UNIVERSITA' DEGLI STUDI DI PADOVA

Dipartimento di Scienze Chimiche

Scuola di Dottorato di Ricerca in Biochimica e Biotecnologie

Indirizzo: Biotecnologie

XX ciclo

Structural characterization of Cag proteins from the Pathogenicity Island of *Helicobacter pylori*

Direttore della Scuola: Ch.mo Prof. Lorenzo Pinna

Supervisore: Ch.mo Prof. Giuseppe Zanotti

Dottorando: Alessandro Angelini

31 Gennaio 2008

UNIVERSITY OF PADUA

Department of Chemistry

Doctorate School of Biochemistry and Biotechnology
Specialization: Biotechnology

Thesis

Structural characterization of Cag proteins from the Pathogenicity Island of *Helicobacter pylori*

Alessandro Angelini

Headmaster:	Professor Lorenzo Pinna (*Department of Biological Chemistry, University of Padua,Padua, Italy*)
Supervisor:	Professor Giuseppe Zanotti (*Department of Chemical Sciences, University of Padua, Italy*)
Assistant Supervisor:	Dr. Laurent Terradot (*Macromolecular Cristallography group, ESRF, Grenoble, France*)
Referees:	Professor Sine Ydun Larsen (*Department of Chemistry, University of Copenhagen, Copenhagen, Denmark*)
	Professor Dušan Turk (*Department of Biochemistry, Molecular and Structural Biology, Jozef Stefan Institute, Ljubljana, Slovenia*)

The research described in this thesis was upported by the Italian Ministries of University and Research (MIUR) with the financial aid grant from Ing. Aldo Gini Foundation.

For my mother

Contents

Summary

Helicobacter pylori chronically infects the gastric mucosa of millions of people annually worldwide: it has been estimated that over 50 % of the world population carries this infection. *H. pylori* has been associated with the development of several diseases, like chronic gastritis, gastric and duodenal ulcer, gastric adenocarcinoma and mucosa-associated lymphoma (Chapter 1).

The complete genome sequence of three different strain isolates of *H. pylori* (J99, 26995 and HPAG1) is known. The strains that contain a foreign DNA region of about 47 kb, called *cag* (*cytotoxin associated genes*) Pathogenicity Island (*cag*-PAI), cause the most severe form of virulence. Some of the *cag*-PAI genes encode proteins with homology to components of the plant pathogen *Agrobacterium tumefaciens* VirB/D4 system, which is considered the prototype of bacterial type IV secretion systems (TFSSs). T4SSs are multi-subunit machines that can deliver proteins and/or DNA into host cells and thereby influence host cell functions. In particular, the *cag* TFSS translocates the CagA effector protein into host epithelial cells. After entering the host cell, CagA induces cellular modifications, such as elongation and spreading of host cells.

The aim of this structural genomic project is to determine the three-dimensional structure of most of the proteins encoded by the *cag*-PAI, a task that will allow to elucidate the function and the organization of the entire T4SS of such a relevant pathogenic bacterium (Chapter 2).

To date, the three dimensional structure of only two proteins of *cag*-PAI are known: one is Cagα, an ATPase located at the level of the inner membrane, and the other one is CagZ, a protein essential for the translocation of CagA, whose three-dimensional crystal structure has been determined by our group.

In particular, we focalized our attention on four Cag proteins (CagA, CagD, CagS and CagV) and on another protein called *Hp*YbgC that does not belong to the *cag*-PAI but that has been found to be directly involved into the interaction with the CagA toxin.

The research described in this thesis was mostly carried out at the Department of Chemistry, University of Padua and Venetian Institute of Molecular Medicine (VIMM), Padua.

The strategy adopted to study CagD, CagS and CagV proteins included bioinformatic analyses, amplification of the selected *cag* genes, starting from purified *H. pylori* chromosomal DNA (strain CCUG 17874), parallel cloning and expression of a protein using a variety of vectors containing different tags and/or fusion partners as well as several *E. coli* host strains.

The soluble recombinant proteins obtained were then purified, concentrated, crystallized and their structures determined by X-Ray diffraction (Appendix B).

To date, a complete success was obtained only in the case of CagD and CagS proteins.

The structure of CagD, which has been determined by a MAD experiment on a Seleno-methionine derivative, is described in detail in the Chapter 4 and some functional hypotheses are discussed. The polypeptide chain folds as a single domain, composed of five β-strands and three α-helices. Three-dimensional homology searches did not reveal structural homologues that could help us to identify its function. However, very recently it has been found that CagD is secreted out of the bacterial membrane and identified in the medium. Preliminary analyses confirmed the idea that CagD could be secreted out of the membrane but it is still unclear whether the Type IV secretion system is responsible for this export or whether CagD is liberated by another mechanism.

The crystal structure of CagS, which represents a new protein fold, has been also determined at 2.3 Å resolution using MAD technique (Chapter 3). The protein is an all-α structure, composed of ten α-helices arranged in a single, compact "peanut-shaped" domain. Specific characteristics of the protein structure include the presence of a methionine cluster and a highly charged external surface. Primary sequence analyses do not show any strong CagS similarities with proteins of known function, except for some weak similarities with components involved in the peptidoglycan biosynthesis, belonging to the FemABX family of enzymes.

Concerning CagV, it is known to be an essential and conserved component of the TFSS directly involved into the translocation of CagA toxin.

In the Chapter 5, we describe the production, characterization, crystallization and preliminary diffraction analysis of two C-terminal periplasmic domains of the membrane protein CagV ($pCagV^{59-252}$ and $pCagV^{68-252}$). Since CagV has been shown to be homologue to the VirB8 proteins, extensive molecular-replacement strategies to phase $pCagV^{59-252}$ and $pCagV^{68-252}$ were performed using both VirB8 homologue structures from *A. tumefaciens* and *B. suis* respectively. Since, till now, all the Molecular replacement attempts failed and we were not able to obtain a solution which

refines, structure determination will continue using multiple isomorphous replacement (MIR) or anomalous scattering methods (SAD and MAD).

Finally, in the Chapter 6 has been described the expression and purification studies of several CagA toxin domains. Previous experiments in fact demonstrated that full-length CagA is improperly folded, being more unstable and completely degraded when expressed in *E. coli*. Moreover, the large size of the CagA protein and the weak homology sequence with other proteins with known structure and function made it difficult to predict which regions of the protein constitute well-folded domains. So to study this multidomain protein we decided to adopt an alternative strategy that consists of dissect randomly the protein into several domains making all the constructs in a single experiment by progressive deletion of its N-terminus and express them using an innovative high-throughput screening technology called ESPRIT (Expression of Soluble Proteins by Random Incremental Truncation). Using this library-based construct screening method we could obtain several soluble C-terminal domains of CagA which are now used for large culture expression and purification. Mostly of this work was carried out at the Partnership of Structural Biology laboratory (PSB) in Grenoble, under the supervision of Dr. Laurent Terradot (ESRF).

Moreover, a central domain of CagA ($CagA^{392-733}$), which was identified by protein-protein interaction studies, has been purified, characterized in solution and several crystallization trials on it alone and in complex with other partners (CagF and *Hp*YbgC) were performed.

The last chapter (Chapter 7) concerns a distinct project, developed in the last year. The protein that is the subject is called *Hp*YbgC and was found to interact specifically with the $CagA^{392-733}$ central domain by two different studies. To gain insight into its function, the recombinant protein was overproduced, purified, its crystal structure have been determined by Molecular Replacement and its enzymatic activity characterized. The structure of *Hp*YbgC shows that it is a member of the hot-dog family of proteins, with a εγ tetrameric arrangement. Comparison with structures containing hot-dog folds indicates that YbgC from *E. coli* is the closest homologue. Moreover, enzymatic assays performed with the purified protein showed that *Hp*YbgC is an acyl-CoA thioesterase that favors long chain acyl-CoA substrates.

Sommario

Il batterio Gram-negativo *Helicobacter pylori* è oggi universalmente riconosciuto come la causa principale di gastrite, ulcera gastrica e duodenale e carcinoma dello stomaco.
E' stimato dalla World Health Organization che circa il 50 % della popolazione mondiale incorre in questa infezione (Capitolo 1).
I ceppi batterici risultati patogeni, definiti di tipo I, sono caratterizzati dalla presenza di un cluster di circa 30 geni, noto come "Isola di Patogenicità" *cag* (*cytotoxin associated genes*) (*cag*-PAI). Questo cluster codifica per un Sistema di Secrezione di tipo IV (TFSS), omologo a quello del patogeno vegetale *Agrobacterium tumefaciens* e di altri batteri Gram-negativi. I TFSS sono veri e propri macrocomplessi che si sviluppano attraverso le membrane e che hanno la funzione di trasferire DNA da o verso l'ambiente esterno o traslocare proteine effettrici. In *H. pylori* la sola proteina che è nota essere secreta da questo sistema è CagA, un fattore di virulenza antigenico codificato dalla *cag*-PAI medesima. Questa, una volta iniettata nella cellula epiteliale gastrica dell'ospite, è fosforilata ad una tirosina e interferisce con la normale cascata dei segnali cellulari della cellula ospite. Poco è invece noto sulle altre proteine codificate dalla *cag*-PAI. Esperimenti sistematici di knock-out hanno dimostrato che l'assenza di molte di queste proteine abolisce la traslocazione di CagA (Capitolo 2).
Le strutture molecolari delle proteine codificate da *cag*-PAI, note fino ad ora, erano soltanto due: una è Cagα, una ATPasi localizzata al livello della membrana interna. L'altra è CagZ, una proteina essenziale per l'esporto della tossina CagA e per l'induzione dell' interleuchina-8 (IL-8), la cui struttura è stata determinata all'interno del nostro laboratorio.
Il progetto di dottorato è stato effettuato presso il gruppo di Biologia Strutturale del Prof. Giuseppe Zanotti (Dip. Di Scienze Chimiche, Università degli Studi di Padova ed Istituto Veneto di Medicina Molecolare). Tale progetto ha come obiettivo la produzione in forma ricombinante di alcune proteine codificate dalla *cag*-PAI, la loro cristallizzazione e successiva determinazione della struttura tridimensionale mediante tecniche di diffrazione di raggi X al fine di ottenere informazioni sull'organizzazione e funzionamento del sistema di secrezione di tipo IV di *H. pylori* (Appendice B).

Nel corso di questi tre anni di Dottorato mi sono dedicato nello specifico, allo studio di alcune proteine codificate dalla *cag*-PAI: CagA, CagD, CagS e CagV.
Alcune di queste (CagD, CagS e CagV) sono state clonate, prodotte per via ricombinante in *E. coli*, purificate per mezzo di tecniche cromatografiche e cristallizzate..
Per quanto riguarda la proteina CagD, misure di diffrazione su cristalli prodotti utilizzando la proteina contenente Selenometionine hanno permesso di determinarne la struttura tridimensionale utilizzando la tecnica denomianta MAD (Multiple Anomalous Dispersion). La struttura di CagD, descritta dettagliatamente nel Capitolo 4, presenta cinque foglietti β antiparalleli e tre α eliche. Poichè tale proteina non presenta alcuna significativa omologia strutturale con proteine a funzione nota, sono tutt'ora in corso esperimenti atti ad identificarne la precisa funzione e localizzazione. Precedenti studi avevano dimostrato che CagD veniva secreta dal batterio all'esterno e risultati preliminari sembrerebbero confermare tale evidenza. Resta comunque da elucidare il meccanismo con cui tale proteina viene secreta e il suo preciso ruolo nella secrezione della tossina CagA o induzione di IL-8.
Per quanto concerne CagS, analogamente a CagD, la sua struttura tridimensionale è stata determinata mediante un esperimento MAD condotto su cristalli prodotti utilizzando la proteina contenente Selenometionine (Capitolo 3). Tale proteina è caratterizzata dalla presenza di dieci α eliche organizzate in modo tale da costituire un dominio compatto. Sulla sua funzione e localizzazione sono state fino ad ora formulate solo alcune ipotesi. CagS infatti rappresenta un nuovo motivo di struttura terziaria e non presenta alcuna significativa omologia con proteine a funzione nota, se non qualche similarità con proteine coinvolte nella biosintesi del peptidoglicano (FemABX).
Nel Capitolo 5 è invece descritta la produzione, purificazione, cristallizzazione ed i preliminari studi di diffrazione effettuati su due domini periplasmatici solubili della proteina di membrana CagV ($pCagV^{59-252}$ e $pCagV^{68-252}$). Tale proteina costituisce un componente strutturale del TFSS di *H. pylori*, essenziale per la traslocazione della tossina CagA. Poiché la proteina CagV presenta delle similarità di funzione e di sequenza, seppur bassa, con le proteine VirB8 di altri batteri gram-negativi, si è dapprima tentato di determinarne la struttura tridimensionale utilizzando la tecnica della sostituzione molecolare utilizzando come template le strutture cristallografiche delle due proteine omologhe VirB8 di *A. tumefaciens e B. suis.* Tuttavia, i numerosi tentativi effettuati fino ad ora sono risultati vani. In futuro altre tecniche, quali MAD, MIR/SIR e SIR/SAR verranno esplorate al fine di determinarne la struttura.

Nel corso del secondo e terzo anno di Dottorato ho effettuato dei periodi di ricerca all'estero presso il Macromolecular Crystallography Group dell' European Synchrotron Radiation Facility (ESRF) con la supervisione del Dr. Laurent Terradot.
Durante tale periodi di ricerca mi sono occupato dell'identificazione, espressione e purificazione di alcune porzioni solubili della tossina CagA che verranno in seguito utilizzate per studi strutturali (Capitolo 6).
In particolare, utilizzando una tecnica innovativa denominata ESPRIT (Expression of Solubile Proteins by Random Incremental Truncation) è stato possibile costruire una libreria di frammenti genici della proteina CagA i quali si differenziano l'uno dall'altro per poche basi a partire dall'estremità 5' del gene. Questo approccio sistematico, supportato dell'utilizzo di sistemi robotizzati, ci ha permesso di individuare diversi domini C-terminale solubili della tossina CagA. Tali domini sono stati poi espressi in piccola scala in *E. coli* e purificati per mezzo di cromatografia di affinità. I vettori plasmidici codificanti i frammenti proteici purificabili sono stati sequenziati. In seguito ad accurati allineamenti ed analisi di sequenza, alcuni di questi domini sono stati selezionati per le successive prove di espressione e purificazione su larga scala.
Un altro domino solubile della tossina CagA è stato invece individuato sulla base di studi di interazione. Tale dominio centrale della proteina CagA, denominato $CagA^{392-733}$, è stato clonato in un opportuno vettore di espressione, prodotto in *E. coli*, purificato utilizzando tecniche cromatografiche e successivamente sottoposto a numerosi test di cristallizzazione, fino ad ora privi di successo (Capitolo 6).
Per quel che riguarda la proteina *Hp*YbgC, la sua struttura tridimensionale è stata risolta mediante diffrazione ai raggi X e sulla base di omologie di struttura con altre proteine note, è stato possibile identificare la funzione, ulteriormente confermata per mezzo di saggi enzimatici (Capitolo 7).
In particolare tale enzima si è rivelato essere attivo nei confronti di substrati a catena lunga che costituiscono inoltre la maggiore componente della membrana di *H. pylori*.
Poiché *Hp*YbgC, così come CagF, è stato dimostrato interagire con il dominio centrale della tossina CagA, studi di gel filtrazione analitica e dynamic light scattering cosi come prove di co-cristallizzazione del dominio della tossina CagA in complesso con le *Hp*YbgC e CagF sono state effettuate (Capitolo 6).

Chapter 1

General Introduction

Contents

1.1 *Helicobacter pylori*

The bacterium *Helicobacter pylori* colonizes the stomach of more than half of the world's population, with the highest rates in developing countries, making it one of the most successful bacterial pathogens (Rothenbacher and Brenner, 2003). The presence of microbes in the human stomach was firstly described by Bizzorero at the end of the XIX century (Bizzorero *et al*. 1893) and after ninety years, in 1983, *H. pylori* was isolated and characterised (Marshall and Warren, 1984; Nobel Prize in 2005), thereby becoming the focus of intense research activity. This small (3.5 μm x 0.5 μm), slow growing, microaerophilic, non spore-forming and spiral-shaped gram-negative rod bacteria [Fig. 1.1] is transmitted from human-to-human possibly by the fecal–oral or oral–oral route (Blaser, 1998). The circular genomes of three strains, *Helicobacter pylori* 26695, J99 and HPAG1, were completely sequenced using a random shotgun approach. The genome of *H. pylori* strain 26695 contains 1,667,867 base pairs. It is 24 kb larger than the J99 strain, which contains 1,643,831 base pairs, and 71 kb larger than HPAG1, which contains 1,596,366 base pairs (Tomb *et al*., 1997; Alm *et al*., 1999; Oh *et al*., 2006). All the genomes had similar average lengths of coding sequences, coding density and the bias of initiation codons. The DNA of *H. pylori* has a GC content of 35-38 %, which is similar to that *Campylobacter* species. However, comparison of 16S ribosomal RNA showed that *H. pylori* was different from *Campylobacter* but similar to *Wolinella succinogenes* which has a GC content of 42-49 %. *H. pylori* was finally placed into its own genus, *Helicobacter,* after the analysis of the ultra structure, the fatty acid composition and biochemical tests which proved different for *H. pylori* and *W. succinogenes* (Tomb *et al*., 1997; Mobley *et al*., 2001; Blaser *et al*., 2004). The extremely high recombination rate, in combination with a high sequence diversity suggested, a panmictic population

Figure 1.1. This scanning micrograph shows numerous *Helicobacter pylori* bacilli. (Adapted from the Meridian Bioscence Europe web site).

structure (Suerbaum *et al*., 1998 and 2007). Geographic subdivisions were described that correlate with the continent of origin: this bacterium probably has accompanied humans for tens of thousands of years during migrations (Kersulyte *et al*., 2000; Suerbaum *et al*., 2002). For these reasons, *H. pylori* was recognised as one of the most common and successful pathogens and was also hypothesized that *H. pylori* colonization could have provided benefits to its human carriers and hence provided a selective advantage during long periods of the human history (Suerbaum *et al*., 2002 and 2007). We now know that infection is associated with a spectrum of disease outcomes. Epidemiological studies suggested that *H. pylori* is mainly transmitted vertically within families and does not spread epidemically (Kivi *et al*., 2003). It is usually acquired in early childhood and accompanies its human host for decades in the absence of a targeted antimicrobial therapy (Suerbaum *et al*., 2002). While many *H. pylori* infected individuals are clinically asymptomatic, most will exhibit some degree of gastritis. Approximately 10% of the infected subjects will develop more severe gastric pathologies, like peptic ulcer disease and atrophic gastritis. Gastric adenocarcinoma and lymphoma of the mucosa-associated lymphoid tissue (MALT) develops in approximately 1% of infected individuals, a correlation that prompted the World Health Organization to declare *H. pylori* the first bacterial class 1 carcinogen in 1994 (Blaser, 1998; Peek *et al*., 2002; Suerbaum *et al*., 2002).

1.2 Resistance to acid

The human stomach is an unique ecological niche characterised by a very acid pH, a condition lethal for most microbes. *H. pylori* is not an acidophilus bacterium and is able to survive in the acid lumen only for a short period, sufficient to enter the highly viscous mucosa, reach the gastric epithelium, find nutrient and multiply (Suerbaum *et al*., 1999). To survive in the acid lumen of the stomach and colonise the gastroenteric tract, *H. pylori* dedicates a gene cluster (*ureAB and ureEFGHI*) to the biosynthesis of **Urea Amidohydrolase** (EC 3.5.1.5) also named **Urease**. It is a nickel-containing enzyme, which is able to hydrolyze the urea, available in the stomach, to generate ammonia and carbamate, which successively spontaneously decomposes to yield another molecule of ammonia and carbonic acid, increasing the pH (Weeks *et al*., 2000). The essential role of Urease as a virulence factor is shown by the fact that Urease-defective *H. pylori* mutants cannot colonise the stomach (Eaton *et al*., 1991). Moreover, Urease is one of the main antigens recognised by the human immune

response to *H. pylori,* even if the nature of this immune response after infection is not fully clear (Del Giudice *et al*., 2001). The amount of Urease produced by the bacterium varies with the culture conditions and may reach as much as 10 % of total bacterial protein (Montecucco *et al*., 2001). Urease is a dodecamer composed of two different subunits: six **UreA** (HP0073) (26.5 kDa) and six **UreB** (HP0072) (60.3 kDa), to form a huge complex whose molecular mass was estimated to be ~600 kDa (Ha *et al*., 2001). Transmission electron microscopy (TEM) indicated that the enzyme is arranged as a double ring of 13 nm diameter and a three-fold symmetry (Ha *et al.*, 2001; PDB accession codes: 1E9Z and 1E9Y) [Fig. 1.2]. It is a cytoplasmic enzyme but it was also found on the surface of the bacteria, due to lysis of some organisms (Phadnis *et al*., 1996; Marcus *et al*., 2001). It is still unclear whether the Urease reaches the bacterial external site by a process of active secretion and/or it is released after a genetically programmed cell-lysis, defined "altruistic autolysis", and consequently tightly absorbed into the envelope of still living bacteria (Dunn *et al*., 1998). It was proposed that external urease generates a cloud of ammonia around the bacteria that protects it against the acidity of the stomach, whereas the internal urease defends the bacterium against gastric acidity by increasing periplasmic pH and membrane potential and stimulating protein synthesis at acidic pH (Scott *et al*., 1998; Stingl *et al*., 2002). Urease activity in the bacterial cytoplasm is regulated by a proton-gated urea channel, **UreI** (HP0071) (Weeks *et al*., 2000). The urea channel UreI is a six helices transmembrane protein regulated positively by protons, opening at acidic pH values to allow more urea to enter into the cytoplasm for a maximal production of ammonia and carbamate, and closing at neutral pH to avoid over-alkalinixation, which is lethal to *H. pylori* (Clyne *et al*., 1995). UreE (HP0070), UreF (HP0069), UreG (HP0068) and UreH (HP0067) are instead accessory components which probably participate as chaperones during the Urease enzyme assembly and/or in the activation/inhibition of its catalysis. In addition to Urease, *H. pylori* also possesses other ammonia-producing enzymes, including the **Acylamide**

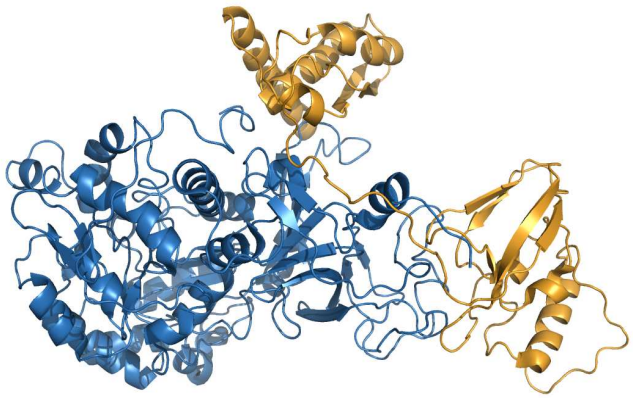


Figure 1.2. Ribbon representation of the monomeric unit of the *H. pylori* Urease. Each unit consists of the α-subunit (blue) and the β-subunit (orange). (Ha *et al*., 2001) [PDB accession codes 1E9Z and 1E9Y].

Amidohydrolase (EC 3.5.1.4) **AmiE** (HP0294), and the **Formamide Amidohydrolase** (EC 3.5.1.49) **AmiF** (HP1238) (Skouloubris *et al*., 1997, 2001; Hung *et al.*, 2007). In response to growth at mild acidic conditions, genes encoding these proteins are up-regulated (van Vliet *et al*., 2004; Bury-Mone *et al*., 2004). Both AmiE and AmiF are paralogue enzymes and belong to members of aliphatic amidase family, that hydrolyze short-chain amides to produce ammonia and the corresponding organic acid. In particular, AmiF belongs to the nitrilase superfamily and demonstrates a restricted substrate specificity in which its only known substrate formamide is hydrolyzed to produce formic acid and ammonia. The crystal structure of AmiF was recently determined (Hung *et al*., 2007; PDB accession codes 2DYU, 2DYV, 2E2K and 2E2L). The structure consists of a homohexamer in which each subunit has an αββα four-layer fold and reveals a Cys-Glu-Lys (CEK) catalytic triad (Hung *et al*., 2007) [Fig. 1.3]. Amidases were previously reported only in environmental bacteria and their presence in *H. pylori* underlines the importance of ammonia to the survival of this organism.

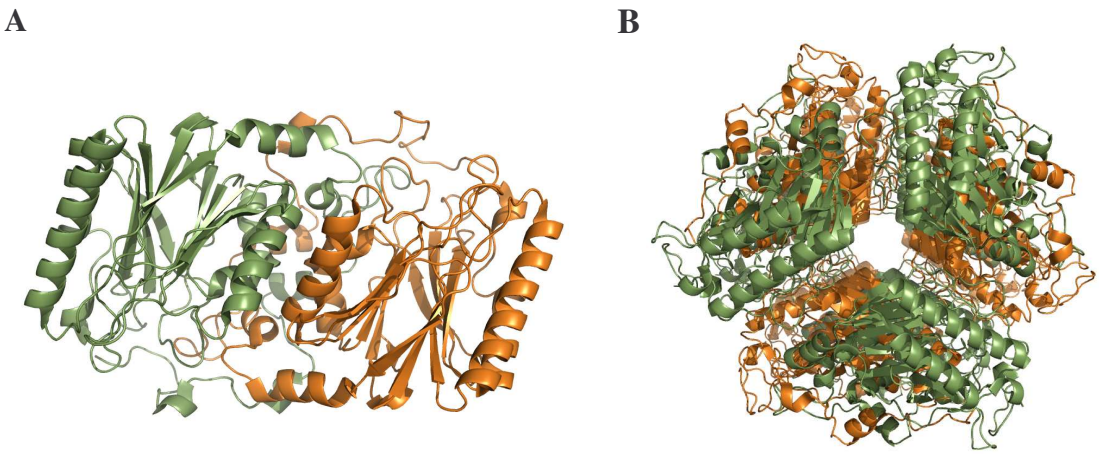


Figure 1.3. Structure of *H. pylori* AmiF. [A] Ribbon representation of the dimeric AmiF structure (ab) viewed down a 2-fold axis. The subunit "a" is depicted in olive-green and the subunit "b" is in orange. [B] Ribbon representation of the hexameric AmiF structure $(ab)_3$ viewed down the 3-fold axis. (Hung *et al*., 2007) [PDB accession codes 2DYU, 2DYV, 2E2K and 2E2L].

H. pylori also produces an **Arginase** (EC 3.5.3.1) encoded by the ***rocF*** (*hp1339*) gene (McGee *et al*., 1999, 2004). RocF is an enzyme which hydrolyzes L-arginine to L-ornithine and urea and is involved in acid resistance, but it is not considered essential for colonization of mice or for Urease activity (McGee *et al*., 1999). The major ammonia-producing pathway, central to the acid resistance of *H. pylori* (Urease, AmiE, AmiF and RocF), was shown to be regulated in response to low pH by the

ArsSR two-component system, which is essential for *H. pylori* colonization in a mouse infection model (Forsyth *et al*., 2002; Pflock *et al*., 2006a). The ArsSR system consists of the histidine kinase **ArsS** (HP0165), responsible for low pH sensing (Pflock *et al*., 2004) and that was recently shown to be essential for acid resistance (Loh & Cover, 2006) and an OmpR-like response regulator ArsR (HP0166), which is phosphorylated and activated. Inactivation of ArsS renders *H. pylori* unable to colonize mice (Panthel *et al*., 2003), while ArsR is essential for growth *in vitro* (Schar *et al*., 2005). Several other genes of the ArsSR two-component system encoding outer membrane proteins, Ni^{2+} storage proteins, detoxifying enzymes involved in the oxidative stress resistance and *H. pylori*-specific proteins of unknown function show a similar mode of regulation (Pflock *et al*., 2006; Josenhans *et al*., 2007).

The metal-dependent regulators **NikR** (HP1338) and **Fur** (HP1027) were also shown to play an important role in the transcriptional regulation of acid resistance. Analysis of the genome sequence indicated that *H. pylori* has a relatively limited capacity for gene regulation, and thus it is possible that the few regulatory proteins present regulate multiple responses and metabolic processes. One well-characterized regulatory protein of *H. pylori* is the Ferric Uptake Regulator (Fur), which controls intracellular iron homeostasis *via* concerted expression of iron-uptake and iron-storage genes (van Vliet *et al*., 2002; Delany *et al*., 2001a and 2001b). The transcription of Fur is repressed at low pH (Bijlsma *et al*., 2002; Bury-Mone *et al*., 2004), and this was shown to be mediated by NikR (van Vliet *et al*., 2004). The crystal structure of NikR from *H. pylori* was recently solved in a nickel-free conformation (apo-NikR) and in two different nickel-bound conformations (Ni1-NikR and Ni2-NikR) (Dian *et al*., 2006) [PDB accession codes: 2CA9, 2CAD and 2CAJ]. It contains two separate domains: a C-terminal tetrameric domain (TD) that coordinates nickel and a N-terminal dimeric DNA binding domain (DBD), that contains the ribbon-helix-helix (RHH) motif which is present in the transcriptional

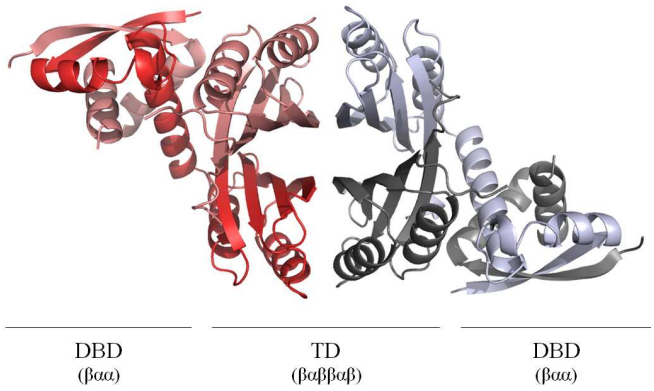


Figure 1.4. Structure of *H. pylori* apo-NikR. Ribbon representation of the tetrameric NikR structure parallel to the crystallographic 2-fold axis. The subunits are coloured in red, salmon, grey and cyan. (Dian *et al*., 2006) [PDB accession codes 2CA9, 2CAD and 2CAJ].

regulators (Dian *et al.*, 2006) [Fig. 1.4]. Unlike its homologues, the *H. pylori*'s NikR acts as a nickel-regulator that directly repress and, more importantly, also activates several genes (Contreras *et al.*, 2003; Dosanjh *et al.*, 2006). NikR appears to act as a global regulator at the centre of a complex cascade that enables the optimal use of the nickel ions under acid-stress conditions, thereby preventing the emergence of toxic nickel concentrations (Dosanjh *et al.*, 2006). Several studies have looked at the effect of acid on global gene expression by *H. pylori*. There were considerably little overlap in the results of these studies, most likely due to different growth conditions, different growth phases of the bacteria and different conditions used to provoke the acid shock in the bacteria. However, in three studies (Merrell *et al.*, 2003; Wen *et al.*, 2003; Bury-Mone *et al.*, 2004) it has been demonstrated that several genes encoding outer membrane proteins are downregulated under acidic conditions, suggesting that alteration of the outer membrane composition is an important aspect in the acid-adaptive response (Pflock *et al.*, 2006b).

1.3 Colonization of the gastric mucosa

Flagellar motility, including chemotaxis, is one of the most predominant colonization and virulence factors of *H. pylori* (Kavermann *et al.*, 2003). *H. pylori* is considered a "good swimmer": propelled by its **flagella** (four to eight arranged at one of its pole) and its **helicoidal-shape**, it is able to travel across the viscous gastric mucus film that covers and protects the gastric epithelial and to establish a persistent infection in the stomach. In this way *H. pylori* gets nutrients and it avoids to be discharged in the intestinal tract by peristalsis. The general organization of the physical structure of the *H. pylori* flagellum conforms to those of the model enteric organisms *E. coli* and *S. typhimurium* (Macnab, 1996). Its organization consists of a basal body, a hook and a flagellar filament. The basal body includes the rings, motor and switching proteins, and usually the flagellar export proteins. The filament, first studied by Geis and colleagues (Geis *et al.*, 1989), extends 3-5 μm from the cell. The flagella are invariably polar and 4-8 per cell. In *H. pylori* more than fifty putative proteins are predicted to be involved in expression, secretion, and assembly of this complex flagellar apparatus (Tomb *et al.*, 1997; Alm *et al.*, 1999). At least twenty of these proteins constitute the structural components of the flagellar basal body, hook, and filament. To date, only components of the filament and of the hook have been characterized in some detail. The filament is a copolymer of the flagellin subunits FlaA and FlaB (Haas *et al.*, 1993;

Leying *et al*., 1992; Suerbaum *et al*., 1993). The FlaA is the predominant subtype and FlaB the minor subtype, localized close to the basis of the flagellum (Kostrzynska *et al*., 1991). Both flagellins have a similar molecular mass (53 kDa) and share considerable amino acid homology (58 % identity), but the respective genes are located at different sites on the *H. pylori* chromosome (Suerbaum *et al*., 1993). Studies with isogenic mutants of either *flaA* or *flaB* revealed that both genes are necessary for full motility (Josenhans *et al*., 1995) and for the establishing a persistent infection in the piglet model (Eaton *et al*., 1996). The *flgE* gene encodes the structural protein of the hook, and the *fliD* gene encodes a hook-associated protein, which is localized at the tip of the flagellar filament and promotes the incorporation of the flagellin monomers into the growing flagellar filament. Mutants in *flgE* are nonmotile and aflagellate (O'Toole *et al*., 1994), and mutants in *fliD* produce truncated flagella and are severely impaired in motility and in their ability to colonize the gastric mucosae of mice (Kim *et al*., 1999). Several other genes are involved in flagella biogenesis, assembly of the movement apparatus and chemotaxis system, but little is known about them (O'Toole *et al*., 2000). In addition, *H. pylori* was shown to possess the enzymatic ability to disrupt the oligomeric strucuture of mucin (Windle *et al*., 2000), which may assist the pathogen to move freely in the mucus layer. The movement of *H. pylori* towards the stomach mucosa is guided by chemotactic factors, which include urea and bicarbonate ions (Yoshiyama *et al*., 1999).

