

Review

## Structural Analysis of Hypothetical Proteins from *Helicobacter pylori*: An Approach to Estimate Functions of Unknown or Hypothetical Proteins

Sung Jean Park <sup>1,†</sup>, Woo Sung Son <sup>2,†</sup> and Bong-Jin Lee <sup>3,\*</sup>

<sup>1</sup> College of Pharmacy, Gachon University, 534-2 Yeonsu 3-dong, Yeonsu-gu, Incheon 406-799, Korea; E-Mail: psjnmr@gachon.ac.kr

<sup>2</sup> College of Pharmacy, CHA University, 120 Haeryong-ro, Pocheon-si, Gyeonggi-do 487-010, Korea; E-Mail: wson@cha.ac.kr

<sup>3</sup> Research Institute of Pharmaceutical Sciences, College of Pharmacy, Seoul National University, San 56-1, Shillim-Dong, Kwanak-Gu, Seoul 151-742, Korea

† These authors contributed equally to this work.

\* Author to whom correspondence should be addressed; E-Mail: lbj@nmr.snu.ac.kr; Tel.: +82-2-880-7869; Fax: +82-2-872-3632.

Received: 9 March 2012; in revised form: 29 May 2012 / Accepted: 1 June 2012 /

Published: 8 June 2012

---

**Abstract:** *Helicobacter pylori* (*H. pylori*) have a unique ability to survive in extreme acidic environments and to colonize the gastric mucosa. It can cause diverse gastric diseases such as peptic ulcers, chronic gastritis, mucosa-associated lymphoid tissue (MALT) lymphoma, gastric cancer, etc. Based on genomic research of *H. pylori*, over 1600 genes have been functionally identified so far. However, *H. pylori* possess some genes that are uncharacterized since: (i) the gene sequences are quite new; (ii) the function of genes have not been characterized in any other bacterial systems; and (iii) sometimes, the protein that is classified into a known protein based on the sequence homology shows some functional ambiguity, which raises questions about the function of the protein produced in *H. pylori*. Thus, there are still a lot of genes to be biologically or biochemically characterized to understand the whole picture of gene functions in the bacteria. In this regard, knowledge on the 3D structure of a protein, especially unknown or hypothetical protein, is frequently useful to elucidate the structure-function relationship of the uncharacterized gene product. That is, a structural comparison with known proteins provides valuable information to help predict the cellular functions of hypothetical proteins. Here, we show the 3D structures of

some hypothetical proteins determined by NMR spectroscopy and X-ray crystallography as a part of the structural genomics of *H. pylori*. In addition, we show some successful approaches of elucidating the function of unknown proteins based on their structural information.

**Keywords:** *Helicobacter pylori*; structural genomics; NMR; X-ray; unknown protein; hypothetical protein; structural homology

---

## 1. *H. pylori* as a Pathogen

*Helicobacter pylorus* is one of the pathogens involved in various gastric diseases such as peptic ulcers, chronic gastritis, mucosa-associated lymphoid tissue lymphoma, and gastric cancer [1–3]. Infection with *H. pylori* is associated with an increased risk of gastric adenocarcinoma and has attracted attention as a cofactor in the pathogenesis of this malignant condition [4]. Moreover, the risk of developing cancer is related to the physiologic and histologic changes induced by a *H. pylori* infection in the stomach [5]. Despite a general decline in the incidence of gastric cancer, it remains the fourth most common cancer and second leading cause of cancer-related deaths worldwide [6]. However, most *H. pylori* infections do not cause cancer. The sporadic distribution of the disease caused by *H. pylori* looks to be dependent on host-related factors: the host (human individual) genetics controlling the inflammatory response, the age when the *H. pylori* infection was acquired, poor nutrition, storage of food, and the pattern of food consumption can be considered as host-related factors [7–9].

In addition, bacterial factors associated with the risk of gastric cancer are also emphasized, and molecular and cell biology approaches aimed at understanding the interaction between *H. pylori* and transforming epithelial cells have been carried out. Since *H. pylori* is a highly heterogeneous bacterial species, both genotypically and phenotypically, and is highly adapted for survival in the gastric niche, it is not easy to figure out the major bacterial factors that are directly associated with etiopathogenesis [10,11]. Based on the current knowledge, several virulence factors such as genes within the *cag* (cytotoxin-associated antigen) pathogenicity island, including the gene encoding the CagA protein, as well as polymorphic variation in the VacA vacuolating exotoxin and the blood group antigen binding adhesions, BabA and SabA, are regarded as possible bacterial factors [6,10,12]. A duodenal ulcer-promoting gene (*dupA*), located in the “plasticity region” of the *H. pylori* genome, was reported as a potential virulence marker [10,13]. Other bacterial factors such as peptidoglycan, lipopolysaccharide(LPS),  $\gamma$ -glutamyl trans-peptidase(GGT), and protease HtrA may be linked to pathogenicity [14].

Although a huge amount of biological data on *H. pylori* has been accumulated, enzymes or proteins of unknown function still make up more than a third of the open reading frames (ORF) of *H. pylori*. An unknown protein could be defined as a protein whose function has not yet been characterized, and a hypothetical protein could be defined as a protein that is supposed to exist in an organism although its existence has not been shown experimentally. Therefore, in a broad sense, hypothetical proteins could be included in unknown proteins. To completely understand the pathogenic mechanism of *H. pylori*, it is very important to elucidate the functions of these unknown proteins. To fill in the

“missing parts list” is accordingly one of the greatest challenges for post-genomic biology, and a tremendous opportunity to discover new biological and pathogenic machinery in *H. pylori*.

## 2. *H. pylori* Genomic Sequence

The sequencing of the *H. pylori* genome started in 1997 with the *H. pylori* strain 26695 [15]. It was isolated from an English patient with chronic gastritis. The chromosome of strain 26695 is circular and composed of 1.67 mega base pairs (Table 1). The average G-C content is approximately 38.9% and the genome has 1590 open reading frames (ORF) that are possibly protein-coding loci [1], together with the RNA coding genes (2 copies of 16S rRNA and 23S rRNA genes, 36 tRNA genes). From the following analysis of the same genome, it was suggested that a smaller number of ORFs is in the sequence of strain 26695 [16].

**Table 1.** Genomes of *H. pylori*. Currently, 36 sub-species have been identified and the genome sizes are from 1.55 mega base pairs to 1.82 mega base pairs. All data were collected and processed from the NCBI genome database [17].

Organism	Gene	Size (Mb)	GC%	Protein (unknown)	Type	Project
<i>Helicobacter pylori</i>	1480	1.57	38.9	1405 (476)	chr <sup>a</sup>	Gyeongsang National University College of Medicine and 21c Frontier Human Genome Functional Research Project <i>Helicobacter pylori</i> 52 genome sequencing project
<i>Helicobacter pylori</i> 2017	1647	1.55	39.3	1593 (525)	chr	Pathogen Biology Laboratory, University of Hyderabad <i>Helicobacter pylori</i> 2017 genome sequencing project
<i>Helicobacter pylori</i> 2018	1655	1.56	39.3	1603 (459)	chr	Pathogen Biology Laboratory, University of Hyderabad <i>Helicobacter pylori</i> 2018 genome sequencing project
<i>Helicobacter pylori</i> 26695	1627	1.67	38.9	1573 (1301)	chr	TIGR (The Institute for Genome Research) Causes gastric inflammation and peptic ulcer disease
<i>Helicobacter pylori</i> 35A	1560	1.57	38.9	1470 (362)	chr	Baylor College of Medicine Reference genome for the Human Microbiome Project
<i>Helicobacter pylori</i> 51	1495	1.59	38.8	1415 (386)	chr	Gyeongsang National University College of Medicine and 21c Frontier Human Functional Genome Research Project Bacterium isolated from duodenal ulcer patient
<i>Helicobacter pylori</i> 83	1656	1.62	38.7	1609 (445)	chr	Baylor College of Medicine Reference genome for the Human Microbiome Project
<i>Helicobacter pylori</i> 908	1646	1.55	39.3	1595 (444)	chr	University of Hyderabad, India <i>Helicobacter pylori</i> 908 genome sequencing project
<i>Helicobacter pylori</i> B38	1571	1.58	39.2	1382 (643)	chr	Institut Pasteur Causes peptic ulcers
<i>Helicobacter pylori</i> B8	1744	1.67	38.8	1702 (736)	chr	CeBitec, Bielefeld University <i>Helicobacter pylori</i> B8 genome sequencing project
	5	0.01	35.9	5 (3)	plsm <sup>b</sup>	
<i>Helicobacter pylori</i> Cuz20	1606	1.64	38.9	1564 (538)	chr	Dept. of Molec. Microbiology, Washington University Medical School, Saint Louis <i>Helicobacter pylori</i> Cuz20 genome sequencing project
<i>Helicobacter pylori</i> F16	1543	1.58	38.9	1500 (494)	chr	The University of Tokyo <i>Helicobacter pylori</i> F16 genome sequencing project

Table 1. Cont.

<i>Helicobacter pylori</i> F30	1522	1.57	38.8	1479 (470)	chr	The University of Tokyo <i>Helicobacter pylori</i> F30 genome sequencing project.
	5	0.01	34.1	5 (1)	plsm	
<i>Helicobacter pylori</i> F32	1533	1.58	38.9	1490 (485)	chr	The University of Tokyo <i>Helicobacter pylori</i> F32 genome sequencing project.
	1	0	36.7	1 (0)	plsm	
<i>Helicobacter pylori</i> F57	1563	1.61	38.7	1520 (498)	chr	The University of Tokyo <i>Helicobacter pylori</i> F57 genome sequencing project.
<i>Helicobacter pylori</i> G27	1570	1.65	38.9	1493 (470)	chr	University of Oregon Strain used extensively in <i>H. pylori</i> research
	11	0.01	34.9	11 (5)	plsm	
<i>Helicobacter pylori</i> Gambia94/24	1646	1.71	39.1	1604 (611)	chr	Berg lab, Washington University Medical School <i>Helicobacter pylori</i> Gambia94/24 genome sequencing project
	1	0	37.4	1 (1)	plsm	
<i>Helicobacter pylori</i> HPAG1	1573	1.60	39.1	1531 (515)	chr	Washington University (WashU) Isolated from a Swedish patient with chronic atrophic gastritis
	8	0.01	36.4	8 (5)	plsm	
<i>Helicobacter pylori</i> India7	1638	1.68	38.9	1600 (561)	chr	Berg lab, Washington University Medical School <i>Helicobacter pylori</i> Ind7 genome sequencing project
<i>Helicobacter pylori</i> J99	1534	1.64	39.2	1488 (560)	chr	Astrazeneca-Boston Causes gastric inflammation and peptic ulcer disease
<i>Helicobacter pylori</i> Lithuania75	1588	1.62	38.8	1546 (522)	chr	Berg lab, Washington University Medical School <i>Helicobacter pylori</i> Lit75 genome sequencing project
	19	0.02	33.7	19 (12)	plsm	
<i>Helicobacter pylori</i> P12	1624	1.67	38.8	1568 (450)	chr	Max von Pettenkofer-Institut für Hygiene und Medizinische Mikrobiologie, Ludwig-Maximilians- Universität München Clinical isolate
	10	0.01	35.1	10 (2)	plsm	
<i>Helicobacter pylori</i> PeCan4	1597	1.63	38.9	1555 (529)	chr	Dept. of Molec. Microbiology, Washington University Medical School, Saint Louis <i>Helicobacter pylori</i> PeCan4 genome sequencing project
	8	0.01	32.9	8 (0)	plsm	
<i>Helicobacter pylori</i> Puno120	1567	1.62	38.9	1525 (518)	chr	Washington University Medical School <i>Helicobacter pylori</i> Puno120 genome sequencing
	15	0.01	35.8	15 (13)	plsm	
<i>Helicobacter pylori</i> Puno135	1615	1.65	38.8	1573 (532)	chr	Washington University Medical School Genome sequence of <i>Helicobacter pylori</i> strain Puno135
<i>Helicobacter pylori</i> SJM180	1623	1.66	38.9	1581 (558)	chr	Dept. of Molec. Microbiology, Washington University Medical School, Saint Louis <i>Helicobacter pylori</i> SJM180 genome sequencing project

Table 1. Cont.

<i>Helicobacter pylori</i> SNT49	1557	1.61	39	1515 (495)	phage	Washington University Medical School
	4	0	37.4	4 (3)	plsm	Genome sequence of <i>Helicobacter pylori</i> SNT49
<i>Helicobacter pylori</i> Sat464	1544	1.56	39.1	1502 (504)	chr	Dept. Molec. Microbiology, Washington University Medical School in Saint Louis
	6	0.01	33.5	6 (4)	plsm	<i>Helicobacter pylori</i> Sat464 genome sequencing project.
<i>Helicobacter pylori</i> Shi470	1647	1.61	38.9	1568 (593)	chr	Washington University Medical School Clinical isolate from the Amazon River region
<i>Helicobacter pylori</i> SouthAfrica7	1585	1.65	38.4	1543 (555)	chr	Berg lab, Washington University Medical School <i>Helicobacter pylori</i> SouthAfrica7 genome sequencing project
	29	0.03	33.7	29 (19)	plsm	
<i>Helicobacter pylori</i> v225d	1625	1.59	39	1541 (555)	chr	The Pathosystems Resource Integration Center (PATRIC)
	9	0.01	32.9	9 (7)	plsm	<i>Helicobacter pylori</i> v225 genome sequencing
<i>Helicobacter pylori</i> B45	27	0.02	37.3	27 (26)	chr	Karolinska Institute
<i>Helicobacter pylori</i> 98-10	1566	1.57	38.8	1527 (1527)	S/C <sup>c</sup>	Vanderbilt University School of Medicine Gastric cancer strain
					S/C	
<i>Helicobacter pylori</i> B128	1770	1.65	38.8	1731 (1731)	S/C	Vanderbilt University School of Medicine Gastric ulcer strain
<i>Helicobacter pylori</i> HPKX_438_AG0C 1	2939	1.82	39.5	2898 (1564)	S/C	Washington University Medical School Clinical isolate
<i>Helicobacter pylori</i> HPKX_438_CA4C1	3962	1.57	39.2	3925 (1548)	S/C	Washington University Medical School Isolate from a patient with gastric carcinoma
Total	<b>59,776</b>	-	-	<b>57,872</b> <b>(23,261)</b>	-	-

<sup>a</sup> Chromosome; <sup>b</sup> Plasmid; <sup>c</sup> S/C: Scaffolds or Contigs.

