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

# Protein HP1028 from the human pathogen *Helicobacter pylori* belongs to the lipocalin family

**Nicola Barison,[a]‡ Laura Cendron,[a]§ Valentina Loconte,[a] Elizabeth A. Proctor,[b] Nikolay V. Dokholyan[b] and Giuseppe Zanotti[a,c]***

[a]Department of Biomedical Sciences, University of Padua, Viale G. Colombo 3, 35131 Padova, Italy, [b]Department of Biochemistry and Biophysics, University of North Carolina, Chapel Hill, North Carolina, USA, and [c]Venetian Institute of Molecular Medicine (VIMM), Via Orus 2, 35129 Padova, Italy

‡ Present address: Max Planck Institute for Infection Biology, Charitéplatz 1, D-10117 Berlin, Germany.
§ Present address: Department of Biology, University of Padua, Viale G. Colombo 3, 35131 Padova, Italy.

Correspondence e-mail: giuseppe.zanotti@unipd.it

Received 10 January 2013
Accepted 25 March 2013

**PDB Reference:** HP1028, 4inn

*Helicobacter pylori* is a bacterial pathogen that causes severe diseases, including gastritis, ulcers and gastric cancer. Although this bacterium has been extensively studied, the physiological functions of a large number of the proteins encoded by its genome are unknown. HP1028 is a protein that is relevant to colonization and to the survival of the bacterium in the stomach, but its function is not clearly understood. Bioinformatics studies suggest that HP1028 is a monomeric protein that is secreted in the *H. pylori* periplasm. The crystal structure of HP1028 has been determined at 2.6 Å resolution using the SAD method. The three-dimensional structure of the protein reveals that it belongs to the lipocalin family, a group of proteins that bind and transport (often hydrophobic) small molecules. The structure of HP1028, together with the possible localization of the mature protein in the bacterial periplasm and the position of the *hp*1028 gene in the bacterial genome, point to a role in *H. pylori* chemotaxis.

## 1. Introduction

*Helicobacter pylori* is a Gram-negative bacterial pathogen that affects about half of the human population. Although infection is mostly asymptomatic, *H. pylori* is responsible for severe gastroduodenal pathologies, including gastritis, ulcers and eventually gastric adenocarcinoma and MALT lymphoma (Blaser, 1998; Peek, 2002; Rothenbacher & Brenner, 2003). Despite the identification of the bacterium dating back to 1984 (Warren, 2006), its pathogenesis remains poorly understood at the molecular level. Indeed, it is estimated that the function of approximately 40–50% of the ∼1550 genes encoded within its genome is hypothetical or completely unknown. The *H. pylori* *hp*1028 gene is paradigmatic of this situation. Theoretical analysis based on bioinformatics data (Shannon *et al.*, 2006) identified a possible link between the *hp*1028 gene and the cytotoxicity-associated PAI genes. In addition, the HP1028 gene product was predicted to share structural similarity with the B-chain of ebulin, a ricin-like toxin, suggesting that the HP1028 protein may play a role in *H. pylori* pathogenesis. Moreover, the role of *hp*1028 has been analysed in a study aimed at identifying the *H. pylori* genes that are responsible for stomach colonization (Baldwin *et al.*, 2007). A C57BL/6 mouse infection model was used to query a collection of 2400 transposon mutants in two different bacterial strain backgrounds (NSH79 and NSH57). The analysis reported 223 putative genes involved in stomach colonization, including previously described *H. pylori* virulence genes, some of which had been implicated in virulence in other pathogenic bacteria, and 81 unknown genes encoding hypothetical proteins. Among the unknown genes, knockout of *hp*1028 showed

strong phenotypes in both of the *H. pylori* strain backgrounds tested, confirming the relevance of *hp*1028 to *H. pylori* pathogenicity.

As far as gene organization is concerned, *hp*1028 is connected *via* gene-cluster operon edges to *hp*1029, *fliY* (*hp*1030) and *fliM* (*hp*1031), which encode key flagellar switch proteins. An alternate sigma factor ($\sigma^{28}$) gene *fliA* (hp1032) is also present in this operon. *fliA* is involved in the transcription of FlaA, the major flagellar subunit required for both motility (Josenhans *et al.*, 1995) and gastric colonization (Mankoski *et al.*, 1999). *hp*1027, which encodes a protein (ferric uptake regulator protein Fur) involved in the regulation of urease expression in response to nickel (van Vliet *et al.*, 2001; Dian *et al.*, 2011), is also located close to the *hp*1028 gene. Although *hp*1028 shares no significant sequence similarity with proteins of known function, sequence-alignment analysis with *BLASTp* (http://www.blast.ncbi.nlm.nih.gov/Blast.cgi) identified a conserved domain, named COG4731, that is present in bacteria such as the Burkholderiales, Campylobacterales, Caulobacterales, Rhizobiales and Spirochaetales. The function and the structure of the COG4731 domain, which belongs to the DUF2147 superfamily, are unknown.

Here, we report the three-dimensional structure of HP1028 determined by X-ray crystallography. The crystal structure provides new suggestions about the possible function of the protein.

## 2. Materials and methods

### 2.1. Cloning of HP1028

The *hp*1028 gene was cloned from *H. pylori* strain CCUG17874. Since analysis with the *SignalP* bioinformatics tool (http://www.cbs.dtu.dk/services/SignalP/) identified a signal peptide at the N-terminus and predicted a proteolytic site between the 16th and 17th amino acids of the protein sequence, the recombinant construct was designed to exclude the signal peptide and to add an N-terminal His tag flanked by a cleavage sequence for TEV protease (Fig. 1). *hp*1028 was PCR-amplified from *H. pylori* strain CCUG17874 genomic DNA using the primers 5′-**CACC**GTAGAGTTGCCTGGAATTTATC-3′ (forward; topoisomerase recognition site in bold) and 5′-CTATTAGATAGGGCTATCTTTAATG-3′ (reverse) and inserted into the pET151/D-TOPO expression vector (Invitrogen). In addition, since the amino-acid sequence of the protein contained only one internal methionine residue, in order to solve the phase problem mutants with one and two additional methionine residues (pET151-*hp1028-L129M* and pET151-*hp1028-L129M-A110M*, respectively) were produced using the QuikChange site-directed mutagenesis kit (Stratagene). The mutated amino acids, L129M and A110M, were selected as the most variable amino acids in a multiple alignment of all of the HP1028 sequences available from different *H. pylori* strains.