1.4 Interaction of *H. pylori* with gastric epithelial cells

H. pylori adheres strongly to human gastric epithelial cells using fucosylated glycoproteins and sialylated glycolipids as cellular receptors (Mahdavi *et al*., 2002; Boren *et al*., 1993). This adhesion is probably mediated by several proteins and glycolipids which play an essential role in the development of the disease (Hessey *et al*., 1990).

The best-characterized *H. pylori* adhesin molecules that have been described to date are two outer membrane proteins (omp): BabA and SabA. **BabA** (HP1243) is a 75 kDa surface exposed protein, which mediates strong binding of the organism to the fucosylated Lewis antigens Le^b and H-1. The latter are abundantly expressed on the surface of gastric epithelial cells and in gastric mucus of people with blood groups ABO (Ilver *et al*., 1998; Linden *et al*., 2002). **SabA** (HP0725) mediates a weaker binding of the organism to sialyated Le^x and Le^a antigens structures that are

upregulated during inflammation (Mahdavi *et al*., 2002). The BabA and SabA sequences show strong allelic variation, and different alleles encode proteins with different binding modes. They might help *H. pylori* to adapt its adhesive properties to different conditions as well as to vary the expression of these adhesion molecules by multiple mechanisms. For example, Le^b-non-binding strains usually possess silent *babA* gene sequences, which can be activated by recombination into the *babB* locus. The BabB/A chimeric adhesin has been shown to be expressed at lower levels and subjected to phase variation, conferring a high versatility to *H. pylori* phenotype into the habitat of the human stomach (Backstrom *et al*., 2004). Amino acid polymorphisms have been detected among BabA proteins expressed by different strains (Hennig *et al*., 2004), and BabA expression has been shown to be modulated by the types of receptors available in different populations (Aspholm-Hurtig *et al*., 2004). Heterogeneity among *H. pylori* strains in expression of the BabA protein may be a factor that contributes to different clinical outcomes among *H. pylori*-infected individuals. SabA has been shown to mediate the binding of the organism to sialylated glycoconjugates expressed in inflamed gastric tissues (Mahdavi *et al*., 2002), the extracellular matrix protein laminin (Walz *et al*., 2005), neutrophils (Unemo *et al*., 2005) and erythrocytes (Aspholm *et al*., 2006). SabA expression can also be switched on or off and the SabA-positive status has been inversely related to the ability of the stomach to secrete acid (Yamaoka *et al*., 2006). Although adherence facilitates better access to nutrients and delivery of effector molecules, the ability of the organism to modulate adherence may aid escape from sites where host defense responses are the most vigorous.

Other adhesins of *H. pylori* identified to date include the **Outer membrane proteins AlpA/AlpB** (HP0912/HP0913) (Odenbreit *et al*., 1999) and **HopZ** (HP0009) (Peck *et al*., 1999), which are known to mediate the attachment to epithelial cells by defining a macromolecular complex on the bacterial surface with other outer membrane proteins. Recently, **OipA** (HP0638), a phase-variable outer membrane protein associated with more virulent strains of *H. pylori*, was identified as a potential colonization factor for the organism (Yamaoka *et al.*, 2000; Dossumbekova *et al*., 2006). An OipA-positive status was associated with the presence of duodenal ulceration and gastric cancer, high *H. pylori* density and severe neutrophil infiltration (Yamaoka *et al*., 2006). Sequencing of OipA in *H. pylori* strains revealed that the OipA 'on' genotype is linked to bacterial virulence determinants, such as functional *vacA*, *babA* and, most strongly, *cagA* genotypes (Dossumbekova *et al*., 2006). OipA mutagenesis resulted in reduced bacterial adherence to gastric epithelial cells but did not alter IL-8 secretion *in vitro*

(Dossumbekova *et al*., 2006). OipA may be linked to gastroduodenal diseases owing to its association with other virulence factors or increased bacterial adherence and colonization.
Surface-located heat shock proteins in *H. pylori* were also postulated to play a role in adhesion to gastric cells (Yamaguchi *et al*., 1996 and 1998).

1.5 Virulence factors

***cag*-PAI** and **CagA** (HP0547): *Helicobacter pylori* strains isolated are classified as cag^+ and cag^-, depending on the presence of the *cag* Pathogenicity Island (*cag*-PAI), a chromosome region of about 40 kb containing approximately 30 genes, which plays a central role in *H. pylori* pathogenesis (Censini *et al*., 1996) [Fig. 1.5]. The *cag*-PAI is characterised by a different nucleotide composition with respect to the bacterial genome and is flanked by transposable elements allowing to think that it has been probably acquired by a horizontal transfer process (Censini *et al*., 1996; Tomb *et al*., 1997; Hacker *et al.*, 2000). The role of the *cag*-PAI in pathogenesis is illustrated by the fact that in Mongolian gerbils cag^- bacteria cause only mild asymptomatic inflammation of the stomach, whereas cag^+ organisms are associated with more severe diseases (such as ulcers and adenocarcinomas) (Ogura *et al*., 2000). These data are consistent with many epidemiological studies reporting that, in humans, severe gastric diseases are always associated with infection by cag^+ strains (Parsonnet *et al*., 1997; Witherell *et al*., 1997; Ley *et al*., 2001).
Some of the *cag*-PAI genes encode proteins with homology to components of the *Agrobacterium tumefaciens* VirB/D4 system, which is considered the prototype of bacterial type IV secretion systems (TFSSs) (Christie *et al*., 2005). T4SSs are multi-subunit machines that can deliver proteins and/or DNA into host cells and thereby influence host cell functions (Christie *et al*., 2005). In particular, the *cag* TFSS translocates the cytotoxin-associated gene A (CagA) effector protein into host epithelial cells (Odenbreit *et al*., 2000). After entering the host cell, CagA (128-145 kDa) becomes phosphorylated by cellular kinases and binds to several host target proteins. These interactions, some of which are phosphorylation-dependent and some are phosphorylation-independent, induce multiple cellular events that contribute to cellular responses. They include the morphogenetic changes that are characteristic of cell infection with cag^+ *H. pylori* strains, and may ultimately lead to malignant transformation (Tummuru *et al*., 1993; Covacci *et al*., 1993; Hatakeyama *et al*. 2005).

CagA is therefore termed a bacterial oncoprotein (Hatakeyama, 2004 and 2006). In addition, the TFSS mediated contact between bacteria and epithelial cells was shown to lead to the delivery of peptidoglycan fragments into the host cell, which are recognized by the intracellular pattern receptor, NOD1, leading to the activation of pro-inflammatory signalling pathways and increased interleukin-8 (IL-8) secretion (Crabtree *et al*., 1995; Viala *et al*., 2004).

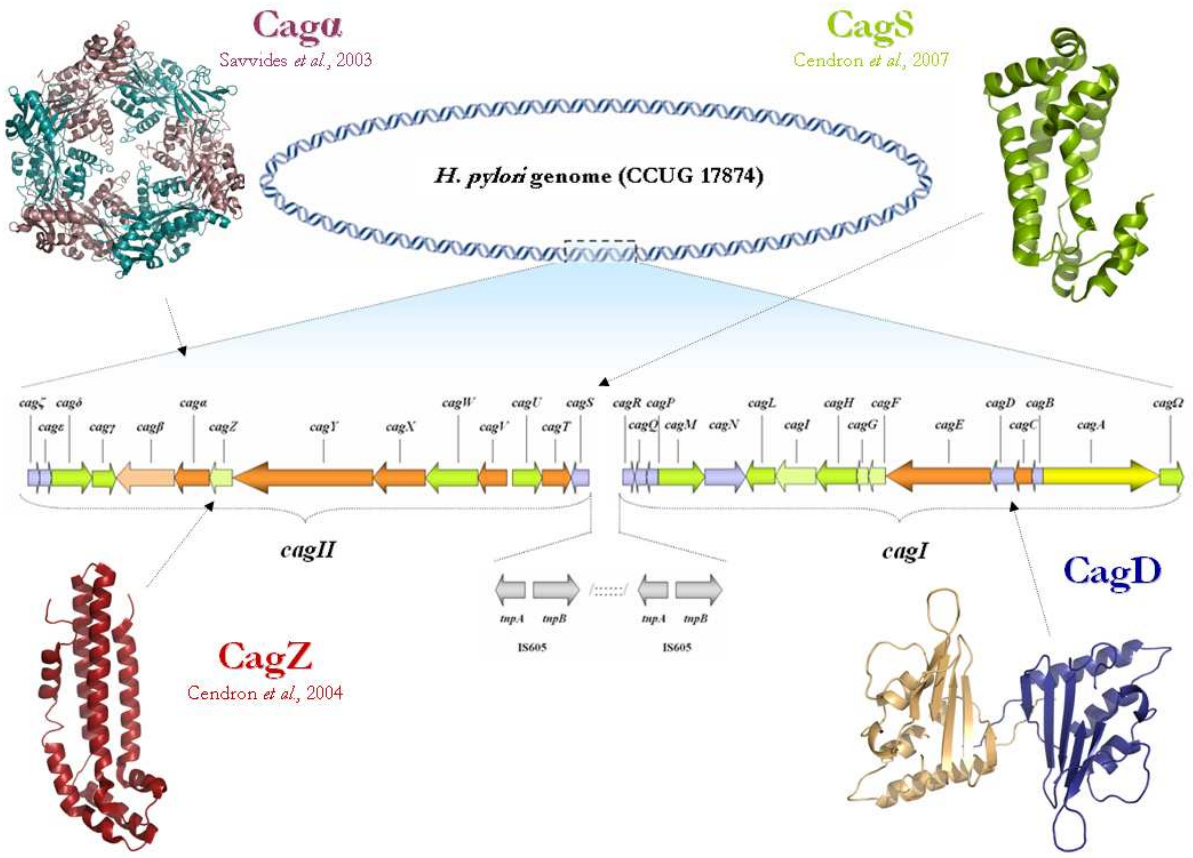


Figure 1.5. Arrangement of *cag*-PAI genes in *H. pylori* CCUG 17874 strain and ribbon representations of Cag protein structures solved till now: CagD (Chapter 4), CagS (Cendron *et al*., 2007 and Chapter 5; PDB accession code: 2G3V), CagZ (Cendron *et al*., 2004; PDB accession code: 1S2X) and Cagα (Yeo *et al*., 2000; Savvides *et al*., 2003; PDB accession codes: 1NLY, 1OPX, 1G6O, 1NLZ).

Since the aim of this project is to determine the three-dimensional structure of some of the proteins encoded by the *cag*-PAI, its role will be extensive discussed in the following chapters.

VacA (HP0887): The Vacuolating cytotoxin A (VacA) is a secreted protein of *H. pylori* that induces the formation of large cytoplasmic vacuoles in gastric cultured cell lines (Leunk *et al*., 1988; Cover *et al*., 1992) The *vacA* is a polymorphic gene and both

active and inactive forms of the toxin exist (Telford *et al*., 1994). VacA is not only an important antigen in the human response to *H. pylori* (Del Giudice *et al*., 2001), but also an important virulence factor which confers a strong competitive advantage to *H. pylori* wild-type strains with respect to VacA-defective mutants in the colonisation of the stomach (Salama *et al*., 2001). Strains of *H. pylori* that express active forms of the toxin are in fact associated with more severe cases of disease (Atherton *et al*., 1995; van Doorn *et al*., 1998). The mature toxin has a molecular weight of 88 kDa, and two domains, p33 and p55. The latter are cleaved after a two-step bacteria secretion process involving a N-terminal signal peptide of 33 amino acids, which directs secretion from the cytoplasm to the periplasm, and a large C-terminal domain, around 45 kDa, which seems to act like an autotransporter directing export across the outer membrane (Telford *et al*., 1994; Montecucco *et al*., 2001). After the cleavage, the two domains, p33 and p55, remain however attached and the p55 domain mediates host cell binding (Reyrat *et al*., 1999), whereas the p33 domain, together with the N-terminus of p55, causes vacuolation when expressed in the cytoplasm of cells (Telford *et al*., 1994; de Bernard *et al*., 1998). Up to a few months ago, the only three-dimensional structure information about VacA was available through cryo-

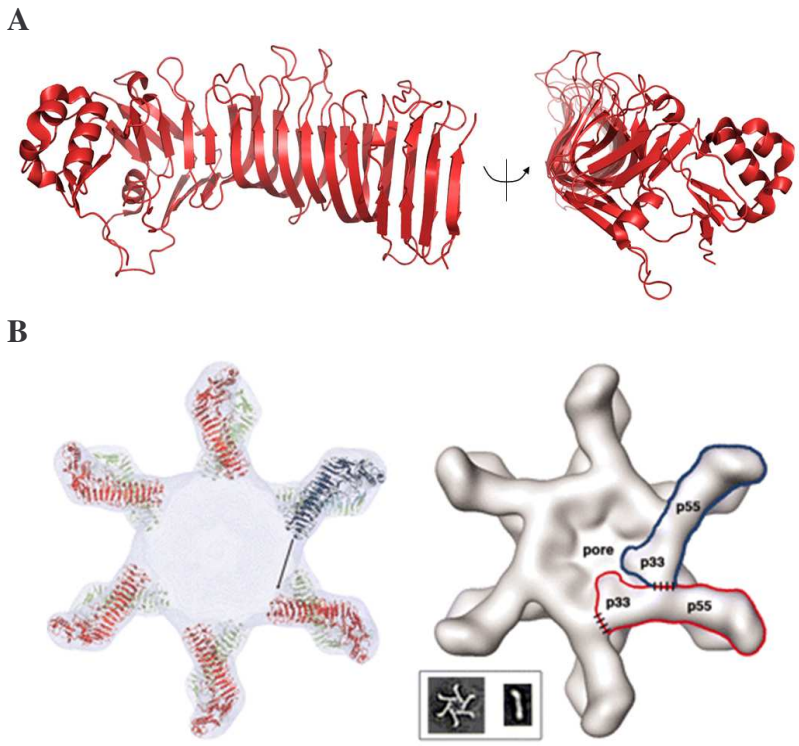


Figure 1.6. Crystal structure of VacA p55 domain. A) The VacA p55 fragment adopts a β-helix structure that is composed of three parallel β-sheets connected by loops of varying length and structure. The C-terminal domain has a mixture of α/β secondary structure elements, not previously observed in an autotransporter passenger domain structure. On the right, the molecule is rotated by 90°. B) Docking the p55 crystal structure into a 19-Å cryo-EM map of the VacA dodecamer results in a model for oligomerization. On the lrft, twelve p55 subunits are shown docked into a 19-Å cryo-EM map of a VacA dodecamer. An arrow is shown to indicate the space that the blue molecule will occupy if p33 extends the β-helix structure of p55. On the right the proposed oligomerization model in which p33 interacts with the N-terminal portion of p55 from the neighboring subunit. Regions of contact between p33 and p55 are depicted with dashed lines (Adapted from Gangwer *et al*., 2007) [PDB accession code 2QV3].

electron microscopy (cryo-EM) studies, which revealed that the secreted toxin has a strong tendence to oligomerize into rosettes (Lupetti *et al*., 1996). The structure resembles "flowers" in which a central ring is surrounded by peripheral "petals" [Fig. 1.6 B]. Very recently, the crystal structure of p55 VacA domain was determined (Gangwer *et al*., 2007; PDB accession code 2QV3) [Fig. 1.6 A]. The VacA p55 domain has an important role in mediating VacA binding to the host cell. The structure is predominantly a right-handed parallel β-helix, a feature that is characteristic of autotransporter passenger domain, but unique among known bacterial protein toxins (Gangwer *et al*., 2007).

The oligomeric form of VacA has little vacuolating activity, but it is activated by exposure to acidic conditions (Cover *et al*., 1997). Low pH dissociates the VacA oligomer into monomers, enabling the toxin to bind the apical portion of epithelial cells and to insert into lipid membranes, forming a hexameric anion-specific channel of low conductance (Cover *et al*., 1997; Szabo *et al*., 1999). A wide range of cellular effects have been attributed to VacA and it has been shown to act on both epithelial and immune cells (Cover *et al*., 2005). Such activities include the induction of apoptosis (Galmiche *et al*., 2000), alteration of the process of antigen presentation (Molinari *et al*., 1998), inhibition of T cell activation and proliferation (Boncristiano *et al*., 2003; Gebert *et al*., 2003; Sundrud *et al*., 2004). However, the physiological relevance and role in the infectious process of some of the effects of VacA that have been demonstrated *in vitro* remain to be clarified *in vivo* (Lu *et al*., 2005b; Schmees *et al*., 2006).

HP-NAP (HP0243): *H. pylori* neutrophil-activating protein (HP-NAP) is a 150 kDa protein oligomer composed of identical 15 kDa subunits. It promotes the adhesion of human neutrophils to the endothelial cells and stimulates the production of reactive Oxygen Radicals (ROIs) (Yoshida *et al*., 1993; Evans *et al*., 1995). Purified HP-NAP, free of LPS, is a highly immunogenic protein, both in mice and humans, responsible for the attraction of neutrophils and monocytes by chemotaxis at the site of *H. pylori* infection (Yoshida *et al*., 1993; Satin *et al*., 2000). It binds to a specific receptor and seems to act through a cascade of intracellular activation events, which are completely prevented by pertussis toxin: it includes the increase of cytosolic Ca^{2+} and the phosphorylation of proteins, leading to the assembly of functional NADPH oxidase on the neutrophil plasma membrane (Satin *et al.*, 2000; Montecucco *et al*., 2001). Moreover, HP-NAP shows similarities to some bacterial DNA binding proteins (Dps). These proteins are highly conserved among geographically distinct strains and the few

non-conservative differences within the amino acid sequence do not affect their whole structure. Surprisingly, HP-NAP does not bind to DNA fragments like the other Dps proteins, but it is able to store up to 500 atoms of iron (Tonello *et al.*, 1999). Its atomic structure, which has been determined in our laboratory (Zanotti *et al.*, 2002; PDB accession code: 1JI4), reveals a four helix bundle protein that oligomerizes to form a dodecamer with a central internal iron-containing negative cavity. Since iron is not an essential nutrient for *H. pylori* growth and HP-NAP synthesis is not regulated by iron concentration or influenced by other metals presence, several researchers have hypothesized that HP-NAP was originally an iron-binding or iron-regulated protein, which has evolved to function as a leukocyte activator able to induce mucosal damage, through neutrophils activation, and an easy release of nutrients for the bacterial survival and colonization (Montecucco *et al.*, 2001).

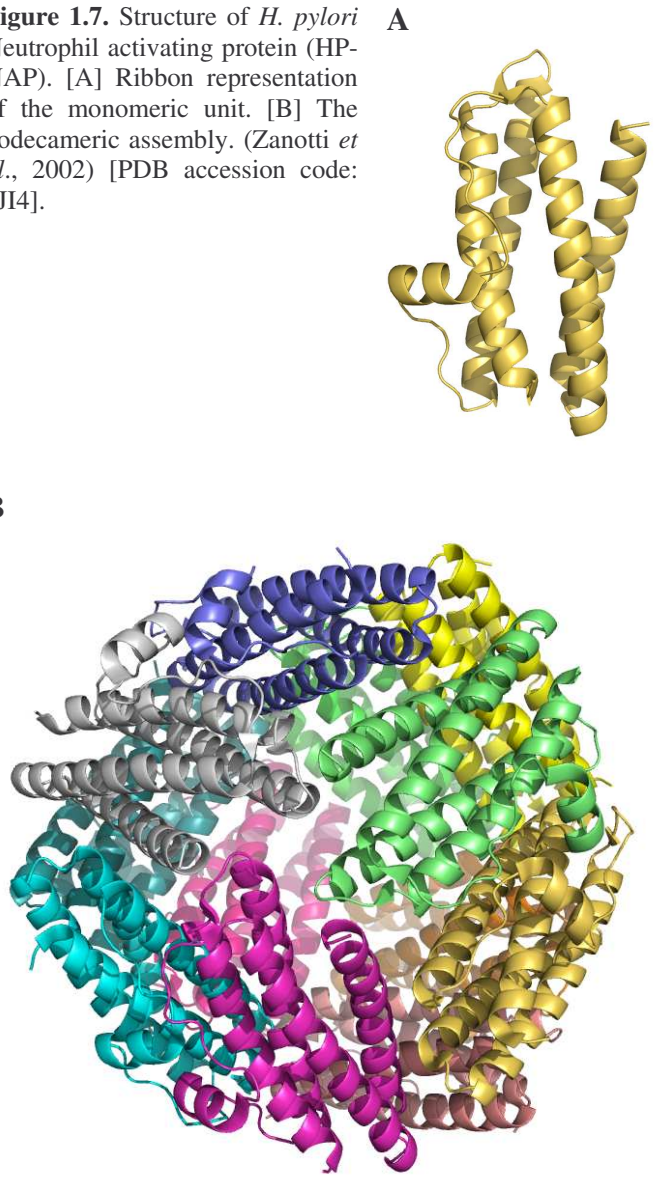


Figure 1.7. Structure of *H. pylori* Neutrophil activating protein (HP-NAP). [A] Ribbon representation of the monomeric unit. [B] The dodecameric assembly. (Zanotti *et al.*, 2002) [PDB accession code: 1JI4].

Chapter 2

Production and characterization of Cag proteins from the Pathogenicity Island of *Helicobacter pylori*

Contents

This chapter has been adapted from:
Alessandro Angelini [1,2,*], Laura Cendron [1,2,*], Anke Seydel [1,2,*], Nicola Barison [1,2], Roberto Battistutta [1,2] and Giuseppe Zanotti [1,2]; Production and characterization of Cag proteins from the pathogenicity island of *Helicobacter pylori*. (Manuscript in preparation).

[1)] Department of Chemistry, University of Padua, Institute of Biomolecular Chemistry, CNR Padua, Via Marzolo 1, 35131 Padua, Italy,
[2)] Venetian Institute of Molecular Medicine (VIMM), Via Orus 2, 35129 Padua, Italy,
[*)].These authors equally contributed to this work.

2.1 Introduction

Many of the defined *Helicobacter pylori* virulence factors, including flagella, the acid neutralizing enzyme urease, the neutrophil activating protein HP-NAP and the vacuolating cytotoxin VacA, are found in all strains (Monack *et al.*, 2004). On the contrary, the *cag* Pathogenicity Island (*cag*-PAI) is present only in Type I strains, isolated from patients with severe diseases (Censini *et al*., 1996). The *H. pylori* Pathogenicity Island was originally named *cag* (*cytotoxin-associated gene*) since it was thought to be associated with the expression of VacA. It was later shown that both factors, VacA and *cag*-PAI, are independent of each other, even though *cag*-negative strains often do not express VacA (Tummuru *et al.*, 1994). At the genetic level, the *cag*-PAI is a 40-kb chromosomal region, which was horizontally transferred from an unknown source, possibly about one million years ago, into the *H. pylori* chromosome. It is flanked by a 31-bp direct repeat, which contains the recombination site and corresponds to the last nucleotides of the glutamate racemase (*glr*) gene (Censini *et al*., 1996 Akopyants *et al*., 1998) (Fig. 2.1). Different numbers of insertion sequences are associated with the *cag*-PAI, depending on the strain. Often, however, the *cag*-PAI is split into two sub-regions, *cag*I and *cag*II, interrupted by two IS605 insertion elements (Fig. 2.1). As a consequence, strains with many insertions often resemble, with respect to virulence attributes, type II strains (less virulent) more than type I (more virulent, *cag*-PAI-containing) strains (Jiang *et al.*, 1996). In some cases, the *cag*-PAI can even be lost completely, due to DNA transfer of an empty site from a *cag*-deficient strain into a type I strain and subsequent homologous recombination (Kersulyte *et al*., 1999). Following the discovery of *H. pylori*, it was noted that this pathogen has an extraordinary genetic heterogeneity generated by mutation and/or recombination between different strains due to natural transformation, conjugation, or phage transfer (Suerbaum *et al*., 2007). It has also been shown that individuals can be colonized by multiple strains (Taylor *et al*., 1995), and strains have been shown to change during chronic infection (Kersulyte *et al*., 1999; Kuipers *et al*., 2000). The highly plastic region *cag*-PAI represents a perfect paradigm in the *H. pylori* genome. Not only strains differ significantly in the number of *cag*-PAI genes they carry, but the island can be lost, either completely or partially, during chronic infections of humans (Bjorkholm *et al*., 2001; Kraft *et al*., 2006) or experimentally infected animals (Sozzi *et al*., 2001). Many patients have been shown to simultaneously carry *cag*-positive and *cag*-negative bacteria (Figura *et al*., 1998). It has therefore been proposed that a dynamic population of *cag*-positive and *cag*-negative strains can exist in one human

individual that can expand and contract depending on the physiological and immune status of the host, and on the requirements of different niches within the stomach (Covacci *et al*., 1998). Since the first description of the *cag*-PAI (CCUG 17874) in 1996 (Censini *et al*., 1996), the full *cag*-PAI has been sequenced in three other complete genome sequences (*H. pylori* 26695, J99 and HPAG1).

Depending on the clinical strain, there are from 27 to 30 open reading frames predicted within the *cag* region (Fig. 2.1). One of these open reading frames encodes the immunodominant antigen CagA, which is localized to the 3′ end of the island (Covacci *et al*., 1993; Tummuru *et al*., 1992). CagA was identified as the first protein of the *cag*-PAI and appeared to be a major virulence factor (Parsonnet *et al*., 1997). Moreover, CagA is so far the only *H. pylori* macromolecule known to be translocated into the gastric epithelial cells by the Type IV Secretion System (TFSS) codified by *cag*-PAI (Odenbreit *et al*., 2000). Inside the host cell, CagA toxin induces multiple changes which cause the development of ulcers and cancer (Bourzac *et al.*, 2005). Biopsies from patients with severe gastric diseases including chronic active gastritis, peptic ulcer disease, mucosa-associated lymphoid tissue lymphoma, and gastric cancer contain the *cag*A gene in more than 90% of all the cases, establishing a direct correlation of the presence of the *cag*A gene with diseases (Blaser, 1998; Covacci *et al.*, 1997 and 1998).

Moreover, infection of cultured gastric epithelial cells with *cag*-PAI positive strains results in interleukin-8 (IL-8) induction and a dramatic cellular elongation. The discovery of mutations that can block CagA delivery but not IL-8 induction has led to the hypothesis that a second secreted effector molecule could be responsible for IL-8 induction. Recent experiments have implicated peptidoglycan, acting through the cytoplasmic receptor Nod1, as the secreted factor that induces IL-8 (Viala *et al*., 2004), while others have hypothesized that in some strain backgrounds CagA itself is responsible for some IL-8 induction (Brandt *et al*., 2005).

Besides CagA, the *cag* Pathogenicity Island (*cag*-PAI) of *H. pylori* encodes proteins with homologies to structural and functional components of TFSS of other gram-negative bacteria (Akopyants *et al*., 1998; Censini *et al*., 1996; Fischer *et al*., 2001) (Fig. 2.1). TFSS are multicomponent membrane-spanning transport systems ancestrally related to the conjugation system and can be divided according to their function into three categories: DNA transfer, DNA uptake and release, and effector translocation. The TFSS of *H. pylori* belongs to the last group. This group is shared by several prominent plant and animal pathogens that deliver effector molecules (such as protein-coupled DNA or proteins) to their target eukaryotic cell *via* what is thought to

be a needle-like surface organelle. Other bacteria that codes for a TFSS include *Agrobacterium tumefaciens*, *Brucella suis*, *Bartonella henselae* and *Legionella pneumophila* (Yeo *et al.*, 2004; Christie *et al.*, 2005).

Of the 30 genes carried by the *cag*-PAI, 18 genes are required for translocation of CagA into gastric epithelial cells (Fischer *et al*., 2001; Bourzac *et al*., 2005) and 9 of them have weak homology to genes that encode components of the VirB/D4 type IV secretion system of *Agrobacterium tumefaciens*, the best-studied model of the TFSS and that has became the standard reference of comparison for effector translocation TFSS in gram-negative bacteria. The remaining genes are unique to the *H. pylori cag*-PAI, and very little is known about their expression and function. Previous studies have detected the expression *in vivo* of eight *cag*-PAI proteins (CagA, CagD, CagM, CagS, CagX, CagZ, Cagα and Cagδ) (Jungblut *et al*., 2000; Backert *et al*., 2005; Busler *et al*., 2006; Smith *et al*., 2007). Three of them (CagY, CagX and CagT) were also detected by electron microscopy and reported to be associated with the TFSS structure (Rohde *et al*., 2003; Tanaka *et al*., 2003).

Deletion studies have been an important first step in understanding the functions of individual members of the *cag*-PAI, but many questions about the functions of individual *cag*-PAI genes remain unanswered. The *cag*-PAI genes whose deletion does not alter CagA delivery and IL-8 secretion in tissue cultured cells may encode delivered effector proteins with unknown activities in infection. Other pathogens that employ a T4SS in their infection process, such as *A. tumefaciens*, *L. pneumophila*, and *B. henselae*, deliver multiple effector proteins into host cells (Luo *et al*., 2004; Schulein *et al*., 2005; Vergunst *et al*., 2005). For instance, eight secreted effectors of *L. pneumophila* have been identified, but only one of them was found to be required for growth of this pathogen within infected cells, suggesting that the effectors have more subtle or redundant functions in infection (Luo and Isberg, 2004). In *B. henselae* seven contiguous genes located within a TFSS-encoding PAI were predicted to produce proteins with a common C-terminal domain that mediated their secretion into host cells, which was validated with the Cre/lox system. Interestingly, the combined set of effectors is required for host cell cytoskeleton changes and one of these effectors is phosphorylated within the host cells, reminiscent of *H. pylori* CagA (Schulein *et al.*, 2005). Moreover, mutations that resulted in intermediate and variable levels of CagA delivery and IL-8 induction are promising candidates for proteins important in T4SS assembly and function. These genes may encode accessory proteins required for substrate translocation (such as chaperones), may encode additional substrates themselves or simply encode proteins with redundant functions in TFSS.

To date, crystal structures of some TFSS components have been published, including homologues of hexameric NTPases VirB11 (Cagα from *H. pylori*; Yeo *et al*., 2000; Savvides *et al*., 2003) and VirD4 (TrwB from *E. coli* R388; Gomis-Ruth *et al*., 2001), the VirB5 protein (TraC from pKM101 plasmid; Yeo *et al*., 2003), VirB8 (form *B. suis* and *A. tumefaciens*; Terradot *et al.*, 2005, Bailey *et al*., 2006), and VirB10 (ComB10 from *H. pylori*; Terradot *et al.*, 2005). The NMR structure of a complex between a domain of the VirB9 homologue TraO bound to a VirB7-like domain of TraN from pKM101 plasmid have been also recently described (Bayliss *et al.*, 2007). The model of the hexameric structure for *A. tumefaciens* VirB4 has been recently proposed (Middleton *et al*., 2005). Till now, apart the ATPase Cagα VirB11 homologue, high-resolution structure data of Cag proteins are available only for CagD, CagS and CagZ proteins (Cendron *et al*., 2004 and 2007).

The goal of the study is to determine most of the Cag protein structures by X-ray crystallography to elucidate the role of *cag*-PAI in the *H. pylori* pathogenisis and also furnish valuable information for vaccine production and provide insights into the mechanism of bacterium pathogenesis. In this review we will describe our strategy to produce recombinant proteins in *Escherichia coli* encoded by *cag*-PAI for the determination of protein structures by X-ray crystallography. Nineteen open reading frames were chosen for expression in *E. coli*, initially selected on the base of a series of predicted properties, in particular the secondary structure. The major obstacle turned out to be poor expression and low solubility. Of the nine soluble proteins, four were crystallized and three structures have been determined.

Figure 2.1. Graphical representation of *cag*-PAI from *H. pylori* strain CCUG 17874 showing relative gene size and orientation. Most of the *cag* genes are probably involved in the assembly of the type IV secretion system that translocates the protein CagA into the gastric epithelial cells. Seven genes (marked in orange) show similarity to components of the VirB/D4 type IV secretion system of the plant pathogen *A. tumefaciens*. Proteins encoded by the island are involved in two major processes, the induction of interleukin-8 (IL-8) production by gastric epithelial cells and the translocation of CagA from the bacterium into host cells. All genes depicted by orange and green arrows are essential for CagA translocation, whereas lighter shades of red and green indicate genes that are not involved in this process but are however important for the IL-8 secretion. The purple arrows indicate genes that are not required for both translocation of CagA and IL-8 secretion. Based on mutagenesis study (Fischer *et al*., 2001) the gene products essential for CagA delivery (Cag^{PY}) and IL-8 induction are indicated with "+", whereas the non essential gene are indicated with "-". The protein which have been detected *in vivo* by 2D-gel electrophoresis and Mass Spectrometry (MS) analysis (Busler *et al*., 2006; Smith *et al*., 2007) are indicated with "x". Interactions among *cag*-PAI encoded proteins have been detected using a yeast two-hybrid system (Busler *et al*., 2006) and electron microscopy (Rohde *et al*., 2003; Tanaka *et al*., 2003).

