS unit has been constructed. This sheds light on the functioning of the phosphatidylserine and phosphatidylinositol synthases involved in many physiological processes that are homologous to P-DIPS. This work provides fresh insights into the synthesis of the unusual thermoprotective compound DIP in hyperthermophiles.

## 1. Introduction

Di-*myo*-inositol-1,1′-phosphate (DIP) is a thermoprotective osmolyte that occurs in a number of thermophilic archaea (Müller *et al.*, 2005; Roberts, 2005), including the genera *Archaeoglobus*, *Methanococcus*, *Pyrococcus* and *Thermococcus*, as well as thermophilic bacteria belonging to the genera *Aquifex*, *Rubrobacter* and *Thermotoga* (Santos & da Costa, 2001; Scholz *et al.*, 1992; Ciulla *et al.*, 1994; Rodionov *et al.*, 2007). Based on initial studies in *M. igneus* (Chen *et al.*, 1998) and subsequent use of genomic comparative methods and *Thermotoga maritima* as a model organism (Rodionov *et al.*, 2007), we determined the four-step biosynthetic pathway (Fig. 1) for this unusual compound: (i) synthesis of L-*myo*-inositol-1-phosphate (L-I-1-P) from D-glucose-6-phosphate by $NAD^+$-dependent *myo*-L-I-1-P synthase (mIPS), (ii) coupling of the L-I-1-P with CTP to form CDP-inositol (CDP-I) by

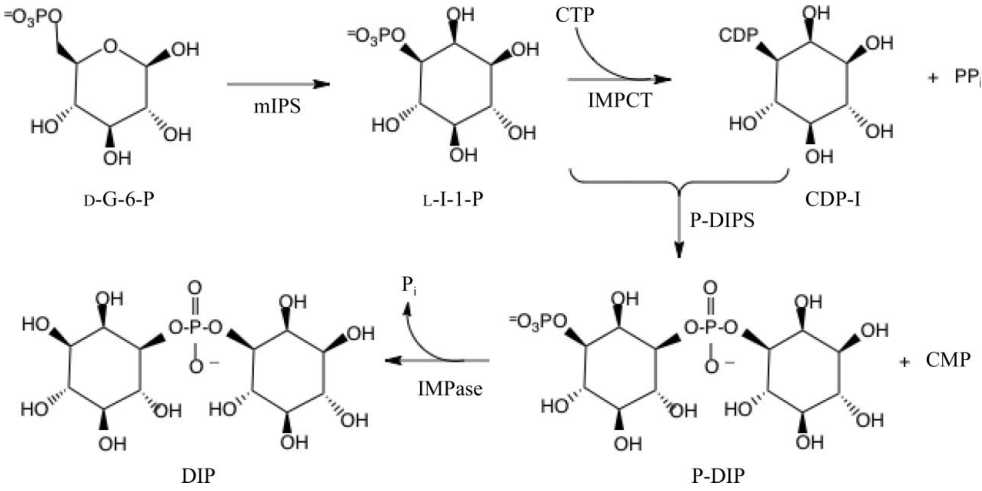



**Figure 1**
Biosynthetic pathway for di-*myo*-inositol-1,1′-phosphate (DIP).

inositol monophosphate cytidylyltransferase (IMPCT), (iii) generation of phosphorylated DIP (P-DIP) by condensation of CDP-I with L-I-1-P *via* a phospho-DIP synthase (P-DIPS) and (iv) hydrolysis of P-DIP to DIP by an inositol monophosphatase (IMPase) (Rodionov *et al.*, 2007; Martins *et al.*, 1996; Majerus, 1992; Fig. 1).

Two of these enzymes, mIPS and IMPase (Chen *et al.*, 1998; Drøbak, 1992), previously identified in *T. maritima*, are not solely utilized in DIP biosynthesis. Both enzymes are well conserved and examples of each from hyperthermophiles have been characterized biochemically and structurally (Stieglitz *et al.*, 2007; Neelon *et al.*, 2005; Stec *et al.*, 2000). The other two enzymes, IMPCT and P-DIPS, have been cloned and expressed in *Escherichia coli* and their activities confirmed by *in vitro* reconstitution of the entire P-DIP pathway in the crude lysate of *E. coli* (Rodionov *et al.*, 2007). In many archaea, IMPCT and P-DIPS are fused to form a single protein (Rodionov *et al.*, 2007; Gonçalves *et al.*, 2012). However, in the bacterium *T. maritima* the genes for these two enzymes are separate. Here, we present a crystal structure of *T. maritima* IMPCT, a soluble protein belonging to a well characterized family of sugar nucleotidyltransferases (Kleczkowski *et al.*, 2011).

The family of nucleotide-diphospho-sugar transferases is large and diverse (Supplementary Fig. S1[1]). It also encompasses a wide range of quaternary structures as well as single-domain and multidomain proteins (Kleczkowski *et al.*, 2011; Frey, 1992; Rodrigues *et al.*, 2007). They all share a common fold for the nucleotide/sugar-binding domain: an $\alpha/\beta$ structure with a conserved metal ion-binding site. According to SCOP there are 12 distinct functionalities that have representative structures in the PDB. All of them share the three-layer $\alpha/\beta/\alpha$ structure: a mixed $\beta$-sheet of seven strands (order 3214657 with strand 6 antiparallel to the rest) surrounded by five helices (Kim *et al.*, 2010). The active site is located in such a manner that the flexible fold can accommodate larger and smaller nucleotides as well as larger and smaller sugar moieties. The active site is surrounded by mobile loops that participate in binding and diphosphate exchange. A kinetic analysis of the *S. enterica* cytidylyltransferase suggested that its catalytic mechanism proceeds *via* a bi-bi ping-pong reaction (Zuccotti *et al.*, 2001) with CTP binding first to the enzyme.

The diversity in tertiary and quaternary structure in this family is particularly notable. Initial reports suggested a monomeric enzyme, while the structures of D-glucose-1-phosphate cytidylyltransferase from *Salmonella enterica*, *Yersinia pseudotuberculosis* and *Streptomyces glaucescens* (Lindqvist *et al.*, 1994; Thorson *et al.*, 1994; Beyer *et al.*, 1998) indicated a tetrameric enzyme. Recently, the X-ray crystallographic structure of *S. typhimurium* glucose-1-phosphate cytidylyltransferase complexed with CDP-glucose was also determined (Koropatkin *et al.*, 2005). The crystal structure showed a different arrangement of an integrated hexamer with active sites formed by residues contributed by two subunits. Interestingly, while the prokaryotic (Thoden & Holden, 2007) and the eukaryotic forms of UDP-glucose pyrophosphorylases (UGPases) catalyze the same type of reaction, they are completely unrelated in amino-acid sequence and three-dimensional structure (Flores-Díaz *et al.*, 1997; Mollerach *et al.*, 1998). In recent years, the bacterial UGPases have become targets for drug design given their biological roles in various Gram-negative bacteria (Kim *et al.*, 2004).