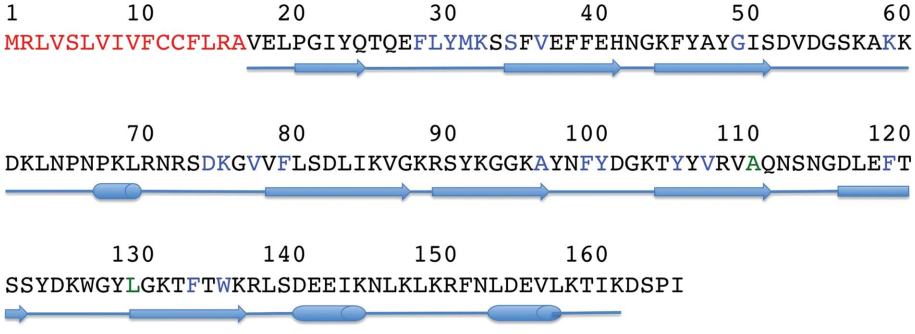



**Figure 1**
The amino-acid sequence of HP1028. The positions of secondary-structure elements are shown as arrows ($\beta$-strands) and cylinders ($\alpha$-helices). Residues in red (1–16) represent the predicted export signal sequence. Residues that were eventually mutated to methionine are shown in green. Amino acids shown in blue are those that line the inner cavity (see text for further details).

### 2.2. Expression and purification

*Escherichia coli* BL21(DE3) cells harbouring pET151-*hp1028* or the mutants were grown at 310 K in 3 l LB medium containing 100 µg ml$^{-1}$ ampicillin until an $OD_{600}$ of 0.6 was reached. Protein expression was induced by adding 1 m*M* isopropyl $\beta$-D-1-thiogalactopyranoside (IPTG) at 303 K. Cells were harvested by centrifugation and suspended in buffer *A* (30 m*M* Tris–HCl pH 7.5, 250 m*M* NaCl, 10 m*M* galactose, 0.05% $NaN_3$, 55 µ*M* $C_{12}E_8$). Resuspended cells were lysed by incubation with 1 mg ml$^{-1}$ lysozyme and sonication in the presence of protein inhibitors (Protease Inhibitor Cocktail, Sigma–Aldrich). After centrifugation (25 min at 40 000*g*), the soluble fraction was loaded onto a 1 ml HisTrap column (GE Healthcare) equilibrated with buffer *A*. After extensive washing with buffer *A* supplemented with 40 m*M* imidazole, the protein was eluted using an imidazole gradient (40–500 m*M*). The His tag was cleaved by enzymatic digestion at 277 K overnight by adding 1.5 mg $His_6$-tagged TEV protease, 0.5 m*M* EDTA and 1 m*M* DTT to the eluted fractions. TEV protease as well as any unprocessed $His_6$-HP1028 were removed by loading the protein solution onto a 1 ml HisTrap column and collecting the flowthrough. The protein was concentrated by ultrafiltration (Vivaspin 15R 5000 MWCO, Sartorius) and loaded onto an XK 16/20 Sephacryl S-100 HR size-exclusion column (GE Healthcare). The monodispersity and the molecular radius of the protein were evaluated by dynamic light scattering (Zetasizer Nano ZS, Malvern Instruments Ltd). All purification steps were verified by gel electrophoresis under denaturing conditions. For the expression of selenomethionine-derivatized proteins, a methionine-auxotroph *E. coli* strain (B834) was used according to a previously described procedure (Cendron *et al.*, 2009). Selenomethionine-labelled protein was purified analogously to native mature HP1028 with a similar yield.

### 2.3. Crystallization and data collection

The purified HP1028 protein was concentrated to 35 mg ml$^{-1}$ for crystallization trials. Crystallization conditions

were identified by the sitting-drop isothermal vapour-diffusion method with an Oryx8 crystallization robot (Douglas Instruments). Preliminary crystals of native protein were obtained in two different conditions with screening solutions 9 (0.2 *M* lithium sulfate, 0.1 *M* Tris–HCl pH 8.5, 30% PEG 4000) and 35 (20% 2-propanol, 0.1 *M* sodium citrate pH 5.6, 20% PEG 4000) from Structure Screen I (Molecular Dimensions Ltd, UK). While the crystals from the latter solution diffracted to a maximum resolution of 4 Å, the crystals obtained using solution 9 belonged to space group $P2_12_12_1$ and diffracted to a maximum resolution of 3.3 Å. The mutations introduced to enrich the methionine content (L129M and L129M/A110M) resulted in an improvement in the quality and the resolution (maximum resolution 2.6 Å) of the diffraction patterns and were used to solve the phase problem and refine the structure. Crystals were cooled in liquid nitrogen after soaking them in a cryoprotectant solution consisting of the precipitant with 10% glycerol or 10% PEG 400 added. A fluorescence emission scan at the selenium edge was performed to optimize the energy for anomalous measurements. Diffraction data were measured on beamline ID14-4 at the European Synchrotron Radiation Facility (ESRF), Grenoble (ADSC Q315r X-ray detector, 1° oscillation per image, 100 K nitrogen stream). All data sets were integrated using *MOSFLM* (Leslie, 2006) and merged with *SCALA* (Evans, 2006) as implemented in *CCP*4 (Winn *et al.*, 2011). Crystals of the L129M variant had unit-cell parameters $a = 34.24$, $b = 79.15$, $c = 131.67$ Å. They contained two molecules in the asymmetric unit, with a $V_M$ of 2.54 Å$^3$ Da$^{-1}$ and an estimated solvent content of ∼51%. Data-collection statistics are given in Table 1.