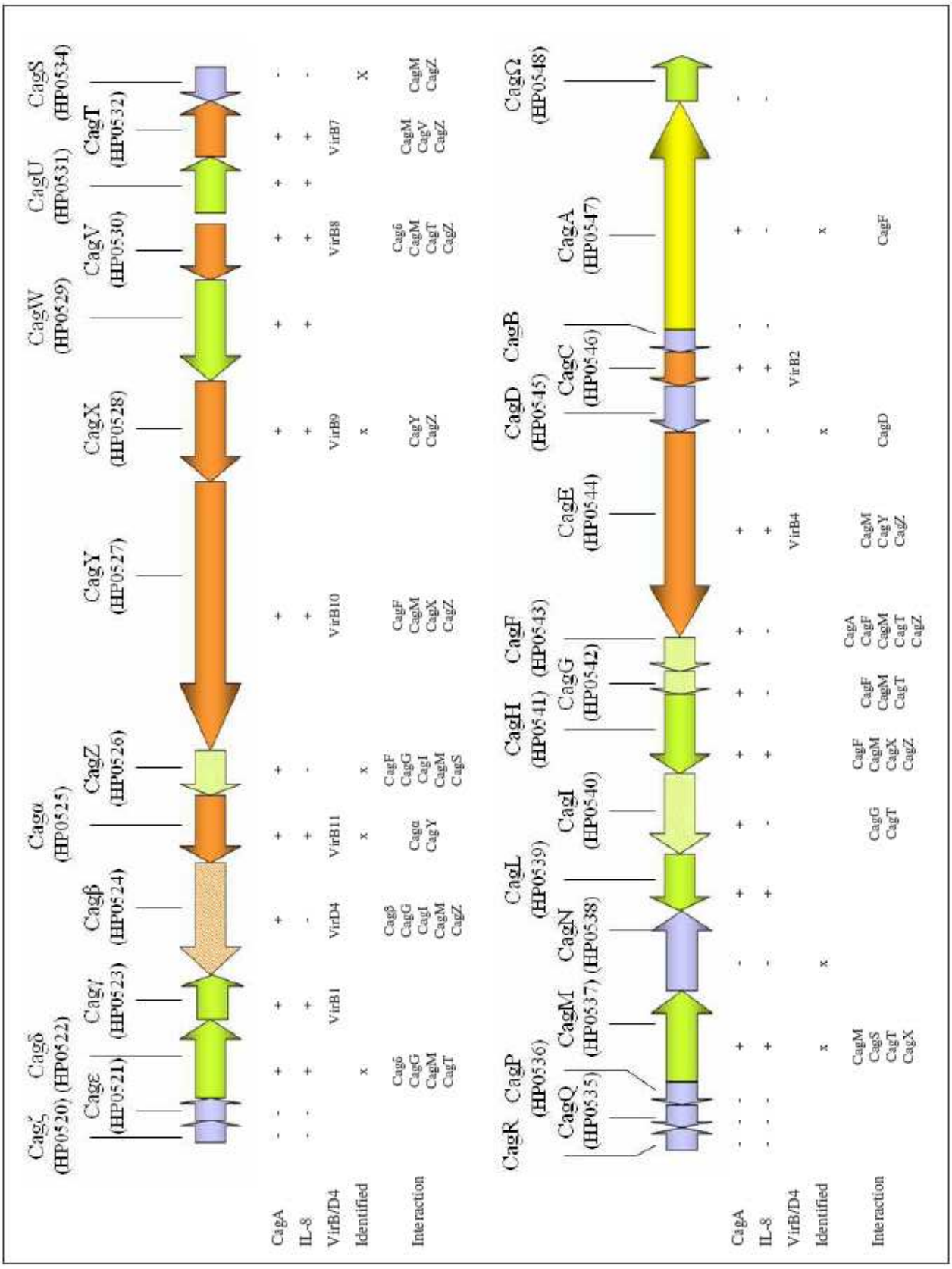


Figure 2.1. Graphical representation of *cag*-PAI from *H. pylori* CCUG17874 strain.

2.2 Interactions among Cag proteins

Based on what is known about the TFSS in other bacterial species (Cascales *et al*., 2003; Yeo *et al*., 2004), it may be presumed that numerous protein-protein interactions occur among *H. pylori* Cag proteins in order to form a type IV secretion apparatus. The protein-protein interaction map of *H. pylori* was the first reported for a prokaryotic organism (Rain *et al*., 2001). Surprisingly, no interactions between Cag proteins were detected, while several others were found to connect them to members of other cell pathways or proteins with unknown functions. Such examples are the interaction between Cagα, a VirB11 homologue, and HP1451, CagY, a VirB10 homologue, with HP0149, CagM and HP0261, Cagδ and HP1414, CagI and HP1489 and between CagA with HP0496 (Rain *et al*., 2001). The role of these open reading frames (HP1489, HP1451, HP1414, HP0496, HP0261 and HP0149) is often unknown and they are not considered component of the *H. pylori* TFSS. Some of these interactions could be not confirmed *in vitro* on isolated proteins, for example CagY-HP0149, Cagδ-HP1414 and CagI-HP1489, owing to the toxicity or insolubility of one or both partners (Terradot *et al*., 2004), whereas other interactions were confirmed by pull-down assays, for example the interaction between CagM and HP0261 and CagA with HP0496 (Terradot *et al*., 2001; Angelini *et al*., Chapter 7). Moreover, very recently, one interaction (Cagα-HP1451) was structurally characterized (Hare *et al*., 2007). The biological significance of those interactions is at the moment unclear, but these results suggest that the assembly and function of the Cag type secretion system of *H. pylori* might depend or require other previously unidentified proteins. This makes more difficult to understand the complex functioning of the *H. pylori's* TFSS.

A very recent study has identified some interactions among Cag proteins using a yeast two-hybrid system and biochemical analyses (Busler *et al*., 2006). Several of the interactions between *H. pylori* Cag proteins correspond to homologous interaction between components of the *A. tumefaciens* VirB/D4 TFSS. For example, homotypic interactions were detected for both *H. pylori* Cagβ and Cagα, similar to what has been observed for the *A. tumefaciens* homologues (VirD4 and VirB11 respectively) (Yeo *et al.*, 2004). The C-terminal portion of *H. pylori* CagY has been shown to interact with CagX, similarly to the *A. tumefaciens* homologues (VirB10 and VirB9, respectively) (Das *et al*., 2000).

2.3 Project design

Significant amount of protein, usually of the order of milligrams, are required for structural studies and, generally, the protein must be active, soluble, stable, highly pure and concentrated. We identify two major bottlenecks in the protein production: the first is protein insolubility or aggregation and the other is protein crystallization.
Combined bioinformatics and literature informations were used to guide initial attempts to identify the optimal boundaries (starting and ending residues of the constructs) of a protein target or protein domain. The target proteins were ranked according to several predicted properties. Each sequence was analyzed to estimate the following features: amino acids content, molecular weight, isoelectric point, absorption coefficients at 276-280 nm, presence of cysteines and stability index. The bioinformatic analysis allowed us to predict if the protein contains transmembrane domains or putative signal secretion. Moreover, amino acid sequence similarity with other proteins, whose properties have been better investigated or which belong to similar apparatus or functional related bacteria facilitated us in the planning strategy. Finally, programs for secondary and tertiary structure prediction of protein were used, as well as bioinformatic tools to predict the presence of folded and unfolded regions in the protein of interest (Prilusky *et al*., 2005). All the softwares used are available from the ExPASy server (ExPASy ProtParam tools; www.expasy.org). The strategy adopted for cloning, protein expression, purification and crystallization tests of Cag proteins is briefly described in the Appendix B.

2.4 Results and Discussion

- VirB/D proteins

The Vir system in *A. tumefaciens* can be considered the prototype of the type IV apparatus (TFSS). Its essential constituents are the structural proteins labeled from VirB1 to VirB11 and VirD4 (Christie *et al*., 2005). Several *H. pylori* proteins encoded by the *cag*-PAI were identified as homologues of VirB proteins (Censini *et al*., 1996; Rohde *et al*., 2003). In particular, the *H. pylori* conserved TFSS components include the inner membrane-localized ATPases CagE/23 (HP0544) (Gomis-Ruth *et al*., 2001; Middleton *et al*., 2005) and Cagα (HP0525), corresponding to VIRB4 and VIRB11, respectively. They possibly provide the energy for substrate translocation (Savvides *et*

al., 2003; Hare *et al*., 2007). The crystal structure of Cagα has been determined (Yeo *et al*., 2000): it assembles as a homo-hexamer (Fig. 2.2), harboring large cavities able to accommodate an unfolded or partially folded polypeptide chain, and it is supposed to participate in the translocation of substrates from the cytoplasm through the inner membrane. A similar function may be played by Cagβ/5 (HP0524), which is the *E. coli* TrwB homologue. In addition to a complex transmembrane transport system, the membrane-associated and surface-exposed pilus is an essential structural and functional component of the TFSS (Christie *et al*., 2005). Analysis of *H. pylori*'s TFSS has also revealed some unique properties, especially in the composition of the TFSS pilus. Several VirB homologues which in *A. tumefaciens* have been characterized as membrane associated proteins, in *H. pylori* resulted to localize in the external filament of the TFSS. For example, in *A. tumefaciens* VirB7 and VirB9 are outer membrane-localized proteins. In *H. pylori* the respective homologues, CagT and CagX, were localized by immunogold electron microscopy along the length of the pilus (Tanaka *et al*., 2003), while a second group described CagT distribution in a ring at the base of the pilus (Rohde *et al*., 2003). The *cag* genes that we have cloned and expressed are individually described in the following of the chapter.

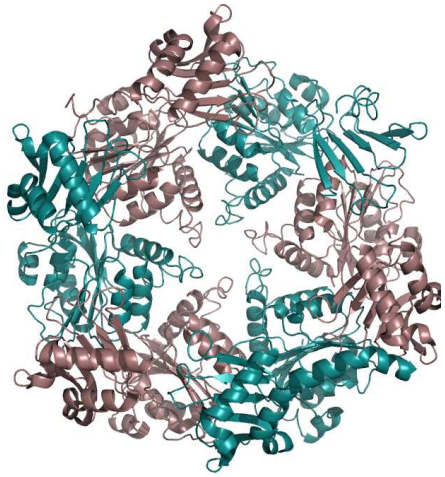


Figure 2.2. Ribbon representation of the exameric Cagα structure (Yeo *et al*., 2000; Savvides *et al*., 2003; PDB accession codes: 1NLY, 1OPX, 1G6O, 1NLZ).

CagT/12 (HP0532). CagT represents the VirB7 lipoprotein homologue of *A. tumefaciens* TFSS (Censini *et al.*, 1996). *cagT* gene codes for a protein of about 280 amino acids and its deletion abolishes the translocation of the toxin CagA and suppresses the induction of interleukin IL-8 (Fischer *et al.*, 2001) [Fig. 2.1].
The *cagT* gene was cloned and expressed in *E. coli*. To avoid difficulties due to poor solubility and/or post translational modifications, it has been cloned to obtain an amino acid sequence starting after Cys 21. Indeed, the CagT N-terminal signal sequence shows the typical topology of the VirB7 lipoproteins: a N-terminal mostly hydrophobic leader sequence, followed by a few positively charged residues which precedes a fully conserved Cysteine residue. The latter is processed and modified to insert the lipid anchor. The mature sequence $CagT_{\Delta SS}$ (22-164 aa) is mostly hydrophilic, without any detectable features typical of transmembrane domains.

CagT was cloned and purified using several tags: the Maltose Binding protein (MalE) fusion peptide (MalE-CagT$_{\Delta SS}$), the Hexahistidine tag (His-CagT$_{\Delta SS}$) and the glutathione S-transferase (GST) recombinant tag (GST-CagT$_{\Delta SS}$) (Table 2.1). The *cag*T gene, deprived of the first 60 nucleotide, was amplified and cloned into pET20b, adding a pelB leader peptide, pET21b and pGEX-4T-1. The pMal-p2-X expression vector was also tested to introduce a signal sequence at the N-terminal of MalE tag, to allow the recognition by the *E. coli* Sec system and mediate the export into the periplasm of the entire recombinant fusion protein.

Preliminary expression trials of the recombinant protein, *E. coli* BL21 (DE3) cells, carrying the MalE-CagT$_{\Delta SS}$ encoding plasmid, revealed, by SDS-PAGE, a very low expression level and a clear degradation of the recombinant MalE-CagT$_{\Delta SS}$. To reduce the protein degradation pattern we decrease the temperature during protein expression.

Firstly the recombinant protein MalE-CagT$_{\Delta SS}$ was purified using an amylose affinity resin (New England Biolabs) and CagT$_{\Delta SS}$ protein was separated from its MBP tag by Factor Xa proteolysis. Several incubation conditions and times of reaction were tested in order to avoid, as much as possible, the degradation of CagT$_{\Delta SS}$, since the treatment with Factor Xa strengthen this process, already observed during the first steps of expression and purification. The results of the hydrolysis, together with Western blotting analyses using antibodies anti-MalE, evidenced that the unstable portion of the recombinant MalE-CagT$_{\Delta SS}$ is the sequence of CagT itself. Most probably, the recombinant CagT$_{\Delta SS}$ does not fold properly with the expression strategy adopted, or the protein is intrinsically unstable if its native partners are absent. Similar results were observed using glutathione S-transferase (GST) tag. In this case, the GST-CagT$_{\Delta SS}$ construct resulted completely unstable, progressively degrading at the C-terminus until the accumulation of the GST tag itself, as proven by Western Blotting analysis (data not shown). Finally, the His-tagged CagT fragment was not detected at all in the expression trials (as in the case of the vector pET20-CagT$_{\Delta SS}$), or was obtained with a reasonable expression yield but mostly insoluble (in the case of pET21b-CagT$_{\Delta SS}$ vector) [Table 2.1].

CagX/8 (HP0528). CagX shows significant sequence similarities with VirB9, one of the most conserved among the VirB proteins, and together with CagT, CagV and CagY is one of the major components of the structural core of *H. pylori* T4SS. Mutagenesis studies have demonstrated its essential role both for the translocation of CagA and the triggering of interleukin-8 production (Stein *et al*., 2000; Fischer *et al*., 2001; Selbach *et al.*, 2002) [Fig. 2.1]. As mentioned before, Tanaka and colleagues

identified by electron microscopy several copies of CagX in the TFSS filamentous macromolecular protrusions on the surface of *H. pylori*, mainly concentrated at the basement and along the first tract of the pilus, adjacent to the envelope (Tanaka *et al.*, 2003). VirB9 protein from *A. tumefaciens* was proved to interact with VirB7 by detergent extraction studies and peptide linkage mapping (Ward *et al.*, 2002; Krall *et al*., 2002). VirB7 and VirB9 resulted to be linked together by two disulphide bridges to form heterodimers. In the attempt to elucidate the behavior of their homologues in *H. pylori* T4SS, both discrepancies and homologies have to be considered. CagX sequence includes only one Cysteine residue, located at the N-terminus. This segment represents a well-predicted export signal, hydrolyzed in the mature protein form. If this is the case, no Cysteine is left to form a disulphide bridge, neither with other copies of CagX, nor with CagT, the *H. pylori* VirB7 homologue. However, both CagT and CagX have been recognized as major components on some filamentous protrusions of the bacterial envelope, suggesting that they co-localize in the secretion apparatus (Tanaka *et al*., 2003; Rohde *et al*., 2003).

The native CagX protein contains 522 amino acids with a predicted molecular weight of 60.7 kDa. The first 26 amino acids, expressed at the N-terminus, are predicted with high confidence to define an export signal, processed in the mature form $CagX_{\Delta SS}$ (27-522). The isoelectric point is strictly basic (pI 9.40), like the majority of the Cag proteins. In particular the sequence presents a cluster rich in positive and negative charges, from residue 139 to 173.

To elucidate its function and structure, the amplified *cag*X gene was cloned into a pGEX-4T-1 expression vector (Table 2.1) and produced using *E. coli* BL21 (DE3) strain. The expression and solubility levels were checked by SDS-PAGE, where the recombinant protein GST-$CagX_{\Delta SS}$ revealed a good yield but a strong tendency to precipitate with the insoluble components. Several additives (detergents, glycerol, alcohols, chaotropic agents) have been tested to increase the solubility, but the extraction yields obtained in those conditions were unsatisfactory. The small amount of soluble GST-$CagX_{\Delta SS}$ recombinant protein was firstly purified using Glutathione Sepharose 4B Fast flow affinity resin, treated with Thrombin, and then concentrated and applied to a size exclusion chromatography. This step revealed a strong tendency to unspecific aggregation, thus producing a non homogeneous sample, not appropriate for crystallization trials.

Owing to the limited yields obtained and the problems encountered in these first steps, different expression vectors and purification strategies were tried. The gene was inserted in other expression vectors (pET20b and pET21b) to obtain different

recombinant constructs. The first introduces a N-terminus pelB export signal, recognized by *E. coli*, and the second produces $CagX_{\Delta SS}$ without any affinity tags, except for a N-terminal T7 sequence for the recognition by specific antibodies. Preliminary expression trials, using *E. coli* cell carrying the pET20b-*cag*$X_{\Delta SS}$ construct, caused a progressive and rapid death of the bacteria after induction with IPTG, indicating that the pelB-$CagX_{\Delta SS}$ protein is toxic for *E. coli* in the tested conditions. Other expression tests, performed using the pET21b-*cag*$X_{\Delta SS}$, gave better results, but the conventional strain *E. coli* BL21 (DE3), harboring the vector of interest, produced a small amount of T7tag-$CagX_{\Delta SS}$ protein. For this reason, we tested the expression of T7tag-$CagX_{\Delta SS}$ in several strains, obtaining slightly higher yields using the *E. coli* BL21 (DE3) Codon Plus strain. However, the tagged protein obtained by this way accumulated in the inclusion bodies. The refolding conditions tested (FoldIt Screen, Hampton Research), did not produce reasonable yields when tested in a larger scale.

CagY/7 (HP0527). CagY protein presents a C-terminal domain homologue to the inner membrane-localized VirB10 protein from *A. tumefaciens*. It is assumed to form a sheath around the inner core of the pilus which is induced upon host cell contact (Rohde *et al*., 2003). However, *H. pylori* CagY is a multi domain protein, much more complex than *A. tumefaciens* VirB10. It is predicted to have a molecular weight of 219 kDa, whereas the VirB10 family is composed of proteins with molecular weights of about 45 kDa. Indeed, CagY sequence homology with VirB members corresponds only to a C-terminal domain, which represents less than 20 % of the entire amino acid sequence. Moreover, the C-terminal portion of CagY protein has been demonstrated to interact with CagX, similarly to the *Agrobacterium* homologues VirB10 and VirB9, respectively (Das et al., 2000) [Fig. 2.1]. The *cagY* gene contains two long, repetitive regions, at the 5′ end and in the central portion, while both are absent in the other *virB10* family components, and moreover these regions undergo peculiar variations in length and in amino-acid composition of the resulting proteins (Aras *et al*., 2003). Isolates from different human hosts and bacteria recovered from experimentally colonized mice and rhesus monkeys varied considerably in the number of repeats, always leading to new in-frame combinations of nucleotide repeats, resulting in CagY proteins of different length (Aras *et al*., 2003). As the antibody response against this protein in human hosts was low, the authors postulate that the almost infinite potential for variation in the *cagY* gene by intergenomic or intragenomic recombination or

deletion serves the purpose of evading the host antibody response, while preserving cag apparatus function.
Starting from these evidences, we decided to isolate and characterize the CagY C-terminal VirB10 homologous domain. The region coding for residues from 1690 to 1927 ($CagY_{\Delta N}$) was cloned in pET151 vector, by the TOPO directional cloning technique. The resulting His_6-$CagY_{\Delta N}$ construct was expressed in BL21 (DE3) Star *E. coli* cells, in a small amount and after the first lysis and purification steps, it has undergone rapid and wide degradation. Most likely, the cloned fragment do not correspond to a proper folded domain and, starting from these results, new fragments should be selected for further studies.

CagC/25 (HP0546). The major component of the multimeric T pilus in *A. tumefaciens* is the processed and cyclized pilin protein VirB2, which was found to participate in the translocation of the T-DNA-protein complex (Lai *et al*., 2002). Up to now, neither a VirB2-like pilin subunit, associated with the *cag* pathogenicity island, nor a pilus forming major component has been detected in the *H. pylori* complete genome sequence so far (Tomb *et al*., 1997). Recently, the CagC protein has been identified as a VirB2-orthologous pilin protein of the Cag apparatus (Andrzejewska *et al*., 2006). The CagC protein has been characterized to play a relevant role in the Cag secretion apparatus and in the interaction with the host. Immunofluorescence and electron microscopy studies have suggested that CagC is predominantly localized in the membrane fractions and might be part of a multimeric structure on bacterial surfaces. CagC is exposed at different sites on bacterial surfaces, within larger amorphous appendages, in wild-type bacteria but not in a *vir*B4/*cag*E mutant, indicating that its surface targeting is dependent upon a functional *cag*-PAI TFSS. Finally, *cag*C deletion mutants displayed diminished abilities to induce IL-8 up-regulation in gastric epithelial cells in comparison to the wild-type strain carrying an intact *cag*-PAI (Fischer *et al*., 2001). Moreover, the mutants lost the ability to translocate the effector protein CagA into host cells (Odenbreit *et al*., 2000). All these results suggest that the protein is a surface-associated VirB2-like pilin subunit, functionally linked to the *H. pylori* Cag apparatus (Andrzejewska *et al*., 2006). The 115 amino acids full-length CagC protein, with a predicted molecular mass of 13 kDa, contains a putative secretion system signal at the N-terminus (SignalP, Bendtsen *et al*., 2004) and the hydrophobicity plot indicates that it contains two hydrophobic domains (amino acids 35-47 and 72-95), similarly to other well characterized VirB2-like pilins.

To try to characterize its structure and function, we generated several recombinant CagC constructs. Firstly, we have cloned the full-length protein, with its secretion signal sequence plus a N-terminal hexa-histidine tag. After the preliminary expression tests, confirming a low level of expression, we decided to clone the protein removing the first 29 amino acids corresponding to the leader peptide sequence. The molecular weight of the putative mature $CagC_{\Delta SS}$ protein after cleavage was calculated to be 9.7 kDa. Three different constructs have been generated: two of them harboring a hexahistidine tag at the N- and C-terminus (pET28b and pET20), respectively, and a third a MBP-tag (pMal) at the N-terminus [Table 2.1]. These constructs (His_6-$CagC_{\Delta SS}$, $CagC_{\Delta SS}$-His_6 and MBP-$CagC_{\Delta SS}$) were then used to perform some protein expression trials, but in all cases the level of protein expression was too low and detectable only by immunoblot analysis (data not shown). To improve the amount of protein production, co-expression with molecular chaperones and folding modulators could be useful, as well as new expression systems, like the cell-free protein synthesis.

CagZ/6 (HP0526). CagZ is a 24 kDa protein with a very acidic pI (5.1). We have cloned, over-expressed and purified the protein and determined its three-dimensional crystal structure by the MAD technique at a maximum resolution of 1.9 Å (Cendron *et al*., 2004). The protein consists of a single compact L-shaped domain (Fig. 2.3), composed of seven α-helices including about 70 % of the total residues. The presence of a disordered C-terminal tail and the nature of the molecular surface suggest that CagZ may participate in the interaction of effector proteins with one or more components of the *H. pylori* TFSS on the cytoplasmic side of the inner membrane. Moreover, preliminary evidences reveal that CagZ could be weakly associated with the external membrane (Tasca and Montecucco, personal communication). All the features evidenced in our *in vitro* studies, the overall 3D structure and its involvement in the secretion of the toxin CagA, as proven by mutagenesis studies (Fischer *et al.*, 2000), suggest that CagZ might constitute the functional homologue of VirB5 for *H. pylori* T4SS. However, against this hypothesis, CagZ does not show any signal peptide in its N-terminal portion, even if it could be exported to the outer membrane by a not yet recognized kind of process, or even the TFSS itself.

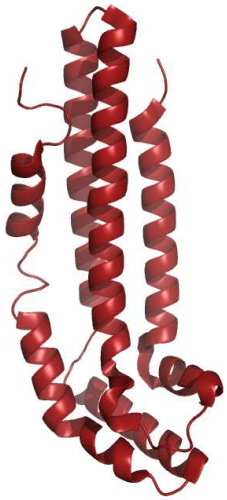


Figure 2.3. Ribbon representation of the exameric CagZ structure (Cendron *et al*., 2004; PDB accession code: 1S2X).

CagW/9 (HP0529). By sequence comparison, a VirB8 homologue seems to lack in the *H. pylori* Cag pull of proteins. Preliminary sequence analysis suggested that CagW protein might represent a homologue of the VirB8 (Covacci and Rappuoli, 2000). Since there is no obvious sequence similarity, this hypothesis was probably based both on its essential role in the CagA translocation and IL-8 induction, and the localization of *cagW* gene, within the core components of the apparatus and immediately upstream the *cagX* gene (*hp0528*, *virB9* homologue). While the most characterized VirB8 proteins from *A. tumefaciens* and *B. suis* were shown to behave as bitopic membrane proteins (Das and Xie, 1998; Kumar *et al*., 2000), the transmembrane helices content of CagW protein, as predicted by bioinformatic analysis, suggests that this protein adopts a polytopic conformation with five to six transmembrane domains in addition to its N-terminal signal sequence. Thus, the CagW protein might define a functional homologue of VirB6, which is the only component of the VirB/D4 system with more than three transmembrane helices (Das and Xie, 1998; Buhrdorf *et al*., 2003). Since CagW is an integral membrane protein that is known to be hard to produce and purify, till now we did not work on it.

CagV/10 (HP0530). The CagV protein is composed of 252 amino acids, corresponding to a molecular weight of 29.1 kDa and a theoretical isoelectric point of 9.52. Mutagenesis studies have demonstrated its essential role both for the translocation of CagA and the triggering of interleukin-8 production (Stein *et al*., 2000; Fischer *et al*., 2001; Selbach *et al.*, 2002) [Fig. 2.1].

Several experiments demonstrated that CagV could be play an analogous role of VirB8 ptoteins in *H. pylori* TFSS. Generally, VirB8 family components varies in length from about 202 to 260 amino acids (Baron, 2006) and display only a limited sequence similarity, but some more or less conserved motifs can be distinguished (Cao and Saier, 2001). Like VirB8 proteins, CagV is a bitopic membrane protein with a short N-terminal cytoplasmic exposed domain, a transmembrane helix and a large C-terminal periplasmic domain (Das and Xie, 1998; Kumar et al., 2000). Its sequence is mostly hydrophilic, except for residues from 38 to 57, which define a putative transmembrane sequence. Moreover, the presence of only one Cysteine suggests the possibility of dimerization events in the native form.

Although full-length CagV protein (252 residues) could be expressed in a soluble form in *Escherichia coli* as a fusion protein with glutathione S-transferase (GST-CagV), it tended to become insoluble upon removal of the GST. To increase its solubility and reduce the aggregation of the sample, proved by the chromatographic gel filtration

profile, the purification was performed in the presence of neutral detergent $C_{12}E_8$, selected since it resulted highly monodisperse and particularly effective with other membrane proteins. However, the full-length CagV sample resulted very inhomogeneous and difficult to separate from GST tag, most likely due to the absence of the native *H. pylori* membrane environment and of the interacting proteins, during CagV expression and folding, even after proteolysis. With the aim of selecting more stable, monodisperse and soluble domains, two C-terminal fragments (His_6-pCagV$^{59-252}$ and His_6-pCagV$^{68-252}$) of CagV were successfully cloned, produced and purified. Before crystallization experiments, the purity and homogeneity of the proteins in solution were confirmed by SDS–PAGE, size exclusion chromatography and dynamic light scattering techniques. After extensive screening to identify the proper conditions to grow diffraction quality crystals, a native data set of His_6-pCagV$^{59-252}$ and His_6-pCagV$^{68-252}$ were collected to a maximum resolution of 2.7 Å and 3.0 Å respectively (Chapter 5).

Cagγ/4 (HP0523). A last gene suggested to function like an *A. tumefaciens* Vir gene is *cag*γ, which, like *vir*B1, is thought to be a transglycosylase that breaks down the peptidoglycan layer to facilitate assembly of the TFSS across the bacterial cell wall (Rohde *et al*., 2003). Cagγ is unable to complement an *A. tumefaciens vir*B1 mutant (Hoppner *et al*., 2004), but it is required for maturation of pilus-associated CagY/VirB10 and CagT/VirB7 proteins.

VirB1-like proteins are believed to act as specialized lytic transglycosylases, which locally degrade the peptidoglycan, facilitating the assembly of the TFSS. Since the peptidoglycan imposes a structural constraint to the assembly of a macromolecular secretion system, most likely specialized enzymes are required to enlarge holes in the peptidoglycan by controlled local degradation (Koraimann, 2003). Analyses of bacterial genome sequences revealed that many, if not all, TFSS gene clusters encode their own murein-degrading enzymes, which belong to the lytic transglycosylase family of proteins, which are widely distributed among gram-negative bacteria and cleave the *β*-1,4 glycosidic bond between *N*-acetylmuramic acid (MurNAc) and *N*-acetylglucosamine (GlcNAc). These small proteins (usually between 150 and 250 aa residues) have been termed 'specialized' lytic transglycosylase because of their defined role in macromolecular transport (Koraimann, 2003). They contain the characteristic SLT sequence motif (see the PFAM database, accession code PF01464; http://www.sanger.ac.uk/Software/Pfam/).

In view of the fact that deletion of genes encoding VirB1 from *Agrobacterium tumefaciens* and *Brucella suis* generally have an attenuating effect on TFSS-related functions, the VirB1 proteins are consider not essential TFSS components (Höppner et al., 2005). The *Helicobacter pylori* VirB1 homologue Cagγ is an exception to this rule, as it was shown to be essential for bacterial virulence (Odenbreit *et al.*, 2001; Rohde *et al.*, 2003). The Cagγ is in fact essential for both CagA translocation into the human host cells and IL-8 induction (Fischer *et al.*, 2001) and displayed peptidoglycanase/muramidase activity in zymogram analyses (Koraimann, 2003).

A. tumefaciens and *B. suis* VirB1, like Cagγ from *H. pylori*, possess an N-terminal signal sequence and are therefore directed to the periplasmic space, by the general secretion pathway, where they carry out their enzymatic activity. However, Cagγ overall sequence identities to VirB1 proteins from *A. tumefaciens* and *B. suis*, respectively, are below 20 % (Zahrl *et al.*, 2005). Both VirB1 from *A. tumefaciens* and *B. suis* have a C-terminal extension which is only present in the VirB1 class of specialized lytic transglycosylases. In T-DNA transfer of *A. tumefaciens,* VirB1 is processed and a peptide, called VirB1*, representing this C-terminal extension, is secreted into medium or loosely interacts with the bacterial cell surface (Baron *et al.*, 1997). VirB1 from *A. tumefaciens* was also shown to interact with itself and several other VirB proteins, VirB4, VirB8, VirB9, VirB10, VirB11 (Ward *et al.*, 2002). Similarly, VirB1 from *B. suis* has been shown to interact with core components VirB8, VirB9 and VirB11 (Höppner *et al.*, 2005).

The region of *hp0523* encoding the full-length Cagγ and the predicted periplasmic domain (17-169) of the protein Cagγ without signal peptide ($Cag\gamma_{\Delta SS}$) were PCR-amplified and cloned in several expression vectors for both periplasmic and cytoplasmic production of native Cagγ protein using N- and C-terminally tagged fusion constructs (His-tag, GST-tag; and MBP-tag; Table 2.1). The numerous attempts to overexpress the full-length Cagγ as well as the truncated $Cag\gamma_{\Delta SS}$ form in the cytoplasm of *E. coli* resulted in the formation of inclusion bodies which could only be solubilized using chaotropic agents (Urea and Guadinium-HCl). In only one case we were able to obtain the soluble protein, when fused to the periplasmic maltose-binding protein (MBP), but the amount of protein obtain was to low to perform structural studied. To obtain large amount of pure protein, denaturation and renaturation experiments were performing without success.

- Non-VirB/D proteins

CagF/22 (HP0543). CagF is composed of 268 residues, with a predicted molecular weight of 31.8 kDa and shows a very acidic isoelectric point of 4.5. Interestingly, only eight proteins in the whole *cag*-PAI have an acidic pI, whereas all the rest of the other Cag proteins, except one, show a clear prevalence of basic residues, conferring a strongly basic isoelectric point. Most of these acidic proteins are codified in the *cag*I region, constituting a continuous cluster from *cag*F to *cag*L. The deletion of *cag*F gene causes the complete abolishment of the secretion of CagA toxin, but it does not interfere with interleukin-8 induction (Fischer *et al*., 2001 and 2003). Moreover, it has been shown that CagF interacts with CagA and that this interaction is essential for the latter's translocation (Couturier *et al*., 2006; Chapter 6). The amino acid sequence of CagF does not show any significant homology with VirB/D proteins of *A. tumefaciens* or TFSS components of other bacteria. Comparing CagF amino acid sequence with a non-redundant database of proteins, only weak homologies have been found. One of the most reliable alignments identifies some similarities with a glycosyl transferase (Q9L7A2) of *Haemophilus ducreyi* (Sun *et al*., 2002).

However, localization studies showed that CagF localizes at the bacterial cytoplasmic membrane, most probably with a periplasmic or, at least, a partially accessible domain of 17.5 kDa and a membrane buried portion of about 14 kDa (Seydel *et al*., 2002; Couturier *et al*., 2006).

Moreover, very recently, two different studies demonstrated that CagF interacts with CagA toxin in the bacterial cell and it has been suggested that CagF may function as a secretion chaperone like molecule that recruits CagA to the TFSS (Couturier *et al*., 2006; Pattis *et al*., 2007; Chapter 6.3).