*T. maritima* IMPCT appears to be a close relative of archaeal sugar-nucleotide transferase enzymes from the family Halobacteriaceae and several bacterial genera, particularly *Rhizobium* and *Desulfovibrio*, as indicated by sequence similarity (Supplementary Fig. S1). The *T. maritima* genome also contains similar but distinct enzymes in this family. In a previous paper (Rodionov *et al.*, 2007), we reported the cloning and the preliminary kinetic evaluation of the IMPCT enzyme. Here, we report the crystal structure of the *T. maritima* inositol-1-phosphate cytidylyltransferase, which catalyzes the synthesis of an important intermediate in the synthesis of di-*myo*-inositol-1,1′-phosphate (DIP). Unexpectedly, we found an arsenoribose that allowed us to localize the active site and speculate on the nature of the catalytic mechanism. Analysis of the protein molecule and its crystal-packing pattern allowed the identification of the preferred faces for interaction with the membrane-localized P-DIPS. Analysis of the sequence of P-DIPS allowed tentative modeling of the entire protein and proposal of the interactions and function of the complex.

[1] Supplementary material has been deposited in the IUCr electronic archive (Reference: DZ5287). Services for accessing this material are described at the back of the journal.

## 2. Methods

### 2.1. Protein expression and purification

The IMPCT gene, amplified from the entire DIP operon identified previously (Rodionov *et al.*, 2007), was cloned into the pBAD-TOPO expression vector (Invitrogen). Expression was performed in *E. coli* strain BL21 with arabinose induction. Cells were grown to an $OD_{600}$ of 0.8–1.0 at 310 K in LB medium. IPTG was added to 0.2 m*M* and the culture was incubated with shaking for 6–12 h at 293 K. The cells were harvested by centrifugation and resuspended in lysis buffer (20 m*M* HEPES pH 7.0 with 100 m*M* NaCl, 0.03% Brij-35, 2 m*M* $\beta$-mercaptoethanol and 2 m*M* phenylmethylsulfonyl fluoride) with the addition of a protease-inhibitor cocktail (Sigma). Lysozyme was added to 1 mg ml$^{-1}$. After 60 min incubation on ice, the cell suspension was frozen in liquid nitrogen. After thawing and sonication, cell debris was removed by centrifugation at 20 000 rev min$^{-1}$ for 2 h. Tris–HCl pH 8.0 was added to the supernatant to a 50 m*M* final concentration and the supernatant was loaded onto an Ni–NTA agarose (Qiagen) column. The protein was eluted from the column with imidazole. The estimated final yield of protein was ~10 mg protein per gram of wet cell weight. The protein was stored in Tris–HCl with 100 m*M* NaCl.

### 2.2. Crystallization of IMPCT

Crystallization conditions were established by vapor diffusion using the sitting-drop method with sparse-matrix screens from Hampton Research. We did not initially obtain any crystals; we left the crystallization trays at 295 K for an extended period of time. Repetitive inspections did not reveal any crystals until several months later. Large hexagonal prism crystals appeared from sitting drops equilibrated against 1.8–2.0 *M* ammonium phosphate, 100 m*M* LiCl buffered at pH 8.5 with 100 m*M* Tris. The crystals grew to maximum dimensions of 0.5 × 0.5 × 0.5 mm. They belonged to space group $P3_221$, with unit-cell parameters $a = b = 133.1$, $c = 43.36$ Å, and contained a single monomer in the asymmetric unit with a Matthews coefficient of 3.32 Å$^3$ Da$^{-1}$, which corresponds to solvent content of around 67%.

### 2.3. Structural analysis of IMPCT

X-ray data were initially measured at 293 K using a Rigaku FR-E rotating-anode generator and an R-AXIS HTC area-detector system. The initial indexing and data collection were carried out using the *CrystalClear* software from Rigaku and the collected data were reduced using *HKL*-2000 and *SCALEPACK* (Otwinowski & Minor, 1997). The final high-resolution data set was collected at 100 K; the crystals were cooled by sudden exposure to an $N_2$ cold stream. Prior to cooling, the crystals were stabilized by transfer into mother liquor containing 15% glycerol. The X-ray data-collection and refinement statistics are presented in Table 1. The structure was solved by the molecular-replacement method as implemented in *Phaser* (McCoy *et al.*, 2007). Several template structures were used, but a successful solution was obtained using an unpublished structure of glucose-1-phosphate cytidylyltransferase from *Bacteroides fragilis* (PDB entry 2qqx) obtained from Dr Karen Allen at Boston University. Refinement was carried out using *REFMAC*5 (Murshudov *et al.*, 2011) and model building was carried out using *Coot* (Emsley *et al.*, 2010). The large number of model-building sessions allowed a complicated pattern of disorder to be resolved as well as proper identification of the solvent components. The final model included an unusual arsenate derivative, several phosphate ions and glycerol molecules. Additionally, we identified several Na and Cl atoms, together with 250 water molecules that completed the structure. The isotropic refinement converged at values of 0.16 and 0.20 for $R$ and $R_{free}$, respectively, while continuation of the refinement with anisotropic ADPs resulted in $R = 0.12$ ($R_{free} = 0.17$) for all measured X-ray data from 24 to 1.8 Å resolution.

**Table 1**
Data-collection and refinement statistics for the structure of *T. maritima* IMPCT (PDB entry 4jd0).

Values in parentheses are for the highest resolution shell.

| | |
|---|---|
| Data collection | |
| Unit-cell parameters (Å, °) | $a = b = 133.08$, $c = 43.37$, $\gamma = 120$ |
| Space group | $P3_221$ |
| Resolution (Å) | 1.80 (1.85–1.80) |
| Total reflections | 38573 (1876) |
| Completeness (%) | 94.0 (66.4) |
| Average $I/\sigma(I)$ | 19.5 (3.0) |
| $R_{merge}$ | 0.045 (0.41) |
| Multiplicity | 6.0 (2.5) |
| Refinement | |
| Resolution | 24–1.80 (1.85–1.80) |
| $R$ | 0.115 (0.34) |
| $R_{free}$ (5% of data) | 0.169 (0.39) |
| R.m.s.d., bond distances (Å) | 0.020 |
| R.m.s.d., bond angles (°) | 1.92 |
| $B$ values (Å$^2$) | |
| Average | 29.2 |
| Ligands | 42.3 |
| Arsenoribose | 31.3 |
| Water | 51.2 |
| Water molecules | 240 |
| Disordered residues | 36 |
| Ramachandran plot, residues in (%) | |
| Favored regions | 98.0 |
| Allowed regions | 2.0 |
| Outlier regions | 0.0 |

### 2.4. Characterization of the As metabolite

In order to confirm the presence and identity of an unusual small molecule that was detected in our electron-density maps, we subjected the crystallization sample to evaluation by MS. The low-mass and high-mass MS confirmed the size of our protein construct and revealed several molecules in the low-molecular-mass range. A dominant peak was detected at ~273 Da (Supplementary Fig. S2). The molecular mass perfectly agreed with that of the modeled compound arsenoribose. The molecular-mass measurements were carried out in semiautomatic mode using a MALDI-TOF/TOF mass spectrometer (Bruker Daltronics AutoFlex II) in the proteomics facility at the SBMRI.