**Table 1**
Data-collection and refinement statistics.

Values in parentheses are for the highest resolution shell. 1° rotations were used. One crystal was used to collect diffraction data for SAD anomalous phasing.

| | |
|---|---|
| Data collection | |
| Space group | $P2_12_12_1$ |
| Wavelength (Å) | 0.9795 |
| Unit-cell parameters (Å) | $a = 34.24$, $b = 79.15$, $c = 131.67$ |
| Resolution range (Å) | 50.71–2.60 (2.74–2.60) |
| $R_{merge}$ | 0.108 (0.225) |
| $\langle I/\sigma(I)\rangle$ | 19.0 (9.5) |
| Completeness (%) | 100 (99.9) |
| Multiplicity | 13.4 (13.9) |
| Unique reflections | 11665 (1656) |
| Anomalous SAD phasing | |
| Figure of merit (centric/acentric) | 0.22/0.07 |
| Phasing power | 0.796 |
| Final score after density modification | 2.78 |
| Refinement | |
| $R_{work}/R_{free}$ | 0.1938/0.2477 |
| No. of protein atoms | 2421 |
| No. of solvent atoms | 64 |
| No. of ligand atoms | 22 |
| Mean $B$ value (Å$^2$) | 26.95 |
| R.m.s.d. from ideal values | |
| Bond lengths (Å) | 0.013 |
| Bond angles (°) | 1.97 |
| Ramachandran plot (%) | |
| Favoured regions (%) | 92.3 |
| Allowed regions (%) | 6.6 |
| Generously allowed regions (%) | 1.2 |
| Disallowed regions (%) | 0 |
| Overall $G$-factor | −0.1 |

### 2.4. Structure determination and refinement

Several anomalous data sets were measured, but owing to the low quality of the data and the low ratio of anomalous signal the phasing power was too poor to allow correct chain tracing. Crystals were generally not single and showed anisotropic diffraction patterns. For the latter reason, the data were truncated to 2.6 Å resolution. Attempts to solve the structure by molecular replacement using a theoretical model built by molecular dynamics (see §2.5) were unsuccessful. Finally, after reprocessing the diffraction data sets and using release 2.8.1 of the *SHARP*/*autoSHARP* suite (Vonrhein *et al.*, 2007; Schiltz & Bricogne, 2008), we were able to obtain interpretable maps from a single data set using the SAD approach. All four Se atoms (two for each monomer) of the L129M variant were correctly located. Despite the low figure of merit of the initial phases, subsequent cycles of density modification allowed the software *Buccaneer* (Cowtan, 2010) to build a model that was essentially correct for molecule *A* with the exception of a short strand at the N-terminus, whilst some errors were present in molecule *B*.

Refinement proceeded smoothly using *CNS* (Brunger, 2007) in the first stages and *REFMAC*5 in the final steps. Noncrystallographic symmetry restraints (tight for the main chain and loose for the side chains) were used throughout the refinement process. The graphical software *Coot* (Emsley *et al.*, 2010) was used to display and to manually correct the model. Solvent molecules were introduced using the automatic procedure in *REFMAC*5 and were manually corrected. In the final stages of the refinement a long density shape was visible in the inner cavity (see below). This density extended continuously from the cavity of one molecule of the asymmetric unit to the other. Given the shape of the density, it was interpreted as a fragment of PEG used as a cryoprotectant or in the crystallization experiments. The final model contained two protein chains, 64 solvent molecules and one PEG molecule, with a crystallographic $R$ factor of 0.1938 and an $R_{free}$ (Brünger, 1997) of 0.2477. The stereochemistry of the final model was checked using *PROCHECK* (Laskowski *et al.*, 1996).

### 2.5. Modelling using discrete molecular dynamics (DMD)

A homology model for HP1028 was constructed using major urinary protein 2 from mouse (PDB entry 2nne; Redondo *et al.*, 2008) as a template. Major urinary protein 2 has 17% identity and 32% similarity to HP1028, as well as 99% coverage, as determined by the NCBI *PSI-BLAST* tool (Altschul *et al.*, 1997). Using the alignment performed by *PSI-BLAST*, necessary deletions or insertions to the template protein were identified. The Eris Suite (Yin *et al.*, 2007*a*,*b*) and Medusa force fields (Ding & Dokholyan, 2006) were used to impose the sequence of HP1028 onto the structure of major urinary protein 2, and regions that were determined to be deletions in HP1028 were removed. The protein backbone was

resealed in the DMD simulations (Shirvanyants *et al.*, 2012) by applying peptide-bond constraints to the backbone of the residues on both sides of the deletion. Sequences determined to be insertions in HP1028 were introduced by removing two residues on each side of the insertion site from the template structure and creating a fragment containing the insertion and these four residues, which were then sealed into the template structure by applying peptide-bond constraints to the proper site in a manner similar to that used for deletions. Upon long equilibration ($10^6$ time steps, or approximately 50 ns) of the new HP1028 model, the structures obtained after energetic equilibriation were clustered and the centroid structure of the largest cluster was selected as the model. The resulting model was further minimized using *Chiron*, a computational tool for resolving clashes (Ramachandran *et al.*, 2011), and was verified using *Gaia*, a tool for model validation (Kota *et al.*, 2011).