The *cag*F gene was amplified by PCR, cloned in a pGEX-4T-3 vector (Table 2.1) and expressed in *E. coli* BL21 (DE3) cells. The recombinant GST-CagF product showed a good level of expression and about 60 % of it could be recovered in the lysate supernatant. However, the use of detergents was necessary in order to avoid aggregation or non-specific oligomerisation during all the purification steps. The use of detergents, such as lauryldimethylamino-*N*-oxide (LDAO), allowed the isolation of the dimer as the main and more stable component, despite the presence of minor assemblies components of more than 600 kDa. Firstly, the recombinant protein was purified using a Glutathione Sepharose affinity resin (see Chapter 6) and separated from its GST tag by thrombin proteolysis, using directly the recombinant fusion product immobilized by the resin. The sample obtained was further purified by gel filtration chromatography, with a Superdex 200 10/30 prepacked column (GE

healthcare). Circular dichroism spectra show a clear prevalence of α-helices, around 49 %, compared with a β-sheets contribution of 12.5 %, coherently with the secondary structure prediction of PROFsec software (Expasy server; Rost, 2004).
Sparse matrix crystallization tests using several standard screens were performed with the dimeric CagF. However, diffraction quality crystals have not been obtained yet, most likely owing to degradation phenomena, occurring during long term storage of the purified samples, as revealed by SDS-PAGE and Western blotting analysis.
The degradation pattern observed suggests again the presence of a more stable globular domain, which migrates in SDS-PAGE with an apparent weight of 20 kDa, and an appendage. The N-terminal sequence of the two main fragments proves that the larger one includes the N-terminal end of the wild type CagF, whilst the other is a mixture. Several software for the prediction of protein topology indicate the presence of a putative transmembrane helix, which involves residues from 150 to 170, followed by another mostly hydrophilic pattern of about 100 residues. The experimental observations and the predictions based on the primary sequence seem to indicate that the protein spans the outer membrane, with a periplasmic smaller C-terminal domain and a N-terminal portion, exposed on the surface. The observed strong tendency to aggregate in high molecular weight oligomers might imply that, *in vivo,* CagF associates to form homo- or, most likely, hetero-oligomers with other components of the secretion apparatus. CagF would be exported to the outer membrane via a Sec-independent mechanism, since its N-terminal end does not include a secretion signal. Like other Cag proteins, which might be exported, its sequence would contain a not yet identified signal for the proper localization in or out of the bacterial cell.

Since the accumulated experimental data and *in silico* predictions suggest that the protein comprises at least two domains. We thus decided to clone two separate domains of CagF (Table 2.1); however the C-terminal selected fragment (171-268, $CagF_{\Delta N}$-His_6) resulted almost insoluble, while the N-terminal one (1-149, His_6-$CagF_{\Delta C}$) has been cloned, expressed and purified successfully, but still shows a marked degradation, indicating that it does not correspond to a globular well folded and stable domain, at least in the tested conditions.

CagL/18 (HP0539). CagL protein is composed of 237 amino-acid, with a predicted molecular weight of 26.9 kDa and a theoretical isoelectric point of 8.3. Deletion of *cagL* significantly diminishes the ability of *H. pylori* to induce secretion of IL-8 by host cells (Fischer *et al.*, 2001), indicating that CagL is essential for pathogenesis [Fig. 2.1]. Initially, CagL was postulated to play a chaperone role for some Cag proteins

(Rohde *et al.*, 2003), but very recent experiment have indicated that CagL, a protein highly conserved among pathogenic *H. pylori* strains, is the only *cag*-PAI-encoded gene product that contains an arginine-glycine-aspartate (RGD) motif, which serves as a recognition site for the host integrins (Kwok *et al*., 2007). CagL, in fact, was shown to be able to act bifunctionally: to hijack host integrin receptors for virulence factors injection into eukaryotic cells as well as to activate host tyrosine kinases. Through activation of integrin $\alpha_5\beta_1$, the pilus-located adhesin and sensor molecule, CagL seems to ensure phosphorylation of translocated CagA directly at the site of injection at focal adhesions and subsequently to promote downstream signaling in the host cell (Kwok *et al.*, 2007).

The analysis of the CagL amino acid sequence by using SignalP (Bendtsen *et al.*, 2004) revealed that CagL contains a predicted N-terminal secretion signal peptide with a predicted cleavage site between amino acid 20 and 21. The gene portion of cagL corresponding to amino acid residues 21–237 of the protein was amplified by PCR and ligated into pET-28a vector, obtaining His_6-$CagL_{\Delta SS}$. Previous attempts to clone and purify the entire protein, including the signal sequence, as a GST tagged construct gave rise to aggregation and misfolding. The His_6-$CagL_{\Delta SS}$ recombinant protein was then expressed in *E. coli* BL21(DE3) cells. Notwithstanding its very low solubility, we were able to prepare recombinant CagL in reasonable amounts. His_6-$CagL_{\Delta SS}$ was purified by metal-chelate affinity chromatography (GE Healthcare), followed by gel filtration using a Superdex 75 10/30 column (GE Healthcare) which allowed to isolate an homogeneous and stable sample, corresponding to a dimer. The folded conformation of the purified CagL was confirmed by circular dichroism using a J-720 spectropolarimeter (Jasco Instruments). Preliminary crystallization trials yielded very thin plates which, up to now, do not diffract. Attempts to improve the crystal quality are under way.

CagM/16 (HP0537). Another example of a non-VirB homologue, CagM is a protein of 376 amino acids. The deletion of the corresponding gene causes a significant loss of the CagA toxin translocation, but not a total abolishment. The data concerning its involvement in the processes which trigger the inflammatory response indicate that *cag*M deletion mutant reduces the IL-8 production by more than 60 %, if compared to the *H. pylori* wild type (Fischer *et al*., 2001). Inactivation of this gene results in decreased activation of the nuclear factor-κB (NF-κB) and mitogen-activated protein kinase (MAPK) signal-transduction cascades, which regulate pro-inflammatory cytokine production, mimicking *in vitro* the observed increase of IL-8 expression and

mucosal inflammation in human gastric tissue (Sharma *et al*., 1995 and 1996; Glocker *et al*.,1998; Keates *et al*., 1999). Moreover, very recently, it has been demonstrated (Viala *et al*. 2004) that pathogenic *H. pylori* strains are recognized by epithelial cells *via* the intracellular host molecule Nod1, which shows a specific affinity for the Gram-negative peptidoglycan. When activated by *H. pylori* infection, the cytoplasmic messenger Nod1 might up-regulate the NF-κB transcriptional activity. The contact with bacterial peptidoglycan is not achieved by *H. pylori* internalization, but the muropeptides might be delivered into the host cells by a bacterial type IV secretion system encoded in the *H. pylori cag*-PAI. The investigators proved that *cag*M product is necessary for the induction of NF-κB in HEK293 cells by a luciferase reporter, whereas the toxin CagA, like what previously observed by other methods, does not affect this kind of response. Using a Δ*cag*M mutant to abrogate the type IV secretion system function, they proved a direct correlation between the Cag TFSS and the delivery of peptidoglycan, responsible for the epithelial cells sensing (Viala *et al*., 2004). Moreover, systematic mutagenesis studies demonstrated that *cag*M inactivated mutant causes the complete loss of CagT expression, which indicates a causal relationship. Most of the knock-out mutants of the other *cag* genes do not affect the transcriptional activity of *cag*T, supporting the idea of a correlation between CagM and CagT. For this reason, some investigators have proposed that CagM could act as a chaperone for CagT in the early stages of TFSS assembling (Fischer *et al*., 2001). Finally, transcription profile analyses, using the bacterial mRNA isolated from infected human gastric mucosa, suggested that CagM is one of the most abundant Cag proteins in the tested conditions (Boonjakuakul *et al*., 2004).

CagM carries a N-terminal signal sequence for the export in the periplasm and some experimental evidences suggest the idea that it might oligomerize in the native form. Sequence comparison against a non-redundant database from Bacteria and Archaea detects no VirB homology. Interestingly, some similarities have been found between the CagM sequence and a DNA repair protein, Rad50, from the termoacidophilus *Picrophilus torridus,* and also with a Glucan 1,6-Alpha-Glucosidase, from murine respiratory pathogen *Micoplasma pulmonis*, whose hydrolase activity on glycosyl bonds has been inferred by homology.

CagM has a molecular mass of 43.7 kDa and a strongly basic isoelectric point, around 9.3. A N-terminal secretion signal is predicted whereas no transmembrane sequences were identified by bioinfomatic analysis of its primary sequence. However, three main hydrophobic patches, interrupted by hydrophilic sequences of about fifty amino acids,

characterize the hydrophobicity pattern and justified the use of detergents for its purification.

Many attempts to produce and purify CagM were performed, using several tags: the Maltose Binding protein (MalE-CagM$_{\Delta SS}$), the Hexahistidine tag (His-CagM$_{\Delta SS}$) and the glutathione S-transferase (GST) protein (GST-CagM$_{\Delta SS}$). In all the recombinant constructs, CagM protein was cloned after the first 20 amino acids. The nucleotide sequence of CagM corresponding to the sequence from 21 to 376 was amplified and cloned into the following expression vectors: pET28, pGEX-4T-3 and pMal-p2-X. The most promising results were obtained using the MalE-CagM$_{\Delta SS}$ and His-CagM$_{\Delta SS}$ constructs which are described below.

The MalE-CagM$_{\Delta SS}$ recombinant protein has been expressed in *E. coli* BL21 (DE3) cells and purified using amylose affinity resin. After the tag cleavage, using Factor Xa protease, CagM$_{\Delta SS}$ was loaded in a MonoS cationic exchange resin (GE Healthcare), to separate the released MalE (pI 5.1) from CagM$_{\Delta SS}$ (pI 9.3). The fractions containing CagM$_{\Delta SS}$ were then further purified by gel filtration (Superdex 200 HR 10/300; GE Healthcare). By this last step, CagM$_{\Delta SS}$ was eluted as a single peak, corresponding to a molecular mass of about 250 kDa and with a reasonable level of pureness. To characterize its putative oligomerization, the purified CagM$_{\Delta SS}$ was analyzed by analytical gel filtration chromatography, loading the sample into a Biosep-SEC-S2000 (Phenomenex) analytical prepacked column, before and after a short treatment with many detergents (CHAPSO, C-HEGA-11, DM30, LDAO, n-hexyl-b-D-glucoside, n-dodecyl-b-D-maltoside, Triton-X100). In all cases, CagM was eluted as a single peak with a retention volume corresponding to a molecular weight of about 250 kDa. These results suggest the hypothesis that CagM associates in a pentameric or hexameric form, and likely interacts with some detergent molecules. In particular, concerning the profile of the elution peak, the best results were observed adding the detergent DM30.

The CD spectrum allowed to estimate the secondary structure content of the CagM protein, purified and treated as described before. The CD pattern argued for a structure composed by 34 % of α-helices and 25 % of β-sheets (CDNN software, version 2.1, Bohm *et al*., 1992).

Sparse matrix crystallization tests were performed with CagM in its oligomeric form, applying the vapor diffusion technique, both with hanging and sitting drop methods and using several standard screens. Three of the most promising conditions were selected and tested with a protein sample previously incubated with different detergents (Detergent Kits I and II, Hampton Research). Despite single crystals of reasonable size have not been obtained yet, promising micro-precipitates were

observed with reagents such as ammonium sulphate, di-sodium and di-potassium hydrogen phosphates and lithium sulphate monohydrate, especially when pH was buffered between 6.5 and 7.5. However, further characterizations are necessary to clarify the nature of the oligomerization state and the homogeneity of the sample obtained, as well as other crystallization trials are in progress using different detergents.

Concerning His_6-$CagM_{\Delta SS}$ construct, it was expressed in a massive amount by *E. coli* BL21 cells, but was found to be present mainly in the inclusion bodies in all the growing conditions tested. For this reason, a refolding approach was attempted, using the so-called on-column refolding method which takes advantage from the capability of the His_6 tag to bind to Ni^{2+}-NTA resin both in denaturing and native conditions (Li *et al*., 2004; Oganesyan *et al*., 2005). A promising amount of protein was recovered in the renaturing buffer. The gel filtration and CD results showed a CagM behavior similar to that observed for the protein purified by the MalE tagging, thus confirming the tendency to oligomerize, at least *in vitro*. As for the previously described construct, many crystallization trials were performed, without obtaining diffraction quality crystals.

CagN/17 (HP0538). CagN is an unknown function protein composed of 306 amino-acids with a predicted molecular weight of 34.7 kDa. The theoretical isoelectric point is slightly basic (pI 7.6), as for the majority of Cag proteins. Analysis of the CagN amino acid sequence by SignalP software (Bendtsen *et al*., 2004) revealed that CagN contains a N-terminal secretion signal peptide, a predicted cleavage site between amino acid 24 and 25, but no other recognizable motifs or structure domain. A *cagN* deletion mutant was reported by Fischer and colleagues to have a variable CagA and IL-8 induction (Fischer *et al*., 2001). Recently, Bourzac and colleagues have performed some experiments to verify if CagN could be delivered into host cells and alter CagA function or it could be a component of T4SS itself. They found that CagN is not a secreted effector of T4SS and appears to be localized close to the bacterial membrane. Mass spectrometry analysis of CagN protein, isolated from *H. pylori* cultures, revealed that in the membrane fractions it is present as a shorter fragment of 22 kDa, being processed at the C-terminus after Ser216. They were also able to demonstrate that this processing event is independent of other cag-PAI proteins and is only partially blocked by mutations of the predicted cleavage site. Finally, they have shown that a cagN mutant does not affect the CagA-induced AGS cells elongation ("humming-bird phenotype"), observed as a consequence of *H. pylori* most virulent

strains infection. On the basis of these results, they concluded that CagN could have a redundant function in the T4SS or it could be required for the delivery of other secreted effectors that have not been identified, yet.

Before the discovery of the C-terminus processing of mature CagN our strategy implied the production and purification of full-length *cag*N, excluding the only N-terminus secretion signal, well predicted by many software. The nucleotide sequence of CagN corresponding to amino acids from 25 to 306 was amplified by PCR and cloned into a pET28 vector containing a hexahistidine tag at the N-terminal followed by a Thrombin protease cleavage site (Table 2.1). The protein has been then produced in *E. coli* BL21 (DE3) strain and purified using IMAC. After the tag cleavage, CagN was further purified by gel filtration (Superdex 200 10/30, GE Healthcare). The results obtained in this direction clearly show strong aggregation phenomena and weak solubility of the construct.

After Bourzac and colleagues studies, we decided to focus our efforts toward the characterization of the CagN protein without the processed C-terminus domain. The *cagN* nucleotide sequence, corresponding to residues 25 to 216, has been sub-cloned into the pET21 vector containing a hexahistidine at the C-terminus and further expressed and purified as previously described. The purified protein has been characterised by gel filtration, demonstrating that CagN clearly behaves as a dimer in solution. The deconvolution of CD spectra allowed to estimate the secondary structure content, revealing a marked prevalence of beta-sheets (30 %) with respect to α-helices (12 %).

Sparse matrix crystallization tests using different protein concentrations and several standard screens were performed with $CagN_{\Delta SS+\Delta C}$ (25-216), but, till now, we were only able to obtain microcrystal precipitate. Indeed, further analysis of aminoacid sequence and predicted topology of CagN evidenced the presence of highly flexible and/or intrinsically disordered regions, thus impairing the probability to obtain single crystals.

CagG/25 (HP0542). The ability of *H. pylori* to adhere to gastric epithelial cells is an important initial step for virulence. Previous studies using isogenic, experimental knock-out mutants of several *cag* genes showed marked reduction in the ability of *H. pylori* to induce IL-8 secretion from gastric epithelial cells (Censini *et al*., 1996). Using the Mongolian gerbil model to examine the effect of some *cag*-PAI genes on gastric inflammation *in vivo*, Saito and colleagues demonstrated that *cagG* mutants do not produce an inflammatory response or increase proliferative activity, most likely

related to their reduced ability to colonize gerbil gastric mucosa (Saito *et al*., 2005). The *cagG* gene is not a *vir* homologue gene and has a weak homology to the flagellar motor switch protein gene or toxin co-regulated pilus biosynthesis protein gene (Censini *et al.*, 1996; Tomb *et al*., 1997). The current idea is that loss of the *cagG* gene also results in the loss of CagA translocation and subsequent phosphorylation (Censini *et al*., 1996; Fischer *et al*., 2001). Recent reports suggest that isolate *H. pylori* strains lacking *cagG* genes abolished gastric inflammation *via* a reduced ability to adhere to KATOIII epithelial cell lines (Mizushima *et al*., 2002). Moreover, they almost completely abolish the IL-8 induction (Censini *et al.*, 1996; Hsu *et al*., 2002; Saito *et al*., 2005). Direct adherence of live *H. pylori* to gastric epithelial cells is in fact necessary for IL-8 production (Rieder *et al*., 1997).

CagG has a molecular mass of 15.2 kDa and an acidic isoelectric point, (pI 5.3). A N-terminal secretion signal is predicted whereas no transmembrane sequences were identified by bioinfomatic analysis of its primary sequence.

Many attempts to produce and purify CagG were performed, using several tags such as the Maltose Binding protein (MalE-$CagG_{\Delta SS}$) and the Hexahistidine tag (His_6-$CagG_{\Delta SS}$ and $CagG_{\Delta SS}$-His_6). After removing the first 27 amino acids, the nucleotide sequence of CagG corresponding to the sequence from 28 to 142 was amplified and cloned into the following expression vectors: pET28, pET20 and pMal-p2-X.

These constructs (His_6-$CagG_{\Delta SS}$, $CagG_{\Delta SS}$-His_6 and MBP-$CagG_{\Delta SS}$) [Table 2.1] were then used to perform some protein production trials, but in all cases the level of protein expression was too low and detectable only by immunoblot analysis (data not shown) since the protein was shown to be toxic leading to bacterial death.

To improve the amount of protein production, co-expression with molecular chaperones and folding modulators could be useful, as well as new expression systems, like the cell-free protein synthesis.

CagS/13 (HP0534). CagS is a 199 amino acid protein with a predicted molecular weight of 23 kDa and a prevalence of basic amino acids, which gives rise to an isoelectric point of 8.80. It is encoded by a well-conserved gene within the *cag*-PAI which is located immediately after the cluster of cag genes whose putative products show homologies with the VirB proteins that define the structural core of TFSS. A very recent study by 2D gel electrophoresis gave the first experimental evidences that CagS is expressed in *H. pylori* cultures in vivo (Busler *et al*., 2006). Primary sequence analysis do not show any strong similarities with proteins of known function, except for some weak similarities with components involved in the peptidoglycan

biosynthesis, belonging to the FemABX family of enzymes (Hegde *et al*., 2001; Cendron *et al*., 2007). The crystal structure of CagS have been recently described (Cendron *et al*., 2007; Chapter 3). It forms a compact single domain protein, with ten α-helices which are held together mainly by hydrophobic forces, with few hydrophilic inter-helical interactions. The results of the crystallographic model do not show any clear evidence of architectural similarity with other known structures, except for a weak structural homology with the phosphotransfer domain (HPt) of CheA, a histidine protein chinase that controls chemotaxis response in bacteria (Mourey *et al*., 2001; Quezada *et al*., 2005)

About the other no VirB homologue proteins, little is known about their localization and function that can not be inferred from sequence homology or bibliographic information (Table 2.1). However some of them were selected by bioinformatic analyses, cloned into appropriate expression vectors and preliminary expression experiments were performed. Till now all these candidates (CagI/19, HP0540; CagU/11, HP0531; Cagδ/3 HP0522 and Cagζ/1, HP0520) gave strong problems concerning expression, solubility and stability. In particular, the CagI recombinant construct His-CagI$_{\Delta SS}$ (20-383 aa) was expressed in a massive amount by *E. coli* BL21 cells, but was found to be present mainly in the inclusion bodies in all the growing conditions tested and a clear pattern of degradation has been shown, already present during the expression of the protein in *E.coli*. About the other three proteins studied, they were successfully cloned (His-Cagζ, Cagζ –His, GST-CagU, CagU-His and His-Cagδ; Table 2.1) but unfortunately their production interfere with the survival of *E. coli* cells. It has been in fact demonstrated that membrane proteins, like for example CagU and Cagδ, are in general highly toxic products that cause significant defects in bacteria growth leading to bacterial death and dramatically decrease expression capabilities. To get around these troubles, new clones have been planned as well new expression strain and systems (cell free protein synthesis).

2.5 Conclusion

To date, we were able to express only 13 over 19 protein cloned. The soluble proteins were nine (9/13) and four out of eight highly purified proteins yielded crystals following crystallization trials. Of the crystals obtained, the crystal structure of three of them (CagD, CagS, CagZ) have been determined, in all cases with the MAD technique using the selenomethionine derivatives.

Recombinant proteins produced in *E. coli* are often insoluble or misfolded. One reason for that in the case of Cag proteins is that their natural binding partners are absent. One method to circumvent this problem is the co-expression of the genes encoding such proteins together with genes encoding potential interaction partners. It has been shown by several researchers that *E. coli* coexpression of two or more proteins of a multiprotein complex enhances yield, solubility, and proper folding. Moreover, protein-protein interaction is often labile and the subunits of unstable complexes can be lost during the various purification steps. Moreover, many Cag proteins are membrane proteins, which makes quite hard to study multi-holigomeric complexes.

For these reason, to test interactions of Cag proteins we are planning to use several biochemical methods like dot-blot, pull-down assay or affinity precipitation, analytical gel filtration, bicistron expression, isothermal titration calorimetric (ITC) (Privalov *et al.*, 2007). Once identified, experiments of coexpression will be performed in an attempt to produce soluble and stable multiprotein products.

Moreover, to structurally characterize the protein complexes, since these protein-protein interaction seem to be reversible and each protein could interact with multiple structurally distinct partners, new techniques like the small angle X-ray scattering (SAXS) combined with high resolution crystal structures could help to efficiently establishe architectures, assemblies, conformations, and unstructured regions for proteins and protein complexes in solution. (Tsutakawa et al., 2007)

Table 2.1: Proteins of *Helicobacter pylori cag*-PAI

Protein name: name of the open reading frame in the *H. pylori* CCUG 17874, J99 and 26695 strains; Function: information on function, localization of the protein. Homologue: name of the homologue in *A. tumefaciens*, if existing. Constructs: name of the construct used in the project: SS (Signal Secretion sequence); His^{N}- (N-terminal His_{tag}) [pET28b vector, Novagen; pET151/D-Topo cloning, Invitrogen]; -^{C}His (C-terminal His_{tag}) [pET20, pET21 (Novagen)]; MBP- (Maltose-Binding Protein) [pMal vector, NEB]; GST- (Glutathione S-transferase) [pGEX vector, GE Healthcare); ΔN (Without a N-terminal Domain); ΔC (Without a C-terminal Domain); ΔSS (Without a Signal Secretion sequence). Status: current status of the project: Cn (Cloned); Ep (Expressed); Sp (Soluble protein); Pp (Purified protein); Cy (Crystals); Sv (Structure solved).

Table 2.1. Proteins of *Helicobacter pylori cag*-PAI.

Protein Name	Function	Homologue	Constructs	Status	Comment
CagA/26 (HP0547)	Toxin				
CagB					
CagC/25 (HP0546) [ΔSS: 30-115 aa]	Pilin	VirB2	SS-CagC-His $CagC_{\Delta SS}$-His His-CagC MBP-$CagC_{\Delta SS}$	Cn	Toxic: expression no detectable
CagD/24 (HP0545) [ΔSS: 33-209 aa]	Periplasmic protein		His-$CagD_{\Delta SS}$	Sv	
CagE/23 (HP0544)		VirB4			
CagF/22 (HP0543) [ΔC: 1-149 aa] [ΔN: 171-268 aa]	CagA Chaperone		GST-CagF His-$CagF_{\Delta C}$ $CagF_{\Delta N}$-His	Pp	Crystallisation trials
CagG/21 (HP0542) [ΔSS: 28-142 aa]	Adhesin		His-$CagG_{\Delta SS}$ $CagG_{\Delta SS}$-His MBP-$CagG_{\Delta SS}$	Ex	Low expression and protein degradation
CagH/20 (HP0541)	Hypothetical Membrane protein				
CagI/19 (HP0540) [ΔSS: 20-383 aa]	Hypothetical Membrane protein		His-$CagI_{\Delta SS}$	Ex	Inclusion bodies and degradation
CagL/18 (HP0539) [ΔSS: 21-237 aa]	Adhesin		His-$CagL_{\Delta SS}$ GST-$CagL_{\Delta SS}$	Pp	Crystallisation trials
CagM/16 (HP0537) [ΔSS: 21-376 aa]	Hypothetical Membrane protein		His-$CagM_{\Delta SS}$ GST-$CagM_{\Delta SS}$ MBP-$CagM_{\Delta SS}$	Pp	Aggregation
CagN/17 (HP0538) [ΔSS: 25-306 aa] [ΔSS + ΔC: 25-216 aa]			His-$CagN_{\Delta SS}$ $CagN_{\Delta SS+\Delta C}$-His	Pp	Crystallisation trials
CagP/15 (HP0536)					
CagQ/14 (HP0535)					
CagR					
CagS/13 (HP0534)			CagS-His	Sv	PDB: 2G3V

CagT/12 (HP0532) [ΔSS: 22-164 aa]	Lipoprotein	VirB7	GST-CagT$_{\Delta SS}$ MBP-CagT$_{\Delta SS}$ CagT$_{\Delta SS}$-His CagT$_{\Delta SS+\Delta(??-??)}$-His	Ex	Aggregation
CagU/11 (HP0531)	Hypothetical Membrane protein		GST-CagU CagU-His	Cn	Low expression and Insoluble
CagV/10 (HP0530) [ΔN1: 59-252 aa] [ΔN2: 68-252 aa] [ΔN3: 98-252 aa]	Hypothetical Membrane protein	VirB8	GST-CagV His-CagV His-CagV$_{\Delta N1}$ His-CagN$_{\Delta N2}$ His-CagN$_{\Delta N3}$	Cy	Crystallization trials and Se-Met preparation
CagW/9 (HP0529)	Hypothetical Membrane protein				
CagX/8 (HP0528) [ΔN: 27-522 aa]	Hypothetical Membrane protein	VirB9	GST-CagX $_{\Delta N}$ CagX $_{\Delta N}$ PelB-CagX $_{\Delta N}$	Ex	Insoluble
CagY/7 (HP0527) [ΔN: 1690-1927 aa]	Hypothetical Membrane protein	VirB10	His-CagY$_{\Delta N}$	Cn	Low expression and degradation
CagZ/6 (HP0526)			GST-CagZ	Sv	PDB: 1S2X
Cagα (HP0525)	ATPase	VirB11			PDB: 1NLY
Cagβ/5 (HP0524)	ATPase	VirD4			
Cagγ/4 (HP0523) [ΔSS: 17-169 aa]	Lytic transglycosylase	VirB1	GST-SS-Cagγ MBP-SS-Cagγ PelB-Cagγ -His SS-Cagγ -His His-SS-Cagγ MBP-Cagγ$_{\Delta SS}$ Cagγ Cagγ$_{\Delta SS}$	Ex	Insoluble
Cagδ/3 (HP0522)	Hypothetical Membrane protein		His-Cagδ$_{\Delta N}$		Toxic: expression no detectable
Cagε/2 (HP0521)					
Cagζ/1 (HP0520)	Hypothetical Membrane protein		His-Cagζ Cagζ -His		Toxic: expression no detectable

Chapter 3

The crystal structure of CagS from the *Helicobacter pylori* pathogenicity island

Contents

This chapter has been adapted from:
Laura Cendron [1,3], Elisabetta Tasca [2,3], Tommaso Seraglio [1], Anke Seydel [1,3], Alessandro Angelini [1,3], Roberto Battistutta [1,3], Cesare Montecucco [2,3] and Giuseppe Zanotti [1,3]; (2007). The crystal structure of CagS from the *Helicobacter pylori* pathogenicity island. *Proteins.* 69(2):440-3.

[1)] Department of Chemistry, University of Padua, Institute of Biomolecular Chemistry, CNR Padua, Via Marzolo 1, 35131 Padua, Italy,
[2)] Centre of Biomembranes, CNR, and Department of Medical Science, University of Padua, Via G. Colombo 3, 35121 Padua, Italy,
[3)] Venetian Institute of Molecular Medicine (VIMM), Via Orus 2, 35129 Padua, Italy.

Coordinates have been deposited at the Protein Data Bank (http://www.rcsb.org) for immediate release upon publication (PDB ID code 2G3V).

3.1 Introduction

Helicobacter pylori chronically infects the gastric mucosa of the majority of the human population (Covacci *et al*., 1999; Montecucco *et al*., 2001) and is associated to the development of severe gastroduodenal diseases in a sizeable minority of patients (Moss *et al*., 2003). Although the molecular and cellular pathogenesis are not well understood, it has been demonstrated that *H. pylori* strains harboring the *cag* pathogenicity island (*cag*-PAI) are more pathogenic (Covacci *et al*., 1999). The *cag*-PAI is a 37 kb gene cluster that encodes a type IV secretion system (T4SS), similar to the VirB/D4 T4SS of the plant pathogen *Agrobacterium tumefaciens* and other Gram-negative bacteria, plus a yet undefined number of proteins, including the secreted toxin CagA (cytotoxin associated gene A) (Covacci *et al*., 1999; Censini *et al*., 1996). Until now, six out of the twenty-nine putative proteins encoded by *H. pylori* cag-PAI have been recognized to be VirB/D4 structural or functional homologues. Little is known about the function of the proteins encoded within the *cag* region, especially those without VirB/D4 homology. Systematic knock-out experiments in *H. pylori* have demonstrated that the absence of most of the *cag*-PAI genes abolish CagA translocation and phosphorylation, as well as the secretion of IL-8 (Fischer *et al*., 2001; Selbach *et al*., 2002). One approach towards the elucidation of the function of the Cag-PAI proteins is the determination of their three-dimensional structures, which are known only for Cagα (HP0525), which is an ATPase located on the inner membrane (Yeo *et al*., 2000), and for CagZ (HP0526), a 23 kDa protein involved in the translocation of CagA (Cendron *et al*., 2004).

CagS (Cag13, HP0534 or JHP0482) is a 23 kDa protein encoded by a well-conserved gene within the *cag*-PAI. The gene is located immediately after the cluster of *cag* genes whose putative products show homologies with the VirB proteins that define the structural core of T4SS. It retains this location even after the frequent insertion events by IS605 sequences and is observed in several unrelated *H. pylori* strains, suggesting a relevant role for it. A very recent study by 2D Difference Gel Electrophoresis gave the first experimental evidences that CagS is expressed in *H. pylori* cultures *in vitro* (Busler *et al*., 2006).

Primary sequence analyses do not show any strong CagS similarities with proteins of known function, except for some weak similarities with components involved in the peptidoglycan biosynthesis, belonging to the FemABX family of enzymes (Hegde *et al*., 2001). Here, we describe the determination of the crystal structure of CagS, which represents a new protein fold.

3.2 Materials and Methods

Cloning, expression and purification

The *cag*S gene (also referred to as *cag*13, HP0534 or JHP0482) was amplified from genomic DNA of *Helicobacter pylori* G27 strain and inserted into the pET-21b vector (thus adding a C-terminal His-tag), generating construct pET21-*cag*S. *E. coli* BL21 (DE3) harbouring pET21-*cag*S was grown at 37 °C in LB medium supplemented with ampicillin (100 μg ml^{-1}) with vigorous shaking; at an OD_{600nm} of 0.6, expression of CagS-His_6 was induced by the addition of IPTG (0.5 mM final concentration), and the cultures incubated for a further 4 h at 30 °C under mild shaking. The bacteria were harvested by centrifugation and stored at -80 °C.

The frozen bacteria were suspended in 50 ml of phosphate buffer (0.05 M Na_2HPO_4, 0.5 M NaCl, pH 7.3), supplemented with a protease inhibitor cocktail (Complete Mini EDTA Free, Roche), and disrupted twice using a French press. Lysates were centrifuged to sediment the insoluble fraction. The recovered soluble fraction (containing at least 50 % of the total expressed protein) was loaded onto a Ni^{II}-NTA Agarose column (GE Healthcare) equilibrated with phosphate buffer. After extensive washing in the same buffer supplemented with 0.02 M imidazole, the protein was eluted in a single peak by applying an imidazole gradient (0.02–0.50 M). Further purification was obtained by size exclusion chromatography with a Superdex 75 16/60 column (GE Healthcare) equilibrated with Hepes buffer (0.03 M Hepes pH 7, 0.15 M NaCl). The eluted fractions containing the CagS-His_6 protein were collected and concentrated to 8 mg ml^{-1} in Hepes buffer.

The Se-Met-substituted protein was produced by metabolic inhibition of methionine. *E. coli* BL21 (DE3) harbouring pET21-*cag*S were grown and induced as before. The bacteria were suspended in Tris buffer (0.03 M Tris pH 7.3, 0.15 M NaCl) and disrupted twice by French press. The Se-Met derivative of CagS-His_6 was mostly insoluble and found almost entirely in inclusion bodies, most likely because of the high number of methionines (15 in 207 amino acids). The inclusion bodies were washed, recovered by centrifugation, and solubilised at room temperature overnight in denaturing buffer (6 M GdCl, 0.03 M Tris pH 7.3, 0.15 M NaCl) to obtain a concentration of about 5 mg ml^{-1} of denatured CagS-His_6. The residual aggregates and solid particles were removed by prolonged centrifugation and filtration (0.45 μm). The supernatant was loaded on a Ni^{II}-NTA Agarose resin (GE Healthcare), previously equilibrated with denaturing buffer. After extensive washing with denaturing buffer,

Se-Met CagS-His_6 was renatured in the column by slow equilibration with Tris buffer supplemented with 0.1 % Triton X-100, washed with a high ionic strength Tris buffer (0.5 M NaCl), and finally with Tris buffer with 5 mM β-cyclodextrin. Se-Met CagS-His_6 was eluted by applying an imidazole gradient (0-0.5 M imidazole). The soluble fraction was eluted in a single peak and further purified by size exclusion chromatography using the same conditions as for the native protein. The resulting solution was concentrated to obtain an 8 mg ml^{-1} stock solution for crystallization trials. Se-Met incorporation into the CagS-His_6 recombinant protein was confirmed by mass spectrometry (ESMS), which yielded a molecular mass of the prevalent species coherent with a high amount of selenium atoms oxidation. Circular dichroism spectrum of the native protein in the 195-260 nm region shows a clear prevalence of α-helix secondary structures. The Se-Met CagS-His_6 CD spectrum superposes well with that of the CagS-His_6 protein, indicating that both refolded and native proteins possess the same secondary structure.