Ongoing studies have found genes that were missing in previous analyses, as in the case of SecE. A general secretion machinery is widely present in bacteria, which functions in the secretion of outer membrane proteins to extracellular environments [18]. From the first annotation results, it was thought that strain 26695 had only a partial general secretion machinery because it lacked SecE [15]. A new small open reading frame between *nusG* and *rmpG* (HP1203-HP1204) in the genome sequences was found using an *ab initio* server, GeneMark, Glimmer, and BlastX [19]. It has a high homology and structural similarity to the SecE protein in related bacteria implying that strain 26695 has a complete general secretion machinery. In addition, small RNA genes are universally present in bacteria [20]. The tmRNA gene (*ssrA*) has been found in *H. pylori*, encoding a functional RNA molecule and a small peptide involved in the quality control of translation [21]. In addition, the *H. pylori* strain contains a sRNA gene encoding the RNA component of RnaseP and the 4.5S RNA gene which is involved in secretion [22,23].

In 2008, the adaptations of *H. pylori* to a rarely captured event in the evolution of its impact on a host biology were characterized by defining the impact of these adaptations on an intriguing but poorly characterized interaction between this bacterium and gastric epithelial stem cells [24]. *H. pylori*

HPKX\_438\_AG0C1 and HPKX\_438\_CA4C1 were isolated from a single patient who progressed from ChAG (chronic atrophic gastritis) to adenocarcinoma using a population-based endoscopy study. ChAG-associated Kx1 and Cancer-associated Kx2 genomes were analyzed to examine the adaptation of *H. pylori*, respectively. Micro-arrays gave a comprehensive view of the genome diversity of the *H. pylori* pathogen. This was performed with information on the origin of the *hspA* together with *glmM* alleles revealing that *H. pylori* infection may be acquired by more diverse routes than previously expected [25]. According to cluster analysis, isolates from family D belonged to three different strains, those from family L consisted of two strains, and those from family A were grouped into at least 5 strains. Strains from family D and family L differed by the presence/absence of 24 to 42 CDSs (coding sequences). In family A, one strain was difficult to define due to the small differences in gene profiles between neighboring branches.

In 2009, the complete genome sequence of *H. pylori* G27 was reported [26]. The G27 strain was originally isolated from an endoscopy patient from Italy [27]. The genome consists of a single circular chromosome with about 1.65 mega base pairs (Table 1) that is AT rich (61.6%), contains 1515 ORFs, and is similar in size and composition to the other published *H. pylori* genomes of strains 26695, J99, and HPAG [15,16,28]. The G27 strain contains 58 genes that are not found in 26695, J99, or HPAG, as defined by a blastp hit. The majority of these G27-specific genes are predicted to encode hypothetical proteins [26].

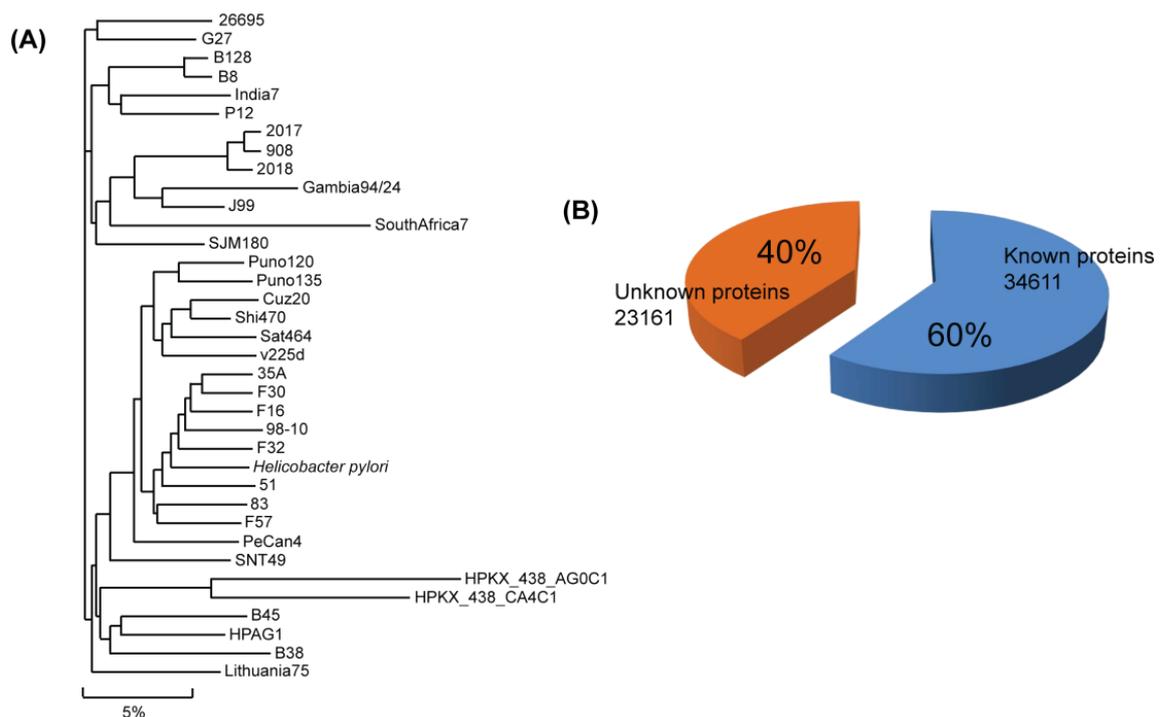
In the same year, the genome sequences of two *H. pylori* strains were analyzed [29]. *H. pylori* strain 98-10 was isolated from a patient with gastric cancer and strain B128 was isolated from a patient with gastric ulcer disease. Strain 98-10 was most closely related to *H. pylori* strains of East Asian origin and strain B128 was most closely related to strains of European origin. Strain 98-10 contained multiple features characteristic of East Asian strains, including a type s1c *vacA* allele and a *cagA* allele encoding an EPIYA-D tyrosine phosphorylation motif.

Very recently, several genome sequences of different strains were reported accelerating *H. pylori* genomic and proteomic research [30–38]. Strain 908 is a close relative strain of J99 [39] and was isolated from an African patient living in France, who suffered from duodenal ulcer disease [40]. The B8 strain consists of about 1.67 mega base pairs and a small plasmid of about 6000 base pairs carrying nine putative genes. Interestingly, the B8 strain contains coding sequences, 293 of which are strain-specific, coding mainly for hypothetical proteins with unknown functions [31]. Similarly, the P12 strain contains plasticity zones, encoding for the type IV secretion system and having the typical properties of genomic islands [32]. Another sequenced genome, the Shi470 strain known as the Shiimaa village strain was more Asian- than European-like genome-wide, indicating Amerind ancestry. This strain contains two unique *cagA* virulence genes and a novel allele of gene hp0519 encoding host tissue interaction protein [33]. There are several *H. pylori* populations such as *hpAfrical*, *hpEurope*, *hspEAsia*, and *hspAmerind* because this bacterium has colonized the stomach since early in human evolution and diverged with ancient human migrations [41–43]. One of these populations, the *hspAmerind* strain V225d, was cultured from a Venezuelan Piaroa Amerindian subject and identified. The V225d strain is *cag*-positive encoding a multifunctional effector protein injected into host cells by the *cag* type IV secretion system [34]. Two strains, 2017 and 2018, are the chronological subclones of strain 908 and cultured from the antrum and corpus, respectively. Using comparative genomic analysis [35,37], these two strains are almost identical and descended from the genome of strain

908 [30,36]. The B45 strain was sequenced from a gastric mucosa-associated lymphoid tissue (MALT) lymphoma patient and induced an integrated prophage in this strain by UV irradiation [38].

The Comprehensive Microbial Resource (CMR) is a free tool that allows researchers to access all of the publicly available bacterial genome sequences completed to date [44] (Figure 1). Currently, it provides genomic sequences of three strains of *Helicobacter pylori* (26695, HPAG1, J99).

**Figure 1.** Genome sequence and proteins of *H. pylori*. In the phylogenetic tree, a total of 36 sub-species are branched with a total of about 60,000 genes (A); and among the translated proteins, the biological functions of 40% of the proteins are unidentified (B).



### 3. Structural Reports on *H. pylori* Proteins

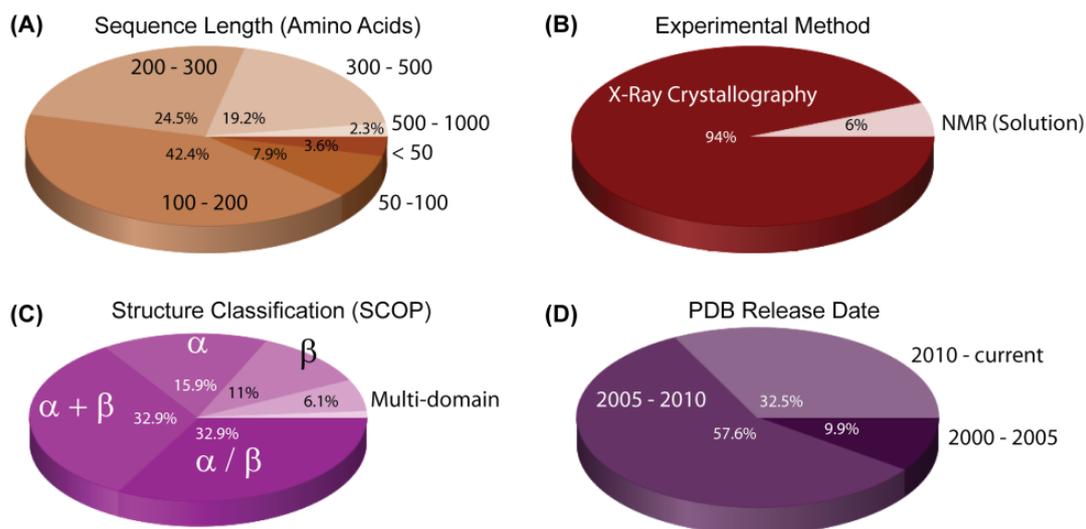
As in the case of other genomic research, Structural Genomics Initiatives are mainly responsible for determination of *H. pylori* protein structures. These initiatives, together with the structure determination of known proteins, have made enormous strides in the elucidation of unknown protein structure of *H. pylori* [15,16,24–26,28–38,45–47]. The available structural data have already led to the identification of potentially new drug targets [48] and has been helpful in assigning functions to proteins of which the functions were previously unknown [49,50].

The increase in structure determination for *H. pylori* has been triggered by the sequencing of the *H. pylori* 52 and 26695 genomes [15,25,45,47]. The genome sequences and their protein structures yielded many clues to help understand the pathogenesis of *H. pylori*. Approximately 14% of Lyase structures have been determined and represent the largest proportion of any functional class of which the structures have already been solved (Table S1).

The sequencing of the genome led to a dramatic increase in the number of known structures for *H. pylori* proteins deposited in the Protein Data Bank (PDB) (Figure 2). The first *H. pylori* protein structure was determined in 2001 (PDB ID: 1G6O) [51]. In the following four years, 32 more

structures were reported (Figure 2). After several sub-species genome sequences of *H. pylori* became publicly available, the number of structures determined after 2005 increased sharply and at an increasing rate.

**Figure 2.** Statistics of protein structures from *H. pylori*. All data were collected and processed from PDB on 14 February 2012 [52]. The dominant properties of the presented data are 100–300 kDa in size, X-ray diffraction as the experimental method, alpha and beta structural motifs, and release date from 2005–2010.



Usually, protein solubility is one of the main bottlenecks in structure determination [53]. In the case of *H. pylori*, methods have already been developed that remedied this problem, such as the development of customized expression strategies for *H. pylori* proteins in *Escherichia coli* [54]. The increase in determined structures is also due to the development of improved methods for high-throughput X-ray crystallography. However, the major driving force for this increase was the availability of genome-wide sequence data in the early 2000s.

There are currently 79,356 structures in the PDB as of 14 February 2012, of which 0.35%, a total of 279, are structures of *H. pylori* proteins. Of these proteins, 28 are unknown in function, which represents 10.03% of the determined *H. pylori* structures (Table 2).

**Table 2.** Unknown protein structures from *H. pylori*. A total of 28 unknown protein structures were elucidated using X-ray diffraction and NMR method. All data were collected and processed from PDB database [52].