### 2.5. Modeling of the PDIP and complex formation

A series of models of P-DIPS were created in *I-TASSER* (Zhang, 2008) using homologous proteins based on the following structures: PDB entries 1t33 (Midwest Center for Structural Genomics, unpublished work), 2nx4 (Midwest Center for Structural Genomics, unpublished work), 3mpv (Takenoya *et al.*, 2010), 2lat (Gayen & Kang, 2011), 2hyj (Midwest Center for Structural Genomics, unpublished work), 1oyz (Northeast Structural Genomics Consortium, unpublished work), 2pbx (DeSilva *et al.*, 2007), 2ww9 (Becker *et al.*, 2009), 2qkw (Xing *et al.*, 2007) and 1anv (Kanellopoulos *et al.*, 1996). The model representing the best agreement with the predicted transmembrane elements as determined by *SPLIT*4 (Juretić *et al.*, 2002) was selected. Subsequently, the lattice packing analysis of IMPCT was used to identify the most likely interacting surfaces. The model of the complex was created by docking the two proteins manually to fulfill three requirements: (i) close proximity of the N- and C-termini of the proteins, which are fused together in many species, (ii) preservation of the packing surfaces observed in the crystal structures of *A. fulgidus* and *T. maritima* IMPCT and (iii) provision of a pathway for substrate migration from the active site of the IMPCT module to that of the P-DIPS module.

## 3. Results and discussion

### 3.1. Overall IMPCT architecture

The general architecture of *T. maritima* IMPCT is organized around an extended seven-stranded $\beta$-sheet, as expected for this structural class of enzymes (Fig. 2*a*). However, there are some small variations. In our expression construct we included nine additional residues at the N-terminus. In the crystal structure the N-terminal fragment that contains the linker and the His tag extends well beyond the compact fold of the enzyme, with four residues of the linker clearly visible in electron-density maps and forming numerous contacts with the neighboring molecule. This arrangement might partially explain the very long time needed for crystallization. The first strand transverses the molecule and constitutes the structural foundation for binding of the base moiety of the cytosine nucleotide. The loop following this first strand is very flexible and is usually disordered in all nucleotidyltransferase crystal

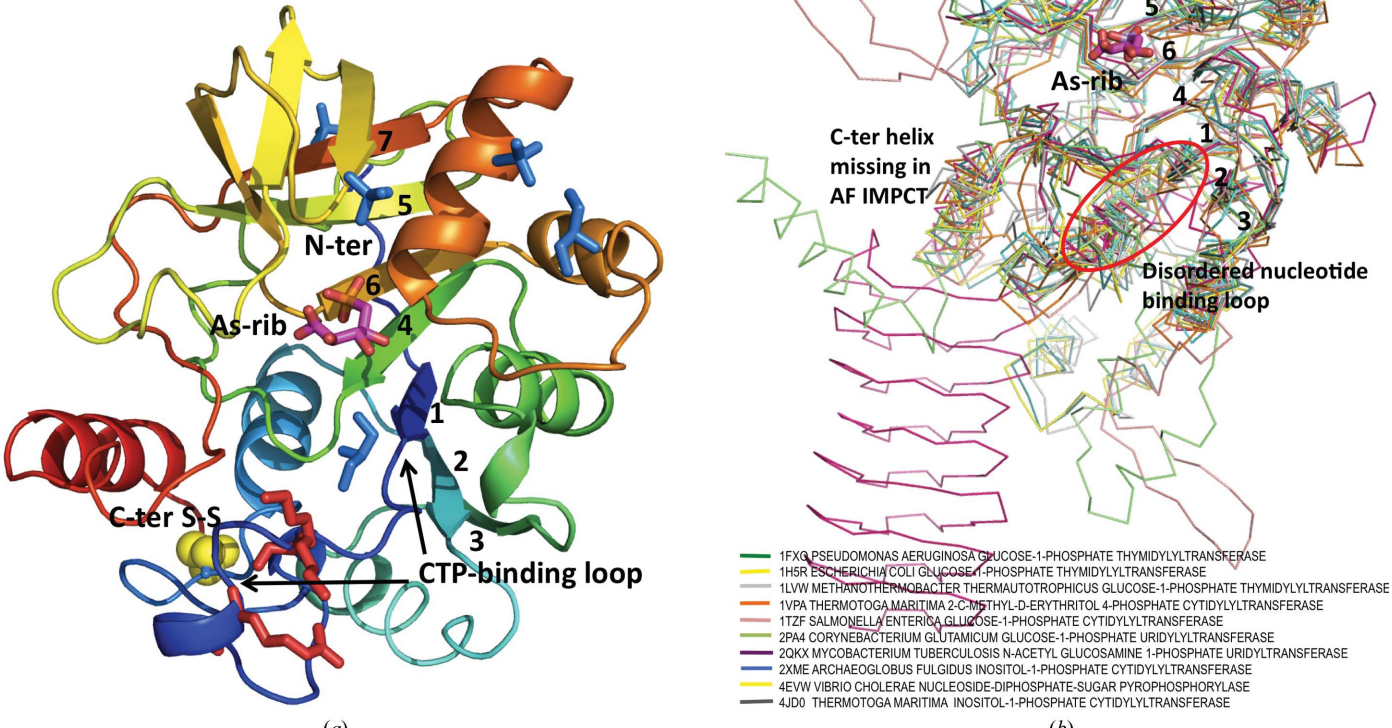



**Figure 2**
*T. maritima* IMPCT crystal structure. (*a*) Ribbon representation in which the color changes from blue at the N-terminus to red at the C-terminus. The arsenoribose (As-rib) in the active site is marked in magenta, the residues suggested to aid in membrane binding are shown in red, the disulfide is shown in yellow, the solute molecules are shown in blue and the seven $\beta$-strands are numbered. (*b*) Structural comparison of the structure of IMPCT from *T. maritima* (gray) with nine representative sugar nucleotidyltransferases: PDB entries 1fxo (green; *Pseudomonas aeruginosa* glucose-1-phosphate thymidylyltransferase; Blankenfeldt *et al.*, 2000), 1h5r (light yellow; *E. coli* glucose-1-phosphate thymidylyltransferase; Zuccotti *et al.*, 2001), 1lvw (light gray; *Methanothermobacter thermautotrophicus* glucose-1-phosphate thymidylyltransferase; Northeast Structural Genomics Consortium, unpublished work), 1vpa (orange; *T. maritima* 2-*C*-methyl-D-erythritol 4-phosphate cytidylyltransferase; Joint Center for Structural Genomics, unpublished work), 1tzf (pink; *S. enterica* glucose-1-phosphate cytidylyltransferase; Koropatkin & Holden, 2004), 2pa4 (light green; *Corynebacterium glutamicum* glucose-1-phosphate uridylyltransferase; Thoden & Holden, 2007), 2qkx (magenta; *M. tuberculosis N*-acetyl glucosamine 1-phosphate uridylyltransferase; Zhang *et al.*, 2009), 2xme (blue; *Archaeoglobus fulgidus* inositol-1-phosphate cytidylyltransferase; Brito *et al.*, 2011) and 4evw (yellow; *Vibrio cholerae* nucleoside-diphosphate-sugar pyrophosphorylase; Northeast Structural Genomics Consortium, unpublished work).

structures that have no nucleotide bound (Fig. 1*b*). This protein has a flexible design, as shown by the number and the extent of multiple side-chain conformations (36 residues, constituting ∼15% of the amino acids). Additionally, the C-terminal region shows some divergence from a canonical fold. The C-terminal helix is shorter and bent so that its last turn can provide a close contact between the C-terminal Cys residue and a Cys residue in the body of the protein. Cys247 and Cys32 form a disulfide bridge that rigidifies the structure and provides structural support for this very thermostable macromolecule.