Growth of *H. pylori*

H. pylori strain G27 was grown for a day at 37 °C on *Campylobacter agar base* (Oxoid) containing 7 % lysed horse blood and supplemented with cefsulodin (6 μg ml^{-1}), amphotericin (5 μg ml^{-1}, vancomycin (10 μg ml^{-1}), harvested with a loop, resuspended and washed in 0.1 vol. of 25 mM Tris pH 7.9. For the localisation experiment, bacteria were passed twice through a French press, and the lysates were ultracentrifuged for an hour to separate membranes from soluble parts. For the trypsin experiment, aliquots were incubated for 30 min on ice with increasing amounts of trypsin (0-50 μg ml^{-1}). Protease inhibitors were added to stop the reaction, and samples were examined by SDS-PAGE immunoblotting with the C-terminus antibody.

SDS-PAGE and Western immunoblotting

Protein samples were resuspended in loading buffer and separated by SDS-PAGE using standard methods (Ausubel *et al.*, 1987). Following SDS-PAGE, the resolved proteins were transferred electrophoretically onto a nitrocellulose membrane (Hybond™, GE Healthcare) and incubated with primary antibody in blocking reagent for 1 h at room temperature. The membrane was washed three times in PBS-Tween (0.15 % Tween 20), and then incubated for 1 h with secondary antibody (anti-mouse IgG or anti-rabbit IgG, coupled to horseradish peroxidase) diluted in blocking reagent. The membrane was washed as above and developed using the SuperSignal® chemilumenescent substrate (Pierce). To be able to detect CagS, polyclonal antibodies

were produced against two synthetic peptides, one for the N-terminus (14-26: ANSKDKKEKLIES), and the other for the C-terminus (187-199: DLEKYMEEKGVQNA). The sequences were chosen with the help of *ProSci antigens design assistance* to maximize the antigenicity coupled with hydrophobicity and solvent accessibility.

Protein crystallization

Tiny regular crystals were obtained in 3-4 days after incubating 1-2 µl of 5 mg ml^{-1} CagS-His_6 solution mixed with an equal volume of precipitant (0.2 M KBr, 25 % PEG 2000 MME, 0.1 M Tris pH 8.5) and left to equilibrate at 25 °C against the same precipitant. The crystals obtained were difficult to reproduce, most probably due to protein degradation. The best results yielded crystals of about 0.02 x 0.02 x 0.08 mm, which diffracted till 2.9 Å resolution. The Se-Met CagS-His_6 derivative crystals were grown in the same conditions, after small adjustments of protein and precipitant concentrations. Se-Met derivative protein produced higher quality crystals, which diffracted to 2.3 Å resolution.

Data collection, structure determination and refinement

X-Ray diffraction data were collected at the ESRF beamline ID29 (Grenoble, France) at 100 °K, without cryoprotectant. Crystals of a native data set measured at 2.9 Å belong to space group $I4_1$, with cell parameters **a** = **b** = 123.1 Å, **c** = 55.7 Å, and two molecules per asymmetric unit. A multiwavelength anomalous dispersion experiment was performed on different zones of one single Se-Met crystal. Three sets of data around the selenium K-edge and a fourth at 1Å were collected and the corresponding statistics are summarized in Table I. Data were processed and scaled with MOSFLM (Leslie, 1992) and SCALA (CCP4 suite, 1994). The Se-Met protein crystals belong to space group $P4_3$ (after structure determination), with **a** = **b** = 123.7 Å, **c** = 55.7 Å. The asymmetric unit contains four molecules, with a solvent content of about 46 % of the crystal volume and a total of 56 Se atoms. The correct solution was obtained using the HKL2MAP software (Schneider *et al*., 2002; Pape *et al*., 2004). The four molecules in the asymmetric unit were built manually by the XtalView graphic program (McRee, 1999). A two-fold non-crystallographic axis relates the monomers A and B to C and D, respectively. Only 9 out of 14 Se atoms per chain were assigned, since 5 of them belong to disordered N or C-terminal portions. With the exception of the final cycles of refinement, only one chain was refined and the other three were generated by transformation matrices. By analyzing the statistics of the intensities, a partial

merohedral twinning was detected with a twinning fraction of 0.22 and the twinning operation (h,-k,-l). All refinement was consequently carried out using the twin option contained in the package CNS (Brunger *et al*., 1998). Several cycles of model building and refinement reduced the crystallographic R factor to 21.9 % (R_{free} 27.9 %) for all data from 39.4 Å to 2.3 Å resolution. Residues from 21 to 187 for each chain are clearly visible in the electron density. Some electron density, not connected with the rest of the molecule, is visible in the region close to the loop connecting helix H to I. Two helical turns were built into this region, but we were not able to define its exact amino acid sequence (see the Results section for discussion).

3.3 Results and Discussion

CagS forms a compact single domain protein, with an all-α structure. Ten α-helices, labeled from A to J, can be distinguished (Fig. 3.1A).

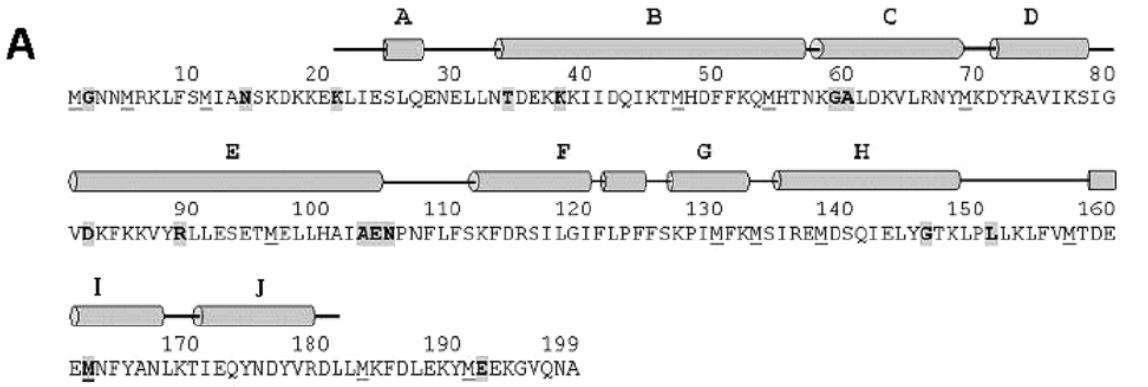


Figure 3.1A. CagS amino acid sequence and secondary structure. The Methionines are underlined. The positions mutated in the sequenced strains of *H. pylori* are in bold and enclosed in grey rectangles. The locations of the ten helices (from A to J) are indicated by cylinders. Numbering refers to the wild type CagS sequence.

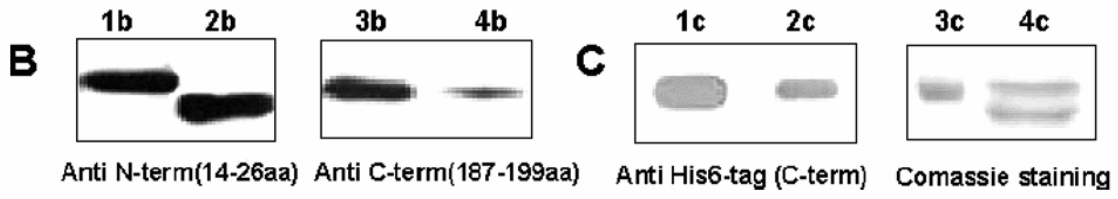


Figure 3.1B and **3.1C**. Western blotting analysis with anti N-term (14 – 26 aa: 1b and 2b) and anti C-term (187-199aa: 3b and 4b) anti His6-tag at the C-terminus of the recombinant protein (1c and 2c), and Comassie Blue staining (3c and 4c) of freshly purified recombinant CagS (1b, 3b, 1c and 3c) and after a long-term storage at 4 °C (2b, 4b, 2c and 4c).

Our model starts at residue K21, since the first twenty amino acids are flexible or disordered and could not be located in the electron density, with the possible exception of a few helical turns. Helices B, E, F and H, arranged in an up-and-down topology, form the structural core, whilst the short helices C, D and G on one side, and A, I and J on the other represent a sort of appendices, conferring a "peanut shape" to the overall structure (Fig. 3.2).

The ten helices that form the protein are held together mainly by hydrophobic forces, with few hydrophilic inter-helical interactions. Wild type CagS has an unusually high methionine content, fourteen in a sequence of 199 amino acids (*H. pylori* strain G27).

Nine are visible in our model, while two are at the C-terminus and three are present within the flexible N-terminal fragment. A peculiar feature of the molecule concerns four methionines that define a cluster in the 3D structure: M69, M130, M133, M138 (Fig. 3.3).

The model shows a highly charged surface, with 24 positively and 24 negatively charged residues. In particular, the tertiary structure defines a negatively charged region including several glutamate and aspartate residues (E24, E28, E30, E104, E160, E161, E172, D176, D180) confined to the portion of the molecule involving α-helix A, the nearby loop, the first and last turns of helices E and F, and the C-terminus helices I and J. Since the electron density maps allows only the sequence from K21 to L181 to be modeled, the lysine-rich N and C terminus could not be assigned in the charge distribution described here, but is in accordance with the basic isoelectric point of CagS. However, these lysine rich un-modeled N and C terminus appendages might define some positively charged

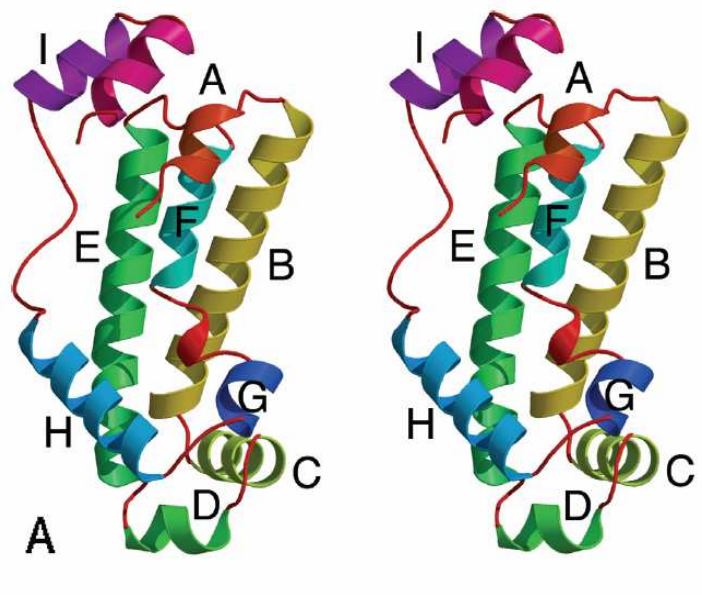


Figure 3.2. Stereo view of a cartoon of CagS monomer. Helices are labeled from A to J, from the N- to the C-terminus. All molecular drawings were obtained with Molscript program and rendered with Raster3D.

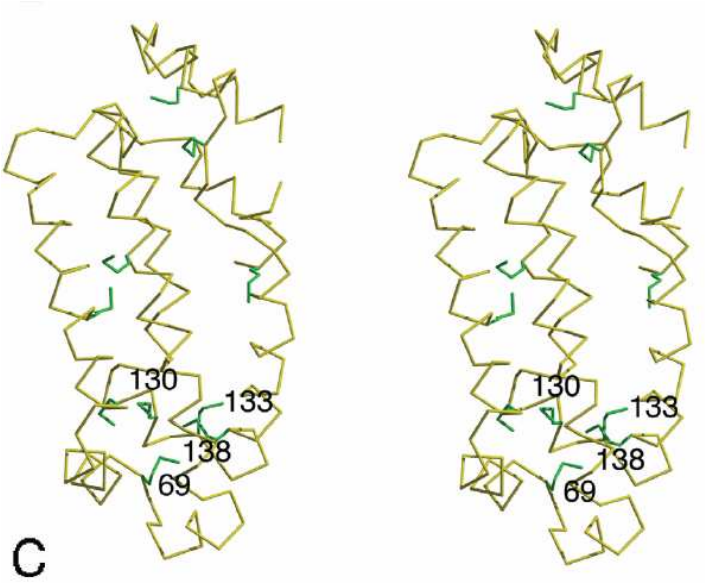


Figure 3.3. Stereo view of a portion of the electron density map. The region of the protein we were not able to connect to the rest of the chain is shown in blue; the Cα chain-trace of CagS is shown in red. The electron density, calculated as an omit-map, is contoured at 3s level.

brunches playing a potential role in Cag proteins interactions. Some electron density is still visible close to the central cavity of the "peanut". This density was tentatively interpreted in terms of two helical turns, but there is no continuity in the density that allows to connect it to either the N- or the C-terminus, that are close together (Fig. 4).
Analyses with anti-His_6-tag antibodies and with 2 others produced against two distinct fragments at the N-terminus (amino acids 14-26, Fig. 3.1A left) and the C-terminus (amino acids 187-199, Fig. 3.1A right) of CagS indicate that the recombinant protein degrades at its C-terminus, loosing the six Histidines appendage and part of the last twenty residues (Fig. 3.1B, 3.1C). This suggests that the short piece of chain that we traced in the electron density but which is not connected to the rest of the polypeptide chain probably belongs to the N-terminus.

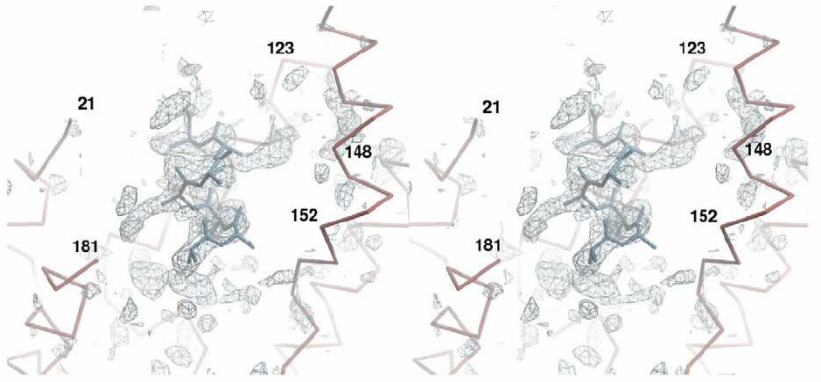


Figure 3.4. Stereo view of Cα atoms showing the positions of the Met residues of CagS. Met side chains are in green. The four methionines that form a cluster are labeled.

The results of the crystallographic model of CagS do not show any clear evidence of architectural similarity with other known structures. No significant similarities were found using classical structure comparison servers, such as DALI and SCOP (Holm *et al.*, 1993; Murzin *et al.*, 1995). The results obtained using ProFunc (Laskowski *et al.*, 2005), suggest only a weak structural homology of this protein with the phosphotransfer domain (HPt) of CheA, a histidine protein kinase that controls chemotaxis response in bacteria (Mourey *et al.*, 2001; Quezada *et al.*, 2005). However, the CagS structural proximity with Hpt domains needs further experimental support and is currently under investigation.

Table 3.1. Statistics on data processing and structure determination. Numbers in parentheses refer to the highest resolution bin.

Space group and cell parameters; $P4_3$: a = b = 123.7 Å, c = 55.7 Å

[a] Defined as $R_{symm} = S\ |I_{hi} - \langle I \rangle| / S\ I_{hi}$, where I_{hi} is the intensity observed in the ith source.

[b] 9.5% of reflections were used in test set.

[c] Defined as $R_{cryst} = S\ |\ |\mathbf{F}_{obs}(hkl)| - |\mathbf{F}_{calc}(hkl)|\ | / S\ |\mathbf{F}_{obs}(hkl)|$.

Data collection

Datasets	Peak	Inflection point	Remote high	Remote low
Wavelength (Å)	0.9792	0.9794	0.9756	1.0000
Resolution range (Å)	62.38-2.30 (2.42-2.30)	61.90-3.00 (3.16-3.00)	61.90-2.50 (2.64-2.50)	50.97-2.30 (2.42-2.30)
Unique reflections	38158 (5413)	17173 (2490)	29619 (4323)	38297(5569)
Multiplicity	5.7 (3.8)	6.1 (6.2)	7.3 (7.4)	4.8 (4.9)
Completeness (%)	99.4 (97.4)	99.8 (100)	99.7 (100)	99.9 (100)
$<I/\sigma(I)>$	4.7 (1.2)	2.2 (1.5)	5.3 (1.9)	5.0 (1.7)
R_{symm} [a)]	11.4 (45.6)	20.0 (43.6)	10.6 (36.4)	11.1 (41.3)
Mean $I/\sigma(I)$	13.0 (3.3)	10.3 (4.6)	17.0 (6.1)	12.8 (3.1)
Anomalous completeness	98.8 (94.7)	98.8 (98.5)	99.6 (100.0)	99.7 (99.9)
Phasing statistics				
Phasing resolution (Å)		99.00–2.80		
CC (weak)		56.7 (46.2)		
CC E_{calc} –E_{obs} (%)		27.6		
Refinement statistics				
Protein atoms		5556		
Solvent atoms		238		
Resolution used (Å)		40-2.30 (2.44-2.3)		
Reflections (R_{free} set) [b)]		38258 (3691)		
R_{cryst} [c)]		0.219 (0.238)		
R_{free}		0.279 (0.304)		
Rmsd from ideal values				
Bond lengths (Å)		0.009		
Bond angles (Å)		1.3		
Ramachandran plot statistics. Percentage of polypeptide torsion angles in regions:				
Most favoured		87.2		
Additionally allowed		12.0		
Generously allowed		0.8		
Overall G factor		0.3		

Chapter 4

The crystal structure of CagD from the Helicobacter pylori pathogenicity island

Contents

4.1 Introduction

Helicobacter pylori infects the stomach of about half the world's population and is associated with a spectrum of gastric diseases. A wide range of Gram-negative bacterial pathogens, including *H. pylori*, translocate virulence factors into host target cells by multisubunit transport apparatus collectively known as "type-IV secretion systems" (TFSS's) (Cascales *et al*. 2003; Backert *et al.*, 2006). These TFSSs, which typically consist of a membrane- spanning secretion channel and an extracellular pilus, act as molecular syringes. The *H. pylori* TFSS has attracted considerable interest because it is associated with the development of severe diseases, ranging from gastritis to peptic ulcer and gastric cancer (Backert *et al*., 2006; Amieva *et al*., 2004; Peek *et al*., 2002; Covacci *et al*., 2000; Hatakeyama *et al*., 2006).

The *H. pylori* TFSS, encoded by a 40-kilobase *cag* Pathogenicity Island (*cag*-PAI), injects the CagA oncoprotein into host cells, resulting in activation of NF-kB and induction of potent pro-inflammatory chemokines such as interleukin (IL)-8. Translocated CagA undergoes tyrosine phosphorylation by Src, leading to actin-cytoskeletal rearrangements, scattering and elongation of infected host cells in cell culture (Backert *et al*., 2006; Amieva *et al*., 2004; Peek *et al*., 2002; Covacci *et al*., 2000; Hatakeyama *et al*., 2006). These phenotypic changes resemble those of malignant cellular transformation and have been the subject of intensive studies.

Depending on the clinical strain, *cag*-PAI comprises from 27 to 31 genes. Eight of them present variable homologies with the TFSS of *Agrobacterium tumefaciens*, the best-studied model of the TFSS. The remaining genes are instead unique to the *H. pylori cag*-PAI, and a function has been proposed for only a few of them (Buhrdorf *et al*., 2003; Rohde *et al*., 2003; Tanaka *et al*., 2003; Couturier *et al*., 2006; Bourzac *et al*., 2006; Andrzejewska *et al*., 2006; Pattis *et al*., 2007).

Mutational studies have been an important first step in understanding the role played by each individual members of the *cag*-PAI, in the CagA translocation and interleukin-8 (IL-8) induction processes (Fischer *et al*., 2001). However, these systematic deletion studies of *cag*-PAI genes identified only four classes of mutants: those that had no effect on CagA delivery or IL-8 induction, those that blocked both CagA delivery and IL-8 induction, those that blocked CagA delivery but not IL-8 induction, and those that had an intermediate and variable effect on both (Fischer *et al*., 2001; Selbach *et al*., 2002; Bourzac *et al*., 2006).

In particular, mutations that resulted in intermediate and variable levels of CagA delivery and IL-8 induction are promising candidates for proteins important in T4SS

assembly and function. In fact, these genes (*cag*D, *cag*G and *cag*N) may encode accessory proteins required for substrate translocation (such as chaperones) or may even encode additional substrates themselves (Fischer *et al*., 2001; Bourzac *et al*., 2006). In particular, the *cag*D locus is found in a majority of clinical isolates, but little is known about its role. Disruption of the *cag*D gene was reported to have an intermediate and variable CagA delivery and IL-8 induction phenotype (Fischer *et al*., 2001).

Very recently it has been also found that CagD is secreted out of the bacterial membrane and identified in the medium (Smith *et al*., 2007). The mode by which CagD is enriched in the extracellular fraction is still unclear. CagD enrichment does not appear to be dependent on *cag*-PAI or mediated by the TFSS, since the CagA effector was present in the cell-associated samples but it was not enriched in the extracellular fraction. Considering that the cag-PAI was functional in the strain examined, it does not appear to be induced under the assayed growth conditions (Smith *et al*., 2007). This result was not unexpected since induction of the TFSS apparatus requires the bacterium contact with the host cell (Matthysse *et al*., 1987; Smith *et al*., 2007). Primary sequence analysis reveals that CagD contains a predicted signal sequence for the type II secretion, which may account for its enrichment in the extracellular medium.

Here, we report a detailed characterization of the three-dimensional structure of CagD.

4.2 Materials and Methods

Cloning

The nucleotide sequence corresponding to amino acids from 33 to 209 was amplified by PCR starting from *H. pylori* genomic DNA (strain CCUG 17874), using a high fidelity thermostable DNA polymerase (Deep Vent DNA polymerase; New England Biolabs) and two flanking primers: *hp0545*forward (5'-GCACATATGAATGACAAAGAAGCC-3') and *hp0545*reverse (5'-CTGAAGCTTATAGATATACCGCTTCAC-3'). These last were built to introduce *Nde*I and *Hind*III restriction sites at the 5' and 3' ends, respectively. The amplified *cag*D (*hp0545*) gene was double digested with *Nde*I and *Hind*III enzymes and inserted into a pET28b expression vector (Novagen), encoding an in frame His_6-tag followed by a Thrombin cleavage site at the N-terminus of the recombinant construct (His_6-CagD).

Mutagenesis

Since the amino acid sequence of CagD contains only two Met residues, one of which located at the N-terminus, a mutant with two additional Met residues was produced in order to solve the phase problem, using QuikChange® Site-directed Mutagenesis Kit (Stratagene). The mutated amino acids, V83M and V144M, were selected as the most variable amino acids in a multiple alignment of all the CagD sequences available from different *H. pylori* strains (data not shown). CagD double mutant was obtained by consecutive PCR reactions using a high fidelity thermostable DNA polymerase (*Pfu*Turbo DNA polymerase; Stratagene). Single clones were then sequenced to confirm the occurrence of the desired mutations.

Expression and Purification

E. coli BL21 (DE3) cells (Invitrogen) harvesting the plasmid pET28b-*cag*D (pET28b-*HP05454*) were grown at 37 °C in 2 l LB medium containing 100 μg ml^{-1} of kanamicin. The expression was induced at 0.7-0.8 OD_{600} by adding 1 mM isopropyl-b-D-thiogalactopyranoside (IPTG) and prolonged for 4 h at 30 °C, under vigorous shaking. The cells were harvested by centrifugation at 11.000 g for 20 min, resuspended in Buffer A (30 mM Tris pH 8.0, 150 mM NaCl) supplemented with protease inhibitor cocktail (Roche, Complete-mini EDTA-free) and disrupted by a French pressure cell (1050 psi). The soluble fraction, obtained after 25 min

centrifugation at 40.000 g, was applied onto a 1 ml His-Trap column (GE Healthcare), equilibrated with Buffer A at about 1-2 ml min^{-1}. After two extensive washes with 5 % and 10 % Buffer B (30 mM Tris-HCl, pH 8.0, 150 mM NaCl, 500 mM Imidazole), the protein was eluted in a single peak at 200 mM Imidazole, by applying a linear gradient from 10 % to 100 % Buffer B. The fractions containing the pure His_6-CagD were pooled, concentrated by ultrafiltration (5,000 MWCO, Millipore), diluted into Buffer A and incubated over-night at 4 °C with Thrombin protease to remove the N-terminal His_6-tag. The cleaved product was further purified by gel filtration (Superdex 75 Prep grade 16/60, GE Healthcare), equilibrated with buffer A. SeMet derivative was obtained for the CagD V83M-V144M double mutant by growing *E. coli* B384 auxotroph strain, harboring the plasmid pET28b-V83MV144M*cag*D, in a SeMet enriched medium, as previously described (Cendron *et al*., 2005). The SeMet derivative of CagD V83M-V144M double mutant was purified as described above, obtaining analogous yield and behavior.

The oligomeric state of purified CagD in solution was examined by analytical size exclusion chromatography using a Superdex 75 HR 10/30 column (GE Healthcare), equilibrated with Buffer A, and dynamic light scattering.

Mass spectrometry

Samples of the V83M-V144M-CagD double mutant protein stock and the SeMet derivative crystals, fished and dissolved in buffer A, were purified by reverse phase chromatography (C-18 column, Vydac) and analyzed by ESI Mass Spectrometry, to define the Selenium amount in the SeMet derivative sample and the molecular mass of the crystallized protein.

Protein crystallization

Crystals were obtained by isothermal vapour diffusion technique in two different crystallization conditions (Structure screen II n. 10, Molecular Dimension Limited and PACT Screen n. 34, Qiagen). Despite the first precipitant induced the formation of highly regular hexagonal crystals and the second one rod shaped twinned and/or disordered crystals, only the latter produced a weak diffraction pattern to 3 Å resolution. Despite many optimization trials were performed and slightly different crystallization conditions were discovered, no significant improvements were obtained. The V83M-V144M-CagD double mutant, produced to enrich the sequence in Methionines, allowed not only the introduction of a sufficient number of anomalous scatterers, but also the improvement of the quality of the crystals. The best datasets

were collected at a maximum resolution of 2.2 Å. Moreover, by adding 0.02 M chloride salts of Cu^{II} or $Strontium^{II}$ to the precipitant (PACT Screen n. 34, Qiagen), hexagonal crystals were obtained.

Structure determination

Diffraction data were measured at beamline ID14-3 of ESRF. Crystals of selenomethionine-substituted CagD double mutant, grown in the absence of cations, belong to the C2 space group, with cell parameters corresponding to **a** = 81.2 Å, **b** = 117.9 Å, **c** = 66.2 Å and β = 110.6°. Datasets were processed by using program MOSFLM (Leslie, 1992) and merged by using SCALA (Evans, 1997), included in the CCP4 Program Suite (Collaborative Computational Project, Number 4, 1994; Table 4.1). A V_M value of 2.67 $Å^3$ Da^{-1} is compatible with the presence of 3 monomers per asymmetric unit, corresponding to a solvent content of about 54 % of the cell volume. Eleven of the twelve methionine sites belonging to three untagged molecules were found in the asymmetric unit by the SHELXD software (Schneider and Sheldrick, 2002) and phases calculated and refined by using the SHARP/autoSHARP suite. Residues from Asp51 to Tyr180 were fitted by using the program COOT (Emsley and Cowtan, 2002). The structure was refined against a remote dataset at 2.2 Å resolution using software programs CNS (Brunger *et al*., 1998), REFMAC (Murshudov *et al*., 1999) and SHELX, alternated by several cycles of manual adjustment and rebuilding. The final model, which includes 3108 protein atoms 253 solvent molecules, was refined to a conventional crystallographic R factor of 0.237 (R_{free} = 0.306). Stereochemical and geometrical parameters are as expected at this limit of resolution.
Data of crystals grown in the presence of Cu^{II} were collected at beamline ID23-1 and ID23-2 of ESRF and processed as described before. They belong to the space group $P6_222$, with **a** = 65.03 Å and **c** = 156.16 Å. The V_M value is 2.3 $Å^3Da^{-1}$, corresponding to a solvent content of about 45 %. The structure was solved by molecular replacement, using the program Phaser and refined. Statistics on the final models are reported below in Table 4.1.

Table 4.1. Crystallographic structure determination and model analysis

Data collection				
Datasets	**Peak (crystal I)**	**Remote (crystal I)**	**Remote (crystal II)**	**Cu adduct**
Wavelength (Å)	0.9794	0.9742	0.9740	0.9762
Space group, cell parameters (Å, °)	C2, **a**=81.22, **b**=117.90, **c**=66.52, β=110.6		C2, **a**=81.22, **b**=117.90, **c**=66.52, β=110.6	$P6_222$, **a**= **b**= 65.03, **c**= 156.16 γ= 120
Resolution range (Å)	63.37-2.50 (2.64-2.50)	63.89-2.20 (2.32-2.20)	62.26-2.32 (2.32-2-20)	57.34-2.80
Unique reflections [1]	19,417 (2,852)	28,578 (4,165)	30507 (4107)	
Multiplicity			7.3 (7.4)	
Completeness (%)	96.7 (96.8)	96.3 (96.5)	99.2 (99.9)	99.9
<I/σ(I)>	9.0 (2.2)	7.1 (1.8)	16.6 (4.1)	
R_{symm}	0.067(0.345)	0.066 (0.396)	0.088 (0.411)	
Anomalous phasing	1.66	1.23	-	-
Figure of merit Centric/acentric	0.361/0.125			
Refinement statistics				
Resolution Used (Å)		10 – 2.2		56.3 – 2.75
Reflections (R_{free} set)		27780 (2778)		5322 (252)
R_{cryst}		0.237		0.284
R_{free}		0.306		0.323
Rmsd from ideal values				
Bond lengths (Å)		0.005		0.036
Bond angles		0.020 Å		3.2 °

[*] Values in parentheses are for outer resolution shell

‡ $R_{sym} = \sum_{hkl} |I_{hkl} - \langle I_{hkl} \rangle| / \sum_{hkl} I_{hkl}$

§ R factor = $\sum_{hkl} ||F_o| - |F_c|| / \sum_{hkl} |F_o|$ where $|F_o|$ and $|F_c|$ are the observed and calculated structure factor amplitudes for reflection hkl, applied to the work (R_{cryst}) and test (R_{free}) (5% omitted from refinement) sets, respectively

4.2 Results and Discussion

Protein cloning and expression

Since the full-length CagD sequence (HP0545, JHP0493, Cag24) was predicted to include a N-terminal secretion signal with high confidence (SignalP 3.0; Bendtsen *et al*., 2004), the first 32 amino acids were excluded from the recombinant construct (His_6-CagD). The recombinant protein was expressed in *E. coli* and purified using metal-affinity chromatography followed by gel permeation chromatography. As indicate by both analytical gel permeation chromatography and dynamic light scattering techniques the protein behaves in solution as a dimer. The dimer is characterized by a covalent intermolecular disulfide bridge, as proven by comparison of SDS-PAGE results in denaturing, reducing and non-reducing conditions (data not shown). Moreover it coexists with the corresponding monomer in the first purification steps, but slightly enriches until the complete conversion from monomeric to dimeric state.

The molecular model

The crystal structure of CagD has been solved in two different crystal forms, a monoclinic one that diffracts to a maximum resolution of 2.2 Å, and a hexagonal crystal form that diffracts to 2.6 Å. In both cases CagD is a homo-dimer, where the two monomers are covalently linked by a disulphide bridge (Fig. 4.1B). In the crystal structure of the monoclinc space group three monomers are present in the asymmetric unit. Two of them form a dimer, whilst the third is the monomer of a dimer generated by the crystallographic two-fold axis. In the hexagonal crystal form only one monomer constitutes the asymmetric unit, and a crystallographic two-fold axis generates the second monomer of the dimer.

The electron density for the polypeptide chain of the CagD monomer that is visible in the crystal comprises amino acids from 51 to 180. The polypeptide chain folds as a single domain, composed of five β-strands and three α-helices. The five β-strands are organized in an anti-parallel β-sheet, flanked by the three α-helices on one side.

The five strands, labeled from A to E in Fig 1B, are all contiguous and include residues 83-87 (A), 97-103 (B), 111-118 (C), 122-129 (D) and 131-135 (E). The electron density for them is very well defined, with the exception of the six residues from 89 to 95, a loop that connects strands A and B, which is disordered and that was not included in the final model. The C-terminal portion of the chain includes two

relatively long α-helices (helix II, residues from 139 to 150 and helix III, from 158 to 173), running anti-parallel to each other, and a final strand, from 174 to 180 (Fig. 4.1A). The latter protrudes form the core of the monomer and runs anti-parallel to the same strand of a second monomer, favoring the formation of a dimer. The surface of interaction involves also portion of chains D and E of the two monomers, which are held together not only by the S-S bridge between two Cys176, but also by hydrogen bonds between main chain atoms of the two strands, and among side chains of other residues, for a total of ten H-bonds (Fig. 1B).