## 3. Results and discussion

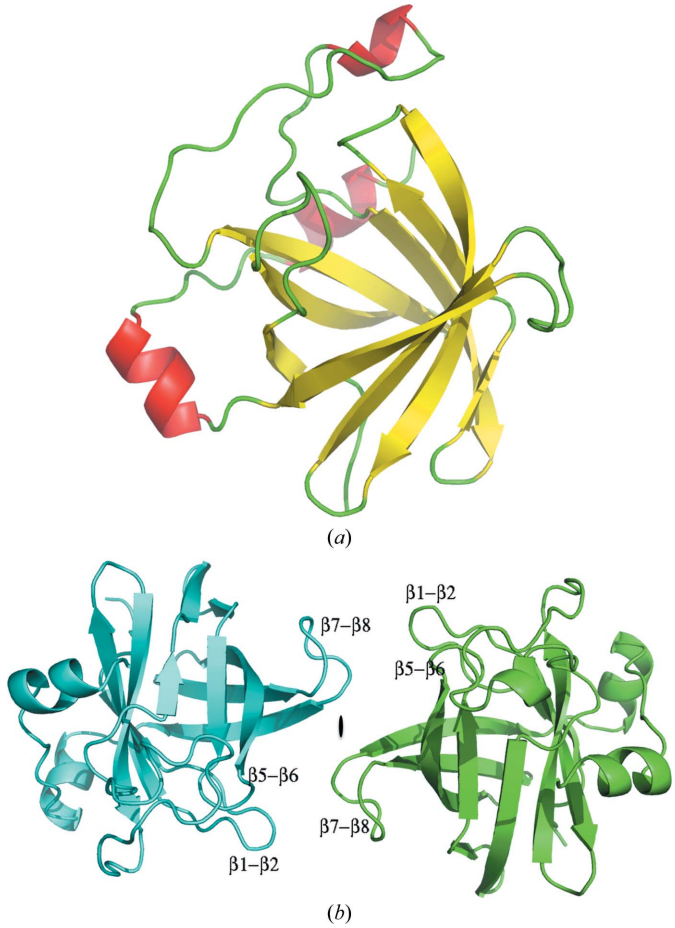
### 3.1. The monomer structure

The crystal structure of HP1028 from *H. pylori* consists of two monomers. The electron density for each monomer is well defined for the entire main chain from amino acids 16 (15 for molecule *B*) to 161 (Supplementary Fig. S1[1]). Since the first 16 amino acids in the sequence were predicted to belong to an export signal peptide and were not included in our construct, all of the amino acids of the recombinant mature protein are visible with the exception of the last two, which are possibly disordered. Side chains are generally well defined, with the exception of some lysine side chains.

The HP1028 monomer shows a classical lipocalin fold (Cowan *et al.*, 1993). The molecule core is a continuous $\beta$-barrel formed by eight antiparallel $\beta$-strands followed by a long stretch (residues 138–163) that packs onto the barrel and includes two short $\alpha$-helices (Fig. 2*a* and Supplementary Fig. S2). Owing to the twist of each $\beta$-sheet, the barrel assumes a calyx shape and forms an internal cavity that hosts the ligands in other lipocalins such as retinol-binding protein (RBP; Cowan *et al.*, 1990; Zanotti, Berni *et al.*, 1993). In the HP1028 structure the inner cavity is defined at one end by the amino-terminal chain and by the bulky side chains of several residues, whilst the entrance at the opposite end is accessible to the solvent. The latter is lined by loops $\beta1$–$\beta2$, $\beta5$–$\beta6$ and $\beta7$–$\beta8$ (Fig. 2*b*).

### 3.2. The binding cavity

The inner cavity formed by the HP1028 $\beta$-barrel, the shape of which is shown in Fig. 3(*a*), differs significantly from that found in other lipocalins. In members of the lipocalin family such as RBP, avidin and PAGP the cavity lies roughly parallel to the barrel axis. In HP1028 the side of the cavity that runs along the axis of the barrel is instead obstructed by bulky residues, specifically Phe79, Phe119 and Trp135. At the same time, different conformations of strands $\beta4$ and $\beta5$ and of the long loop $\beta3$–$\beta4$ allow the cavity to expand in their direction. As a consequence, the long direction of the cavity runs nearly perpendicular to the barrel axis and its bottom corresponds to the upper part of strand $\beta3$ and the final part of loop $\beta3$–$\beta4$ (Fig. 3*b*). The cavity is closed at the bottom by the side chains of the charged residues Lys75, Asp74 and Lys59 and by the main chains of Val77 and Gly49. The lower part of the cavity is lined on one side by hydrophobic residues (Phe99, Phe79, Val36 and Tyr100) and on the opposite side by Ser34 and Lys32. The positive charge of the latter is neutralized by Asp52. The middle and upper parts of the cavity are mostly lined by hydrophobic residues (Trp135, Met31, Ala96, Phe119, Phe133, Val107, Tyr105, Tyr30, Phe28 and Leu29) and some hydrophilic residues (Thr25, Ser121, Asn98 and Gln26). At the opening the cavity becomes larger and is surrounded by charged amino-acid side chains (Asp101, Lys131 and Lys103). The side chain of Tyr123 also hangs over the cavity. This




(*a*)

(*b*)

**Figure 2**
(*a*) Cartoon drawing of HP1028. Secondary-structure elements are shown in different colours. The $\beta$-barrel core of the protein can be seen through the opening of the cavity. (This figure, together with Figs. 3 and 4, was produced using *MacPyMOL*; http://www.pymol.org.) (*b*) The two monomers present in the asymmetric unit are related by a noncrystallographic twofold axis approximately perpendicular to the plane of the paper. The two monomers are essentially identical, with an r.m.s.d. of 0.53 Å. The loops involved in the monomer–monomer interface are labelled.

[1] Supplementary material has been deposited in the IUCr electronic archive (Reference: GM5022). Services for accessing this material are described at the back of the journal.

distribution of polar and hydrophobic amino-acid side chains suggests a specific interaction with an amphipathic ligand.

In the crystal of the HP1028 L129M mutant that was used to solve the structure, a long thin string of electron density is visible inside the protein cavity (Supplementary Figs. S3 and S4). This density spans both molecules present in the asymmetric unit and we have fitted a PEG fragment of 22 atoms to it. PEG 400 was used in the cryoprotectant and is the only molecule among those added during the purification and crystallization procedures that corresponds to the shape of the density. The ligand does not entirely fill the cavity and probably does not properly mimic the physiological ligand. In addition, maps calculated from two other data sets collected from a crystal of the double mutant L129M/A110M showed a cavity that was either empty or filled with smaller blobs of density. In all cases the protein structure is unaltered, suggesting that the lipocalin scaffold is quite rigid and that the transition from the apo to the holo protein does not induce major conformational changes in the $\beta$-barrel. Indeed, in the cases of other lipocalins, such as RBP (Zanotti, Berni *et al.*, 1993), avidin (Livnah *et al.*, 1993; Nardone *et al.*, 1998) and streptavidin (Pazy *et al.*, 2002), differences between liganded and unliganded forms of the protein are also limited to minor rearrangements close to the opening of the cavity.