PDB ID	Chain ID	Chain AA	Structure MW	Macromolecule Name	Classification	Scop Fold	Exp. Method
1MW7	A	240	27161.20	Hypothetical protein HP0162	SG <sup>a</sup> , unknown function	YebC-like	X-ray
1S2X	A	206	23998.70	Cag-Z	Unknown function	STAT-like	X-ray
1Z8M	A	88	10394.30	Conserved hypothetical protein HP0894	SG, unknown function	RelE-like	NMR
1ZHC	A	76	9130.38	hypothetical protein HP1242	Unknown function		NMR

Table 2. Cont.

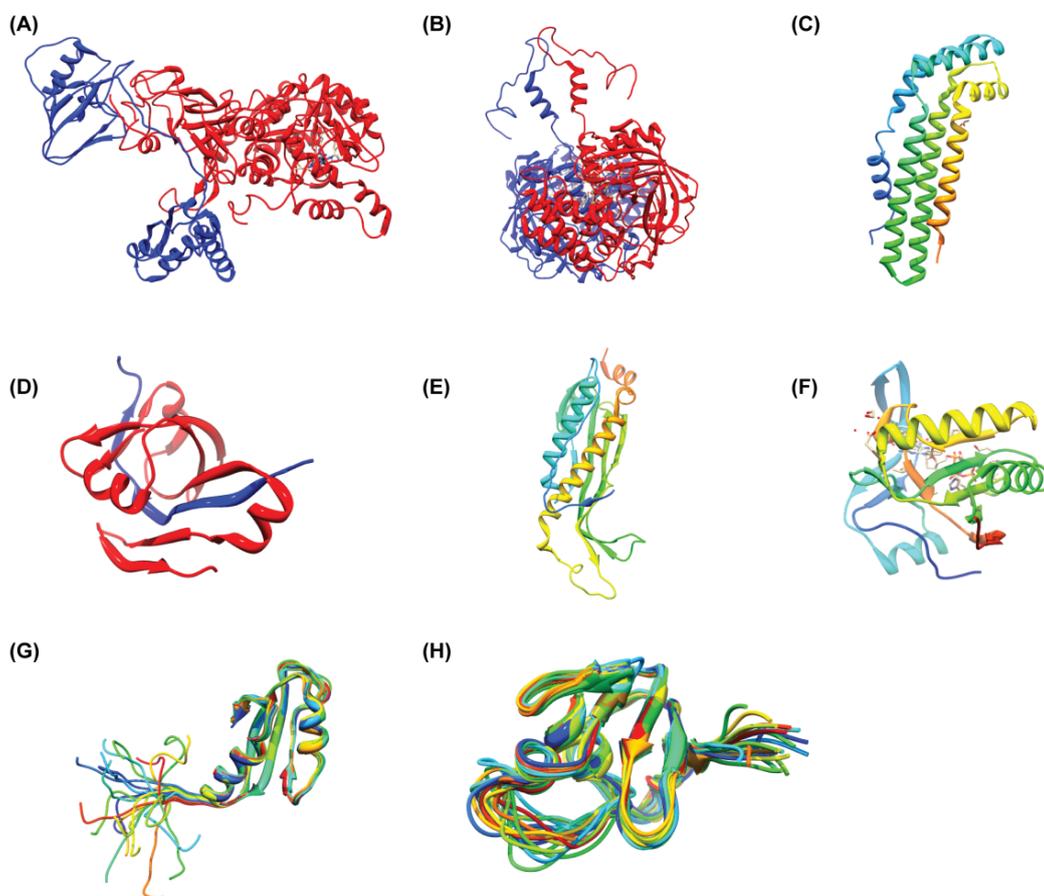
1ZKE	A, B, C, D, E, F	83	56798.00	Hypothetical protein HP1531	SG, unknown function	ROP-like	X-ray
2ATZ	A	180	22049.45	Predicted coding region HP0184	SG, unknown function	Prim-pol domain	X-ray
2BO3	A	94	11101.70	Hypothetical protein HP0242	SG, unknown function	HP0242- like	X-ray
2EVV	A, B, C, D	207	95692.83	hypothetical protein HP0218	SG, unknown function		X-ray
2F6S	A, B	201	47249.90	cell filamentation protein, putative	SG, unknown function	Fic-like	X-ray
2G3V	A, B, C, D	208	104975.36	CAG pathogenicity island protein 13	Unknown function		X-ray
2GTS	A	86	10626.50	hypothetical protein HP0062	SG, unknown function	Ferritin- like	X-ray
2H9Z	A	86	10205.80	Hypothetical protein HP0495	SG, unknown function	Ferredoxin- like	NMR
2I9I	A	254	29526.70	Hypothetical protein	SG, unknown function	Anticodon- binding domain- like	X-ray
2JOQ	A	91	10673.20	Hypothetical protein HP0495	SG, unknown function	Ferredoxin- like	NMR
2K0Z	A	110	12948.60	Uncharacterized protein HP1203	SG, unknown function		NMR
2K6P	A	92	10472.30	Uncharacterized protein HP1423	Unknown function		NMR
2OTR	A	98	11502.60	Hypothetical protein HP0892	SG, unknown function		NMR
2OUF	A	94	11148.60	Hypothetical protein	SG, unknown function		X-ray
2UVP	A, B, C, D	186	87079.82	HOBA, HP1230	Unknown function		X-ray
2XRH	A	100	11635.31	HP0721	Unknown function		X-ray
3BGH	A, B	236	55233.49	Putative neuraminylactose-binding hemagglutinin homolog	SG, unknown function		X-ray
3CWX	A, B, C	176	62332.80	protein CagD	Unknown function		X-ray
3CWY	A	176	20841.15	protein CagD	Unknown function		X-ray
3F42	A, B	99	22671.87	protein HP0035	SG, unknown function		X-ray
3FX7	A, B	94	23207.80	Uncharacterized protein, HP0062	Unknown function		X-ray
3KWL	A	514	60116.00	Uncharacterized protein	Unknown function		X-ray
3MLG	A, B	189	43924.40	Uncharacterized protein	Unknown function		X-ray
3MLI	A, B, C, D	100	47758.96	Uncharacterized protein HP0242	Unknown function		X-ray

<sup>a</sup> Structural genomics.

A complete list of *H. pylori* protein structures deposited in the PDB is given in the Supporting Information Table S1. The predominant method used to determine these structures was X-ray crystallography, which accounts for 261 of the total number of *H. pylori* structures currently determined (Figure 2). A further 18 were elucidated by solution-state NMR spectroscopy. Most

structures are of individual proteins, although many are bound by small molecule ligands such as substrate analogues and only 11 protein-DNA complexes have been determined (Figure 3, Table S1).

**Figure 3.** Several 3D structures from *H. pylori*. Urease subunit  $\alpha$  and  $\beta$  (**A**, pdb code: 1E9Y), Kat catalase (**B**, pdb code: 1WQL) are multiple domain structures with multiple chains. Aspartate 1-decarboxylase adopts a dominant  $\beta$  structure (**D**, pdb code: 1UHD). The structures of unknown proteins are shown with different variations of their structural domains (**C**, pdb code 1S2X, all  $\alpha$ ; **E**, pdb code: 2I9I,  $\alpha/\beta$ ; **F**, pdb code: 2ATZ,  $\alpha + \beta$ ; **G**, pdb code: 2H9Z,  $\alpha + \beta$ ; **H**, pdb code: 2K6P, RNA binding motif). Structures of **G** and **H** are solved by NMR. All structures were displayed using UCSF Chimera with ribbon presentation method [55].



#### 4. Unknown Proteins in *H. pylori* and Estimation of Their Function

The most typical approach of predicting the function of an unknown protein is to use sequence similarity by finding a similar protein of known function [56]. Based on sequence-similarity, a predictor assigns the known function to the inferred protein. Actually, the functions of enzymes tend to be conserved if they share more than a 40%–50% sequence identity. The sequence-based approach is reasonable, however, approximately 50% of the unknown proteins from a newly sequenced genome could not be assigned to their function using only sequence-similarity approaches [57] (Figure 1). The low efficiency of the sequence-similarity search may be partly caused by gene sequences that are quite new and genes that have not yet been characterized in other bacterial systems. To overcome the

weakness of sequence-similarity searches, several trials were employed using so called “similarity free” methods [57]. The methods use physicochemical properties and secondary structure of proteins. Bioinformatics developed the methods and there have been successful cases for characterizing function or structure [58–60]. However, the methods need to be improved since similarity-free methods still depend to a certain extent on similarity.

Another approach to identify function is to use 3D structures. This approach often succeeds in cases where sequence-based methods fail. This may be due to the idea that in many cases evolution retains the folding pattern long after the sequence similarity becomes undetectable. Structural similarity searches use the global fold of the protein [61–64] or detect the functionally important regions of the protein [65–69]. Since structures diverge more slowly than sequences, a sequence comparison may be less sensitive than a structure comparison [70]. However, the structural comparison still has the limitation of false positives being reported and needs to be improved to overcome overestimation of statistical significance like sequence-similarity searches [70]. This means that experimental confirmation is still required for exact assignment of function to an unknown protein.

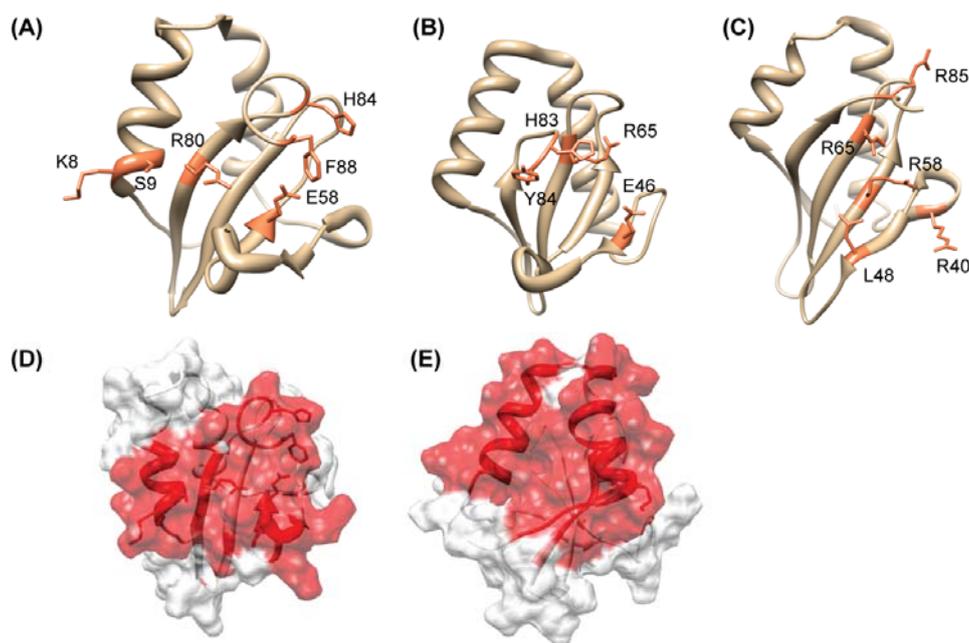
Some examples of functional elucidation of unknown proteins from *H. pylori* are provided below. For estimation, we generally conducted four steps: (i) structure determination; (ii) sequence homology search using PSI-BLAST [71]; (iii) structural homology search using the web server DALI [62]; and (iv) experimental confirmation of the function.

#### 4.1. HP0894-HP0895: Toxin-Antitoxin System in *H. pylori*

The high-quality NMR structure of HP0894 was reported [72]. The HP0894 structure (PDB ID: 1Z8M) has two  $\alpha$ -helices, two  $3_{10}$ -helices, and four  $\beta$ -strands ( $\alpha$ - $\alpha$ - $3_{10}$ - $\beta$ - $3_{10}$ - $\beta$ - $\beta$ ). The  $\beta$ -Strands form a four-stranded anti-parallel  $\beta$ -sheet (Figure 4). BLAST conserved domain search [73] showed that HP0894 contains the conserved domain DUF332 (Domain of Unknown Function), which is equivalent to COG 3041 in the National Center for Biotechnology Information Database of Clusters of Orthologous Groups. However, in the Pfam database [74], HP0894 belongs to the plasmid stabilization system protein family (PF05016). From the sequence homology search, we were able to get a hint of the function. However, a search for structural homologs with a  $Z$  score higher than 3.0 using the programs DALI showed that HP0894 is structurally similar to *Pyrococcus horikoshii* Archaeal RelE (PDB code: 1WMI,  $Z$  score = 7.8, pairwise RMSD = 2.8 Å), *E. coli* YoeB (PDB code: 2A6Q,  $Z$  score = 8.8, RMSD = 2.9 Å), and Guanyloribonuclease (PDB code: 1RGE,  $Z$  score = 3.3, pairwise RMSD = 3.4 Å). These proteins are both ribonucleases, have a similar number of residues as HP0894 (around 90), share a similar  $\beta$ -sheet topology with HP0894, and have a comparable location for two of their helices (Figure 4). As expected, they have no detectable sequence homology with HP0894 in PSI-BLAST searches and Blast2 (pairwise comparison) analyses. The structural homology search revealed HP0894 may have potential ribonuclease activity and represents the toxin-antitoxin (TA) system like RelE [75]. Generally, in a TA system, toxin expression induces arrest of cell growth, whereas the antitoxin neutralizes the toxin by a direct protein-protein interaction [76]. Both proteins of the toxin-antitoxin system are encoded within a single operon, with the toxin gene usually located directly downstream of the antitoxin gene [77]. Thus, we hypothesized: (i) HP0894 is a toxin molecule in *H. pylori*; (ii) there should be an antitoxin molecule that interacts with HP0894; and (iii) it should be

near the gene location for *hp0894* on the chromosome, if an antitoxin molecule exists. Actually, we found that HP0895 (hypothetical protein) is an antitoxin molecule [78] locating upstream of the *hp0894* gene.