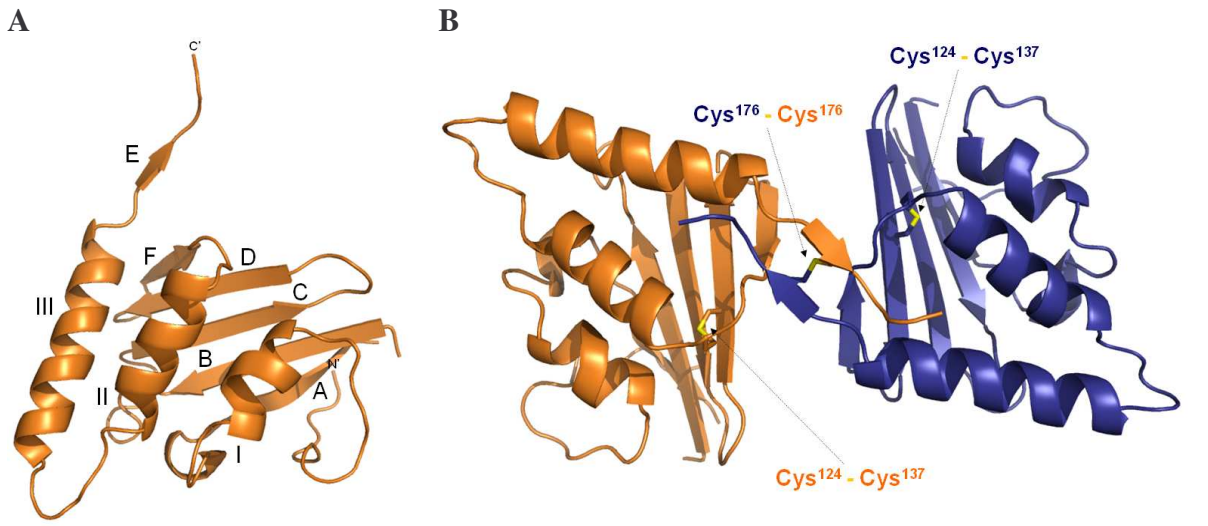


Figure 4.1. Crystal structure of CagD. (A) Ribbon representation of the CagD donomer (51-180 aa) with secondary structure elements labeled. (B) Dimeric arrangement of CagD with the two subunits coloured in orange and blu, respectively, whereas the S-S bridges are coloured in yellow.

A second disulphide bridge is present between Cys124, belonging to strand D, and Cys137, close to the beginning of the second α-helix (Fig. 4.1B). More complex is the architecture of the N-terminal part of the protein: strand A is connected to the α-helix I (residues 65-74) by a turn and a very short α-helix (residues 77-81). Finally, a long stretch (residues from 55 to 64) reverses the direction of the polypeptide chain, in such a way that residues from 51 to 56 form a strand which runs antiparallel to β-strand A. Electron density stops around residue 50, and, moreover, for the first ten residues, from 51 to 60, is not so well defined as it is for the rest of the molecule, suggesting a slightly higher mobility of this part of the protein with respect to the molecular core.
Mass spectrometry data of a sample obtained by dissolving some crystals have demonstrated that the protein we have crystallized undergoes a proteolytic cleavage at residue 50 in one of the two monomers, a phenomenon that was also observed *in vivo*

(data not shown). Since one third of the monomers in the crystal are equivalent (and they are all equivalent in the hexagonal space group form), we have to conclude that the first 50 amino acids, that are present in only one of the two monomers, are disordered and cannot be seen in the electron density.

It interesting to notice that the formation of the dimer produces a sort of large crevice in between the two monomers (Fig. 4.2). The truncated N-terminus is on the same side of the model, so that we could speculate that these 50 amino acids, when present, could partially fill this cavity. In fact, some unexplained electron density at a contour level of 1σ is visible in this crevice. This holds only for the dimer present in the asymmetric unit, whilst in the dimer generated by the crystallographic symmetry this effect is possibly cancelled by the symmetry.

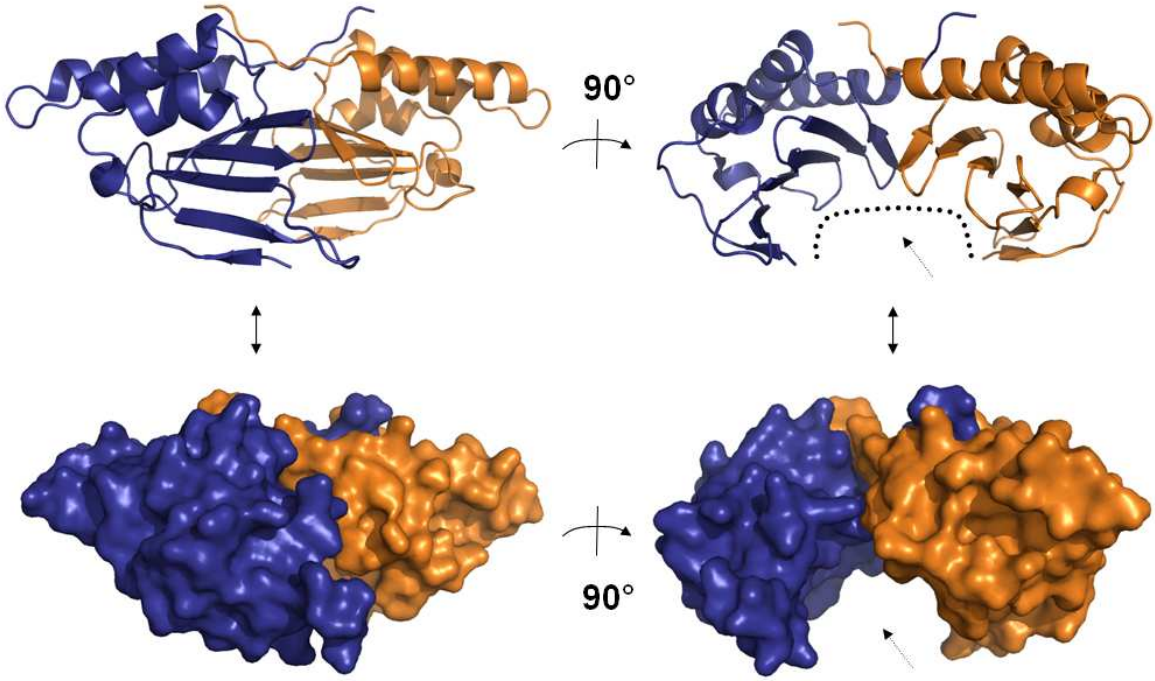


Figure 4.2. On the left, ribbon diagram and surface rendered image of CagD dimer in the same orientation. On the right, a ribbon diagram and a surface rendered image of CagD dimer equivalent to a 90° rotation from the left view. The deep groove present between the two monomers is indicated with an arrow.

The Copper adduct

When crystallized in the presence of Cu^{II} or Zn^{II}, crystals grow with a different morphology. We were able to measure a data set of the crystal grown in the presence of Cu^{II}, despite at relatively low resolution (2.75 Å). The refined molecular model shows that the protein is unchanged with respect to the other form. The only significant difference is represented by the presence of a Cu^{II} ion close to the

molecular surface, in a small cavity formed by the loop connecting strands B and C, the loop connecting helices II and III and the turn of helix connecting strand A to helix I. The ion shows a distorted tetrahedral coordination, given by two His residues, His79 and His160, by a water molecule and one oxygen of Glu106 of a symmetry-related molecule in the crystal (Fig. 4.3).

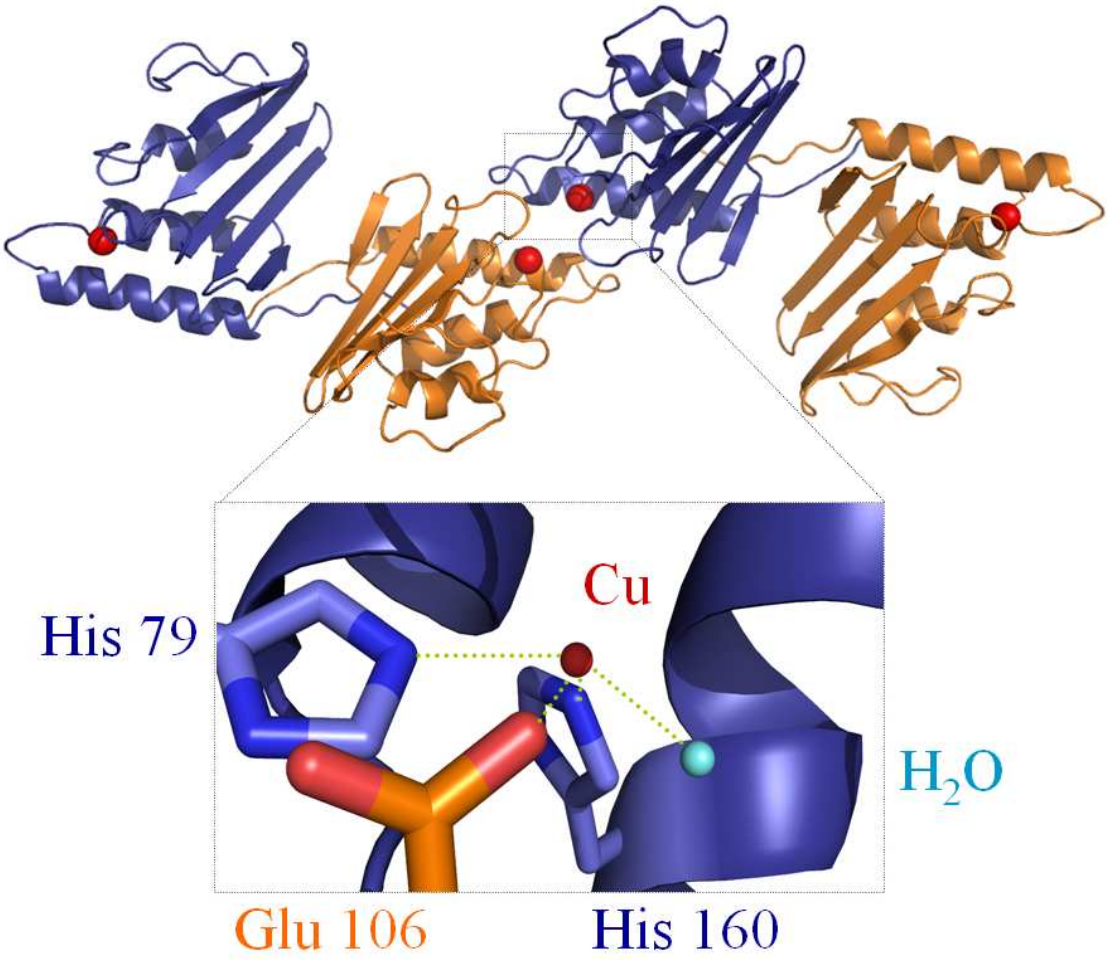


Figure 4.3. Two CagD dimers Lower panel: the copper-binding domain. A red sphere indicates the approximate position of a Cu ion modeled to adopt a distorted tetrahedral coordination geometry between one molecole of water (cyan), His79 (blu) and His160 (blu) of one monomer, and Glu106 (orange) of a simmetry-related molecole in the crystal.

This interaction strongly influences the crystal packing: since the Cu^{II} ion position is on the opposite side of the molecule with respect to the dimerization surface, long rows of dimers are present in the hexagonal crystal form (Fig. 4.3 and 4.4).

This aggregation takes place also at relatively concentrated solutions, as demonstrated by the electron micrograph reported (data not shown). It is hard to say if this cationic site, represented by the two histidine residues, has some functionality on the isolated

molecule or if its function is related to the aggregation process. Moreover, we do not know if it has some relevance *in vivo*. To verify it, several studies have been planned and are still under development in order to characterized the CagD oligomerization in the presence of Cu^{II} *in vitro* (SAXS, ITC and Electron Microscopy) and *in vivo* (Immunofluorescence, 1D and 2D native gel).

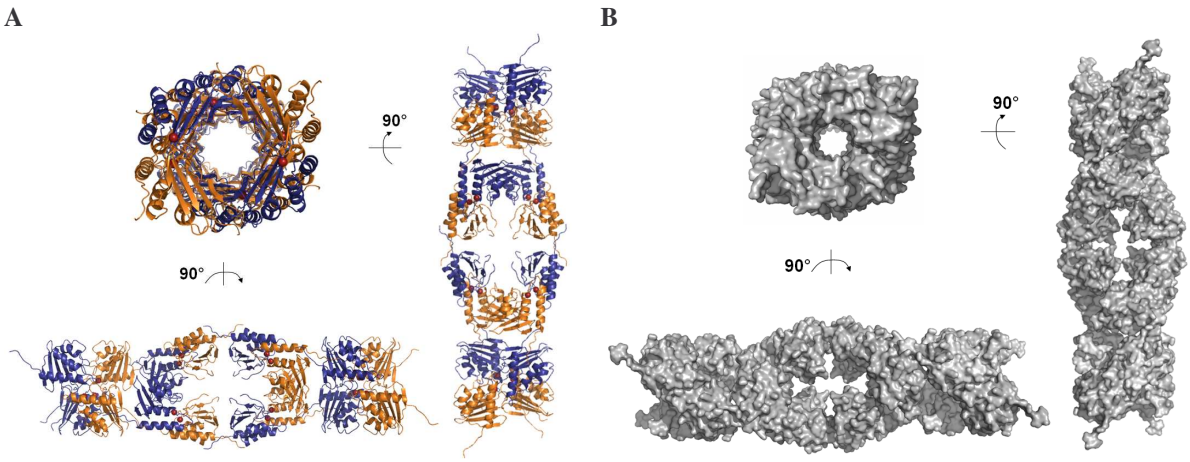


Figure 4.4. Long rows of dimers that are present in the hexagonal crystal form.

Localization of CagD

Recently, by growing two *H. pylori* strains in a defined serum-free medium and directly analyzing the proteins present in the culture supernatants by lC-MS/MS, it was found that CagD is enriched in the extracellular fraction when compared to soluble cell-associated protein sample (Smith *et al*., 2007).

To verify the presence of CagD in the extracellular supernatant (culture supernatant), some preliminary studies have been achieved in collaboration with Dr. Markus Stein lab of the Department of Medical Microbiology and Immunology, University of Alberta, Canada. Preliminary immunofluorescence studies, using WT strain G27 stained with the anti-CagD antibody as well as the serum from a *H. pylori* infected human patient, demonstrated that CagD protein is not present in the outer membrane but instead seem to be localized in the periplasmic space. On the other hand, some immunoprecipitation and acetone precipitation experiments on overnight *H. pylori* culture samples, using anti-CagD antibody (α-CagD), demonstrated that CagD is enriched in the extracellular milieu. To elucidate its precise role and subcellular localization *in vivo*, new experiments have been planned and are currently underway.

Chapter 5

Expression, purification, crystallisation and preliminary X-ray analysis of CagV C-terminal periplasmic domains, a putative VirB8 homologue protein of the *Helicobacter pylori*'s Type IV Secretion System

Contents

5.1 Introduction

Helicobacter pylori is a bacterial pathogen that colonizes the gastric mucosa of more than half of the human population (Covacci *et al*., 1999). In most individuals the infection induces chronic superficial gastritis, a condition that will remain throughout life. However, in some cases, more severe outcomes of the infection may develop, such as peptic ulcer disease, mucosa-associated lymphoid tissue lymphoma, and gastric carcinoma (Peek *et al*., 2002).

One well established *H. pylori* virulence factor is the presence of a cluster of about 30 genes, known as the *cag* pathogenicity island (*cag*-PAI). The *cag*-PAI encodes a type IV secretion system (TFSS), a multi-molecular complex that mediates the translocation of bacterial factors like the CagA protein and peptidoglycan into the cytoplasm of infected cells (Bourzac *et al*., 2005). The TFSSs are used by several human pathogen gram-negative bacteria, including *Bartonellla henselae*, *Bordetella pertussis*, *Brucella suis*, *Helicobacter pylori* and *Legionella pneumophila*, as well as the plant pathogen *Agrobacterium tumefaciens*, to deliver virulence factors into host cells (Cascales *et al*., 2003). In many cases, the existence of a type IV secretion system (TFSS) is inferred by the presence of twelve proteins that can be identified as homologues of the VirB1-11 and/or ViD4 proteins of the best described *A. tumefaciens* T-DNA transfer system (Christie *et al*., 2005). Fischer and colleagues identified 14 genes within the *cag* pathogenicity island that are necessary for building up the secretion apparatus (Fischer *et al*., 2001), but differently to some other TFSSs, the *cag* system of *H. pylori* does not codify for high homologues to the VirB proteins; in fact, only eight of them display, even if low, sequence similarities to VirB1 (Cagγ), VirB4 (CagE), VirB7 (CagT), VirB9 (CagX), VirB10 (CagY), VirB11 (Cagα), and VirD4 (Cagβ) of *A. tumefaciens*. Recently, on the basis of biochemical and bioinformatics studies, other two proteins were proposed to own a similar role of VirB2 (CagC) and VirB8 (CagV) proteins in *H. pylori* (Andrzejewska *et al*., 2006; Buhrdorf *et al*., 2003).

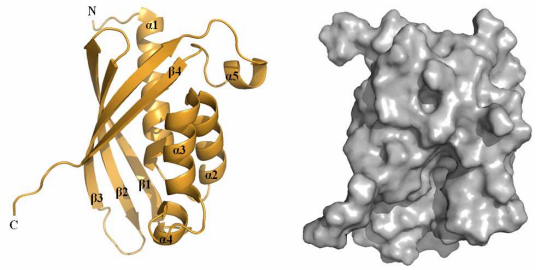
The VirB8-like proteins are essential and conserved components of the TFSSs and provide the structural and functional link between the cytoplasmic and inner membrane-associated energizing proteins, the other core elements, and the outer membrane-localized pilus assembly complex (Baron, 2006; Christie *et al*., 2005).

The crystal structure of the periplasmic domains of two VirB8 proteins, from *A. tumefaciens* (pVirB8^{AT}; residues: 92–231) and from *B. suis* (pVirB8^{BS}; residues: 97–235) were recently determined (Bailey *et al*., 2006; Terradot *et al*., 2005) [Fig. 5.1].

Crystal structure of *A. tumefaciens* pVirB8
(Bailey *et al*., 2006)

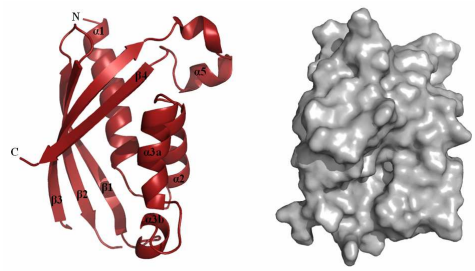
Crystal structure of *B. suis* pVirB8
(Terradot *et al*., 2005)

A



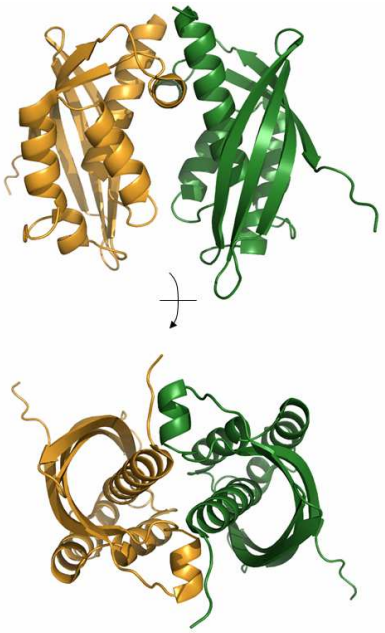
(A) On the left, ribbon representation of pVirB8^{AT} monomer (orange) with secondary structure elements labeled. On the right, surface rendered image of pVirB8^{AT} in the same orientation, showing the deep groove.

B



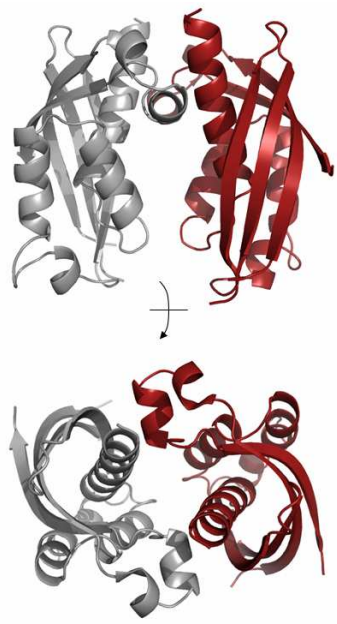
(B) On the left, ribbon representation of pVirB8^{BS} monomer (red) with secondary structure elements labeled. On the right, surface rendered image of pVirB8^{BS} in the same orientation, showing the deep groove.

C



(C) Ribbon diagram of the pVirB8^{AT} dimer rotated 90° with respect to A. Below, ribbon diagram of the same dimer equivalent to a 90° rotation from the above view.

D



(D) Ribbon diagram of the pVirB8^{BS} dimer rotated 90° with respect to A. Below, ribbon diagram of the same dimer equivalent to a 90° rotation from the above view.

Figure 5.1. Structure features of pVirB8^{AT} and comparison with pVirB8^{BS}.

Both periplasmic fragments share 30 % sequence identity and, as expected, the overall fold of the two pVirB8 monomers is very similar, comprising four β-sheets clustered juxtaposed against to five α-helices [Fig. 5.1 and 5.2].

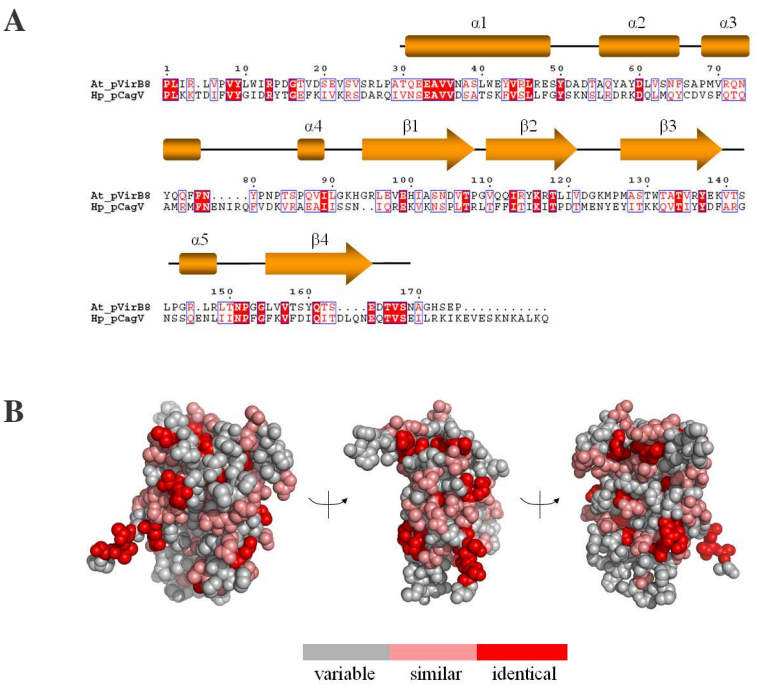
Generally, VirB8 are bitopic membrane proteins with a short N-terminal cytoplasmic exposed domain, a transmembrane helix, and a large C-terminal periplasmic domain (Das and Xie, 1998; Kumar *et al*., 2000).

At the beginning, it was proposed that the CagW (HP0529) protein might be a homologue of the VirB8 (Covacci and Rappuoli, 2000). Since there is no obvious sequence similarity, this suggestion was probably based on the localization of *cagW* gene upstream of the *cagX* gene (*hp0528*), a *virB9* homologue. While VirB8 from *A. tumefaciens* and *B. suis* was shown to be a bitopic membrane protein (Das and Xie, 1998; Kumar *et al*., 2000), the transmembrane helix predictions of CagW, however, suggest that this protein adopts a polytopic conformation with five to six transmembrane domains in addition to its N-terminal signal sequence. Thus, the CagW protein, which is essential for CagA translocation and IL-8 induction, might rather be instead a VirB6 analogue. VirB6 in fact is the only component of the VirB/D4 system with more than 3 transmembrane helices (Das and Xie, 1998; Buhrdorf *et al*., 2003).

In order to determine the presence of a VirB8 homologue protein in *H. pylori,* Buhrdorf and colleagues (Buhrdorf *et al*., 2003) performed some transmembrane topology predictions of all proteins essential for CagA translocation using different programs, such as TMPred or TMHMM (Rost *et al*., 2004; Hofmann and Stoffel, 1993; Krogh *et al*., 2001). They subsequently applied a transposon mutagenesis procedure which confirmed that CagV (HP0530) protein adopts the same membrane topology and also shares several characteristic sequence motifs with the VirB8 proteins of *A. tumefaciens* and *B. suis*.

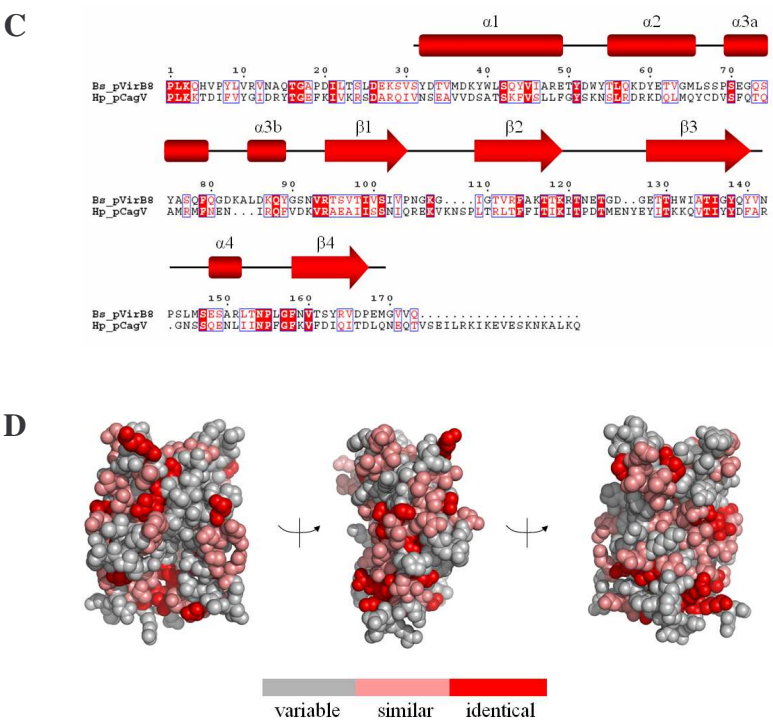
The CagV (HP0530) protein is composed of 252 amino acids, with a molecular weight of 29.1 kDa and a theoretical isoelectric point of 9.52, mainly due to a prevalence of lysine residues. The sequence is mostly hydrophilic, except between residues 38 and 57, which define a putative transmembrane sequence. The presence of only one cysteine moreover suggests the possibility of dimerization events in the native form. Generally, VirB8 length varies from about 202 to 260 amino acids (Baron, 2006) and the homologues display only a limited sequence similarity, but some more or less conserved motifs can be distinguished (Cao and Saier, 2001). In particular, the amino acid sequence of CagV does not display much identity and similarity to the other VirB8 proteins (19/20 % identity and 44/46 % similarity with *A. tumefacines* and *B. suis* VirB8 proteins, respectively). Additionally, CagV (252 aa) is about twenty residues longer than its homologues $VirB8^{AT}$ (231 aa) $VirB8^{BS}$ (235 aa) and in particular it contains an extensional C-terminal charged (K/E/R) domain which is not present in the other VirB8 homologues and that could be a specific feature of CagV.

Sequence alignment of *A. tumefaciens* pVirB8 (pVirB8^{AT}) and corresponding *H. pylori* pCagV$^{59-252}$



(A) Sequence alignment of *A. tumefaciens* pVirB8 (pVirB8^{AT}) and corresponding *H. pylori* pCagV (59-252) homolog. The corresponding secondary structural elements are shown in red and conserved residues are highlighted. White residues in red squares indicate identical residues, whereas the red residues into white squares indicate residues similar.
(B) Space filling models of pVirB8^{BS} (residues 92–231 of structural model). Surface residues are shaded according to the degree of conservation. Residues in pink are similar, whereas residues in red are identical. From the right, the protein is presented in approximately the same orientation as Fig. 6.1 A. In the middle, it is rotated by 90° and on the left by 180°.

Sequence alignment of *B. suis* pVirB8 (pVirB8^{BS}) and corresponding *H. pylori* pCagV$^{59-252}$



(C) Sequence alignment of *B. suis* pVirB8 (pVirB8^{BS}) and corresponding *H. pylori* pCagV (59-252) homolog. The corresponding secondary structural elements are shown in red and conserved residues are highlighted. White residues in red squares indicate identical residues, whereas red residues into white squares indicate residues similar.
(D) Space filling models of pVirB8^{BS} (residues 97–235 of the structural model). Surface residues are shaded according to the degree of conservation. Residues in pink are similar to pCagV, whereas tresidues in red are identical. From the right, the protein is presented in approximately the same orientation as Fig. 6.1 B. In the middle, it is rotated by 90° and on the left by 180°.

Figure 5.2. Sequence alignment and structure analysis of both pVirB8^{AT} and pVirB8^{BS} and pCagV.

For example, this charged domain could be involved into the interaction with other Cag proteins belonging to the *H. pylori* TFSS. However, some conserved motifs can also be found in the CagV protein, but not in CagW, and they are located at different amino acid positions within the VirB8 homologues. For example, the predicted transmembrane helix always precedes the first motif in the VirB8 homologues and this is also the case for CagV, where the predicted transmembrane helix spans amino acids from 38 to 57 and the first conserved motif (PLK) comprises amino acids from 58 to 60. Another conserved region is defined by the sequence motif NPxGxxV, which in the $pVirB8^{AT}$ and $pVirB8^{BS}$ structures seems to be important in stabilizing the position of a loop which is consider essential for the maintenance of the dimer. Both $pVirB8^{AT}$ and $pVirB8^{BS}$ proteins form similar dimers in the crystal, supporting the idea that dimer formation in the crystal reflects self-interactions that are biologically significant (Bailey *et al*., 2006; Terradot *et al*., 2005). From all these data it was predicted that the CagV protein could play an analogous role of VirB8 in the *H. pylori* type IV secretion apparatus.

Mutagenesis studies also demonstrated an essential role of CagV for the translocation of CagA toxin and, at the same time, for the induction of the inflammatory response, in the form of interleukin IL-8 (Fischer *et al*., 2001).

All these data are coherent with a direct involvement of CagV in the constitution of the type four secretion system, likely as a component of the structural core complex together with CagT, CagX, CagY and some ATPases anchored to the inner membrane. Based on biological and genetic data, VirB8 proteins are in fact known to be involved in multiple protein-protein interactions that mediate the assembly of the translocation machinery. Several interaction partners of VirB8 (VirB1, VirB4, VirB5, VirB9, VirB8, VirB10, and VirB11) from both *A. tumefaciens* and *B. suis* were recently discovered by using crosslinking, immuno-electron microscopy and yeast two-hybrid system and some of them were also confirmed biochemically (Paschos *et al*., 2006; Yuan *et al*., 2005; Das *et al*., 2000; Hoppner *et al*., 2005; Ward *et al*., 2002; Kumar *et al*., 2000).

An analysis, performed with the server DALI (Holm and Sander, 1993), of the structures present in the Protein Data Bank suggests that the pVirB8 fold is, in the overall, quite similar to the nuclear transport factor NTF2-like fold (Bailey *et al*., 2006; Terradot *et al*., 2005). The multiple protein–protein interactions of VirB8 are reminiscent of its structural homolog NTF2, which binds to the small GTPase Ran and undergoes multiple transient interactions with components of the nuclear pore during transfer with its cargo to the nucleus (Bayliss *et al*. 2002). Due to its anchoring to the inner membrane, VirB8 is not likely to move extensively inside the T4SS. However,

different binding affinities may drive its sequential interactions with other VirB proteins, thereby coordinating transporters assembly. The proposed function of VirB8 is reminiscent of the PapD chaperone that sequentially interacts with different P-pilus subunits and thereby directs the assembly of this extracellular structure (Sauer *et al.* 2004). P-pilus assembly is based on differential affinities of the chaperone for its binding partners and similar principles could also apply to VirB8 function (Baron, 2006).

Here we describe the production, characterization, crystallization and preliminary diffraction analysis of two periplasmic domains of *H. pylori* VirB8 homologue, CagV (HP0530), corresponding to residues 59–252 (pCagV$^{59\text{-}252}$) and 68-252 (pCagV$^{68\text{-}252}$) [Fig. 5.3].

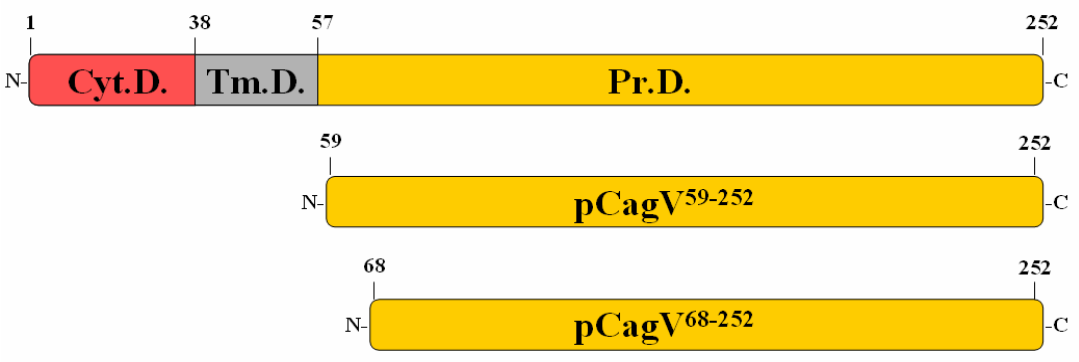


Figure 5.3. Topology of the full-lenght *H. pylori* CagV (HP0530) and the two periplasmic domains selected and studied (pCagV$^{59\text{-}252}$ and pCagV$^{68\text{-}252}$). Cyt.D.: Cytoplasmic Domain (red); Tm.D.: Transmembrane region (grey); Pr.D.: Periplasmic Domain (orange).

5.2 Experimental methods

Cloning

The amino acid sequences of the determined VirB8 periplasmic domain structures from *Agrobacterium tumefaciens* (pVirB8^{AT}; codice gene) and from *Brucella suis* (pVirB8^{BS}; codice gene) were aligned with CagV sequence using the multiple sequence alignment software ClustalW (www.ebi.ac.uk/Tools/clustalw) and used to determine the C-terminal soluble domain of CagV to clone.