### 3.2. Similarity of IMPCT to family members

The arrangement of the entire folding unit with an extended $\beta$-sheet and the similar location of the active site in all known members of the family suggest a modular design that codes for the functionality of this unit (Fig. 2*a*). The superposition of ten well refined models of the representatives of this family is shown in Fig. 2(*b*). The general architecture shows the structural divergence related to the observed sequence similarity. Table 2 presents the r.m.s.d. variations of the corresponding $C^{\alpha}$ positions between five representatives of the family. As expected, the similarity is highest to the *A. fulgidus* IMPCT domain structure (Brito *et al.*, 2011; cloned as the soluble portion of the gene for the IMPCT/P-DIPS fusion protein). The two structures share 37% identity, with an r.m.s.d. of 1.5 Å when comparing the regions that are present in both crystal structures.

The design of the unit allows the accommodation of a large number of different substrate/product combinations but also lends itself to a variety of different quaternary arrangements. These arrangements are modulated by the structural elements added to the base motif, as the core superimposes well. These extensions, which are shown as single-chain traces in Fig. 2(*b*), allow the functioning of the unit as a monomeric, dimeric, tetrameric or even a hexameric species.

The *A. fulgidus* structure does not represent the whole cloned sequence of the fragment. It is missing the final helix (which was only partially represented in the cloned fragment), and a highly flexible loop was not refined in all segments of the *A. fulgidus* IMPCT model. The flexible character of this folding motif (Fig. 2*b*) is underscored by the fact that it is disordered (but visible) in *T. maritima* IMPCT. The mobile fragment of the CTP-binding loop shows weaker density and elevated temperature factors. The need for increased mobility is partially explained by the large size of both of the substrates.

As mentioned above, in *A. fulgidus* IMPCT the construct generated from the fusion protein did not contain the C-terminal helix that constitutes a linker to the P-DIPS module of this multifunctional protein. The lack of this structural element certainly contributed to the success of crystallization, but by the same token made it more difficult to model the entire protein. A relatively high sequence similarity between the proteins suggests that the direction and the general location of the C-terminal helix in *T. maritima* IMPCT provides a good model for the interfacial helix in the *A. fulgidus* protein. Additionally, this structural element suggests the location of the interface between the domains (IMPCT and membrane-bound P-DIPS) of the archaeal multifunctional fusion protein.

**Table 2**
Mutual similarities measured by the r.m.s.d. and the number of residues in the comparison between representatives of the family representing different nucleotides: cytidylyltransferases, uridylyltransferases and thymidylyltransferases.

The computation was carried out using the program *TOPP* from *CCP*4 v.6.2 (Winn *et al.*, 2011).

(*a*) The diagonal (in bold) shows the number of residues in the protein, above the diagonal shows the r.m.s.d. and below the diagonal shows how many $C^{\alpha}$ positions were used in the r.m.s.d. calculation.

| Model r.m.s.d. (Å)/ No. of residues | 4jd0 | 2xme | 1fxo | 1vpa | 2pa4 |
|---|---|---|---|---|---|
| 4jd0 | **241** | 1.59 | 1.70 | 1.71 | 1.64 |
| 2xme | 161 | **222** | 1.60 | 1.71 | 1.44 |
| 1fxo | 154 | 134 | **293** | 1.83 | 1.86 |
| 1vpa | 154 | 111 | 145 | **219** | 1.79 |
| 2pa4 | 169 | 140 | 192 | 144 | **305** |

(*b*) The diagonal (in bold) shows the number of residues in the protein, above the diagonal shows the percentage identity and below the diagonal shows how many identical residues are in the matched fragments in the two proteins.

| Model identity (%)/ No. of residues | 4jd0 | 2xme | 1fxo | 1vpa | 2pa4 |
|---|---|---|---|---|---|
| 4jd0 | **241** | 32.2 | 19.5 | 19.5 | 17.2 |
| 2xme | 52 | **222** | 22.4 | 14.4 | 21.4 |
| 1fxo | 30 | 30 | **293** | 14.5 | 29.9 |
| 1vpa | 30 | 16 | 21 | **219** | 16.0 |
| 2pa4 | 29 | 30 | 63 | 23 | **305** |

### 3.3. IMPCT active site

The refined model of *T. maritima* IMPCT has a pronounced central cavity that, by comparison with other family members, defines the location of the active site. The IMPCT active site is located in a wide cleft formed by the edge of the $\beta$-sheet and the flanking helices. The active site is quite open and is capable of accommodating both substrates: inositol 1-phosphate and CTP. Extensive loops flanking the opening of the cavity show significant mobility, as exemplified by the extensive side-chain disorder and the higher than average temperature factors. In the initial stages of the refinement, after the protein backbone was satisfactorily fitted to the density a strong residual difference electron density was detected in this central cavity.

The starting hypothesis in the modeling of this density was that it represented a phosphorylated sugar. Initially, we attempted to refine several phosphosugars and phosphonate sugars into this density. However, none fitted the stereochemical requirements or the observed density. The initial rounds of refinement revealed a pattern of hydroxyls that did not correspond to any abundant sugars (with the exception of phosphonoribonic acid). Searches for the possible source of this compound suggested that it is unlikely that a common metabolite or a compound derived from the crystallization medium would be present in such an abundance to explain the density. Also, some excessive differences in the temperature

factors suggested that the tetragonal moiety was the key to the identity of this compound.

In order to aid the identification of the small-molecule entity, we carried out MS studies of the crystallization and protein solutions. The MS results indicated the presence of a small molecule with a mass of ∼273 Da. A search of the published literature revealed an arsenoribose to be a compound with this exact mass. This compound refined well in the active site with an occupancy of around 0.6; it was best refined as the *gem*-diol form at the terminal C atom (the aldol form of the arsenoribose). This observation is similar to the recently refined structure of the inositol phosphate synthase (mIPs) from *A. fulgidus* with a trapped catalytic intermediate (Neelon *et al.*, 2011). The authors speculated that the *gem*-diol form protects the sugar in the intermediate steps of catalysis.

Partial occupancy of the arsenosugar and some remaining difference density overlapping with the refined compound suggested that other heavier atoms could overlap with the arsenoribose in specific locations (for instance, the bridging carbon position). Indeed, the placement of $Na^+$ and $Cl^-$ ions at these locations and their refinement with an occupancy of 0.4 explained the difference densities but also fitted the electrostatic profile of the active site very well and provided an explanation for the partial occupancy of the arsenoribose.

The model of the active site with the OMIT map covering the arsenoribose and a sodium ion and a chloride ion overlapping the arsenosugar, as well as an additional sodium ion bound to two hydroxyl groups of the compound, is presented in Fig. 3. The location of this sodium ion is different from the expected position for the $Mg^{2+}$ ion needed for enzymatic activity. Nevertheless, it chelates two sugar hydroxyls pointing towards the open space that is occupied by the CTP and $Mg^{2+}$ in other structures. The sugar moiety is hydrogen bonded to several water molecules. Some of these water molecules are coordinated by an additional $Na^+$ ion directly coordinated by Asp107, which in the active complexes would be the catalytic $Mg^{2+}$ (see below).