**Figure 4.** Comparison of the structural and catalytic residues of HP0894 with those of its structural homologues. **A–C**, ribbon displays of the representative conformer of HP0894; **(A)** *E. coli* YoeB (PDB ID: 2A6R); **(B)** *P. horikoshii* RelE (PDB ID: 1WMI); **(C)** labeled functional or predicted key residues are colored *coral*. The RelE monomer structure was extracted from the aRelE-aRelB complex structure; **(D)** Chemical shift perturbation mapping of the C-terminal peptide of HP0895-binding region on HP0894 (1:1 molar ratio). Ribbon and surface displays of HP0894 structure colored according to chemical shift perturbations. The changes of the residues in obvious slow or fast exchange modes are colored in *red*. **(E)** Chemical shift perturbation mapping of the ssDNA-U [d(ACACUAAGAA)]-binding region on HP0894 (1:4 molar ratio). Residues showing significant chemical shift changes are colored in *red*.



Our experimental data [78] showed that HP0894 and HP0895 forms a stable complex as a large multimer (hexamer,  $((\text{HP0895})_6, (\text{HP0894-HP0895})_6)$ ), and the inhibitory effect of HP0894 on *E. coli* cell growth was neutralized by HP0895. In bacteria, toxins function, or are supposed to function, by inhibiting translation through mRNA cleavage [79]. With a RNA retardation experiment, the *in vitro* RNase activity of HP0894 was confirmed and HP0895 inhibited this RNase activity [78]. A primer extension experiment showed that HP0894-mediated mRNA cleavage occurred predominantly before adenine (A) or guanine (G) residues and we suggested -U:A- and -C:A- sequences are the most preferred cleavage sites [78]. The binding mode between HP0894 and HP0895 was more deeply studied using NMR and CD spectroscopy and we showed the binding interface of HP0894 [78]. Interestingly, HP0316 (hypothetical protein) that has an 85% sequence identity with HP0895 except for 30 residues at the C-terminal tail did not bind to HP0894, suggesting the C-terminal non-conserved



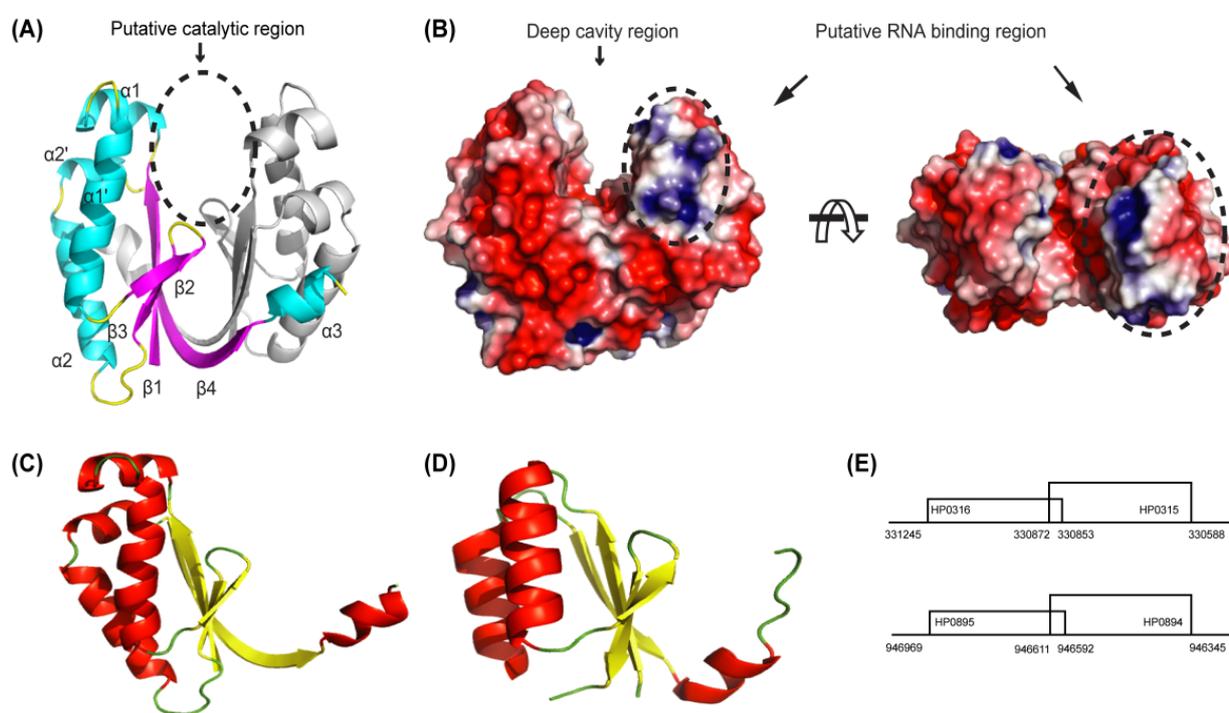
Virulence-associated protein, a product of the *vap* gene in various organisms, may be insufficient in itself, but is a requisite for virulence. The *vap* genes are known as factors or enzyme-producing factors that regulate the expression of true virulence genes or activate virulence factors by translational modification, processing of secretions or that are required for the activity of true virulence factors. Several *vap* genes (*vapA*, *B*, *C*, *D*, *H* and *I*) are known to exist in various organisms [82–84] but how the products of the *vap* genes are related to virulence remains unclear. *H. pylori* strain 26695 has only one type of virulence-associated protein, VapD. Two genes in this strain (HP0315 and HP0967) belong to *vapD* [85]. The exact biological role of the VapD protein has not yet been established, but several suggestions such as toxin, acid tolerance, plasmid stability, *etc.* have been made [86–88]. Here, we summarized the elucidation of the probable function of HP0315 with structural and biochemical studies.

The structure of HP0315 consists of 10 secondary structure elements:  $\beta 1$  (residues 1–8),  $\alpha 1$  (residues 10–17),  $\alpha 1'$  (residues 21–35),  $\beta 2$  (residues 38–41),  $\beta 3$  (residues 44–47),  $\alpha 2$  (residues 53–66),  $\alpha 2'$  (residues 68–73),  $\beta 4$  (residues 75–87) and  $\alpha 3$  (residues 88–93). The monomer has a ferredoxin-like fold. It has the  $\beta 1-(\alpha 1-\alpha 1')-\beta 2-\beta 3-(\alpha 2-\alpha 2')-\beta 4-\alpha 3$  instead of the  $\beta-\alpha-\beta-\beta-\alpha-\beta$  structure of the ferredoxin fold. The dimer of HP0315 is butterfly-shaped (PDB code: 3UI3, Figure 6). The  $\beta 4$  strand and the  $\alpha 3$  helix associate with the adjacent monomer, forming a dimerization interface [89]. This structure is the first structure of a VapD family to our knowledge. A sequence homology search revealed that HP0315 is related to the CRISPR-associated protein Cas2, a novel family of endoribonucleases, suggesting the potential ribonuclease activity of HP0315. The structure-based alignment also yielded a high score from DALI for one of the Cas2 proteins, SSO1404 (PDB code: 2IVY) although the top-scoring proteins were mainly hypothetical unknown proteins. In addition, the interrelationships between VapD and Cas2 proteins were supported by a genomic analysis [90].

The sequence analysis yielded another interesting result: the two genes *HP0315* and *HP0316* exist as an operon, which is a functional unit of genomic DNA containing partially overlapping genes under the control of a single regulatory signal or promoter (gene coordinates: *HP0315* 330872–330588, *HP0316* 331245–330853, Figure 6). As described above, HP0316 has a sequence similarity of 88.9% with HP0895 [78], which might suggest the HP0315-HP0316 system is identical with the HP0894-HP0895 system. In other words, HP0315 might act as a toxin molecule like HP0894 although no sequence and structural similarity exists between them. However, HP0315 did not bind HP0316 and did not affect the cell viability in *in vivo* toxicity experiments [89]. From the sequence/structure analysis and biochemical experiments, HP0315 was speculated to be a ribonuclease but not a toxin even though the gene arrangement is similar to that of a TA system [89]. The RNase activity of HP0315 was confirmed by primer extension and gel retardation experiments, revealing purine-specific endoribonuclease activity [89].

Conclusively, HP0315, a member of the VapD family, has a structural similarity with the Cas2 family and has a gene arrangement similar to the TA system; however, it does not belong to any of them, like an evolutionary intermediate. The exact function of HP0315 has not been determined yet. However, considering the relationship with Cas2 and a TA system, as well as the endoribonuclease activity, HP0315 may be related to either cell maintenance or a defense mechanism against invasion, or possibly both such as Cas2 and/or a TA system.

**Figure 6.** Structure of HP0315 from *H. pylori*. (A) Cartoon representation of the dimer of HP0315 ( $\alpha$ -helices,  $\beta$ -strands and loops are cyan, magenta and yellow, respectively). Dotted circle represents the putative catalytic region located at the deep cavity region. (B) Surface representation of HP0315 showing positive and negative electrostatic potential in blue and red, respectively. The dotted circle represents the putative RNA-binding region. This region would be related to initial binding with RNA, and then a second catalytic reaction would occur around the deep cavity region. (C,D) Structural comparison between HP0315 (C); and the homologue, SSO1404 (PDB code:2ivy) (D).  $\beta$ -Strands are colored in “yellow” and  $\alpha$ -helices in “red”. Both of the structures possess a ferredoxin-like fold. (E) Diagrams of the *hp0315* (*hp0894*) and *hp0316* (*hp0895*) encoding region from the chromosome of *H. pylori*.

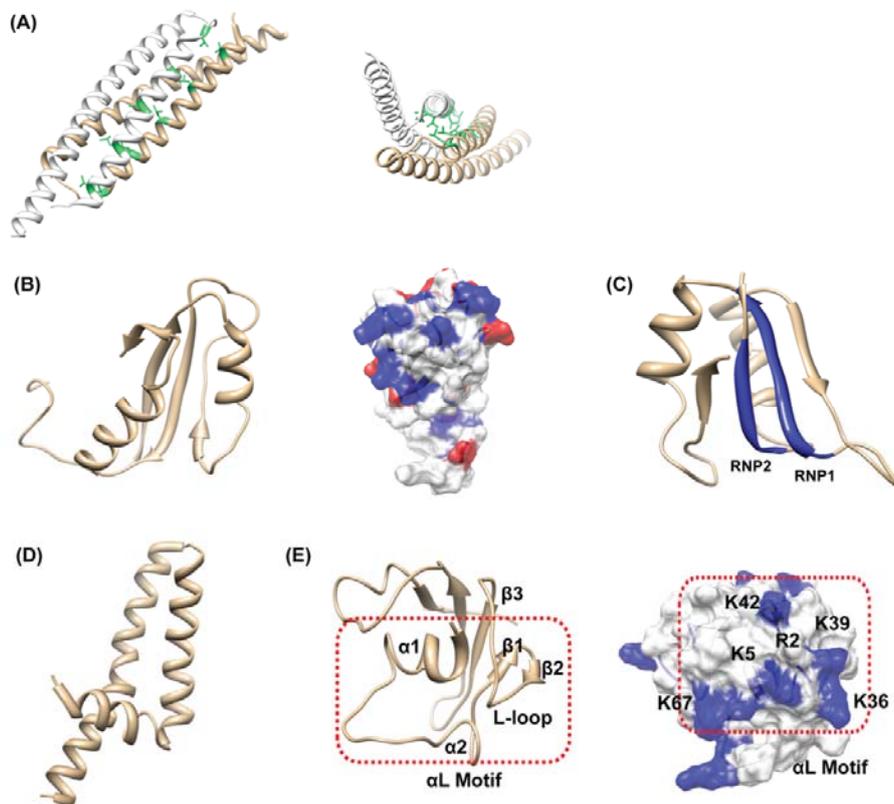


#### 4.3. Others: HP0062, HP0495, HP0827, HP1242, HP1423

The 3D structure of hypothetical protein HP0062 (PDB code: 3FX7) at 1.65 Å resolution was solved [91]. HP0062 is a small protein composed of 86 amino acids but it exists as dimer. The HP0062 monomer folds into a hairpin structure, in which two  $\alpha$ -helices (the *N*- and *C*-helix) are connected by a short loop (Figure 7A) and the *N*-helix displays a modified leucine zipper. The protomers dimerize in an antiparallel arrangement, in which the *N* and *C* helices of one protomer pack against the *N* and *C* helices of the second protomer, forming a four-helix bundle. The two protomers in an asymmetric unit of the orthorhombic crystal are similar, and the topologically equivalent Ca carbons superimpose with a RMSD of 0.79 Å. Actually, the structure of HP0062 was also solved by another group but they reported the protein is monomeric (unpublished, PDB code: 2GTS). Since our gel filtration chromatography revealed the dimeric state of HP0062, it is believed that the biologically relevant form is a dimer [91]. The structural comparison indicated HP0062 has similarity with the coiled-coil segments of over 100 functionally unrelated proteins that are involved in various protein-protein

interactions. Thus, the function of HP0062 is hard to directly estimate from the structural information. Interestingly, HP0062 shows extensively similar characteristics to those of the ESAT-6 family of Gram-positive bacteria; small dimer, helix-hairpin-helix structure, no signal peptide but with WXG motif in the hairpin bend (WRD in HP0062), and gene clusters with a protein with FtsK/SpoIIIE domain [92]. On the other hand, HP0062 also has similar characteristics to those of the TTS (Type Three Secretion) chaperones of Gram-negative bacteria; small dimer, an acidic pI, an overall  $\alpha$ -helical character and a carboxy-terminal amphipathic  $\alpha$ -helix [93]. These results might give a hint for the function of HP0062 as a transport chaperone and/or adaptor protein to facilitate interactions with host receptor proteins.