Therefore, on the base of this bioinformatics analysis (ExPASy Proteomics tools, www.expasy.org) and literature informations (Buhrdorf *et al*., 2003; Terradot *et al*., 2005; Bailey *et al*., 2006), two different C-terminal periplasmic domains called pCagV$^{59\text{-}252}$ (194 aa; 22.7 kDa) and pCagV$^{68\text{-}252}$ (185 aa; 21.6 kDa) [Fig. 5.3] were selected, cloned, expressed, purified and crystallize as follow.

The truncated pCagV$^{68\text{-}252}$ nucleotide sequence was amplified by PCR using as a template the *H. pylori* CCUG 17874 genomic DNA, the forward fw*cagV*$^{68\text{-}252}$ (5'-CGTCATATGGGCATTGATCGATACAC-3' which introduces a *Nde*I restriction site and the reverse primer rv*cagV*$^{68\text{-}252}$ (5'-CGTAAGCTTTTGTTTTAATGCCTTA-3') which introduces a *Hind*III restriction sites respectively. The amplified fragment was digested with *Nde*I/*Hind*III (NEB) restriction enzymes and inserted into a pET28b expression vector (Novagen) which encodes an N-terminal His_6-tag and a Thrombin cleavage site for subsequent affinity-tag removal by protease.

To construct the plasmid pET151-*cagV*$^{59–252}$, the amplified gene fragment encoding amino acids 59-252 of *Helicobacter pylori* CagV (pCagV$^{59–252}$) was cloned into the pET151/D-TOPO vector using the TOPO cloning kit (Invitrogen), which permits directional cloning into expression system and the following forward primer fw*cagV*$^{59\text{-}252}$ (5'-CACCTTGAAGAAAACGGATATATTTG-3') and reverse primer rv*cagV*$^{59\text{-}252}$ (5'-TCATTATTGTTTTAATGCCTTATTTTTTG-3'). The vector adds an N-terminal His_6-tag, a V5 epitope and a Tobacco Etch Virus (TEV) protease cleavage site to the expressed recombinant protein. Correct cloning was verified by sequencing the DNA constructs using standard primers for T7 promoters.

Protein expression, purification and characterization

After confirmation of the sequence, the expression constructs (pET151-*cagV*$^{59-252}$ and pET28-*cagV*$^{68-252}$) were used to transform *Escherichia coli* BL21(DE3) competent cells (Invitrogen). A single colony was spread into Luria Bertani Agar (LBA) plates with/containing 100 μg ml^{-1} Ampicillin (pET151-*cagV*$^{59-252}$) or 50 μg ml^{-1} Kanamycin (pET28-*cagV*$^{68-252}$) at 37 °C for overnight growth. The solid overnight cultures of each constructs were then resuspended in 20 ml LB medium, transferred into 2 l LB broth medium containing 50 mg ml^{-1} Kanamycin for His_6-pCagV$^{68-252}$ construct and 100 mg ml^{-1} Ampicillin for His_6-pCagV$^{59-252}$ and grown at 37 °C with shaking (210 rpm) to an optical density of approximately 0.6 at 600 nm. After induction, with 0.5 mM isopropyl β-D-thiogalactopyranoside (IPTG) the culture were incubated at 30 °C and 210 rpm for about 4 h. The cells were then harvested by centrifugation (11.000 g, 20 min at 4 °C) and stored at -80 °C until further processing. Cell pellets were successively resuspended in 50 ml lysis buffer [30 mM Tris pH 7.5, 200 mM NaCl, 1mg/ml Lysozime and Complete Mini EDTA-free Protease Inhibitor Cocktail tablets (Roche)] and then disrupted by French Press. The lysed cell suspension was cleared of debris by centrifugation for 20 min at 4 °C and 40000 g and the supernatant was loaded onto a 1/5 ml HisTrap HP column (GE Healthcare) which was first equilibrated against buffer A (30 mM Tris pH 7.5, 200 mM NaCl). Following binding, the column was firstly washed with 5 % and with 10 % of buffer B (30 mM Tris pH 7.5, 200 mM NaCl, 500 mM Imidazole) in order to remove unspecific binding *E. coli* proteins. The protein was then eluted by running a linear gradient from 10 to 100 % of buffer B. The pooled peak fractions of His_6-pCagV$^{59-252}$ and His_6-pCagV$^{68-252}$ were incubated with recombinant His-tagged TEV and Thrombin proteases respectively for 16 h at 4 °C to remove the N-terminal His_6 tag and leaving the native protein. The efficiency of the proteolisys was assessed by SDS-PAGE which shown us that in both cases (His_6-pCagV$^{68-252}$ and His_6-pCagV$^{59-252}$) was impossible to cut off the tag. For these reasons, we decided to continue to purify and crystallize the two uncleaved His_6-pCagV$^{68-252}$ and His_6-pCagV$^{59-252}$ domains.

As the final step of the purification procedure, the two soluble proteins were purified by size-exclusion chromatography (HiLoad Superdex 75, 16/60, GE Healthcare) with buffer A and a flow rate of 1 ml min^{-1}.

The proteins were eluted as a single peak with an apparent molecular weight of about 20 kDa, which is consistent with the weight of recombinant His_6-pCagV$^{59-252}$ and His_6-pCagV$^{68-252}$ in their monomeric state. Peak fractions of each pure protein were analyzed with SDS–PAGE, to verify their pureness and stability and following pooled

and concentrated by ultrafiltration (Amicon Ultra 10.000 MWCO, Millipore) to 10 mg ml^{-1} and storage at -80 °C. The proteins concentration was determined by UV/VIS spectroscopy (240-360 nm, b = 1 cm, DU640 UV/VIS spectrophotometer, Beckman Coulter Inc.) using the theoretical absorption coefficients at 280 nm of 0.52 ml mg^{-1} cm^{-1} (11920 M^{-1} cm^{-1}) for His_6-$CagV^{59-252}$ and 0.48 ml mg^{-1} cm^{-1} (10430 M^{-1} cm^{-1}) for His_6-$pCagV^{68-252}$.

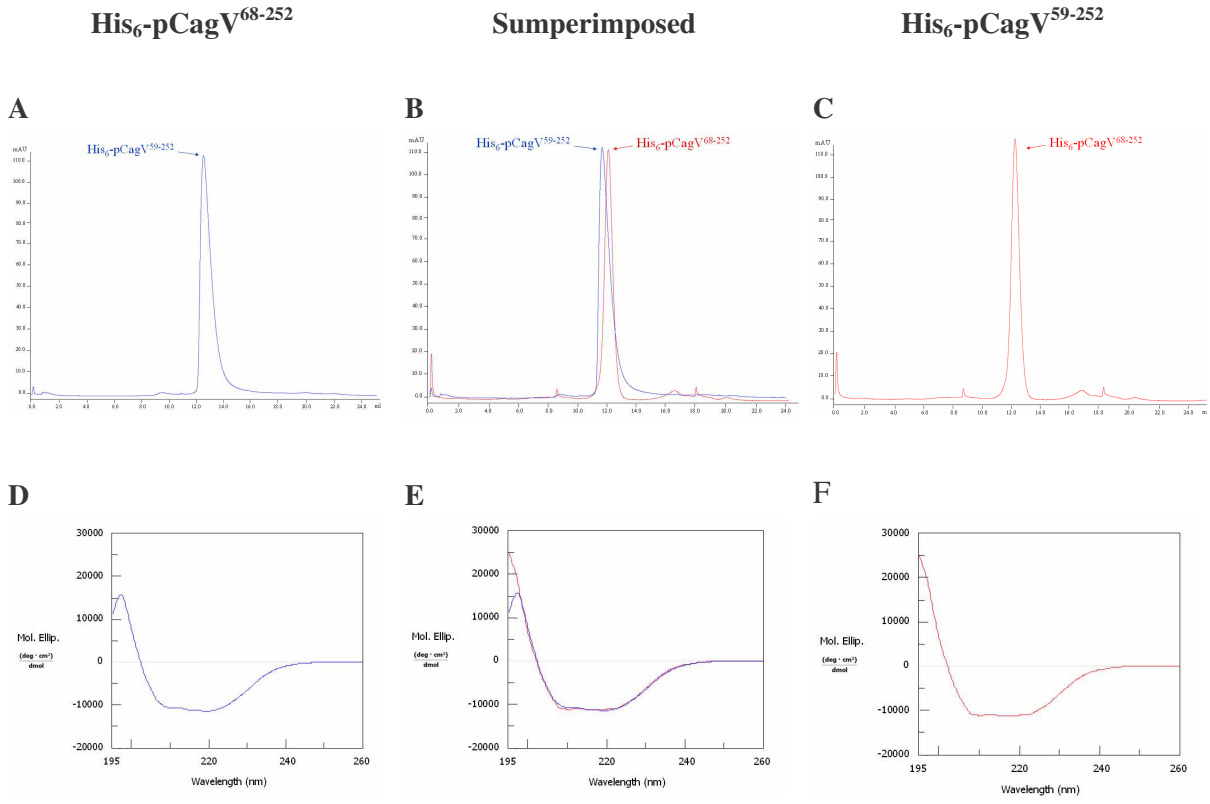


Figure 5.4. Size-exclusion chromatography profiles and CD spectra of His_6-$pCagV^{59-252}$ and His_6-$pCagV^{68-252}$ proteins. The blue line (A and D) refers to His_6-$pCagV^{68-252}$ protein, while the red line (C and F) refers to His_6-$pCagV^{59-252}$ protein. Every gel filtration experiments as well as CD measurements were performed independently and then overlaid (B and E).

CD measurement of the each protein diluited to 1 mg ml^{-1} final concentration were performed using a JASCO-J720 Spectropolarimeter, in a 0.02 cm path length cell. The spectra in the far-UV (195-260 nm) were recorded at a scanning speed of 20 nm min^{-1}. Ten spectra were accumulated in each case and averaged followed by baseline correction by subtraction of the buffer.

Mean residue weight ellipticities were calculated and expressed in units of degree cm^2 $dmol^{-1}$. The circular dichroism spectra were deconvoluted by software program CDNN

(Version 2.1) to estimate the content of secondary structure of His_6-$pCagV^{59-252}$ and His_6-$pCagV^{68-252}$.

The purified His_6-$pCagV^{59-252}$ and His_6-$pCagV^{68-252}$ proteins at several concentrations (7–20 mg ml^{-1} for His_6-$pCagV^{68-252}$, and 2.5-20 mg ml^{-1} for His_6-$pCagV^{59-252}$) and temperatures (4, 10 and 20 °C) were submitted to crystallization screening trials using both sitting- and hanging-drop vapor diffusion techniques and a variety of commercially available screens: Structure Screen I and II from Molecular Dimension, the PACT suite, MbClass I and II from Qiagen. The crystallization experiments were performed manually and also using an Oryx8 crystallization robot (Douglas Instruments). In a typical experiment, 0.1 µl screening solution was added to 0.1 µl protein solution on 96-well crystal plates (Douglas Instruments); the reservoir wells contained 100 µl screening solution. Conditions that yielded crystals were optimized with manually prepared condition using 24-wells Linbro plates and glass cover slides (Hampton Research). Drops containing equal volumes of protein and reservoir solution (1-4 µl of each) were allowed to equilibrate against 400 µl reservoir solution.

Crystallization

Diffraction quality crystals of His_6-$pCagV^{68-252}$ were optimized from MbClass II condition 76 (Qiagen). Crystals were grown using both sitting- and hanging-drop vapour diffusion at 20 °C over a 400 µl reservoir containing 12 % (w/v) PEG 4000 and 200 mM $(NH_4)_2SO_4$ [Fig. 5.5].

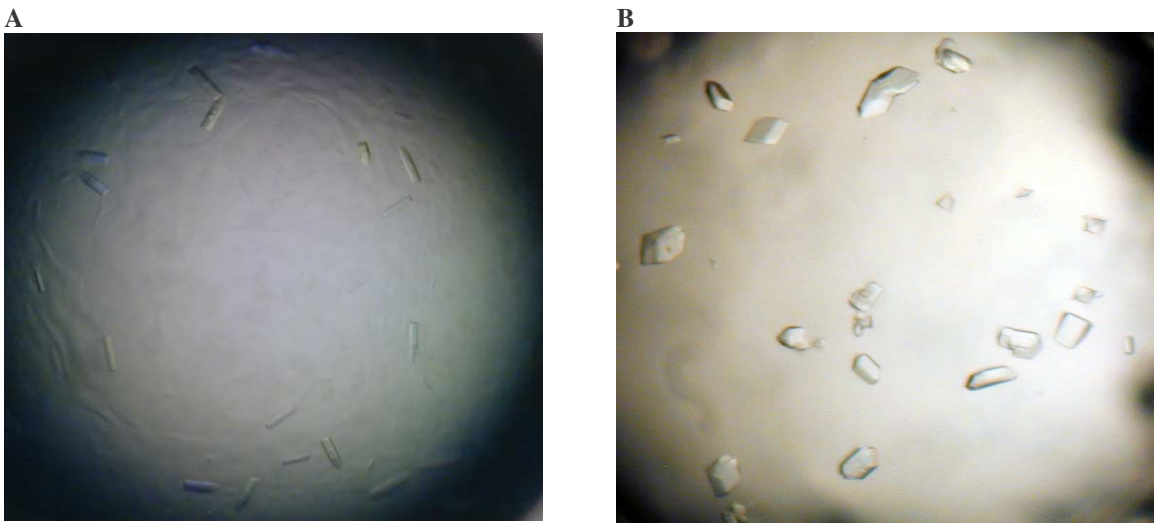


Figure 5.5. Representative crystals forms of His_6-$pCagV^{59-252}$ (A) and His_6-$pCagV^{68-252}$ (B) used in this study are shown.

Each drop contained 1-4 µl of His_6-$pCagV^{68-252}$ (20 mg ml^{-1}) and 1-4 ul reservoir solution. Crystals grew in 3-4 days.

Diffraction quality crystals of His_6-$pCagV^{59-252}$ were instead optimized from Structure Sreen II condition 26 (Molecular Dimension Ltd). Crystals were grown using both sitting- and hanging-drop vapour diffusion at 20 °C over a 400 µl reservoir containing 100 mM MES pH 6.5, 30 % (w/v) PEG 5000 and 200 mM $(NH_4)_2SO_4$ [Fig. 6.5]. Each drop contained 1 µl of His_6-$CagV^{59-252}$ (5 mg ml^{-1}) and 1 µl reservoir solution. Crystals grew in 15 days.

The two crystals obtained shared two common features: all the crystal grew at a neutral pH (pH 6.5-7.5) using polyethyleneglycol (PEG) as a precipitant. The molecular weight of the PEG ranged from 4K to 5K.

Data collecting and processing

The cryocooling solution used to measure diffraction data of His_6-$pCagV^{68-252}$ and His_6-$pCagV^{59-252}$ crystals were 12 % (w/v) PEG 4000, 200 mM $(NH_4)_2SO_4$, 20 % (w/v) EG and 100 mM MES pH 6.5, 30 % (w/v) PEG 5000, 200 mM $(NH_4)_2SO_4$, 20 % (w/v) EG, respectively.

The soaked crystals were then removed from the solution, mounted on a cryo-loop and placed into a stream of nitrogen at 100 K. The His_6-$pCagV^{68-252}$ diffraction data were collected on beamline ID29, at the European Synchrotron Radiation Facility (ESRF), Grenoble, with 1° oscillation per image on an ADSC Quantum-315 detector.

The data set from the His_6-$pCagV^{59-252}$ crystals were instead collected at station ID23-EH2 (ESRF) with 1° oscillation per image on a MAR 225 CCD detector.

All data sets were integrated using *MOSFLM* (Leslie, 1992) and merged with *SCALA* (Evans, 1997) as implemented in *CCP4* (Collaborative Computational Project, Number 4, 1994). The His_6-$pCagV^{59-252}$ crystal belongs to space group $P2_12_12_1$, with unit-cell parameters a = 59.56, b = 103.33, c = 130.33 Å. The His_6-$pCagV^{68-252}$ crystal instead belongs to the $P4_32_12$ space group or its enantiomorph $P4_12_12$, with unit-cell parameters a = 81.68, b = 81.68, c = 139.95 Å. Diffraction data were collected from these two crystals to a resolution of 2.7 Å for His_6-$pCagV^{59-252}$ and 3.0 Å for His_6-$pCagV^{68-252}$. The data-collection statistics are shown in Table 5.1. Molecular-replacement approaches were performed using the programs *MolRep* (Navaza, 1994; Collaborative Computational Project, Number 4, 1994) and *Phaser* (McCoy *et al*., 2005).

5.3 Results and discussion

Although full-length CagV (HP0530) protein (252 residues) could be expressed in a soluble form in *Escherichia coli* as a fusion protein with glutathione S-transferase (GST), it tended to become insoluble upon removal of the GST.

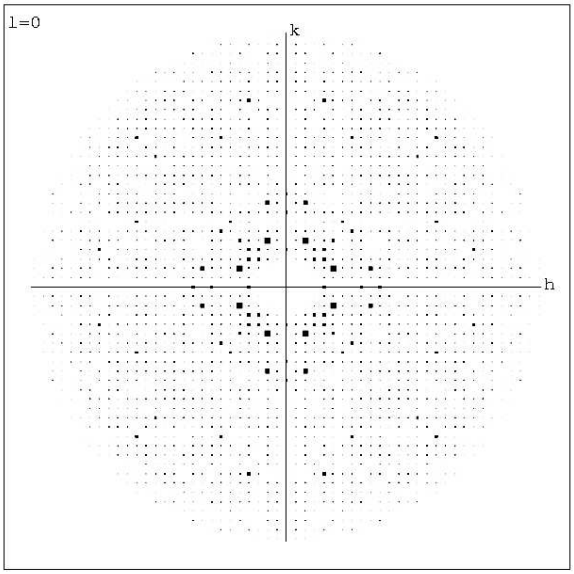
To increase its solubility and reduce the aggregation of the sample, proved by the chromatographic gel filtration profile, the purification was performed in the presence of neutral detergent $C_{12}E_8$, selected since it resulted highly monodisperse and particularly effective with other membrane proteins, simplifying the oligomerization profile. However, possibly due to the absence of the native *H. pylori* membrane environment and of the interacting proteins during CagV expression and folding, the full-length CagV sample resulted very inhomogeneous and difficult to separate from GST tag, even after proteolysis.

To permit proper folding, two C-terminal soluble domains (His_6-$pCagV^{59-252}$ and His_6-$pCagV^{68-252}$) of the CagV (HP0530) were selected, cloned, produced and purified. Prior to crystallization experiments, the high purity and homogeneity of the proteins solution were confirmed by both SDS–PAGE and size exclusion chromatography respectively (Fig. 5.4). After extensive screening of these crystals, a native data set of $pCagV^{68-252}$ was collected on beamline ID29 at ESRF to a maximum resolution of 3.0 Å (Table 6.1). These crystals belong to the tetragonal space group $P4_1$. Based on the molecular weight of the protein (23.996 Da) and the volume of the asymmetric unit, the Matthews parameters (Matthews, 1968) for one, two, three and four molecules in the asymmetric unit are 10.61, 5.30, 2.53 and 2.65 $Å^3$/Da respectively. This suggests the presence of three or four molecules in the asymmetric unit, which corresponds to a solvent content of 65.5 % and 53.3 %, respectively.

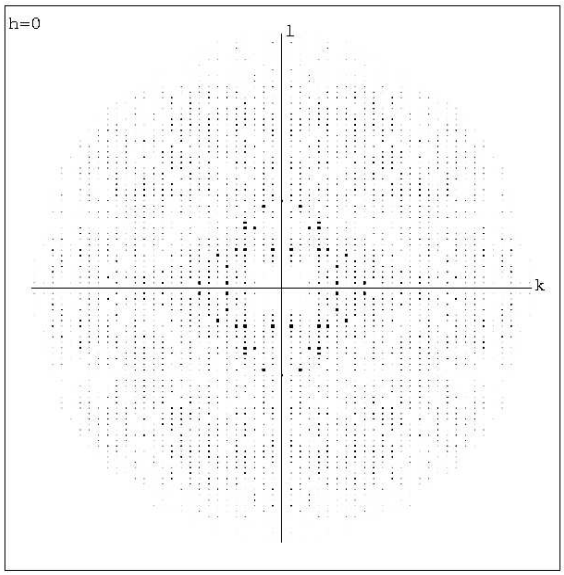
In order to solve the crystal structure of the CagV periplasmic domain, using a molecular replacement (MR) approach, two pVirB8 structures were chosen as a search model with *Phaser* or *MolRep* softwares (McCoy *et al*., 2005; Navaza, 1994; Collaborative Computational Project, Number 4, 1994). One is the X-ray structure of periplasmic VirB8 domain from *Brucella suis* (16 % sequence identity and 43 % smilariry with pCagV; PDB code 2BHM) and the other is the X-ray structure of periplasmic domain VirB8 from *Agrobacterium tumefaciens* (17 % sequence identity and 40 % similarity with pCagV; PDB code 2CC3).

Extensive molecular-replacement strategies to phase $pCagV^{68-252}$using either structures were however unsuccessful.

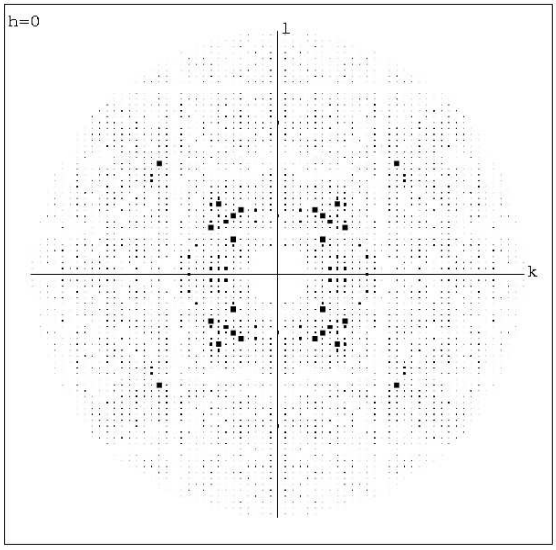
A. His_6-$pCagV^{59-252}$: plane hk0 of the reciprocal space



B. His_6-$pCagV^{59-252}$: plane 0kl of the reciprocal space



C. His_6-$pCagV^{68-252}$: plane 0kl of the reciprocal space



D. His_6-$pCagV^{68-252}$: plane h0l of the reciprocal space

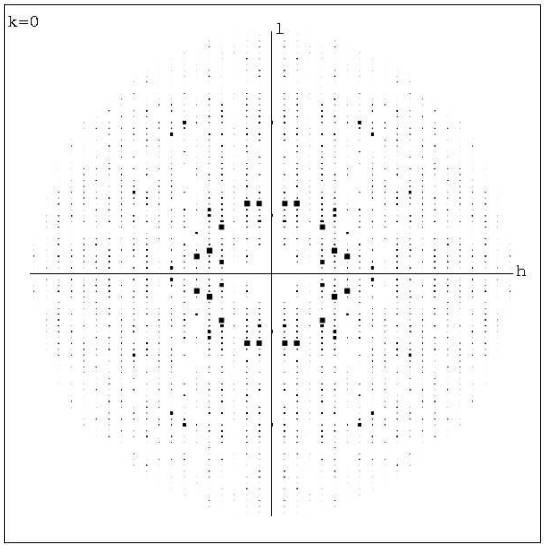


Figure 5.6. Reciprocal space planes of His_6-$pCagV^{59-252}$ (A and B) and His_6-$pCagV^{68-252}$ (C and D).

We therefore decided to produce the protein with some additional residues at the N-terminus of the construct in an attempt to obtain better diffracting crystals. Extensive crystallization screening of the new construct His_6-$pCagV^{59-252}$ led to the identification of a second crystal form with significantly improved diffraction properties. The tetragonal crystals [Fig. 5.5], which grow under conditions that are almost identical to

the previous ones, with the exception of a lower pH, are more reproducible and diffract better. A native data set was collected from this new crystal form on beamline ID23-2 at the ESRF to a resolution of 2.7 Å (Table 5.1).

The crystals belong to the orthorhombic space group $P2_12_12_1$. Based on the molecular weight of the protein (26471 Da, including six additional N-terminal amino-acid residues as cloning artifacts) and the volume of the asymmetric unit, the Matthews parameters (Matthews, 1968) for one, two, three and four molecules in the asymmetric unit are 7.57, 3.78, 2.52 and 1.89 $Å^3$/Da, respectively. This suggests the presence of two or three molecules in the asymmetric unit, which corresponds to a solvent content of 67.5 % and 51.3 %, respectively. Molecular replacement attempts also failed with this second data set.

As yet, we have not been able to obtain a solution which refines. As a consequence, structure determination will continue using multiple isomorphous replacement (MIR) or anomalous scattering methods (SAD and MAD).

Data set	**His_6-pCagV$^{59-252}$**	**His_6-pCagV$^{68-252}$**
Beamline	ID23-2, ESRF	ID29, ESRF
Wavelength (Å)	0.9794	0.9795
Space group	$P2_12_12_1$	$P4_32_12$ (or $P4_12_12$)
Unit cell parameters (Å ,°)	a = 59.56, b = 103.33, c = 130.33	a = 81.68, b = 81.68, c = 139.95
Resolution (Å)	103.14-2.70 (2.85-2.70)	81.65-3.00 (3.16-3.00)
Total No. of reflections	88906 (13235)	136038 (20098)
No. of unique reflections	21057 (3104)	10074 (1420)
Moltiplicity	4.2 (4.3)	13.5 (14.2)
Completeness (%)	92.9 (95.2)	100 (100)
*Rsym (%)	11.3 (36.5)	9.8 (46.7)
I/σ(I)	4.6 (1.7)	4.4 (1.5)

Table 5.1. Data-collection and processing statistics. Values in parentheses are for the highest resolution shell.

$*R_{sym} = \Sigma_{hkl}|I_{hkl} - \langle I_{hkl}\rangle|/\Sigma_{hkl} I_{hkl}$

Chapter 6

Expression, purification and characterization of CagA protein domains

Contents

- Construction of N-terminal deletion variants of CagA
- High-troughput protein expression and purification screening of deletion products of the CagA protein

6.3 Expression, purification, characterization of a CagA soluble central domain and preliminary crystallization and interaction studies with CagF and *Hp*YbgC proteins

6.3.1 Material and Methods

- CagA$^{392-733}$ Protein Expression and Purification
- CagF and *Hp*YbgC protein Expression and Purification
- Protein Characterization
- Analytical Gel Filtration experiments
- Far Western Blotting Assay

6.3.2 Results and Discussion

- Single protein expression, purification and characterization
- Characterization of the CagA$^{392-733}$-CagF interaction
- Characterization of the CagA$^{392-733}$-*Hp*YbgC interaction
- Crystallization and proteolysis experiments

Mostly of this work has been done at the PSB laboratories (Grenoble) in collaboration with: Tommaso Tosi[1] that contributed equally to this work, Philippe Mas[2], Darren Hart[2] and Laurent Terradot[1].

[1]Macromolecular Crystallography Group, European Synchrotron Radiation Facility (ESRF), B.P. 220, 6 rue Jules Horowitz, F-38043 Grenoble Cedex, France.
[2]European Molecular Biology Laboratory (EMBL) Grenoble Outstation, 6 rue Jules Horowitz, BP181, 38042 Grenoble Cedex 9, France.

6.1 Sequence analysis of CagA

Helicobacter pylori strains can be divided into two major subpopulations based on their ability to produce a 128-146 kDa immunodominant protein called cytotoxin associated gene A (CagA) antigen (Covacci *et al*., 1993; Tummuru *et al*., 1993). CagA is a very hydrophilic protein and does not show any transmembrane sequence. The most hydrophilic region includes the repetition of EFKNGKNKDFSK and EPYA and the presence of a stretch of six asparagines (NNNNNN) (Covacci *et al*., 1993) [Fig. 6.1].

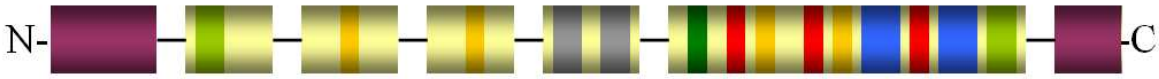
The *cagA* gene, which encodes CagA, is localised at one end of the *cag* pathogenicity island (*cag*-PAI), a 40-kb DNA segment that was most likely incorporated into the *H. pylori* genome by a process of horizontal transfer (Censini *et al*., 1996). The *cag*-PAI DNA segment contains about 31 putative genes (open reading frames), including *cagA*, and those encoding components of a molecular 'syringe' termed the type IV secretion system (TFSS) (Akopyants *et al*., 1998; Fischer *et al*., 2001; Rohde *et al*., 2003) through which macromolecules are injected directly from the inside to the outside of the bacterium in the infected gastric epithelial cell. Clinically, infection with the *cagA* positive *H. pylori* strain was associated with higher grades of gastric mucosal inflammation as well as severe atrophic gastritis and was suggested to play an important role in the development of gastric carcinoma (Hatakeyama *et al*., 2005).

Sequence Diversity in CagA

Diversity in the amino acid sequence of CagA is one of the most important variables in determining the outcome of *H. pylori* infection. The so-called variable region in CagA containing a 102- to 108-nucleotide repeat sequence first became apparent when the gene encoding this protein was sequenced (Covacci *et al*., 1993; Tummuru *et al*., 1993). At the protein level, the molecular size of CagA is therefore variable and depends on the number of repeats. The molecular weight of CagA ranges from 128 kDa (absence of 102 bp repeat) to 146 kDa (four 102 bp repeats in tandem) (Xiang *et al*., 1995; Hatakeyama et al., 2005). The CagA variable region is very complex and contains both smaller, 5- to 15-amino acids, repeat sequences and larger structural variations (Evans *et al*., 1998; Yamaoka *et al*., 1998) which appear to be a result of frequent genetic recombination events and accumulation of mutations within the *cagA* gene. The number, sequence, and position of the variable region can vary, depending on the geographic region from which the strain is isolated. In general, CagA protein

can be sub-classified into two main types, East-Asian CagA and Western CagA, based on its polymorphisms in the C-terminal domain (Kersulyte *et al*., 2000; Hatakeyama, 2004) [Fig. 6.2]. The Comparison of strains with differences in the number of repeats isolated from patients with gastric cancer suggested that strains with more repeats were associated with higher levels of CagA antibody, more severe degrees of atrophy, and reduced survival in a low pH.

Figure 6.1. Schematic representation of the CagA (HP0547) known motifs.



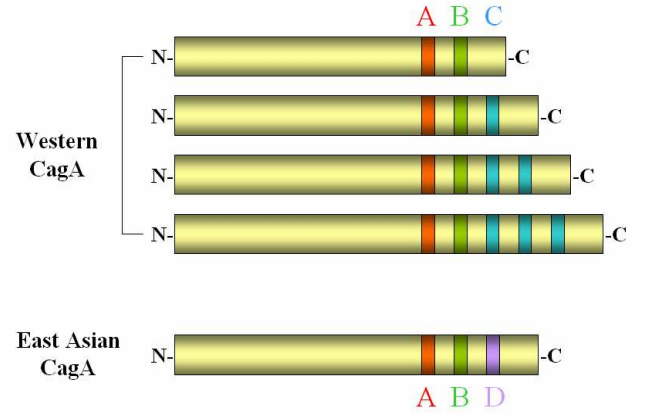
- Tyrosine Phosphorylation Motifs (TPMs): [RK]-$X_{(2,3)}$-[D/E]-$X_{(2,3)}$-[Y] and (EPYA);
- cAMP Phosphorylation Motifs (CPMs) : $[R/K]_{(2)}$-X-[S/T];
- Multimerisation Domain (MDs) : [FPL]-X-[R]-$X_{(3)}$-[V]-X-[DLSKVG] ;
- Membrane-targeting Signal : (EPYA);
- Signals translocation: N- and C-terminal secretion signals;
- Repited sequences: EFKNGKNKDFSK and NNNNN.

MTNETIDQTRTPDQTQSQTAFDPQQFINNLQVAFIKVDNVVASFDPDQKPIVDKNDRDNRQAFDGISQLREEYSNKAIKNPTKKNQYFSDFI
DKSNDLINKDNLIDVESSTKSFQKFGDQRYQIFTSWVSHQKDPSKINTRSIRNFMENIIQPPIPDDKEKAEFLKSAKQSFAGIIIGNQIRTD
QKFMGVFDESLKERQEAEKNGGPTGGDWLDIFLSFIFNKKQSSDVKEAINQEPVPHVQPDIATTTTDIQGLPPEARDLLDERGNFSKFTLGD
MEMLDVEGVADIDPNYKFNQLLIHNNALSSVLMGSHNGIEPEKVSLLYAGNGGFGDKHDWNATVGYKDQQGNNVATLINVHMKNGSGLVIAG
GEKGINNPSFYLYKEDQLTGSQRALSQEEIRNKVDFMEFLAQNNTKLDNLSEKEKEKFQNEIEDFQKDSKAYLDALGNDRIAFVSKKDTKHS
ALITEFNNGDLSYTLKDYGKKADKALDREKNVTLQGSLKHDGVMFVDYSNFKYTNASKNPNKGVGATNGVSHLEAGFNKVAVFNLPDLNNLA
ITSFVRRNLENKLTAKGLSLQEANKLIKDFLSSNKELAGKALNFNKAVAEAKSTGNYDEVKKAQKDLEKSLRKREHLEKEVEKKLESKSGNK
NKMEAKAQANSQKDEIFALINKEANRDARAIAYTQNLKGIKRELSDKLEKISKDLKDFSKSFDEFKNGKNKDFSKAEETLKALKGSVKDLGI
NPEWISKVENLNAALNEFKNGKNKDFSKVTQAKSDLENSVKDVIINQKVTDKVDNLNQAVSVAKAMGDFSRVEQVLADLKNFSKEQLAQQAQ
KNEDFNTGKNSELYQSVKNSVNKTLVGNGLSGIEATALAKNFSDIKKELNEKFKNFNNNNNGLKNSTEPIYAKVNKKKTGQVASPEEPIYTQ
VAKKVNAKIDRLNQIASGLGGVGQAAGFPLKRHDKVDDLSKVGLSASPEPIYATIDDLGGPFPLKRHDKVDDLSKVGRSRNQELAQKIDNLN
QAVSEAKAGFFGNLEQTIDKLKDSTKKNVMNLYVESAKKVPASLSAKLDNYAINSHTRINSNIQNGAINEKATGMLTQKNPEWLKLVNDKIV
AHNVGSVSLSEYDKIGFNQKNMKDYSDSFKFSTKLNNAVKDIKSGFTHFLANAFSTGYYCLARENAEHGIKNVNTKGGFQKS

CagA Tyrosine Phosphorylation Motifs (TPMs)

Phosphorylation of CagA is not only required for much of its activity in eukaryotic cells but also the presence, type and number of phosphorylation sites (TPMs) within CagA produced by different strains of *H. pylori* can account for strain differences in virulence. The sequence motifs of the tyrosine phosphorylation sites indicate which cellular signal transduction cascades are likely to be targeted by different effector proteins. First, the motif determines which kinases can phosphorylate the tyrosine residue. Second, once the tyrosine is phosphorylated, the amino acid sequence determines which host-cell proteins can bind to the motif. Thus, the sequence of tyrosine phosphorylation sites indicates which cellular signal transduction cascades are likely to be targeted by the different effector proteins. The potential phosphorylated sites were predicted using MOTIF (www.motif.genome.ad.jp) and NETPHOS 2.0 (www.cbs.dtu.dk/services/NetPhos) softwares (Evans *et al*., 2001; Puls *et al*., 2002).