### 3.3. Quaternary assembly

The two protein molecules present in the asymmetric unit of the crystal are related to each other by a noncrystallographic twofold axis. The two monomers face each other through the opening of the loops that surrounds the entrance to the protein inner cavity (Fig. 2*b*). The interactions are mostly hydrophobic (Phe28, Leu29, Tyr30, Met31, Tyr105, Tyr123, Trp126, Tyr128, Met129 and Gly130), whilst hydrogen bonds or salt bridges are totally absent, despite the presence of some charged residues (Asp101, Asp124, Lys103 and Lys131) at the interface. Of particular relevance is the presence of a $\pi$-interaction between the phenyl rings of the two Tyr123s (Supplementary Fig. S5). Bioinformatics analysis performed with *PISA* (Krissinel & Henrick, 2007) suggests that the dimer present in the crystal is not suggestive of a physiological oligomer, since the buried surface area after dimer formation is quite modest (814 Å$^2$ from a total surface of 17 079 Å$^2$; CSS = 0.1). These results change significantly if the same analysis is performed in the presence of the ligand: the total buried surface becomes 1930 Å$^2$ (CSS = 0.61) and in addition the presence of a ligand that spans two molecules contributes to the stabilization of the dimer. Since the ligand observed in one of the crystal forms examined is an artefact that possibly favours crystallization but does not represent the physiological state, we are tempted to conclude that the protein in the apo form is present as a monomer *in vivo* and dimerization in the crystal could take place as a consequence of binding of the ligand.

### 3.4. Comparison with other lipocalins

Supplementary Table S2 lists the proteins that are structurally most similar to HP1028 according to the *DALI* server (Holm & Rosenström, 2010). Most of them belong to the lipocalins, a large family of proteins that share a common fold characterized by an eight-stranded up-and-down $\beta$-barrel that

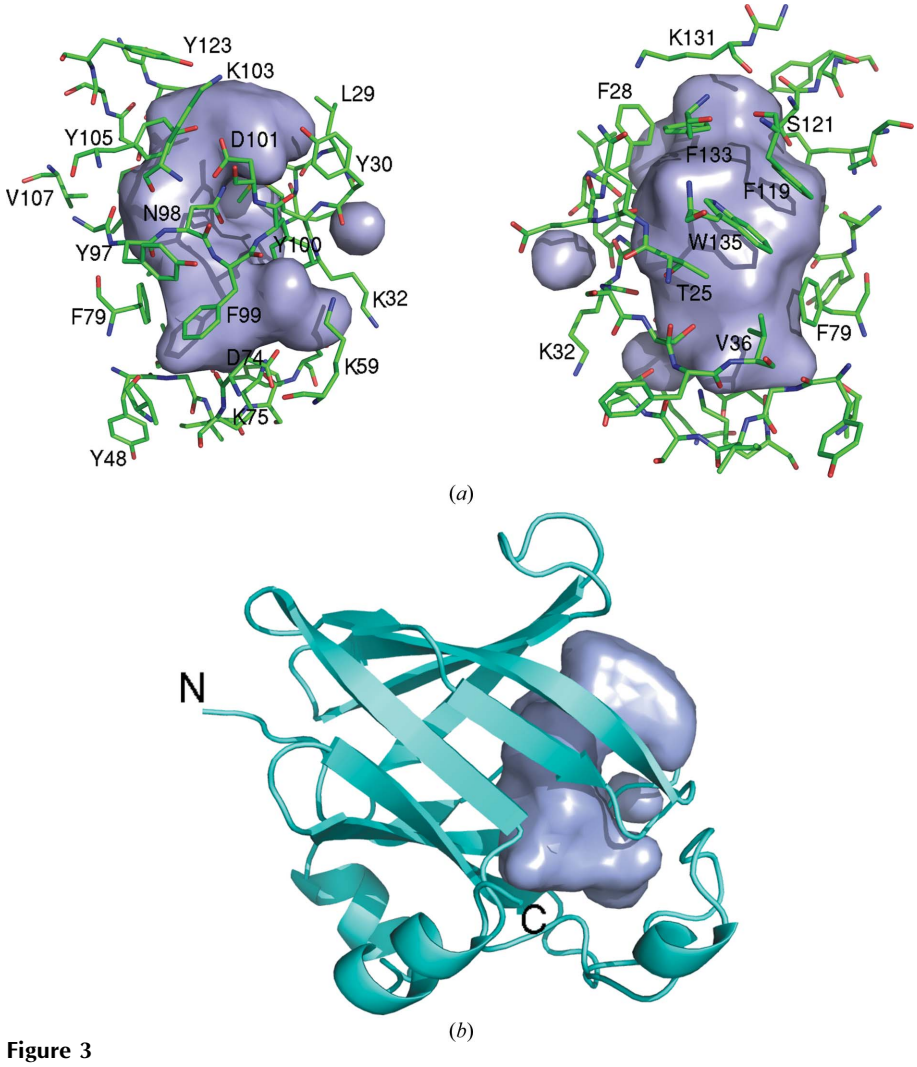



**Figure 3**
(*a*) Surface of the inner cavity of HP1028 as defined by *HOLLOW* (Ho & Gruswitz, 2008). The two faces of the cavity are shown: the partially hydrophilic face (left side) and the more hydrophobic face (right side). (*b*) Cartoon view of HP1028 showing the position of the cavity in the overall structure. The axis of the $\beta$-barrel is approximately horizontal, whilst the long direction of the cavity is roughly vertical. N and C indicate the N- and C-termini of the protein.

generates the internal cavity in which one or more ligands are hosted (Flower, 2000). The prototype of lipocalins is retinol-binding protein (RBP), the transporter of the alcoholic form of vitamin A in plasma (Cowan *et al.*, 1993). Others members of the family presented in Supplementary Table S2 are avidin (Livnah *et al.*, 1993), streptavidin (Katz & Cass, 1997), odorant-binding protein (Bianchet *et al.*, 1996; Tegoni *et al.*, 1996), crustacyanin (Cianci *et al.*, 2001), human tear lipocalin (Breustedt *et al.*, 2005) and PGAP (Cuesta-Seijo *et al.*, 2010). A protein family that is closely related to lipocalins is the fatty acid-binding protein (FABP) family (Sacchettini *et al.*, 1989). However, the FABP binding cavity is made up of ten anti-parallel $\beta$-strands instead of eight. In addition, in FABP the cavity is excluded from the outside solvent and a significant conformational change must take place in order to allow the entrance of the ligand to the cavity. From the structural point of view HP1028 definitely belongs to the lipocalin family, despite the fact that its sequence identity to other members of the family does not exceed 12%. The most similar protein in terms of sequence identity is Der f 13 (Chan *et al.*, 2006), which shares 18% identity over a total of 83 aligned amino acids. However, Der f 13 is a member of the FABP family and its r.m.s.d. with HP1028 is 3.2 Å.