**Figure 7.** (A) Ribbon diagram of the HP0062 dimer is shown. Side and top views of the HP0062, showing the leucine zipper (green); (B) Ribbon drawing of the representative conformer of HP0495. Distribution of the surface charges on two distinct faces of HP0495 is shown. Positively-charged residues are blue, negatively-charged residues are red; (C) Ribbon drawing of the representative conformer of HP0827. Blue colors represent conserved RNP motifs lying side by side; (D) Ribbon drawing of the representative conformer of HP1242; (E) Ribbon drawing of the representative conformer of HP1423. The  $\alpha$ L motif consists of two  $\alpha$ -helices and the loop between  $\beta$ 2 and  $\beta$ 3. Electrostatic potential surface diagrams of HP1423 shows a strong concentration of positive charge in the proposed RNA-binding  $\alpha$ L motif facing outwards.



HP0495 is an 86-residue hypothetical protein with a molecular weight of 10,192.7 Da. The atomic coordinates of the final structure have been deposited in PDB (2H9Z). HP0495 has two  $\alpha$ -helices and four  $\beta$ -strand, forming a ferredoxin-like fold,  $\beta$ 1- $\alpha$ 1- $\beta$ 2- $\beta$ 3- $\alpha$ 2- $\beta$ 4 (Figure 7B). HP0495 is a completely

unknown protein since HP0495 has a restricted sequence homology with unknown proteins from several bacteria [94,95]. The ubiquitous ones like HP0495 merit the highest priority for functional characterization because they have the greatest potential payoff in new biological knowledge. In this case, the structure of HP0495 and structural homology data may be more important and provide a clue for the function. Unfortunately, a structural homology search using DALI indicated that HP0495 has structural homology with a variety of proteins [94]. This should be because the ferredoxin-like fold of HP0495 is abundant in other structures. Twenty proteins had a higher Z-score of 5.0 from DALI analysis including the NikR protein from *Pyrococcus horikoshii* (nickel responsive repressor; PDB code: 2BJ9, RMSD = 2.9 Å), LrpA from *Thermus thermophilus* (transcriptional regulator; PDB code: 1RIS, RMSD = 2.9 Å), S6 protein from *Archaeoglobus fulgidus* (ribosomal protein; PDB code: 1Y7P, RMSD = 2.9 Å), and a hypothetical YbeD protein from *E. coli* (unknown; PDB code: 1RWU, RMSD = 3.6 Å). The structural comparison did not show a clear result. However, the function of HP0495 seems to be related to nucleic acid interaction since its homologues are mainly nucleic acid binding proteins and HP0495 possesses positive surface charges (Figure 7B).

HP0827 is classified as a putative single-stranded (ss)-DNA binding protein 12RNP2 precursor protein. The solution structure of HP0827 (PDB code: 2KI2) has a ferredoxin-like fold,  $\beta$ 1- $\alpha$ 1- $\beta$ 2- $\beta$ 3- $\alpha$ 2- $\beta$ 4 [96]. The four  $\beta$ -strands are arranged in a right-handed twist and form an antiparallel  $\beta$ -sheet that packs against the two  $\alpha$ -helices (Figure 7C). This protein contains one RRM (RNA Recognition Motif) comprised of two ribonucleo-protein motifs (RNP1, Lys/Arg-Gly-Phe/Tyr-Gly/Ala-Phe/Tyr-Val/Ile/Leu-X-Phe/Tyr and RNP2, Ile/Val/Leu-Phe/Tyr-Ile/Val/Leu-X-Asn-Leu). Since the RRM motif is an abundant component in protein structures, only the RRM motif could not tell the exact function of HP0062. Actually, a total of 6,056 RRM motifs can be found in 3541 different proteins in the Pfam database [97]. We could not elucidate the biological function of HP0827 from a structural basis, though the structure may provide information on the putative RNA binding site. Further biological studies may be required for this case.

The HP1242 gene encodes a 76-residue conserved hypothetical protein with a molecular weight of 9111 Da. HP1242 adopts a full helical structure, which is composed of three  $\alpha$ -helices [98]. These correspond to residues 6–14 ( $\alpha$ I), 18–38 ( $\alpha$ II), and 43–75 ( $\alpha$ III). The overall structure of HP1242 represents a coiled-coil-like conformation (Figure 7D). Based on the sequence homology, HP1242 is classified as the DUF (Domain of Unknown Function) 465 family, which has an unknown function. These family members are found in several bacterial proteins, and also in the heavy chain of eukaryotic myosin and kinesin, which are predicted to form coiled coil structures. HP1242 has a structural homology with a variety of proteins including the rop protein (transcription regulation), arfaptin 2 fragment (signaling protein), sensory rhodopsin II fragment (membrane protein complex) and so on [99]. This result indicates that the function of HP1242 could not be evaluated by only a structural comparison.

We also determined the solution structure of HP1423, which has 84 amino acid residues. HP1423 is a hypothetical protein as well. According to the Pfam database, HP1423 belongs to S4 (PF01479) superfamily. The S4 domain is a small domain consisting of 60–65 amino acid residues that probably mediates binding to RNA [100]. The structure of HP1423 is composed of five  $\beta$ -strands and three  $\alpha$ -helices [101]. The topology can be described as  $\alpha$ 1- $\alpha$ 2- $\beta$ 2- $\beta$ 1- $\beta$ 3- $\beta$ 4- $\alpha$ 3- $\beta$ 5 (Figure 7E). Notably, the region, extending from  $\alpha$ 1 through  $\beta$ 3, forms an obvious structural motif, the so called  $\alpha$ L motif,

because of the two  $\alpha$ -helices and the loop between  $\beta$ 2 and  $\beta$ 3 which forms an L-shaped meander (Figure 7E). This structural motif shows a high degree of conservation between different families within the S4 (PF01479) superfamily and may be important for interaction with RNA [100]. The surface region of the  $\alpha$ L motif of HP1423 has a strong concentration of positive charge and the loop between  $\beta$ 4 and  $\alpha$ 3 exposes another positively charged side chain of K67, which may raise the possibility that HP1423 is a RNA binding protein (Figure 7E). The DALI result also showed that HP1423 is structurally similar to proteins that belong to S4 superfamily. The S4 superfamily includes the Hsp15 protein (PDB code: 1DM9-B), ribosomal small subunit pseudouridine synthase A (PDB code: 1VIO-A), 30S ribosomal protein S4 (PDB code: 1FJG-D), and so on. All these homologues contain the  $\alpha$ L motif. However, the distribution of positively-charged residues on the protein surfaces was somewhat different between homologous proteins [101], suggesting that HP1423 may bind to RNA through the  $\alpha$ L motif in a similar but not exactly same manner as the S4 RNA binding proteins.

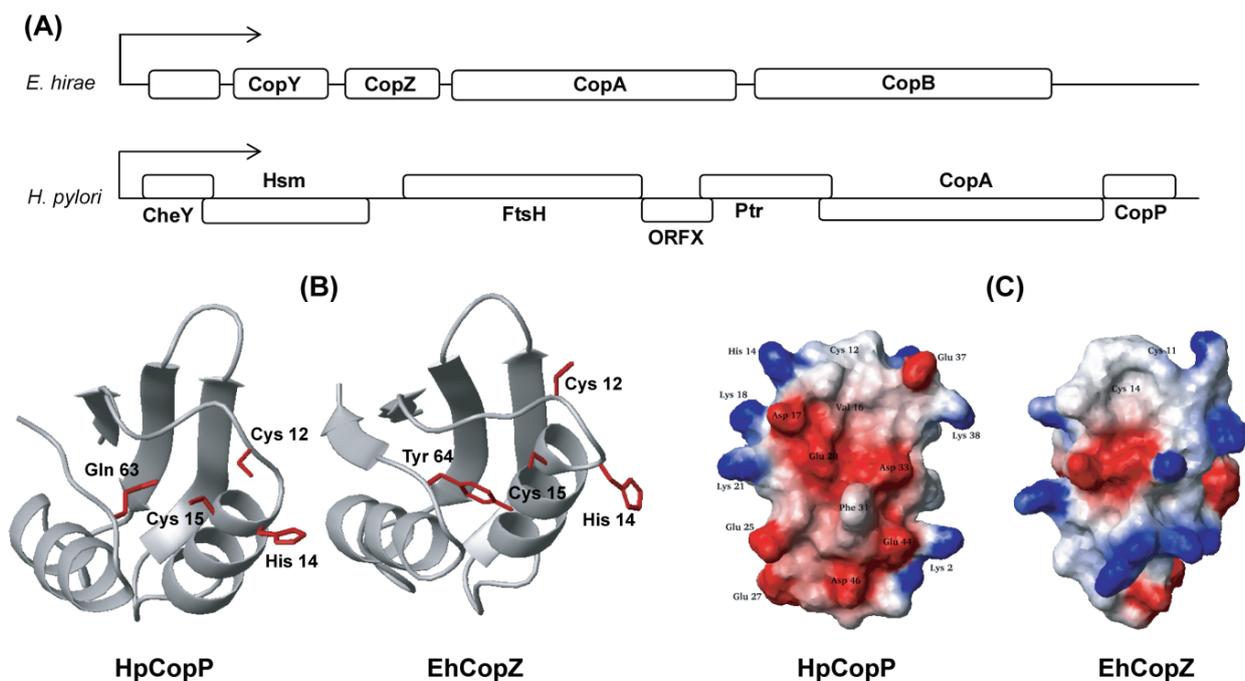
## 5. Different Characteristic with Known Function

Bioinformatics tools have been remarkably developed, providing biologists valuable information for functional elucidation. Nevertheless, prediction of protein function from sequence and structure is a difficult problem, because homologous proteins often have different functions. In addition, the protein that is classified into a known protein, based on the sequence homology, often shows some functional ambiguity since the composition of the operon is quite different from that of the known system. In addition, some of the proteins, which are considered to be well characterized, may have additional functions beyond their listed function [102]. In this regard, it is still worth investigating known proteins from a newly sequenced genome for their cell and biological functions. Here, we present two examples of well-defined proteins that have different characteristics compared to the homologues.

Copper metabolism by copper chaperones has been studied extensively in both eukaryotes and bacteria. In the gram-positive bacterium, *Enterococcus hirae*, the cop operon is composed of four proteins: two integral membrane P-type ATPases, CopA, and CopB which transport Cu(I) into cells under Cu(I) limiting conditions and eliminate Cu(I) under conditions of high Cu(I) levels, respectively [103,104]. The imported copper ions are transferred from CopA to the CopZ chaperone [105–107] and CopY, a gene repressor, is released from the cop operon promoter when Cu(I) is delivered to CopY by the copper chaperone, CopZ (Figure 8A). In the case of the gram-negative bacterium, *H. pylori*, copper homeostasis seems to be maintained by only two proteins CopA and CopP (HP1073). The *H. pylori* cop operon (Figure 8A) is included in a novel stress-responsive operon (sro), which encodes the flagellar motor switch protein CheY, the putative methyltransferase Hsm, the cell division protein FtsH, the putative phosphatidyltransferase Ptr, the heavy metal-binding proteins CopA and CopP, and an open reading frame of unknown function [108]. CopA is a member of the bacterial copper ion ATPase family, and CopP, which is homologous to *E. hirae* CopZ, is a putative copper binding regulatory protein of 66 amino acids [104,108]. CopA of *H. pylori* was identified as a Cu(II) export ATPase [109], which shows that its biological role is more similar to that of *E. hirae* CopB, rather than CopA [110]. Moreover, the CopP gene resides immediately downstream of the CopA gene, while the *E. hirae* CopZ gene resides upstream of the CopA gene. Therefore, the cop operon organization seems to be evolutionarily modified in each bacterium.

Generally, CopZ proteins share a conserved structure,  $\beta\alpha\beta\beta\alpha$  with a similar metal binding region. Interestingly, HpCopP adopts the  $\beta\alpha\beta\beta\alpha$  fold with a missing C-terminal  $\beta$  strand [111]. The overall topologies of the secondary structural components are very similar between the CopZs and HpCopP, while some variations in the loop regions appear (Figure 8). The relationship between the unusual fold and the copper specificity was evaluated [111]. We showed that HpCopP was not adequate for Cu(II) binding since the fold stability decreased in the presence of Cu(II) ion, suggesting that the structure of HpCopP is optimized for the transfer of toxic Cu(I). The absence of the C-terminal  $\beta$ -strand may lead to decreased conformational stability of loop I including the CXXC motif (Cu binding motif), which probably contributes to the disulfide bond formation between the two cysteine residues in the presence of Cu(II) ion. These findings should be helpful in evaluating the copper metabolism related with HpCopA and HpCopP in *H. pylori*.