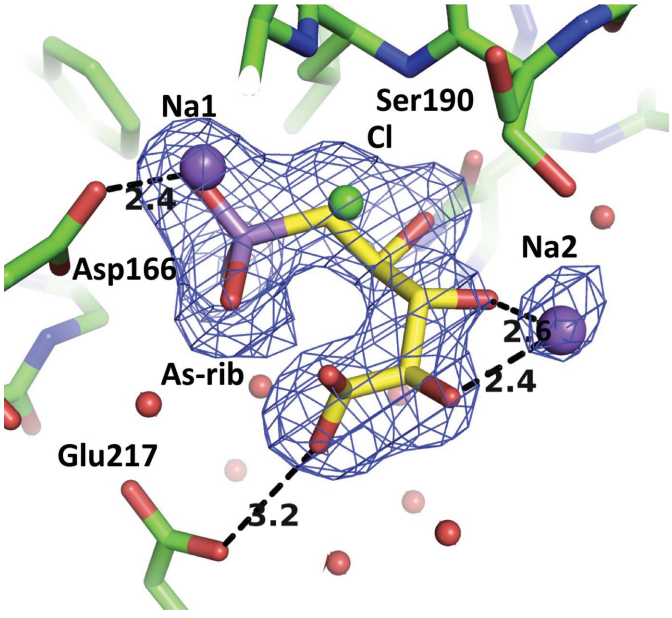



**Figure 3**
The IMPCT active-site view with OMIT map electron density covering the arsenoribose refined at an occupancy of 0.6. The purple and green spheres represent sodium and chloride ions occupying the active site with an occupancy of 0.4. The second sodium ion is connected to two hydroxyls of the ribose entity. The ribose is in the *gem*-diol form with stabilizing hydrogen bonds to the O atoms of the arseno group and Glu217.

The location of the arsenoribose compound confirmed the initial assumptions about the location of the active site, but its presence was a substantial surprise. We cannot account for its presence nor explain its origin. However, on the basis of a recent controversy concerning halophilic organisms from Mono Lake that were suggested to thrive in arsenic-rich environments and to incorporate arsenic in place of phosphorus in various biomolecules (Wolfe-Simon *et al.*, 2011; Elias *et al.*, 2012), we can hypothesize its origin. One of the results precipitated by this controversy was a very careful study of the arsenic resistance of GFAJ-1 and *E. coli* (Erb *et al.*, 2012). This study clearly demonstrated that a large number of arsenic-containing metabolites can be found in *E. coli* upon exposure to even small amounts of arsenic. Furthermore, some of these compounds are ribose derivatives compatible with the compound found in our structure. Additionally, several studies showed that algae are exceptionally resistant to arsenic compounds (Levy *et al.*, 2005). Some popular reagents used in cell culturing (for instance agar) are derived from algae; hence, inadvertent contamination and kinetic selection may have contributed to its presence in our sample.

The active-site cleft contains many well resolved water molecules close to the arsenoribose. Water molecules close to the mobile loop become much less defined in the electron density. Additionally, an elongated difference density that was detected along the loop was tentatively identified as a glycerol molecule, which is a component of the cryoprotectant solution. At the top of the cleft three fragments of difference electron density were identified as phosphates and a second glycerol molecule. Both glycerol molecules were present at a partial occupancy similar to that of arsenoribose. This second glycerol molecule was refined close to the loop containing residues 180–183. A comparison of the IMPCT active site with that of glucose-1-phosphate uridylyltransferase (Thoden & Holden, 2007) shows that the arsenoribose moiety is found at the site that is occupied by glucose in this structure, while the position of the first glycerol molecule overlaps with the ribose sugar moiety of the UDP (Fig. 4*a*).

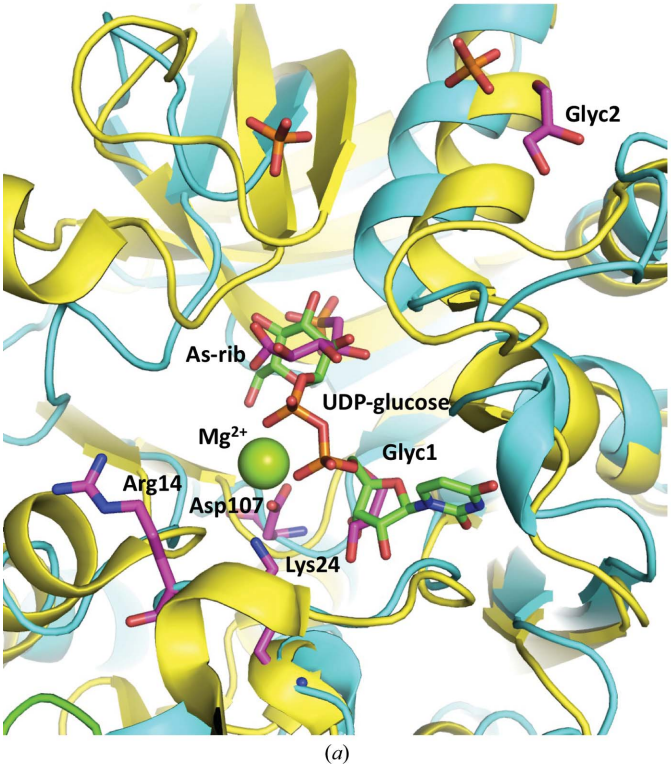
### 3.4. Proposed catalytic mechanism

The crystal structure, and particularly the location of the arsenoribose, when compared with the structures of other nucleotide transferases with substrates/products bound (Zuccotti *et al.*, 2001; Thoden & Holden, 2007) suggest the mechanism of catalysis for this cytidylytransferase from a hyperthermophile. The conserved Asp107 serves as an electrostatic anchor to which the metal ion (most likely $Mg^{2+}$) binds to initiate the catalysis. The binding of a metal ion subdivides the active-site cavity into two distinct

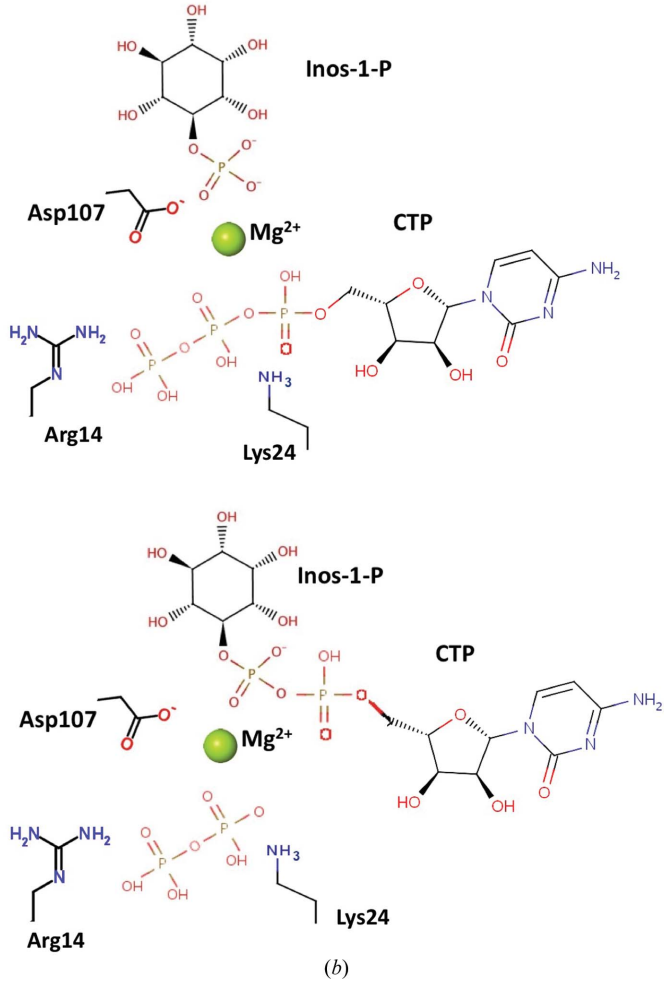
compartments (Fig. 4*a*). A relatively tight upper compartment is where the phosphorylated sugar, a phosphorylated inositol, binds to IMPCT, while the second much more open compartment is flanked by two long mobile loops and constitutes the binding site for a nucleotide triphosphate (here CTP; Frey, 1992).