A superimposition between HP1028, avidin and RBP, with the last two being representative members of the lipocalin family (Fig. 4*a*), shows significant differences between our structure and the other members of the lipocalin family. Whilst overall the $\beta$-barrel cores superimpose quite well, the loops connecting them differ significantly; specifically, an $\alpha$-helical turn is present in the HP1028 loop connecting strands $\beta$3 and $\beta$4. The major difference involves the C-terminal tail, which includes 28 residues (138–165) in HP1028. Indeed, HP1028 reveals two short $\alpha$-helices instead of a long $\alpha$-helix, and the tail itself is oriented in the opposite direction with respect to the other lipocalins.

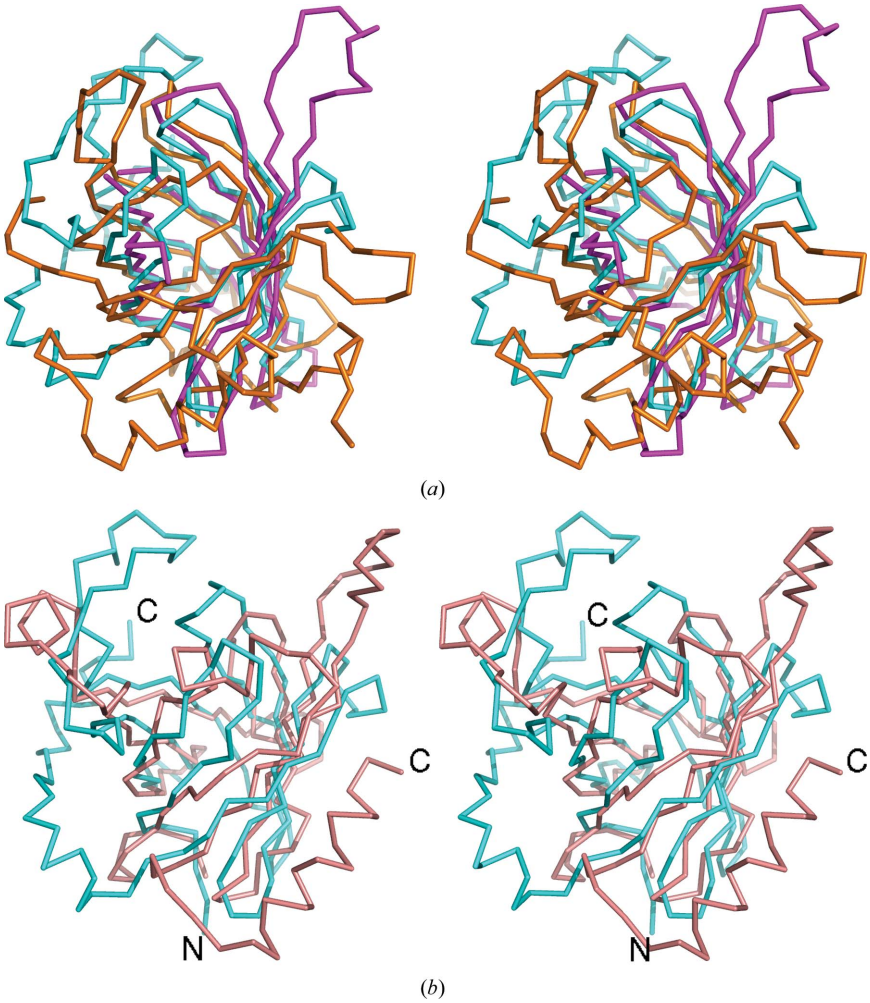



**Figure 4**
(*a*) Stereoview of the superimposition of the C$^{\alpha}$ chain trace of HP1028 (cyan) on RBP (orange, PDB entry 1brq; Zanotti, Ottonello *et al.*, 1993) and avidin (magenta; PDB entry 1ldq; Pazy *et al.*, 2002). The r.m.s.d.s for the superposition of 87 and 84 aligned C$^{\alpha}$ atoms of RBP and avidin on HP1028 are 3.0 and 3.1 Å, respectively. (*b*) Stereoview of the superimposition of the C$^{\alpha}$ chain traces of the HP1028 structure (cyan) and the molecular-dynamics model (salmon). The r.m.s.d. between the two structures is 2.98 Å for 91 aligned C$^{\alpha}$ atoms. The major difference between the two models is in the C-terminal tail regions, whilst the positions of the two N-termini are in rough agreement.

### 3.5. Comparison of the HP1028 structure with the molecular-dynamics model

A superimposition of the final HP1028 structure with the model obtained by discrete molecular dynamics (Shirvanyants *et al.*, 2012) is illustrated in Fig. 4(*b*). The overall r.m.s.d. between equivalent C$^{\alpha}$ atoms of 98 residues of the core molecule is 2.98 Å. In particular, the C-terminal tail points in the wrong direction, since it follows the orientation of the tail in other lipocalins. In addition, we note that residues are generally shifted in secondary-structural elements by two or three positions with respect to the final crystallographic structure, which could be the result of inaccuracies in alignment. Nevertheless, these two aspects do not represent the reason for the failure of the molecular-replacement method to find a correct solution, since an attempt performed *a posteriori* using only the $\beta$-barrel core of the molecule and a polyalanine model also failed. A major drawback is possibly represented by the generally inaccurate positioning of the chain, in particular of the loops that connect the $\beta$-strands. They contribute strongly to the high r.m.s.d. between the theoretical model and the crystal structure. It must be mentioned that this inaccuracy is probably dependent on the incorrect positioning of the amino acids in the secondary-structural elements along the main chain in the starting model. Another reason why the model could not be used successfully for molecular replacement possibly resides in the presence of two molecules in the asymmetric unit, a fact that

makes molecular replacement more difficult, and in the shape of the molecule. For molecules that are quite isotropic, such as lipocalins, finding a solution for the rotation function is more challenging. Nevertheless, molecular-dynamics techniques are improving continuously and hopefully in the near future it will be possible to successfully use models that have been built *ab initio* for molecular-replacement structure solution.