**Figure 8.** Structural comparison between apo-HpCopP and apo-CopZ. (A) The composition of cop ORFs of *H. pylori* and *E. hirae*; (B) The orientation of the two cysteines and one histidine in the CXXC motif of HpCopP is compared with that of EhCopZ. The hydrophobic protection by Tyr 64 in loop V stabilizes the Cu(I)-coordination in EhCopZ. This residue is highly conserved in bacterial proteins, but is replaced with Gln 63 in HpCopP. The side-chain of Gln 63 is not fully exposed to the solvent and points toward the metal binding site in apo-HpCopP. The structures of EhCopZ (PDB ID: 1CPZ) were obtained from the PDB; (C) The electrostatic potential surfaces of HpCopP and EhCopZ are compared to each other. The positively and negatively charged residues are represented in blue and red, respectively.

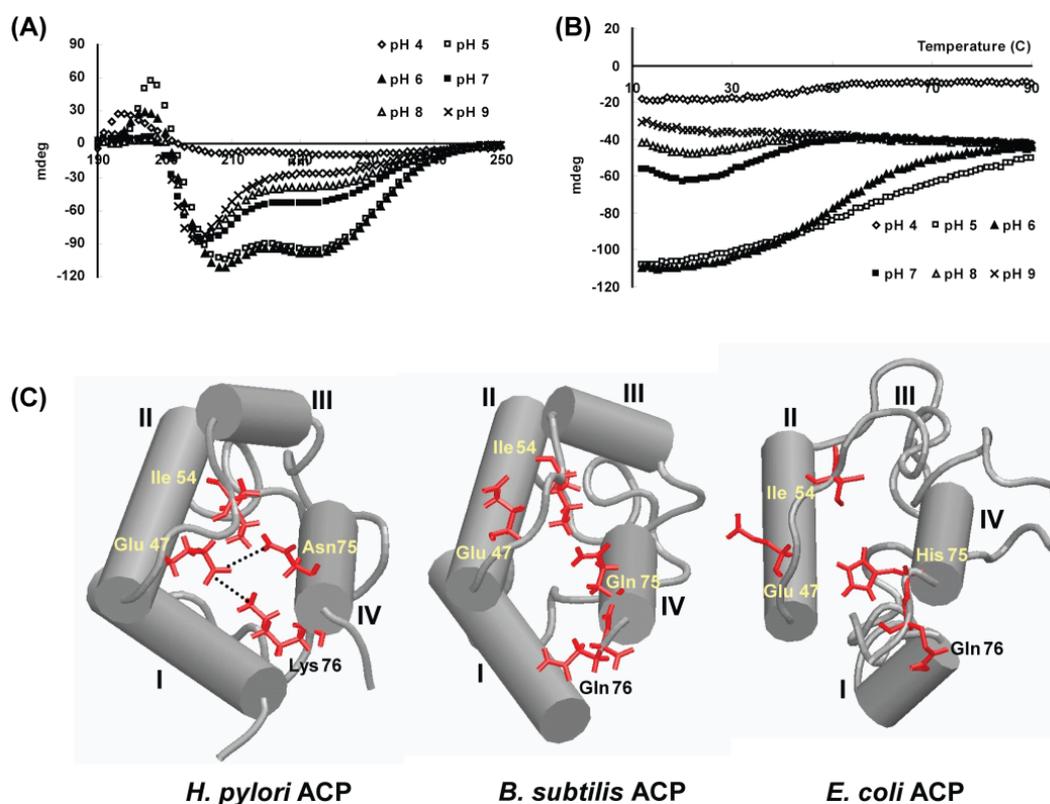


Acyl carrier protein (ACP) found in bacteria is a monofunctional protein, that is, a type II enzyme in fatty acid biosynthesis. All the ACPs are decorated by acyl carrier protein synthase (ACPS) with fatty acids, which are covalently attached as thioesters to the 4'-phosphopantetheine prosthetic group at

highly conserved Ser 36 [112]. Fatty acid binding has little influence on ACP conformation under physiological conditions [113], but it stabilizes ACP against denaturation at alkaline pH [114].

*H. pylori* ACP (HP0559) is composed of 78 amino acids with a pI value of 3.9, and its primary structure is similar with those of homologous ACPs. Like other ACPs, HpACP forms a helical bundle structure through hydrophobic contacts between the helices (Figure 9). However, we found an unusual behavior of HpACP at neutral pH [115]. HpACP exists as a partially unfolded state at neutral pH, which is a unique characteristic of HpACP (Figure 9). In contrast, the overall helical structure of *E. coli* ACP was maintained at pH 7 [116] and *Vibrio harveyi* ACP exhibited a random coil-like conformation at pH 7 [117].

**Figure 9.** Comparison of the *H. pylori* ACP structure with the *B. subtilis* ACP and *E. coli* ACP structures. (A) CD spectra of HpACP recorded at various pHs. At neutral and alkaline pH, the conformational transition of HpACP occurred; (B) Tm curves of HpACP. At acidic pH 6, the temperature curves of the HpACP showed a distinct melting temperature around 50 °C. The unfolding process above neutral pH proceeded through multi-phasic changes, showing at least three stages exist; (C) Schematic representation showing the buried hydrophilic residues Glu 47, Asn 75 and Lys 76 in the energy minimized average structure of HpACP. Putative hydrogen-bonding interactions are indicated by dotted lines. The corresponding residues are compared to those of *B. subtilis* ACP and *E. coli* ACP.



The pH dependent-conformational change of a protein from *H. pylori* is a very interesting feature, considering that the environment of the stomach has a low pH. A few studies showed the relationship between the mutation of various residues and the pH-dependent structural stability. The mutation of Val 43 to Ile in *E. coli* ACP increases the stability to pH-induced expansion in electrophoretic systems,

concomitantly inducing more compact folding [118]. The mutants F50 A and I54 A of *V. harveyi* are incapable of adopting a native conformation with increased hydrodynamic radius at neutral pH [117]. In addition, a few basic residues scattered near the *N*- and *C*-termini, for example, His 75 of *E. coli* ACP, are necessary for ACP to maintain a native conformation at neutral pH [119]. Through our structural analysis, we found that several hydrophilic residues (Glu 47, Asn 75, and Lys 76) play an important role in structural stability. Therefore, we could suggest that, unlike other ACPs, the helical bundle of *H. pylori* ACP is maintained by, not only hydrophobic interactions, but also by hydrophilic interactions and these interactions may be weakened by elevation of the pH because the exchange rate of protons attached to the side chain amide of Asn and Lys may increase [115].

## 6. Concluding Remarks

Mass genomic sequencing has been yielding many protein sequences that cannot be annotated, and structural genomics projects are yielding many protein structures that have unknown functions. Unknown proteins represent up to about half of the proteins in prokaryotic genomes, and much more than this in higher plants and animals [120]. In bacteria such as *H. pylori*, 30–40% of the proteins encoded by typical bacterial genomes have no clear known function [121]. Thus, a major issue of genomic studies may be to narrow the gap between the richness of sequences (and/or structures) and functional characterization as subsequent experimental investigation is costly and time-consuming [122]. Actually, only 54% of *E. coli* gene products have been experimentally investigated so far [123]. Therefore, more robust bioinformatic methods or approaches may be necessary to overcome this situation. Here, we showed several examples of successful cases for elucidating the function of *H. pylori* unknown proteins based on their structural information, which supports the potential of structural comparison for functional identification. It is hoped that the structural comparison can at least act as a guide to the possible function, even though all structures cannot elucidate the actual function.

## Acknowledgements

This study was supported by the National Research Foundation of Korea (NRF) grant funded by Korean government (MEST) (Grant number 20110001207 and 2012R1A2A1A01003569). This study was also supported by a grant of the Korea Healthcare technology R&D Project, Ministry for Health, Welfare & Family Affairs, Republic of Korea. (Grant number: A092006). This research was also supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2011-0011603).

## References

1. Rothenbacher, D.; Brenner, H. Burden of *Helicobacter pylori* and *H. pylori*-related diseases in developed countries: Recent developments and future implications. *Microbes Infect.* **2003**, *5*, 693–703.

2. Wotherspoon, A.C.; Doglioni, C.; Diss, T.C.; Pan, L.; Moschini, A.; de Boni, M.; Isaacson, P.G. Regression of primary low-grade B-cell gastric lymphoma of mucosa-associated lymphoid tissue type after eradication of *Helicobacter pylori*. *Lancet* **1993**, *342*, 575–577.
3. Peek, R.M., Jr.; Blaser, M.J. *Helicobacter pylori* and gastrointestinal tract adenocarcinomas. *Nat. Rev. Cancer* **2002**, *2*, 28–37.
4. Parsonnet, J.; Friedman, G.D.; Vandersteen, D.P.; Chang, Y.; Vogelman, J.H.; Orentreich, N.; Sibley, R.K. *Helicobacter pylori* infection and the risk of gastric carcinoma. *N. Engl. J. Med.* **1991**, *325*, 1127–1131.
5. Ferreira, A.C.; Isomoto, H.; Moriyama, M.; Fujioka, T.; Machado, J.C.; Yamaoka, Y. *Helicobacter* and gastric malignancies. *Helicobacter* **2008**, *13*, 28–34.
6. Yamaoka, Y. Mechanisms of disease: *Helicobacter pylori* virulence factors. *Nat. Rev. Gastroenterol. Hepatol.* **2010**, *7*, 629–641.
7. El-Omar, E.M. Role of host genes in sporadic gastric cancer. *Best Pract. Res. Clin. Gastroenterol.* **2006**, *20*, 675–686.
8. Graham, D.Y. *Helicobacter pylori* infection in the pathogenesis of duodenal ulcer and gastric cancer: A model. *Gastroenterology* **1997**, *113*, 1983–1991.
9. Graham, D.Y.; Lu, H.; Yamaoka, Y. African, Asian or Indian enigma, the East Asian *Helicobacter pylori*: Facts or medical myths. *J. Dig. Dis.* **2009**, *10*, 77–84.
10. Wen, S.; Moss, S.F. *Helicobacter pylori* virulence factors in gastric carcinogenesis. *Cancer Lett.* **2009**, *282*, 1–8.
11. Blaser, M.J.; Atherton, J.C. *Helicobacter pylori* persistence: Biology and disease. *J. Clin. Invest.* **2004**, *113*, 321–333.
12. Mahdavi, J.; Sonden, B.; Hurtig, M.; Olfat, F.O.; Forsberg, L.; Roche, N.; Angstrom, J.; Larsson, T.; Teneberg, S.; Karlsson, K.A.; *et al.* *Helicobacter pylori* SabA adhesin in persistent infection and chronic inflammation. *Science* **2002**, *297*, 573–578.
13. Lu, H.; Hsu, P.I.; Graham, D.Y.; Yamaoka, Y. Duodenal ulcer promoting gene of *Helicobacter pylori*. *Gastroenterology* **2005**, *128*, 833–848.
14. Backert, S.; Clyne, M. Pathogenesis of *Helicobacter pylori* infection. *Helicobacter* **2011**, *1*, 19–25.
15. Tomb, J.F.; White, O.; Kerlavage, A.R.; Clayton, R.A.; Sutton, G.G.; Fleischmann, R.D.; Ketchum, K.A.; Klenk, H.P.; Gill, S.; Dougherty, B.A.; *et al.* The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **1997**, *388*, 539–547.
16. Alm, R.A.; Ling, L.S.; Moir, D.T.; King, B.L.; Brown, E.D.; Doig, P.C.; Smith, D.R.; Noonan, B.; Guild, B.C.; deJonge, B.L.; *et al.* Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* **1997**, *397*, 176–180.
17. NCBI genome database. Available online: <http://www.ncbi.nlm.nih.gov/genome> (accessed on 31 May 2012).
18. Bieker, K.L.; Silhavy, T.J. The genetics of protein secretion in *E. coli*. *Trends Genet.* **1990**, *6*, 329–334.
19. Medigue, C.; Wong, B.C.; Lin, M.C.; Bocs, S.; Danchin, A. The *secE* gene of *Helicobacter pylori*. *J. Bacteriol.* **2002**, *184*, 2837–2840.
20. Wassarman, K.M.; Repoila, F.; Rosenow, C.; Storz, G.; Gottesman, S. Identification of novel small RNAs using comparative genomics and microarrays. *Genes Dev.* **2001**, *15*, 1637–1651.