Upon binding, the triphosphate group of the nucleotide is positioned to interact with a metal ion as well as the phosphate of the inositol. The O atom of the inositol phosphate group that is polarized by this interaction attacks the triphosphate and causes its cleavage to release the pyrophosphate group. Two residues that are highly conserved in these nucleotidyltransferases serve additional roles in the catalytic process. Lys24 stabilizes the developing negative charge in the transition state and Arg14, which is positioned to bind the pyrophosphate, serves as a release agent. In most nucleotidyltransferases the loop including Arg14 is either disordered (high temperature factors or multiple conformations) or is not visible in the electron-density map. In the structure of *T. maritima* IMPCT it is discretely disordered, with the backbone atoms well localized. The mechanism for the formation of CDP-inositol, based on other nucleotidyltransferases (Frey, 1992; Thoden & Holden, 2007), is suggested in Fig. 4(*b*).

Additionally, we refined a glycerol molecule at the putative CTP-binding site of the *T. maritima* IMPCT. The binding of a glycerol molecule imitating the ribose moiety of CTP (as shown by direct superposition in Fig. 4*a*) causes the entire nucleotide-binding loop to be relatively well ordered despite having higher mobility, with the exception of Arg14. The second key residue that is highly conserved in this loop, Lys24, is well ordered. In several species there is an additional negatively charged residue responsible for binding a metal ion (Asp219 in the *T. maritima* enzyme), but it is not universally conserved. The rest of the active site appears to be very loosely defined; the crucial elements are highly variable in this class of enzymes. This arrangement provides the flexibility to accommodate different substrates and products in many




(*a*)




(*b*)

**Figure 4**
Catalytic mechanism of *T. maritima* IMPCT. (*a*) Superposition of *C. glutamicum* glucose-1-phosphate uridylyltransferase (yellow; PDB entry 2pa4) with *T. maritima* IMPCT (blue; PDB entry 4jd0). The green sphere represents the catalytic $Mg^{2+}$ bound by the conserved Asp107. The arsenoribose (magenta) occupies the space occupied by the glucose moiety in glucose uridyldiphosphate (green). The uridyl moiety is in a similar position to those occupied by the nucleotidyl bases in the entire family of nucleotidyltransferases; it is perched at the top of the nucleotide-binding loop. A glycerol molecule (magenta) that is bound at the active site is positioned where the ribose moiety is positioned in the nucleotide. Additionally, three side chains marked in magenta, Asp107, Arg14 and Lys24, are shown in stick representation. These conserved residues are marked in the multiple sequence alignment shown in Supplementary Fig. S1. (*b*) Schematic suggesting the proposed IMPCT catalytic cycle aided by the metal ion coordinated by the conserved Asp107 residue. Lys24 aids in neutralization of the developing negative charge upon cleavage of the triphosphate and Arg14 is responsible for the coordination and release of the pyrophosphate.

oligomeric arrangements as reported for other nucleotidyltransferases.

### 3.5. Identification of the interaction surface of *T. maritima* IMPCT with membrane-bound PDIPS

Although IMPCT is a soluble protein in *T. maritima*, in many other hyperthermophiles it is part of a larger fusion protein with the membrane-bound P-DIPS (Rodionov *et al.*, 2007). In the bifunctional protein the IMPCT enzyme must transfer its product, CDP-I, to the P-DIPS module. Given the structural homology of the *A. fulgidus* IMPCT domain to the *T. maritima* IMPCT enzyme, we can use the more defined structure of the latter to address how IMPCT can interact with the membrane-bound protein and to determine which surface it interacts with. Two aspects of the IMPCT structure provide insight into these interactions. The first is that *T. maritima* IMPCT has a more extended nucleotide-binding loop with additional positively charged residues, Lys13 and Arg16 (Fig. 2*a*). These positively charged residues could facilitate binding to negatively charged membranes. The second structural insight comes from the packing of the IMPCT molecules in the crystal lattice (Supplementary Fig. S3). In the *A. fulgidus* dual-activity fusion variant of the protein (Rodionov *et al.*, 2007; Brito *et al.*, 2011) the polypeptide must continue from the IMPCT unit to the P-DIP synthase unit. Therefore, the last helix of IMPCT must be located in the vicinity of the first helix of P-DIPS, thus providing a spatial constraint. Additionally, for efficient transfer of CDP-I from the IMPCT to the P-DIP synthase active site a physical pathway connecting both active sites should exist. These two criteria provide significant clues and limitations for the construction of models of this complex.

Inspection of the molecular-surface characteristics of the refined model allows a tentative assignment of the interacting surface (Fig. 5). A very interesting packing arrangement of the IMPCT is imposed by its surface characteristics. Despite forming a lattice characterized by the threefold symmetry in arranging protein molecules, the packing of the neighboring molecules forms sheets (Fig. 5*b* and Supplementary Fig. S3). The packing captures certain surfaces of the molecules, leaving others open to solvent. Analysis of the packing interfaces indicates that these surfaces have significant hydrophobic character, as expected for a temporary docking of the IMPCT to the membrane. Such an arrangement would bring the active site into close proximity to the membrane-bound P-DIPS. The proposed model of the arrangement of *T. maritima* IMPCT with the membrane-bound P-DIP synthase unit is depicted in Fig. 5(*a*). Despite a significantly different spatial arrangement of the individual molecules in the crystals of *A. fulgidus* IMPCT (Brito *et al.*, 2011), the packing is reminiscent of that in *T. maritima* IMPCT, leaving the same faces of the molecules open and burying the perimeter of the molecule as suggested above.