### 3.6. Protein function

The product of gene *hp*1028 is annotated in the databases as a putative protein of unknown function. Its crystal structure clearly demonstrates that it belongs to the lipocalin protein family. Lipocalins are characterized by the property of binding a small ligand inside their cavity and this feature generally defines their function. Since HP1028 was expressed in a heterologous system and is present in the crystals either in the apo form or as a complex with ligands that bind during the purification or crystallization process that could mimic the holo form, we could not describe its physiological ligand. Since the cavity surface is amphiphilic, we hypothesize that the ligand is a molecule with both hydrophobic and hydrophilic features. In particular, the portion of the ligand exposed to the entrance of the cavity could be hydrophilic or negatively charged owing to the presence of charged residues in this area. Following the three-dimensional structure similarities identified by reverse-template comparison (using the *ProFunc* server; http://www.ebi.ac.uk/thornton-srv/databases/ProFunc/), a putative ligand could share the features of fatty acids, such as in the case of FABP, of biotin, the substrate of avidin and streptavidin, or of amines such as histamine, serotonin and epinephrine. The latter represent substrates of monomines, which are lipocalins that play the role of biogenic amine scavengers developed by ticks and other blood-feeding arthropods to avoid the immune response (Mans *et al.*, 2008).

A clue to the function of the protein can be suggested by the position of the *hp*1028 gene in the *H. pylori* genome (Fig. 5). *hp*1028 is positioned last in a group of genes involved in control of flagella movement. Specifically, HP1029 is annotated as a flagellar motor switch protein FliM2 in strain H-45, and *hp*1030 and *hp*1031 encode the flagellar switch components FliY and FliM, respectively. All of these genes are under the control of the same promoter, suggesting that their function is linked. Taking together the facts that HP1028 is predicted to be secreted in the periplasm and that it has the property of binding a small ligand, we hypothesize that HP1028 is involved in *H. pylori* chemotaxis. In Gram-negative bacteria, chemoeffectors diffuse into the periplasm and interact with sensory receptors. The signal is then transduced to a two-component system such as CheA–CheW, which regulates chemotaxis (Lertsethtakarn *et al.*, 2011; Sourjik & Wingreen, 2012). The interactions of the chemoeffectors with the receptors can be direct or indirect through periplasmic binding proteins. HP1028 may act as a periplasmic binding protein that detects amphiphilic molecules in the environment.

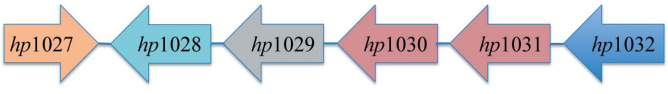



**Figure 5**
Local organization of the *H. pylori* genome around the *hp*1028 gene. Arrows indicate the directions of single genes. Their annotations in UniProtKB (http://www.uniprot.org) are as follows: *Hp*1027, ferric uptake regulator protein Fur; *hp*1029, flagellar motor switch protein FliM2; *hp*1030, flagellar switch component FliY; *hp*1031, flagellar switch component FliM; *hp*1032, FliA.

## 4. Conclusions

The crystal structure of HP1028, which is annotated in the databases as a hypothetical protein of unknown function, demonstrates that it belongs to the lipocalin family, a group of proteins that bind small ligands. Indeed, HP1028 has a binding pocket that contains a small molecule in some crystal forms. Structural analysis of this HP1028 cavity suggests that its physiological ligand, despite still being chemically undefined, is a small amphiphilic molecule. Within the *H. pylori* genome, the *hp*1028 gene is localized next to genes encoding flagellar switch proteins and flagellar transcription factors, suggesting a role of HP1028 in chemotaxis. In addition, bioinformatics analysis of the HP1028 sequence suggests its localization in the periplasm. Taken together, this evidence indicates that HP1028 is a periplasmic binding protein that is involved in the regulation of *H. pylori* chemotaxis. As such, it is required for stomach colonization and persistence, as demonstrated previously (Baldwin *et al.*, 2007). However, the biological function of the protein needs to be confirmed by further studies. A deeper analysis would also clarify the exact nature of its ligand(s) and the importance of HP1028 in *H. pylori* pathogenesis. Notably, the COG4731 domain contained by HP1028 seems to be conserved in several proteins of unknown function from different bacteria. Therefore, understanding the function of HP1028 may provide new insights into the chemotactic mechanisms of bacteria during infection or niche colonization.

We thank the staff of beamline ID14-4 at the European Synchrotron Radiation Facility (ESRF), Grenoble, France for technical assistance during data collection. We are grateful to Alberto Danielli, Gianluca De Bellis and Riccardo Percudani for discussion and for critical analysis of the genome sequence. This work was supported by University of Padua grant 'Progetto di Ateneo 2011' and by PRIN 2010–2011 'Unraveling structural and functional determinants behind *Helicobacter pylori* pathogenesis and persistence'.

## References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). *Nucleic Acids Res.* **25**, 3389–3402.
Baldwin, D. N., Shepherd, B., Kraemer, P., Hall, M. K., Sycuro, L. K., Pinto-Santini, D. M. & Salama, N. R. (2007). *Infect. Immun.* **75**, 1005–1016.
Bianchet, M. A., Bains, G., Pelosi, P., Pevsner, J., Snyder, S. H., Monaco, H. L. & Amzel, L. M. (1996). *Nature Struct. Biol.* **3**, 934–939.