21. Dong, Q.; Zhang, L.; Goh, K.L.; Forman, D.; O'Rourke, J.; Harris, A.; Mitchell, H. Identification and characterisation of *ssrA* in members of the *Helicobacter* genus. *Antonie Van Leeuwenhoek* **2007**, *92*, 301–307.
22. Kazantsev, A.V.; Pace, N.R. Bacterial RNase P: A new view of an ancient enzyme. *Nat. Rev. Microbiol.* **2006**, *4*, 729–740.
23. Vogel, J.; Bartels, V.; Tang, T.H.; Churakov, G.; Slagter-Jager, J.G.; Huttenhofer, A.; Wagner, E.G. RNomics in *Escherichia coli* detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucleic Acids Res.* **2003**, *31*, 6435–6443.
24. Giannakis, M.; Chen, S.L.; Karam, S.M.; Engstrand, L.; Gordon, J.I. *Helicobacter pylori* evolution during progression from chronic atrophic gastritis to gastric cancer and its impact on gastric stem cells. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 4358–4363.
25. Raymond, J.; Thiberge, J.M.; Kalach, N.; Bergeret, M.; Dupont, C.; Labigne, A.; Dauga, C. Using macro-arrays to study routes of infection of *Helicobacter pylori* in three families. *PLoS One* **2008**, *3*, doi:10.1371/journal.pone.0002259.
26. Baltrus, D.A.; Amieva, M.R.; Covacci, A.; Lowe, T.M.; Merrell, D.S.; Ottemann, K.M.; Stein, M.; Salama, N.R.; Guillemin, K. The complete genome sequence of *Helicobacter pylori* strain G27. *J. Bacteriol.* **2009**, *191*, 447–448.
27. Covacci, A.; Censini, S.; Bugnoli, M.; Petracca, R.; Burroni, D.; Macchia, G.; Massone, A.; Papini, E.; Xiang, Z.; Figura, N.; *et al.* Molecular characterization of the 128-kDa immunodominant antigen of *Helicobacter pylori* associated with cytotoxicity and duodenal ulcer. *Proc. Natl. Acad. Sci. USA* **1993**, *90*, 5791–5795.
28. Oh, J.D.; Kling-Backhed, H.; Giannakis, M.; Xu, J.; Fulton, R.S.; Fulton, L.A.; Cordum, H.S.; Wang, C.; Elliott, G.; Edwards, J.; *et al.* The complete genome sequence of a chronic atrophic gastritis *Helicobacter pylori* strain: Evolution during disease progression. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 9999–10004.
29. McClain, M.S.; Shaffer, C.L.; Israel, D.A.; Peek, R.M., Jr.; Cover, T.L. Genome sequence analysis of *Helicobacter pylori* strains associated with gastric ulceration and gastric cancer. *BMC Genomics* **2009**, *10*, doi:10.1186/1471-2164-10-3.
30. Devi, S.H.; Taylor, T.D.; Avasthi, T.S.; Kondo, S.; Suzuki, Y.; Megraud, F.; Ahmed, N. Genome of *Helicobacter pylori* strain 908. *J. Bacteriol.* **2010**, *192*, 6488–6489.
31. Farnbacher, M.; Jahns, T.; Willrodt, D.; Daniel, R.; Haas, R.; Goesmann, A.; Kurtz, S.; Rieder, G. Sequencing, annotation, and comparative genome analysis of the gerbil-adapted *Helicobacter pylori* strain B8. *BMC Genomics* **2010**, *11*, doi:10.1186/1471-2164-11-335.
32. Fischer, W.; Windhager, L.; Rohrer, S.; Zeiller, M.; Karnholz, A.; Hoffmann, R.; Zimmer, R.; Haas, R. Strain-specific genes of *Helicobacter pylori*: Genome evolution driven by a novel type IV secretion system and genomic island transfer. *Nucleic Acids Res.* **2010**, *38*, 6089–6101.
33. Kersulyte, D.; Kalia, A.; Gilman, R.H.; Mendez, M.; Herrera, P.; Cabrera, L.; Velapatino, B.; Balqui, J.; Paredes Puente de la Vega, F.; Rodriguez Ulloa, C.A.; *et al.* *Helicobacter pylori* from Peruvian amerindians: Traces of human migrations in strains from remote Amazon, and genome sequence of an Amerind strain. *PLoS One* **2010**, *5*, doi:10.1371/journal.pone.0015076.

34. Mane, S.P.; Dominguez-Bello, M.G.; Blaser, M.J.; Sobral, B.W.; Hontecillas, R.; Skoneczka, J.; Mohapatra, S.K.; Crasta, O.R.; Evans, C.; Modise, T.; *et al.* Host-interactive genes in Amerindian *Helicobacter pylori* diverge from their Old World homologs and mediate inflammatory responses. *J. Bacteriol.* **2010**, *192*, 3078–3092.
35. Thiberge, J.M.; Boursaux-Eude, C.; Lehours, P.; Dillies, M.A.; Creno, S.; Coppee, J.Y.; Rouy, Z.; Lajus, A.; Ma, L.; Burucoa, C.; *et al.* From array-based hybridization of *Helicobacter pylori* isolates to the complete genome sequence of an isolate associated with MALT lymphoma. *BMC Genomics* **2010**, *11*, doi:10.1186/1471-2164-11-368.
36. Avasthi, T.S.; Devi, S.H.; Taylor, T.D.; Kumar, N.; Baddam, R.; Kondo, S.; Suzuki, Y.; Lamouliatte, H.; Megraud, F.; Ahmed, N. Genomes of two chronological isolates (*Helicobacter pylori* 2017 and 2018) of the West African *Helicobacter pylori* strain 908 obtained from a single patient. *J. Bacteriol.* **2011**, *193*, 3385–3386.
37. Furuta, Y.; Kawai, M.; Yahara, K.; Takahashi, N.; Handa, N.; Tsuru, T.; Oshima, K.; Yoshida, M.; Azuma, T.; Hattori, M.; *et al.* Birth and death of genes linked to chromosomal inversion. *Proc. Natl. Acad. Sci. USA* **2011**, *108*, 1501–1506.
38. Lehours, P.; Vale, F.F.; Bjursell, M.K.; Melefors, O.; Advani, R.; Glavas, S.; Guegueniat, J.; Gontier, E.; Lacomme, S.; Alves Matos, A.; *et al.* Genome sequencing reveals a phage in *Helicobacter pylori*. *MBio* **2011**, *2*, doi: 10.1128/mBio.00239-11.
39. Alvi, A.; Devi, S.M.; Ahmed, I.; Hussain, M.A.; Rizwan, M.; Lamouliatte, H.; Megraud, F.; Ahmed, N. Microevolution of *Helicobacter pylori* type IV secretion systems in an ulcer disease patient over a ten-year period. *J. Clin. Microbiol.* **2007**, *45*, 4039–4043.
40. Prouzet-Mauleon, V.; Hussain, M.A.; Lamouliatte, H.; Kauser, F.; Megraud, F.; Ahmed, N. Pathogen evolution *in vivo*: Genome dynamics of two isolates obtained 9 years apart from a duodenal ulcer patient infected with a single *Helicobacter pylori* strain. *J. Clin. Microbiol.* **2005**, *43*, 4237–4241.
41. Linz, B.; Balloux, F.; Moodley, Y.; Manica, A.; Liu, H.; Roumagnac, P.; Falush, D.; Stamer, C.; Prugnolle, F.; van der Merwe, S.W.; *et al.* An African origin for the intimate association between humans and *Helicobacter pylori*. *Nature* **2007**, *445*, 915–918.
42. Falush, D.; Wirth, T.; Linz, B.; Pritchard, J.K.; Stephens, M.; Kidd, M.; Blaser, M.J.; Graham, D.Y.; Vacher, S.; Perez-Perez, G.I.; *et al.* Traces of human migrations in *Helicobacter pylori* populations. *Science* **2003**, *299*, 1582–1585.
43. Wirth, T.; Wang, X.; Linz, B.; Novick, R.P.; Lum, J.K.; Blaser, M.; Morelli, G.; Falush, D.; Achtman, M. Distinguishing human ethnic groups by means of sequences from *Helicobacter pylori*: Lessons from Ladakh. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 4746–4751.
44. Peterson, J.D.; Umayam, L.A.; Dickinson, T.M.; Hickey E.K.; White, O. The comprehensive microbial resource. *Nucleic Acids Res.* **2001**, *29*, 123–125.
45. Marais, A.; Mendz, G.L.; Hazell, S.L.; Megraud, F. Metabolism and genetics of *Helicobacter pylori*: The genome era. *Microbiol. Mol. Biol. Rev.* **1999**, *63*, 642–674.
46. Merrell, D.S.; Thompson, L.J.; Kim, C.C.; Mitchell, H.; Tompkins, L.S.; Lee, A.; Falkow, S. Growth phase-dependent response of *Helicobacter pylori* to iron starvation. *Infect. Immun.* **2003**, *71*, 6510–6525.

47. Wen, Y.; Marcus, E.A.; Matrubutham, U.; Gleeson, M.A.; Scott, D.R.; Sachs, G. Acid-adaptive genes of *Helicobacter pylori*. *Infect. Immun.* **2003**, *71*, 5921–5939.
48. Cremades, N.; Velazquez-Campoy, A.; Martinez-Julvez, M.; Neira, J.L.; Perez-Dorado, I.; Hermoso, J.; Jimenez, P.; Lanas, A.; Hoffman, P.S.; Sancho, J. Discovery of specific flavodoxin inhibitors as potential therapeutic agents against *Helicobacter pylori* infection. *ACS Chem. Biol.* **2009**, *4*, 928–938.
49. Han, K.D.; Matsuura, A.; Ahn, H.C.; Kwon, A.R.; Min, Y.H.; Park, H.J.; Won, H.S.; Park, S.J.; Kim, D.Y.; Lee, B.J. Functional identification of toxin-antitoxin molecules from *Helicobacter pylori* 26695 and structural elucidation of the molecular interactions. *J. Biol. Chem.* **2011**, *286*, 4842–4853.
50. Han, K.D.; Park, S.J.; Jang, S.B.; Son, W.S.; Lee, B.J. Solution structure of conserved hypothetical protein HP0894 from *Helicobacter pylori*. *Proteins* **2005**, *61*, 1114–1116.
51. Yeo, H.J.; Savvides, S.N.; Herr, A.B.; Lanka, E.; Waksman, G. Crystal structure of the hexameric traffic ATPase of the *Helicobacter pylori* type IV secretion system. *Mol. Cell* **2000**, *6*, 1461–1472.
52. Protein Data Bank. Available online: <http://www.rcsb.org> (accessed on 1 March 2012).
53. Goulding, C.W.; Perry, L.J. Protein production in *Escherichia coli* for structural studies by X-ray crystallography. *J. Struct. Biol.* **2003**, *142*, 133–143.
54. Cussac, V.; Ferrero, R.L.; Labigne, A. Expression of *Helicobacter pylori* urease genes in *Escherichia coli* grown under nitrogen-limiting conditions. *J. Bacteriol.* **1992**, *174*, 2466–2473.
55. Pettersen, E.F.; Goddard, T.D.; Huang, C.C.; Couch, G.S.; Greenblatt, D.M.; Meng, E.C.; Ferrin, T.E. UCSF Chimera—A visualization system for exploratory research and analysis. *J. Comput. Chem.* **2004**, *25*, 1605–1612.
56. Altschul, S.F.; Gish, W.; Miller, W.; Myers, E.W.; Lipman, D.J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410.
57. Kannan, S.; Hauth, A.M.; Burger, G. Function prediction of hypothetical proteins without sequence similarity to proteins of known function. *Protein Pept. Lett.* **2008**, *15*, 1107–1116.
58. Chou, K.C.; Shen, H.B. Cell-PLoc: A package of Web servers for predicting subcellular localization of proteins in various organisms. *Nat. Protoc.* **2008**, *3*, 153–162.
59. Shen, H.B.; Chou, K.C. EzyPred: A top-down approach for predicting enzyme functional classes and subclasses. *Biochem. Biophys. Res. Commun.* **2007**, *364*, 53–59.
60. Dobson, P.D.; Cai, Y.D.; Stapley, B.J. Doig, A.J. Prediction of protein function in the absence of significant sequence similarity. *Curr. Med. Chem.* **2004**, *11*, 2135–2142.
61. Dundas, J.; Ouyang, Z.; Tseng, J.; Binkowski, A.; Turpaz, Y.; Liang, J. CASTp: Computed atlas of surface topography of proteins with structural and topographical mapping of functionally annotated residues. *Nucleic Acids Res.* **2006**, *34*, W116–W118.
62. Holm, L.; Kääriäinen, S.; Rosenström, P.; Schenkel, A. Searching protein structure databases with DaliLite v.3. *Bioinformatics* **2008**, *24*, 2780–2781.
63. Holm, L.; Rosenström, P. Dali server: Conservation mapping in 3D. *Nucleic Acids Res.* **2010**, *38*, W545–W549.
64. Kawabata, T.; Nishikawa, K. Protein structure comparison using the markov transition model of evolution. *Proteins* **2000**, *41*, 108–122.