The sequence analysis, which is supported by comparative genomics (Rodionov *et al.*, 2007), indicates that both the IMPCT and P-DIPS proteins are relatively small, containing around 250 amino acids each. The significant shape similarity to other members of the IMPCT family provides additional credence to the speculation that P-DIPS might be equally well folded and have a structure that is conserved in the family of related proteins. Encouraged by searches for homologous proteins, we decided to model the P-DIPS protein by homology. A search using *FFAS*03 (Jaroszewski *et al.*, 2011) provided several well defined templates (PDB entries 1t33, 2nx4, 3mpv, 2lat, 2hyj, 1oyz, 2pbx, 2ww9, 2qkw and 1anv). Modeling in *I-TASSER* (Zhang, 2008) produced a series of remarkably similar all-helical models. All of the models contained a core five-helix bundle that showed significant hydrophobic characteristics that were independently confirmed by a search for transmembrane helices in *SPLIT*4 (Juretić *et al.*, 2002). What came as an additional bonus is that the signature

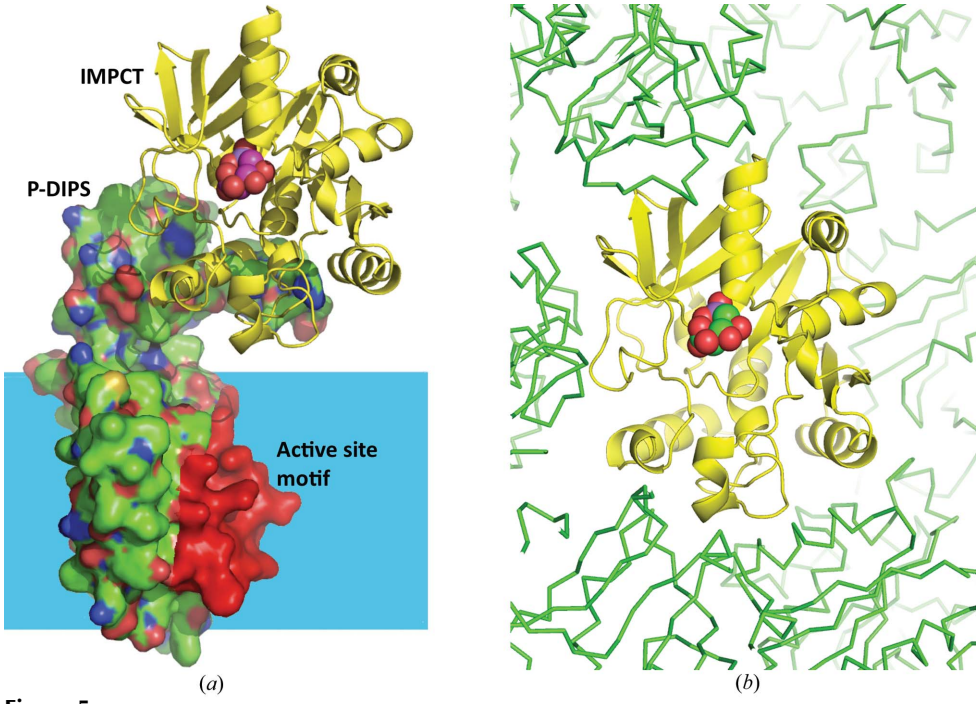



**Figure 5**
Proposed model of the IMPCT–P-DIPS complex formed transiently on the membrane. (*a*) Space-filling model of P-DIP synthase generated in *I-TASSER*. The predicted active site marked in red is embedded in the membrane (blue). The IMPCT molecule is docked taking into account the proximity of the N- and C-termini of both proteins as described in the text and the preferred surfaces deduced from the packing diagram presented in (*b*). (*b*) Ribbon diagram of IMPCT (in yellow) with a space-filling model of the arsenoribose that marks the active site. The molecule is surrounded by a band of symmetry-related molecules in the $P3_221$ lattice, while two opposing faces are open to the solvent and include the active site.

residues for this family, which are suggested to constitute the active site, were all grouped at a small helix that might be mobile and would be expected to orient at the membrane–cytosol interface. Polar and charged residues provide sufficient interactions for the catalytic activity of this fragment. This motif is very similar to that in phospholipid synthases (Nikawa & Yamashita, 1997; Yamashita & Nikawa, 1997). The extra-membrane portion of the model has two or three helices (depending on the model) that are perpendicular to the bundle and are parallel to the surface of the membrane. These helices would provide a suitable docking site for the IMPCT molecule. The detailed ribbon diagram is shown in Supplementary Fig. S4, while the global model of P-DIPS with docked IMPCT is shown in Fig. 5.

The most interesting element of this tentative model is that not only does it explain the mode of communication between two sequential enzymes in a biosynthetic pathway that in many instances are fused together and the route of substrate/product migration, but it also suggests the placement of the active site within the dimensions of the membrane. To avoid exposing polar active-site residues to the hydrocarbon portion of the bilayer, the protein is most likely to aggregate as a dimer or a higher order oligomer. Whether this is a homo-oligomeric arrangement or whether another membrane-localized binding partner is used remains to be determined.

This model is supported by a number of experimental observations made during the reconstitution of the P-DIP biosynthetic pathway in *E. coli*. When cells expressing all three proteins were lysed by sonication, IPS and IMPCT were in the soluble fraction while P-DIPS was in the pellet, as expected. However, when the soluble fraction was heated for 30 min at 353 K (below the growth temperature of *T. maritima*), a large fraction of the IMPCT protein and the enzymatic activity was associated with the pellet. Additionally, *E. coli* cells that only expressed P-DIPS were unable to synthesize P-DIP with added CDP-I and I-1-P even when sonicated and heated. Only after extraction with detergents or with the addition of soluble mIPS and IMPCT and heating of the sample could the activity of P-DIPS be detected. These observations are difficult to quantify, but since the organism makes DIP at temperatures above 353 K this may reflect a conformational change that allows soluble IMPCT to interact with membrane-localized P-DIPS. Altogether, these observations strongly suggest that in *T. maritima* transient membrane binding by IMPCT after a temperature-induced conformational change is needed for P-DIP production.

## 4. Summary

In this report, we have described the structure of a protein that is the product of a recently discovered gene (TM1418A) in *T. maritima* that codes for an enzyme involved in the biosynthesis of DIP. This IMPCT protein belongs to the well described family of sugar nucleotidyltransferases. The structure has a crucial disulfide that provides added thermal stability by covalently linking the C-terminal Cys residue to the bulk of the protein by a disulfide bridge. The finding of an unusual compound (an arsenoribose) in the structure provided a direct clue to the location of the active site. A comparison with known nucleotidyltransferase structures containing bound nucleotides led us to a proposed mechanism for this enzyme. The analysis of crystal packing combined with an examination of the protein surface suggested the mode of interaction of IMPCT with the membrane-bound P-DIPS and its orientation at the membrane surface. The sequence similarity of P-DIPS to known enzymes allowed the construction of a three-dimensional model using *I-TASSER* and the creation of a model of the IMPCT–P-DIPS complex. This model suggested the means of communication between the enzymes and the relative location of the active sites. The exposed nature of the P-DIPS active site in our model suggested that the enzyme should dimerize/oligomerize in the membrane in order to function properly. Many of the conclusions of this work provide a framework to guide future experiments that will provide a clear picture of the unique components of the DIP biosynthetic system.

We would like to thank Dr Karen Allen for sharing her structure prior to publication. Also, we express our gratitude to the laboratory of Dr Andrei Osterman for help with cloning and protein production and to Dr Khatereh Motamedchaboki, the director of the proteomics facility, for performing the mass-spectrometric analyses. This work has been supported by the US Department of Energy Office of Science, Energy Biosciences grant DE-FG02-91ER20025 (to MFR).

## References

Becker, T., Bhushan, S., Jarasch, A., Armache, J.-P., Funes, S., Jossinet, F., Gumbart, J., Mielke, T., Berninghausen, O., Schulten, K., Westhof, E., Gilmore, R., Mandon, E. & Beckmann, R. (2009). *Science*, **326**, 1369–1373.

Beyer, S., Mayer, G. & Piepersberg, W. (1998). *J. Biochem.* **258**, 1059–1067.