Blaser, M. J. (1998). *BMJ*, **316**, 1507–1510.
Breustedt, D. A., Korndörfer, I. P., Redl, B. & Skerra, A. (2005). *J. Biol. Chem.* **280**, 484–493.
Brünger, A. T. (1997). *Methods Enzymol.* **277**, 366–396.
Brunger, A. T. (2007). *Nature Protoc.* **2**, 2728–2733.
Cendron, L., Couturier, M., Angelini, A., Barison, N., Stein, M. & Zanotti, G. (2009). *J. Mol. Biol.* **386**, 204–217.
Chan, S. L., Ong, S. T., Ong, S. Y., Chew, F. T. & Mok, Y. K. (2006). *J. Immunol.* **176**, 4852–4860.
Cianci, M., Rizkallah, P. J., Olczak, A., Raftery, J., Chayen, N. E., Zagalsky, P. F. & Helliwell, J. R. (2001). *Acta Cryst.* D**57**, 1219–1229.
Cowan, S. W., Newcomer, M. E. & Jones, T. A. (1990). *Proteins*, **8**, 44–61.
Cowan, S. W., Newcomer, M. E. & Jones, T. A. (1993). *J. Mol. Biol.* **230**, 1225–1246.
Cowtan, K. (2010). *Acta Cryst.* D**66**, 470–478.
Cuesta-Seijo, J. A., Neale, C., Khan, M. A., Moktar, J., Tran, C. D., Bishop, R. E., Pomès, R. & Privé, G. G. (2010). *Structure*, **18**, 1210–1219.
Dian, C., Vitale, S., Leonard, G. A., Bahlawane, C., Fauquant, C., Leduc, D., Muller, C., de Reuse, H., Michaud-Soret, I. & Terradot, L. (2011). *Mol. Microbiol.* **79**, 1260–1275.
Ding, F. & Dokholyan, N. V. (2006). *PLoS Comput. Biol.* **2**, e85.
Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.
Evans, P. (2006). *Acta Cryst.* D**62**, 72–82.
Flower, D. R. (2000). *Biochim. Biophys. Acta*, **1482**, 46–56.
Ho, B. K. & Gruswitz, F. (2008). *BMC Struct. Biol.* **8**, 49.
Holm, L. & Rosenström, P. (2010). *Nucleic Acids Res.* **38**, W545–W549.
Josenhans, C., Labigne, A. & Suerbaum, S. (1995). *J. Bacteriol.* **177**, 3010–3020.
Katz, B. A. & Cass, R. T. (1997). *J. Biol. Chem.* **272**, 13220–13228.
Kota, P., Ding, F., Ramachandran, S. & Dokholyan, N. V. (2011). *Bioinformatics*, **27**, 2209–2215.
Krissinel, E. & Henrick, K. (2007). *J. Mol. Biol.* **372**, 774–797.
Laskowski, R. A., Rullmann, J. A., MacArthur, M. W., Kaptein, R. & Thornton, J. M. (1996). *J. Biomol. NMR*, **8**, 477–486.
Lertsethtakarn, P., Ottemann, K. M. & Hendrixson, D. R. (2011). *Annu. Rev. Microbiol.* **65**, 389–410.
Leslie, A. G. W. (2006). *Acta Cryst.* D**62**, 48–57.
Livnah, O., Bayer, E. A., Wilchek, M. & Sussman, J. L. (1993). *Proc. Natl Acad. Sci. USA*, **90**, 5076–5080.
Mankoski, R., Hoepf, T., Krakowka, S. & Eaton, K. A. (1999). *J. Med. Microbiol.* **48**, 395–399.
Mans, B. J., Ribeiro, J. M. & Andersen, J. F. (2008). *J. Biol. Chem.* **283**, 18721–18733.
Nardone, E., Rosano, C., Santambrogio, P., Curnis, F., Corti, A., Magni, F., Siccardi, A. G., Paganelli, G., Losso, R., Apreda, B., Bolognesi, M., Sidoli, A. & Arosio, P. (1998). *Eur. J. Biochem.* **256**, 453–460.
Pazy, Y., Kulik, T., Bayer, E. A., Wilchek, M. & Livnah, O. (2002). *J. Biol. Chem.* **277**, 30892–30900.
Peek, R. M. (2002). *Gastroenterology*, **123**, 1739–1740.
Ramachandran, S., Kota, P., Ding, F. & Dokholyan, N. V. (2011). *Proteins*, **79**, 261–270.
Redondo, C., Vouropoulou, M., Evans, J. & Findlay, J. B. (2008). *FASEB J.* **22**, 1043–1054.
Rothenbacher, D. & Brenner, H. (2003). *Microbes Infect.* **5**, 693–703.
Sacchettini, J. C., Gordon, J. I. & Banaszak, L. J. (1989). *Proc. Natl Acad. Sci. USA*, **86**, 7736–7740.
Schiltz, M. & Bricogne, G. (2008). *Acta Cryst.* D**64**, 711–729.
Shannon, P. T., Reiss, D. J., Bonneau, R. & Baliga, N. S. (2006). *BMC Bioinformatics*, **7**, 176.
Shirvanyants, D., Ding, F., Tsao, D., Ramachandran, S. & Dokholyan, N. V. (2012). *J. Phys. Chem. B*, **116**, 8375–8382.
Sourjik, V. & Wingreen, N. S. (2012). *Curr. Opin. Cell Biol.* **24**, 262–268.
Tegoni, M., Ramoni, R., Bignetti, E., Spinelli, S. & Cambillau, C. (1996). *Nature Struct. Biol.* **3**, 863–867.
Vliet, A. H. M. van, Bereswill, S. & Kusters, J. G. (2001). *Ion Metabolism and Transport*, edited by H. L. T. Mobley, G. L. Mendz & S. L. Hazell, ch. 17. Washington: ASM Press.
Vonrhein, C., Blanc, E., Roversi, P. & Bricogne, G. (2007). *Methods Mol. Biol.* **364**, 215–230.
Warren, J. R. (2006). *ChemMedChem*, **1**, 672–685.
Winn, M. D. *et al.* (2011). *Acta Cryst.* D**67**, 235–242.
Yin, S., Ding, F. & Dokholyan, N. V. (2007*a*). *Nature Methods*, **4**, 466–467.
Yin, S., Ding, F. & Dokholyan, N. V. (2007*b*). *Structure*, **15**, 1567–1576.
Zanotti, G., Berni, R. & Monaco, H. L. (1993). *J. Biol. Chem.* **268**, 10728–10738.
Zanotti, G., Ottonello, S., Berni, R. & Monaco, H. L. (1993). *J. Mol. Biol.* **230**, 613–624.