65. Nimrod, G.; Schushan, M.; Steinberg, D.M.; Ben-Tal, N. Detection of functionally important regions in “hypothetical proteins” of known structure. *Structure* **2008**, *16*, 1755–1763.
66. Aloy, P.; Querol, E.; Aviles, F.X.; Sternberg, M.J. Automated structure-based prediction of functional sites in proteins: Applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J. Mol. Biol.* **2001**, *311*, 395–408.
67. Ondrechen, M.J.; Clifton, J.G.; Ringe, D. THEMATICs: A simple computational predictor of enzyme function from structure. *Proc. Natl. Acad. Sci. USA* **2001**, *98*, 12473–12478.
68. Pazos, F.; Sternberg, M.J. Automated prediction of protein function and detection of functional sites from structure. *Proc. Natl. Acad. Sci. USA* **2004**, *101*, 14754–14759.
69. Pettit, F.K.; Bare, E.; Tsai, A.; Bowie, J.U. HotPatch: A statistical approach to finding biologically relevant features on protein surfaces. *J. Mol. Biol.* **2007**, *369*, 863–879.
70. Sierk, M.L.; Pearson, W.R. Sensitivity and selectivity in protein structure comparison. *Protein Sci.* **2004**, *13*, 773–785.
71. Altschul, S.F.; Gish, W. Local alignment statistics. *Methods Enzymol.* **1996**, *266*, 460–480.
72. Han, K.D.; Park, S.J.; Jang, S.B.; Son, W.S.; Lee, B.J. Solution structure of conserved hypothetical protein HP0894 from *Helicobacter pylori*. *Proteins* **2005**, *61*, 1111–1113.
73. Marchler-Bauer, A.; Bryant, S.H. CD-Search: Protein domain annotations on the fly. *Nucleic Acids Res.* **2004**, *32*, W327–W331.
74. Bateman, A.; Birney, E.; Cerruti, L. The Pfam protein families data base. *Nucleic Acids Res.* **2002**, *30*, 276–280.
75. Takagi, H.; Kakuta, Y.; Okada, T.; Yao, M.; Tanaka, I.; Kimura, M. Crystal structure of archaeal toxin-antitoxin RelE-RelB complex with implications for toxin activity and antitoxin effects. *Nat. Struct. Mol. Biol.* **2005**, *12*, 327–331.
76. Gerdes, K.; Christensen, S.K.; Løbner-Olesen, A. Prokaryotic toxin-antitoxin stress response loci. *Nat. Rev. Microbiol.* **2005**, *3*, 371–382.
77. Wilson, D.N.; Nierhaus, K.H. RelBE or not to be. *Nat. Struct. Mol. Biol.* **2005**, *12*, 282–284.
78. Han, K.D.; Matsuura, A.; Ahn, H.C.; Kwon, A.R.; Min, Y.H.; Park, H.J.; Won, H.S.; Park, S.J.; Kim, D.Y.; Lee, B.J. Functional identification of toxin-antitoxin molecules from *Helicobacter pylori* 26695 and structural elucidation of the molecular interactions. *J. Biol. Chem.* **2011**, *286*, 4842–4853.
79. Kamada, K.; Hanaoka, F.; Burley, S.K. Crystal structure of the MazE/MazF complex: Molecular bases of antidote-toxin recognition. *Mol. Cell* **2003**, *11*, 875–884.
80. Han, K.D.; Park, S.J.; Jang, S.B.; Lee, B.J. Solution structure of conserved hypothetical protein HP0892 from *Helicobacter pylori*. *Proteins* **2008**, *70*, 599–602.
81. Terry, C.E.; McGinnis, L.M.; Madigan, K.C.; Cao, P.; Cover, T.L.; Liechti, G.W.; Peek, R.M., Jr.; Forsyth, M.H. Genomic comparison of cag pathogenicity island (PAI)-positive and -negative *Helicobacter pylori* strains: Identification of novel markers for cag PAI-positive strains. *Infect. Immun.* **2005**, *73*, 3794–3798.
82. Cheetham, B.F.; Tattersall, D.B.; Bloomfield, G.A.; Rood, J.I.; Katz, M.E. Identification of a gene encoding a bacteriophage-related integrase in a vap region of the *Dichelobacter nodosus* genome. *Gene* **1995**, *162*, 53–58.

83. Katz, M.E.; Strugnell, R.A.; Rood, J.I. Molecular characterization of a genomic region associated with virulence in *Dichelobacter nodosus*. *Infect. Immun.* **1992**, *60*, 4586–4592.
84. Takai, S.; Hines, S.A.; Sekizaki, T.; Nicholson, V.M.; Alperin, D.A.; Osaki, M.; Takamatsu, D.; Nakamura, M.; Suzuki, K.; Ogino, N.; *et al.* DNA sequence and comparison of virulence plasmids from *Rhodococcus equi* ATCC 33701 and 103. *Infect. Immun.* **2000**, *68*, 6840–6847.
85. Tomb, J.; White, O.; Kerlavage, A.R.; Clayton, R.A.; Sutton, G.G.; Fleischmann, R.D.; Ketchum, K.A.; Klenk, H.P.; Gill, S.; Dougherty, B.A.; *et al.* The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **1997**, *388*, 539–547.
86. Katz, M.E.; Strugnell, R.A.; Rood, J.I. Molecular characterization of a genomic region associated with virulence in *Dichelobacter nodosus*. *Infect. Immun.* **1992**, *60*, 4586–4592.
87. Benoit, S.; Benachour, A.; Taouji, S.; Auffray, Y.; Hartke, A. Induction of vap genes encoded by the virulence plasmid of *Rhodococcus equi* during acid tolerance response. *Res. Microbiol.* **2001**, *152*, 439–449.
88. Galli, D.M.; LeBlanc, D.J. Characterization of pVT736-1, a rolling-circle plasmid from the gram-negative bacterium *Actinobacillus actinomycetemcomitans*. *Plasmid* **1994**, *31*, 148–157.
89. Kwon, A.R.; Kim, J.H.; Park, S.J.; Lee, K.Y.; Min, Y.H.; Im, H.; Lee, I.; Lee, K.Y.; Lee, B.J. Structural and biochemical characterization of HP0315 from *Helicobacter pylori* as a VapD protein with an endoribonuclease activity. *Nucleic Acids Res.* **2012**, *40*, 4216–4228.
90. Makarova, K.S.; Grishin, N.V.; Shabalina, S.A.; Wolf, Y.I.; Koonin, E.V. A putative RNA-interference-based immune system in prokaryotes: Computational analysis of the predicted enzymatic machinery, functional analogies with eukaryotic RNAi, and hypothetical mechanisms of action. *Biol. Direct.* **2006**, *1*, doi:10.1186/1745-6150-1-7.
91. Jang, S.B.; Kwon, A.R.; Son, W.S.; Park, S.J.; Lee, B.J. Crystal structure of hypothetical protein HP0062 (O24902\_HELPY) from *Helicobacter pylori* at 1.65 Å resolution. *J. Biochem.* **2009**, *146*, 535–540.
92. Pallen, M.J. The ESAT-6/WXG100 superfamily—And a new Gram-positive secretion system? *Trends Microbiol.* **2002**, *10*, 209–212.
93. Plano, G.V.; Day, J.B.; Ferracci, F. Type III export: New uses for an old pathway. *Mol. Microbiol.* **2001**, *40*, 284–293.
94. Seo, M.D.; Park, S.J.; Kim, H.J.; Lee, B.J. Solution structure of hypothetical protein, HP0495 (Y495\_HELPY) from *Helicobacter pylori*. *Proteins* **2007**, *67*, 1189–1192.
95. Seo, M.D.; Park, S.J.; Kim, H.J.; Seok, S.H.; Lee, B.J. Backbone 1H, 15N, and 13C resonance assignment and secondary structure prediction of HP0495 from *Helicobacter pylori*. *J. Biochem. Mol. Biol.* **2007**, *40*, 839–843.
96. Jang, S.B.; Ma, C.; Lee, J.Y.; Kim, J.H.; Park, S.J.; Kwon, A.R.; Lee, B.J. NMR solution structure of HP0827 (O25501\_HELPY) from *Helicobacter pylori*: Model of the possible RNA-binding site. *J. Biochem.* **2009**, *146*, 667–674.
97. Bateman, A.; Birney, E.; Cerruti, L.; Durbin, R.; Eddy, S.R.; Griffiths-Jones, S.; Howe, K.L.; Marshall, M.; Sonnhammer, E.L. The Pfam protein families database. *Nucleic Acids Res.* **2002**, *30*, 276–280.
98. Kang, S.J.; Park, S.J.; Jung, S.J.; Lee, B.J. Backbone 1H, 15N, and 13C resonance assignment of HP1242 from *Helicobacter pylori*. *J. Biochem. Mol. Biol.* **2005**, *38*, 591–594.

99. Kang, S.J.; Park, S.J.; Jung, S.J.; Lee, B.J. Solution structure of HP1242 from *Helicobacter pylori*. *Proteins* **2005**, *61*, 1111–1113.
100. Aravind, L.; Koonin, E.V. Novel predicted RNA-binding domains associated with the translation machinery. *J. Mol. Evol.* **1999**, *48*, 291–302.
101. Kim, J.H.; Park, S.J.; Lee, K.Y.; Son, W.S.; Sohn, N.Y.; Kwon, A.R.; Lee, B.J. Solution structure of hypothetical protein HP1423 (Y1423\_HELPY) reveals the presence of alphaL motif related to RNA binding. *Proteins* **2009**, *75*, 252–257.
102. Copley, S.D. Enzymes with extra talents: Moonlighting functions and catalytic promiscuity. *Curr. Opin. Chem. Biol.* **2003**, *7*, 265–272.
103. Odermatt, A.; Suter, H.; Krapf, R.; Solioz, M. Primary structure of two P-type ATPases involved in copper homeostasis in *Enterococcus hirae*. *J. Biol. Chem.* **1993**, *268*, 12775–12779.
104. Odermatt, A.; Solioz, M. Two trans-acting metalloregulatory proteins controlling expression of the copper-ATPases of *Enterococcus hirae*. *J. Biol. Chem.* **1995**, *270*, 4349–4354.
105. Wunderli-Ye, H.; Solioz, M. Effects of promoter mutations on the *in vivo* regulation of the cop operon of *Enterococcus hirae* by copper(I) and copper(II). *Biochem. Biophys. Res. Commun.* **1999**, *259*, 443–449.
106. Pufahl, R.A.; Singer, C.P.; Peariso, K.L.; Lin, S.; Schmidt, P.J.; Fahrni, C.J.; Culotta, V.C.; Penner-Hahn, J.E.; O'Halloran, T.V. Metal ion chaperone function of the soluble Cu(I) receptor Atx1. *Science* **1997**, *278*, 853–856.
107. Banci, L.; Bertini, I.; Ciofi-Baffoni, S.; Del Conte, R.; Gonnelli, L. Understanding copper trafficking in bacteria: Interaction between the copper transport protein CopZ and the N-terminal domain of the copper ATPase CopA from *Bacillus subtilis*. *Biochemistry* **2003**, *42*, 1939–1949.
108. Beier, D.; Spohn, G.; Rappuoli, R.; Scarlato, V. Identification and characterization of an operon of *Helicobacter pylori* that is involved in motility and stress adaptation. *J. Bacteriol.* **1997**, *179*, 4676–4683.
109. Bayle, D.; Wangler, S.; Weitzenegger, T.; Steinhilber, W.; Volz, J.; Przybylski, M.; Schafer, K.P.; Sachs, G.; Melchers, K. Properties of the P-type ATPases encoded by the copAP operons of *Helicobacter pylori* and *Helicobacter felis*. *J. Bacteriol.* **1998**, *180*, 317–329.
110. Solioz, M.; Stoyanov, J.V. Copper homeostasis in *Enterococcus hirae*. *FEMS Microbiol. Rev.* **2003**, *27*, 183–195.
111. Park, S.J.; Jung, Y.S.; Kim, J.S.; Seo, M.D.; Lee, B.J. Structural insight into the distinct properties of copper transport by the *Helicobacter pylori* CopP protein. *Proteins* **2008**, *71*, 1007–1019.
112. Vandem, B.T.; Cronan, J.E., Jr. Genetics and regulation of bacterial lipid metabolism. *Annu. Rev. Microbiol.* **1989**, *43*, 317–343.
113. Jones, P.J.; Holak, T.A.; Prestegard, J.H. Structural comparison of acyl carrier protein in acylated and sulfhydryl forms by two-dimensional <sup>1</sup>H NMR spectroscopy. *Biochemistry* **1987**, *26*, 3493–3500.
114. Cronan, J.E., Jr. Molecular properties of short chain acyl thioesters of acyl carrier protein. *J. Biol. Chem.* **1982**, *257*, 5013–5017.
115. Park, S.J.; Kim, J.S.; Son, W.S.; Lee, B.J. pH-induced conformational transition of *H. pylori* acyl carrier protein: Insight into the unfolding of local structure. *J. Biochem.* **2004**, *135*, 337–346.

116. Schulz, H. On the structure-function relationship of acyl carrier protein of *Escherichia coli*. *J. Biol. Chem.* **1975**, *250*, 2299–2304.
117. Flaman, A.S.; Chen, J.M.; Van Iderstine, S.C; Byers, D.M. Site-directed mutagenesis of acyl carrier protein (ACP) reveals amino acid residues involved in ACP structure and acyl-ACP synthetase activity. *J. Biol. Chem.* **2001**, *276*, 35934–35939.
118. Keating, D.H.; Cronan, J.E., Jr. An isoleucine to valine substitution in *Escherichia coli* acyl carrier protein results in a functional protein of decreased molecular radius at elevated pH. *J. Biol. Chem.* **1996**, *271*, 15905–15910.
119. Keating, M.-M.; Gong, H.; Byers, D.M. Identification of a key residue in the conformational stability of acyl carrier protein. *Biochem. Biophys. Acta* **2002**, *1601*, 208–214.
120. Hanson, A.D.; Pribat, A.; Waller, J.C.; de Crécy-Lagard, V. “Unknown” proteins and “orphan” enzymes: The missing half of the engineering parts list—and how to find it. *Biochem. J.* **2010**, *425*, 1–11.
121. Galperin, M.Y.; Koonin, E.V. “Conserved hypothetical” proteins: Prioritization of targets for experimental study. *Nucleic Acids Res.* **2004**, *32*, 5452–5463.
122. Galperin, M.Y.; Koonin, E.V. From complete genome sequence to “complete” understanding? *Trends Biotechnol.* **2010**, *28*, 398–406.
123. Frishman, D. Protein annotation at genomic scale: The current status. *Chem. Rev.* **2007**, *107*, 3448–3466.

© 2012 by the authors; licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution license (<http://creativecommons.org/licenses/by/3.0/>).