Blankenfeldt, W., Asuncion, M., Lam, J. S. & Naismith, J. H. (2000). *EMBO J.* **19**, 6652–6663.

Brito, J. A., Borges, N., Vonrhein, C., Santos, H. & Archer, M. (2011). *J. Bacteriol.* **193**, 2177–2185.

Chen, L., Spiliotis, E. T. & Roberts, M. F. (1998). *J. Bacteriol.* **180**, 3785–3792.

Ciulla, R. A., Burggraf, S., Stetter, K. O. & Roberts, M. F. (1994). *Appl. Environ. Microbiol.* **60**, 3660–3664.

DeSilva, R. S., Kovacikova, G., Lin, W., Taylor, R. K., Skorupski, K. & Kull, F. J. (2007). *J. Bacteriol.* **189**, 5683–5691.

Drøbak, B. K. (1992). *Biochem. J.* **288**, 697–712.

Elias, M., Wellner, A., Goldin-Azulay, K., Chabriere, E., Vorholt, J. A., Erb, T. J. & Tawfik, D. S. (2012). *Nature (London)*, **491**, 134–137.

Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.

Erb, T. J., Kiefer, P., Hattendorf, B., Günther, D. & Vorholt, J. A. (2012). *Science*, **337**, 467–470.

Flores-Díaz, M., Alape-Girón, A., Persson, B., Pollesello, P., Moos, M., von Eichel-Streiber, C., Thelestam, M. & Florin, I. (1997). *J. Biol. Chem.* **272**, 23784–23791.

Frey, P. (1992). *The Enzymes*, Vol. 20, edited by D. S. Sigman, pp. 142–183. San Diego: Academic Press.

Gayen, S. & Kang, C. (2011). *Biochem. Biophys. Res. Commun.* **409**, 572–576.

Gonçalves, L. G., Borges, N., Serra, F., Fernandes, P. L., Dopazo, H. & Santos, H. (2012). *Environ. Microbiol.* **14**, 691–701.
Jaroszewski, L., Li, Z., Cai, X.-H., Weber, C. & Godzik, A. (2011). *Nucleic Acids Res.* **39**, W38–W44.
Juretić, D., Zoranić, L. & Zucić, D. (2002). *J. Chem. Inf. Comput. Sci.* **42**, 620–632.
Kanellopoulos, P. N., Tsernoglou, D., van der Vliet, P. C. & Tucker, P. A. (1996). *Acta Cryst.* D**52**, 942–945.
Kim, H., Choi, J., Kim, T., Lokanath, N. K., Ha, S. C., Suh, S. W., Hwang, H.-Y. & Kim, K. K. (2010). *Mol. Cells*, **29**, 397–405.
Kim, H., Wu, C. A., Kim, D. Y., Han, Y.-H., Ha, S. C., Kim, C.-S., Suh, S. W. & Kim, K. K. (2004). *Acta Cryst.* D**60**, 1447–1449.
Kleczkowski, L. A., Geisler, M., Fitzek, E. & Wilczynska, M. (2011). *Biochem. J.* **439**, 375–379.
Koropatkin, N. M., Cleland, W. W. & Holden, H. M. (2005). *J. Biol. Chem.* **280**, 10774–10780.
Koropatkin, N. M. & Holden, H. M. (2004). *J. Biol. Chem.* **279**, 44023–44029.
Levy, J. L., Stauber, J. L., Adams, M. S., Maher, W. A., Kirby, J. K. & Jolley, D. F. (2005). *Environ. Toxicol. Chem.* **24**, 2630–2639.
Lindqvist, L., Kaiser, R., Reeves, P. R. & Lindberg, A. A. (1994). *J. Biol. Chem.* **269**, 122–126.
Majerus, P. W. (1992). *Annu. Rev. Biochem.* **61**, 225–250.
Martins, L. O., Carreto, L. S., Da Costa, M. S. & Santos, H. (1996). *J. Bacteriol.* **178**, 5644–5651.
McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.
Mollerach, M., López, R. & García, E. (1998). *J. Exp. Med.* **188**, 2047–2056.
Müller, V., Spanheimer, R. & Santos, H. (2005). *Curr. Opin. Microbiol.* **8**, 729–736.
Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). *Acta Cryst.* D**67**, 355–367.
Neelon, K., Roberts, M. F. & Stec, B. (2011). *Biophys. J.* **101**, 2816–2824.
Neelon, K., Wang, Y., Stec, B. & Roberts, M. F. (2005). *J. Biol. Chem.* **280**, 11475–11482.
Nikawa, J. & Yamashita, S. (1997). *Biochim. Biophys. Acta*, **1348**, 173–178.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Roberts, M. F. (2005). *Saline Syst.* **1**, 5.
Rodionov, D. A., Kurnasov, O. V., Stec, B., Wang, Y., Roberts, M. F. & Osterman, A. L. (2007). *Proc. Natl Acad. Sci. USA*, **104**, 4279–4284.
Rodrigues, M. V., Borges, N., Henriques, M., Lamosa, P., Ventura, R., Fernandes, C., Empadinhas, N., Maycock, C., da Costa, M. S. & Santos, H. (2007). *J. Bacteriol.* **189**, 5405–5412.
Santos, H. & da Costa, M. S. (2001). *Methods Enzymol.* **334**, 302–315.
Scholz, S., Sonnenbichler, J., Schäfer, W. & Hensel, R. (1992). *FEBS Lett.* **306**, 239–242.
Stec, B., Yang, H., Johnson, K. A., Chen, L. & Roberts, M. F. (2000). *Nature Struct. Biol.* **7**, 1046–1050.
Stieglitz, K. A., Roberts, M. F., Li, W. & Stec, B. (2007). *FEBS J.* **274**, 2461–2469.
Takenoya, M., Nikolakakis, K. & Sagermann, M. (2010). *J. Bacteriol.* **192**, 6056–6063.
Thoden, J. B. & Holden, H. M. (2007). *Protein Sci.* **16**, 1379–1388.
Thorson, J. S., Kelly, T. M. & Liu, H.-W. (1994). *J. Bacteriol.* **176**, 1840–1849.
Winn, M. D. *et al.* (2011). *Acta Cryst.* D**67**, 235–242.
Wolfe-Simon, F., Switzer Blum, J., Kulp, T. R., Gordon, G. W., Hoeft, S. E., Pett-Ridge, J., Stolz, J. F., Webb, S. M., Weber, P. K., Davies, P. C., Anbar, A. D. & Oremland, R. S. (2011). *Science*, **332**, 1163–1166.
Xing, W., Zou, Y., Liu, Q., Liu, J., Luo, X., Huang, Q., Chen, S., Zhu, L., Bi, R., Hao, Q., Wu, J.-W., Zhou, J.-M. & Chai, J. (2007). *Nature (London)*, **449**, 243–247.
Yamashita, S. & Nikawa, J. (1997). *Biochim. Biophys. Acta*, **1348**, 228–235.
Zhang, Y. (2008). *BMC Bioinformatics*, **9**, 40.
Zhang, Z., Bulloch, E. M. M., Bunker, R. D., Baker, E. N. & Squire, C. J. (2009). *Acta Cryst.* D**65**, 275–283.
Zuccotti, S., Zanardi, D., Rosano, C., Sturla, L., Tonetti, M. & Bolognesi, M. (2001). *J. Mol. Biol.* **313**, 831–843.