The first motif is defined by the amino acid sequence [RK]-$X_{(2,3)}$-[D/E]-$X_{(2,3)}$-[Y] which can be considered as a potential tyrosine phoshorylation site, despite the fact that at the moment there are no experimental evidence of its phosphorylation by human kinases. Another potential active TPM, corresponding to the sequence QKFGDQRY (highlighted in green, Fig. 6.1), or [K]-(FG)-[D]-(QR)-[Y], is close to Y^{122} in the genome sequence of Tomb *et al*. (Evans *et al*., 2001; Puls *et al*., 2002; Tomb *et al*., 1997). Another potential active TPM, presents only in few strains of *H. pylori,* corresponds to the sequence DKLKDSTKY (highlighted in green, Fig. 6.1), is close to the C-terminus of CagA and is poorly conserved. Since single mutations can abolish a TPM and since a phosphorylation motif within the CagA variable region might be amplified in sequence repeats, it is very likely that potential biological activity of CagA can vary from one strain to another (Evans *et al*., 2001; Puls *et al*., 2002). The other tyrosine phosphorylation site of CagA, which was determined experimentally *in vivo* (Stein *et al*., 2000), is characterized by the presence of a unique Glu-Pro-Ile-Tyr-Ala (EPIYA) motif (highlighted in red, Fig. 6.1), which is present in multiple copyes in the carboxy-terminal variable region of the protein (Evans *et al*., 2001; Puls *et al*., 2002; Higashi *et al*., 2002a and 2002b). From the sequences flanking these EPIYA motifs, four distinct segments, EPIYA-A, -B -C and -D, each of which contains a single EPIYA motif, have been identified in the protein (Higashi *et al*., 2002b; Hatakeyama, 2004; Higashi *et al*., 2005). The representative CagA protein of western *H. pylori* isolates possesses a 32-amino acid EPIYA-A and a 40-amino acid EPIYA-B segment, followed by a 34-amino acid EPIYA-C segment (‘A-B-C’-type CagA). The tyrosine residue that constitutes the EPIYA-C site is the major site of tyrosine phosphorylation in western CagA by Src family kinases and c-Abl in gastric epithelial cells, whereas those present within the EPIYA-A and EPIYA-B segments are only weakly phosphorylated in the cells (Higashi *et al*., 2002a and 2002b; Hatakeyama, 2006; Poppe *et al*., 2007). Most CagA proteins of *H. pylori* isolated in east Asian countries (East Asian CagA) also possess the EPIYA-A and EPIYA-B segments, but not the repeatable EPIYA-C one. Instead, they have a distinct EPIYA-containing sequence, termed the EPIYA-D segment, which is unique to east Asian CagA species (Higashi *et al*., 2002b; Hatakeyama, 2006). East Asian CagA is therefore regarded as ‘A-B-D’-type CagA. The EPIYA-D segment represents the major tyrosine phosphorylation site of East Asian CagA [Fig. 6.2]. Since EPIYA is a repeat sequence in CagA, the number and type of EPIYA motifs is positively correlated with the amount of phosphorylated CagA in the cell and its ability to activate its partners.



CagA Type	Tyrosine phosphorylation	Induction of hummingbird Phenotype
AB	+/-	+/-
ABC	+	+
ABCC	++	++
ABCCC	+++	+++
ABD	+	++++

Figure 6.2. The influence of EPIYA-repeat polymorphism on the pathophysiological activities of CagA. EPIYA-C [▮] and EPIYA-D [▮] sites are major tyrosine phosphorylation sites of Western CagA and East Asian CagA, respectively. EPIYA-D site of East Asian CagA binds SHP-2 more strongly than does EPIYA-C site of Western CagA. As a result, East Asian CagA exhibits greater activity to induce hummingbird cells than Western CagA. Among Western CagA species, those having larger numbers of EPIYA-C exhibit stronger SHP-2 binding activity and greater activity to induce the hummingbird cells than those having less numbers of EPIYA-C. (Adapetd from Hatakeyama *et al*., 2005).

CagA cAMP Phosphorylation Motifs (CPMs)

Analysis of the CagA amino acid sequences revealed that various CagA proteins contain between four and six potential cAMP-dependent phosphorylation sites, that is $[R/K]_{(2)}$-X-[S/T], for example KKXT (highlighted in yellow, Fig. 6.1) (Evans *et al*., 2001). This motif is present only in another Cag protein, CagM, which has only two of these motifs, compared to the four-six potential active motifs in CagA (Evans *et al*., 2001). It is interesting that CagM also has two TPMs and therefore the CPMs in CagM could be biologically relevant. From this analysis it can be concluded that CPMs in CagA are worth attention because of their locations in the protein and their potential role in CagA biological activity (Evans *et al*., 2001). As shown in the Fig. 6.1, there are two potential CPMs near the N terminus of CagA and at least two in the CagA variable region. The best conserved of these corresponds to the sequence KKQS, located 101 residues upstream from Y^{122} in the *H. pylori* 26695 strain genome sequence (Tomb *et al*., 1997) and KKKT five residues upstream from Y^{899}, typically repeated as KKVT (active) or KKVN (inactive) (Evans *et al*., 2001). Although phosphorylation of serine (S) or threonine (T) in CagA was not reported, it could be possible that transient phosphorylation events may escape detection and also that this type of phosphorylation may only be important in selected host cell types.

CagA Multimerisation Domains (MDs)

CagA multimerizes in gastric epithelial cells in a manner independent of its tyrosine phosphorylation. Using a series of CagA mutants, Ren and colleagues identified a conserved amino acid sequence (FPLXRXXXVXDLSKVG) which mediates CagA multimerisation and demonstrated that CagA multimerisation motif (CM), immediately following the EPYA-C and EPYA-D segments, is highly conserved between Western and East Asian CagA proteins (highlighted in blue, Fig. 6.1) (Ren *et al*., 2006). They have also shown that a single Src homology-2 tyrosine phosphatase (SHP-2) molecule binds to a preformed CagA dimer or multimer, demonstrating that CagA multimerisation is a prerequisite of the interaction and subsequent dereguraltion of its partner SHP-2 (Ren *et al*., 2006).

Membrane-targeting Signal

In addition to its role in tyrosine phosphorylation, the EPIYA motif acts as a membrane-targeting signal of CagA in gastric epithelial cells, although the CagA–membrane interaction does not require EPIYA tyrosine phosphorylation (Higashi *et al*., 2005). The presence of at least a single EPIYA motif is necessary and sufficient for the membrane association of CagA (Higashi *et al*., 2005). From these observations, it could be concluded that EPIYA motif has a dual function in membrane association and tyrosine phosphorylation in CagA.

CagA signals translocation

The CagA protein is so far the only effector protein known to be translocated by the *H. pylori* Cag TFSS (Odenbreit *et al*., 2000). Previous studies have shown that CagA translocation depends on a characteristic C-terminal motif, which is also present in other type IV-secreted proteins; in addition, it is influenced by its N-terminal region, which is a novel feature within this group of protein transport systems (Hohlfeld *et al*., 2006). Using *cagA* genes with 3' and 5' deletions of variable lenght and performing site specific mutagenesis, Hohlfeld and colleagues were able to demonstrate that the C-terminal 20 amino acids and a N-terminal fragment greater than 57 amino acids are required for the recognition of CagA as a TFSSs substrate (highlighted in violet, Fig. 6.1) (Hohlfeld *et al*., 2006). Many effector proteins secreted by TFSS contain a domain at their extreme C-terminus which is necessary and sufficient for secretion (Vergunst *et al*., 2005). The C-terminal regions of type IV-secreted proteins present a positively charged and weakly conserved motif containing several arginine or lysine residues ([R/K]X_{3-4}[R/K]). The latter was proposed, by sequence comparison and by

extensive screening mutagenesis, to constitute the translocation signal (Fig. 6.3). Interestingly, the positive charges present at the C-terminal end are neither necessary nor sufficient for CagA translocation, but the replacement of the C-terminal region of CagA with that of other type IV-secreted proteins reconstitutes CagA translocation competence. This suggests that the positive charges are not essential to determine a secretion signal in CagA; it could possibly be considered more as a conformational rather than a linear signal (Hohlfeld *et al*., 2006). Toxins from other bacteria contain N-terminal (sec-dependent) signal sequences, for example the DotA protein of *L. pneumophila* (Nagai and Roy, 2001). This probably means that entry of these proteins into the type IV secretion channel is accomplished from the periplasmic space. Interestingly, the DotA protein does not have a C-terminal translocation signal similar to the one previously described. Thus, N-terminal (sec-dependent) and C-terminal signal sequences might be alternative possibilities used by CagA to gain access to the type IV translocation channel. In this respect, the Cag type IV secretion system seems to diverge from other systems not only with respect to its composition and architecture, but also in terms of substrate recognition and transport. A model has been proposed to explain the involvement of both N- and C-terminal regions in CagA translocation, where the C-terminal signal sequence would be required for recruitment to the secretion apparatus, possibly also for translocation across the cytoplasmic membrane, and the N-terminal part would be instead involved in a later step, such as translocation across the outer membrane (Hohlfeld *et al*., 2006).

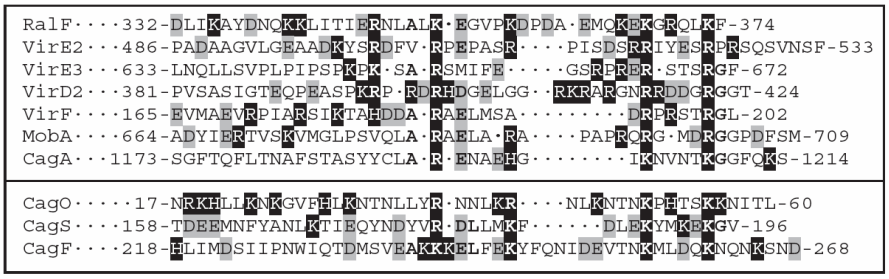


Figure 6.3. Alignment of the C-terminal amino acid sequences of type IV-secreted proteins: CagA (from the *cag*-PAI system of *H. pylori*), RalF (from the Dot/Icm system of *L. pneumophila*), VirD2, VirE2, VirE3 and VirF (from the VirB/D4 sysetm of *A. tumefaciens*) and the MobA relaxase of plasmid RSF1010 are shown. Furthermore, the C-terminal regions of the putative CagO, CagS and CagF proteins are shown. Positively charged amino acids are shaded in black, negatively charged amino acids in grey and residues that are well-conserved are printed in boldface (Adapted from Hohlfeld *et al*., 2006).

Potential similarities between amino acid sequences of CagA and Eukaryotic Proteins

The CagA full-length sequence has no amino acid similarity with eukaryotic proteins involved in the signaling patways, but it may have structural similarities to proteins that can link and activate these pathways. For example, Hatakeyama hypothesized that CagA may share some of the functional properties of Gab proteins, docking molecules that serve as scaffolds to mediate activation of multiple signals downstream of RTK (Hatakeyama, 2003). To elucidate the role and structure of single CagA domains, we have compared their amino acid sequences with sequences of proteins with known structures using BLASTp (www.wxpasy.ch/tools/blast). Like in previous studies, this homology search did not reveal any striking overall homology with other known protein sequences, however some similar boundaries domain with high score were identified and analized [Fig. 6.4 A].

For instance, the extreme N-terminal region of CagA shows some sequence similarity with the C-terminal domain of Yeast Bro1 protein (Kim *et al*., 2005; PDB ID code: 1ZB1), with 26 % identity and 44 % similarity in the regions between residues 21 and 98. The proteins known as Bro1 in yeast and Alix in human are components of the Multivesicular bodies (MVBs), late endosomes that contain vesicles within their lumen that are formed by membrane invagination (Kim *et al*., 2005). Since MVBs have been reported to be extraordinarily rich in the unusual phospholipid 2,2' lysobisphosphatidic acid (2,2' LBPA), and Bro1 and Alix proteins contain a highly elecropositive region, it has been proposed that they could favour the binding to this acidic phospholipids (Kim *et al*., 2005). Since also CagA protein is charachterized by the presence of many positive charged reisdues, a similar role, in the binding to the anionic membrane surface could be proposed for the N-terminal domain of CagA.

Between the N-terminal region and the central region of CagA, a domain of about three-hundred residues which shows sequence similarity with three eukaryotic proteins was identified: the PDZ domain of mouse Harmonin protein (Yamada *et al*., *to be published*; PDB ID code: 1V6B), the PH domain of the human LrbaBGL (Gebauer *et al*., 2004; PDB ID code 1T77) and the liganded sterol carrier protein type 2 (Haapalainen *et al*., 2001; PDB ID code 1IKT) which displayed 23, 29 and 31 % identity and 47 %, 43 and 47 % similarity in the CagA regions between residues 147-218, 249-356 and 337-384, respectively. The PDZ domain is one of the most common protein-protein interaction domains in humans (Jemth *et al*., 2007; Yamada *et al*., to be published). PDZ domains were also found in several *E. coli* proteins (Ponting *et al*., 1997a). The function of these domains in tyrosine phosphatases is to bind to the C-

terminal tetrapeptide or tripeptide (for example, [S/T]-X-[V/L/I], motifs often found on ion channel and receptor proteins (Ponting *et al*., 1997b). Like PDZ domain, the modular pleckstrin homology (PH) domain is present in numerous modular eukaryotic proteins implicated in signal transduction, membrane trafficking, and cytoskeletal dynamic. PH domains belong to one of the largest domain families. They have been most thoroughly investigated as modules that target membranes through recognition of phosphoinositide head groups. A number of other PH domains, however, have been well characterized as protein-interaction modules or as dual phosphoinositide and protein binding modules (Di Nitto *et al*., 2007; Lemmon *et al*., 2007).

Both PH and PDZ domains can bind phosphoinositides and are frequently found in signal transducing proteins. Therefore, the presence of PH- and PDZ-like domains of CagA could provide a direct link between the CagA protein and many of the responses of gastric target cells to *H. pylori* that can be attributed to interactions between CagA and host cell signal transduction networks. The sterol carrier protein type 2 (Scp-2), also known as a non-specific lipid transfer protein, has a domain of about fifty residues (31-78) wich shows 31 % identity and 47 % similarity with a CagA domain included between residues 337 and 384. The similarity with a domain of a protein whose functions is the transfer of fatty acyl-CoA to other enzymes was found also in prokaryotic proteins (see below *E. coli* ACP), supporting the hypothesis that the thioesterase *Hp*YbgC could be a potential partner interacting with CagA central domain (Rain *et al*., 2001; Terradot *et al*., 2004; Chapter 6.3.3 and Chapter 7).

Moreover, like the PDZ and PH domains, Scp-2 domain is also involved into the lipid binding. A database search with the N-terminal and central CagA amino acid sequence has revealed another case of sequence similarity which might indicate a basis for interaction between CagA and host proteins. The CagA region between residues 307 and 376 shows 30 % identity and 47 % similarity with the Shc SH2 domain (Zhou *et al*., 1995; PDB ID code 1TCE). Shc is a widely expressed adapter protein that has been implicated in Ras activation following the stimulation of a number of different cell surface receptors, including those of gowth factors, cytokine and antigens. In particular, the SH2 domain of Shc has been shown to interact with the tyrosine-phosphorylated receptors of EGF and PDGF, as well as the ζ chain of the T-cell recptor (Zhou *et al.*, 1995). Receptors stimulation leads then to tyrosine phosphorylation of Shc and, following binding with another adapter protein, the growth factor receptor bound protein 2 (Grb2), which activates Ras via son of sevenless (Sos) (Zhou *et al*., 1995). Since also CagA has been shown to bind Grb2 (Mimuro *et al*., 2002) and since the CagA–Grb2 complex has been demonstrated to

promote proliferation of gastric epithelial cells through activation of the Ras-dependent extracellular signal-regulated kinases 1 and 2 (Erk1/2) (Hirata *et al*., 2002; Brandt *et al*., 2005; Kim *et al*., 2006), it is beliaveble that also this domain of CagA could be have a similar role of the SH2 Shc domain.

Moreover, in the central region of CagA was possible to find some other similarities with eukaryotic proteins. For instance, between residues 394-441 and 450-519, the CagA sequence shows 37 %, 32 % and 36 % identity and 56 %, 44 % and 60 % similarity with the dimerization domain of human Skap-Hom (Togni *et al*., 2005; Tang *et al*., *To be published*; PDB ID code 1U5E), the human Mo25 domain (Milburn *et al*., 2004 ; PDB ID code 1UPK) and the human Geminin coiled-coil (Saxena *et al*., 2004; PDB ID code 1UII) proteins, respectively.

Several cytosolic adaptor proteins have been identified during the last years which appear to be involved in the reorganization of the cytoskeleton and/or integrin-mediated adhesion after external engagement of immunoreceptors. In T cells, these include the cytosolic adaptor proteins Adap (adhesion and degranulation promoting adaptor protein) and Skap55 (Src-kinase-associated phosphoprotein of 55 kDa) (Togni *et al*., 2005). In contrast to Adap and Skap55, which are expressed in T cells and T lymphocytes, respectively, the cytosolic adaptor Skap-Hom (Skap55 homologue) is an adaptor protein that is more widely expressed within the hematopoietic system (Togni *et al*., 2005). Similarly to Skap55, Skap-Hom has been reported to associate with Adap via its SH3 domain and to represent a specific substrate for the Src family protein tyrosine kinase Fyn. In particular, the Skap-Hom seems to be involved in integrin mediated signaling pathways. Moreover, it has been demonstrated that the enteropathogenic species of *Yersinia enterocolitica* exert resistance to phagocytosis by injecting the virulence factor YopH, a protein tyrosine phosphatase, into host cells via a type III secretion system. In these cells YopH apparently dephosphorylates Adap and Skap-Hom (Togni *et al*., 2005). Either by itself or in concert with other mechanisms, this leads to the abrogation of the phagocytic process and causes cellular detachment. Since CagA, like Skap-Hom protein, has been proposed to act as an adaptor protein involved in reorganization of the cytoskeleton and it has been shown to represent a specific substrate for the Src family protein tyosine kinases, this similarity can be significant and suggests a function for this domain.

Mouse protein 25 (MO25) is a 40-kDa protein that, together with the STE20-related adaptor-α (STRADα) pseudo kinase, forms a regulatory complex capable of stimulating the activity of the LKB1 tumor suppressor protein kinase and activate AMPK (Milburn *et al*., 2004). Although MO25 is topologically similar to Armadillo

helical-repeat proteins (ARPs), the protein does not contain any known motifs. It is charaterised by the presence of a concave positive charged surface that can be expected to bind a ligand that has not yet been defined (Milburn *et al*., 2004). These data suggest that MO25 acts as a scaffolding protein, and an analogous role could be played by the similar CagA domain.

The human Geminin coiled-coil is a cellular protein that interacts with many proteins like Cdt1 and Mcm2-7, inhibiting the replication during S phase. It prevents multiple cycles of replication per cell cycle and prevents episome replication. It also directly inhibits the HoxA11 transcription factor (Saxena *et al*., 2004). The function of Geminin does not seem to be correlated with known functions of CagA, but like other proteins, below described, Geminin forms a long nonglobular coiled-coil homodimer that is characterized by the presence of an array of charged residues (Saxena *et al*., 2004). This long charged cylinder suggests that this domain, like in CagA, can bind several proteins via electrostatic interactions.

The CagA region between residues 610 and 734 shows 30 % identity and 51 % similarity with the Tropomyosin dimeric binding sites for actin.

We know that the translocated CagA protein into the gastric epithelial cells undergoes tyrosine phosphorylation, a process which was demonstrated to be associated with characteristic actin cytoskeletal rearrangements, resulting in a scatter factor-like ('hummingbird') phenotype. Like Geminin, the Tropomyosin model contains multiple coiled-coil regions, each of which is capable of forming homodimers that bind actin.

Tight junctions are composed of three types of transmembrane proteins: claudin, junctional adhesion molecule (JAM) and occludin. Occludin is a transmembrane protein, localized at tight junctions, whose functions are complex, despite yet poorly understood. Current evidence supports a role for occludin in both the formation of the paracellular barrier and in cell signaling. In particular, the C-terminal cytoplasmic domain of occludin has a large positively charged surface that contains the binding site for ZO-1, and this surface is required for proper localization of occludin to cell–cell junctions. The distal cytoplasmic domain of occludin is an important mediator of protein–protein interactions at the tight junction, and it was shown to interact directly with the cytoskeletal protein F-actin, cingulin, and the membrane trafficking protein VAP33, suggesting an important role for this domain. The cytoplasmic tail of occludin is also the target of several kinases, including the c-Src and c-Yes oncogenes, whose cellular activation results in dramatic reorganization of cell–cell junctions in epithelial cells (Li *et al*., 2005; PDB accession code 1WPA). The CagA region between residues 686 and 730 shows 29 % identity and 57 % similarity with the Occludin C-terminal

domain. CagA has several homologies with the C-terminal domain of CagA. Like Occludine, CagA interacts with the scaffold protein ZO-1 and contains many lysine and arginine residues, suggesting that this positive charged surface could be functionally important also in CagA, and perhaps responsible for binding to ZO-1 or other proteins.

A search using the C-terminal CagA amino acid sequence revealed only one case of sequence similarity between CagA and host proteins. The CagA region between residues 955 and 1034 shows 27 % identity and 46 % similarity with the SH2-kinase linker domain of the Human tyrosine kinase ZAP-70 protein (Deindl *et al*., 2007; PDB ID code: 2OZO). ZAP-70 (zeta-chain-associated protein of 70 kDa) is crucial for signalling by the T cell antigen receptor (TCR) and the key step that links TCR activation to ZAP-70 is the phosphorylation by Src family kinases (e.g., Lck).

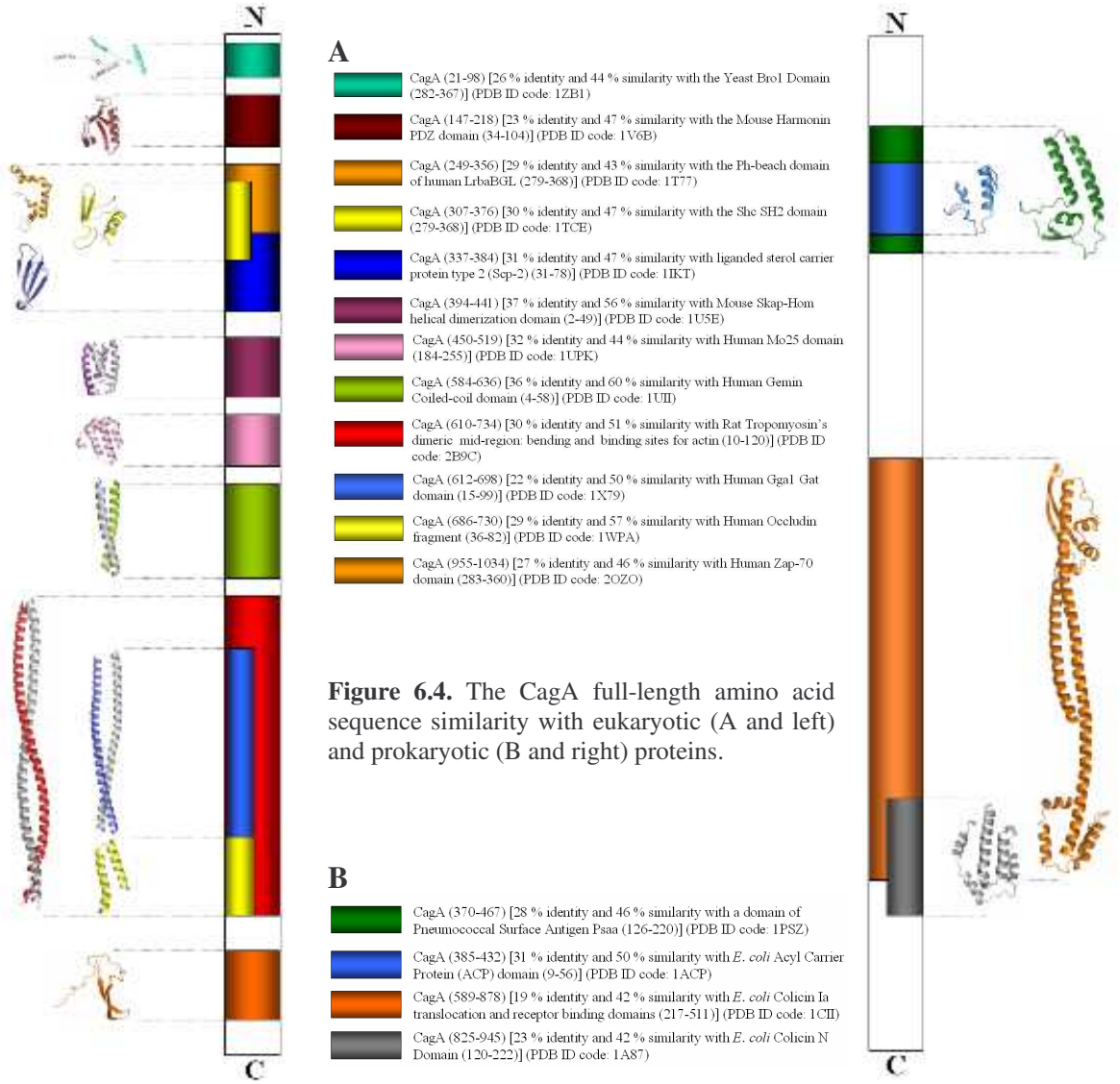


Figure 6.4. The CagA full-length amino acid sequence similarity with eukaryotic (A and left) and prokaryotic (B and right) proteins.

ZAP-70 contains two tandemly arranged SH2 domains (N-SH2 and C-SH2) and a kinase domain. The C-SH2 domain is connected to the kinase domain by a flexible region named SH2-kinase linker. The CagA region that is similar to the SH2-kinase linker contains the C-EPYA motif and like ZAP-70 also CagA is phosphorylated at the tyrosine residue of EPYA motif by a kinase belonging to the Src family. Moreover, the evidence that tyrosine residues of ZAP-70 may serve as SH2 docking sites when phosphorylated, induce us to think that also the CagA region between residues 955 and 1034 could play a similar role when phosphorylated.

Finally, an interesting feature of the primary structure of this protein is the presence of structures of homopolymeric amino acid sequences, most notably polyasparagine. In searches comparing this asparagine-rich region with regions in various protein sequence data-bases, there was strong homology with sequences of nucleotide-binding proteins in both prokaryotic and eukaryotic organisms.

Potential similarities between amino acid sequences of CagA and Prokaryotic Proteins

The homology search did not reveal any striking overall amino acid similarity of CagA full-length seuqnce with prokaryotic proteins, however, as for the eu eukaryotic proteins, some similar boundaries domain with high score were identified and analized [Fig. 6.4 B]. The first four-hundred amino acids of CagA were compared to those in the Protein Data Bank (PDB) by submitting the sequence to the BLAST expasy server. The two most similar polypeptides were found to be Pneumococcal surface antigen PsaA and the *E. coli* acyl carrier protein (ACP), with 28 %, 31 % identity and 46 %, 50 % similarity in the CagA regions between residues 370-467 and 385-432, respectively (Kim *et al*., 1990; Lawrence *et al*., 1998). PsaA is a putative ATP-binding cassette-type (ABC-type) binding protein involved in the uptake of Mn^{2+} and possibly Zn^{2+} and is considered to be both a potential drug target and a candidate vaccine component (Lawrence *et al*., 1998). The *E. coli* ACP is a small (10 kDa) highly acidic protein that is required for a variety of biosyntethic processes (fatty acids, polyketides, post-translational acylation of proteins, quorum sensing substrates). Moreover, a recent study showed that *E. coli* ACP interacts with the Acyl-CoA thioesterase enzyme *Ec*YbgC (Gully *et al*., 2006). Since the *H. pylori* homologue enzyme, *Hp*YbgC, has been shown to interact specifically with a domain of CagA between reisudes 392 and 733 (Rain *et al*., 2001; Terradot *et al*., 2004) we hypothesize that the CagA domain could mimic the ACP fold interacting with *Hp*YbgC enzyme as well as with human thioesterases.

Pore-forming colicins are water-soluble bacteriocins capable of binding to and translocating through the *E. coli* cell envelope. In particular, the ion-channel forming colicins A, B, E1, Ia, Ib and N kill bacterial cells selectively by co-opting bacterial active-transport pathways and forming voltage-gated ion conducting channels across the plasma membrane of the target bacterium (Wiener *et al*., 1997; Vetter *et al*., 1998).
The CagA region between residues 785 and 905 shows 19 % identity and 42 % similarity with a domain of the *E. coli* colicin Ia comprising residues 308-435. The colicin Ia domain includes the receptor binding and the channel forming domains, connected through two very long α-helices (Wiener *et al*., 1997; PDB accession code 1CII). The colicin Ia domain model, similar to the geminin, tropomyosin and occludin domain models, is constituted by a long nonglobular structure.
Moreover, the CagA region including the 825-945 residues shows 23 % identity and 42 % similarity with the pore-forming domain of colicin N, that consists of a bundle of α helices arranged in three layers forming the so called 'globin-like' topology, also found in the pore forming apoptosis regulatory protein Bcl-xL (Vetter *et al*., 1998; PDB accession code 1A87).

6.1.2 *Helicobacter pylori* proteins interacting with CagA

Few years ago the interaction map of *Helicobacter pylori* was described. It was the first prokaryotic interaction map, obtained through the strategy of a variant of the two-hybrid assay (Rain *et al*., 2001). More than two hundred *H. pylori* random genomic fragments were selected and screened against a highly complex library of genome encoded polypeptides; more than one thousand interactions were identified among *H. pylori* proteins (PIMRider database, http://pim.hybrigenics.com; Rain *et al*., 2001).
This protein-protein interaction map revealed some biological pathways and at same time permitted to predict the function of some uncharacterised proteins. Using this method, it was identified the interaction between CagA and HP0496, which was then validated biochemically by a pull-down experiment (Rain *et al*., 2001; Terradot *et al*., 2004). The interaction is domain specific: in fact, the minimum fragment of CagA which was identified to interact with the HP0496 full-length protein is the central domain, including residues from 392 to 733 (Rain *et al*., 2001; Terradot *et al*., 2004). HP0496 does not belong to the *cag*-PAI and shows homology (30 % identity) to thioesterase proteins, especially to YbgC-like protein, a cytoplasmic component of the Tol-Pal system (Angelini *et al*., Chapter 7).

The yeast two-hybrid technique has however many limitations; in fact, the protein of interest are often expressed in a heterolgous system, yeast or *E. coli*, and the mostly binary interactions are observed in a particular compartment, the cytoplasm, whereas we know that many protein components of the *cag*-TFSS of *H. pylori* are membrane localised. For this reason other approaches, based on different physiochemical criteria (pI-, Mr- or affinity-based purification) are now used to analyse the complex, including: co-purification, co-immunoprecipitation and two-dimensional (2D) polyacrilamide gel elecrophoresis under both native and denaturating conditions. Just using a native two-dimensional (2D) polyacrilamide gel elecrophoresis method combined with liquid chromatography mass spectrometry (LC-MS/MS) it was recently found in the *H. pylori* cytoplasmic sample a multiprotein complex including CagA and the DNA gyrase subunit A (GyrA). The complex corresponds to a molecular mass of 475 kDa, suggesting a probable oligomerisation of CagA and/or GyrA in this complex (Pyndiah *et al*., 2007). Since cultures of CagA-negative *H. pylori* strains showed a slower growth than those with CagA/GyrA interaction (Censini *et al*., 1996; Van Doorn *et al*., 1997), Pyndiah and collegues proposed that this interaction could have a favourable effect on the normal DNA replication process, leading to a better development of the bacterial cell (Pyndiah *et al*., 2007). However, Hybrigenics (Rain *et al*., 2001) described an interaction between GyrA and CagI, which includes the small cystein-rich protein B (HcpB), which is a weak beta-lactamase (Rain *et al.,* 2001). At the moment these interactions have not been confirmed biochemically and their biological significance is still unclear.

The specific mechanism by which *H. pylori* translocates CagA into the host gastric epithelial cell is largely unknown. Till one year ago, no confirmed interactions between components of the Cag-TFSS were reported, suggesting that Cag-Cag protein interactions are not easily detected with the two-hybrid methods. Very recently the first interaction between CagA and another Cag protein, CagF, was reported. The interaction between CagA and CagF was firstly identified by co-immunoprecipitation (Couturier *et al*., 2006), suggesting that it is a strong and direct interaction, independent of a functional TFSS and of the presence of other translocation factors (Pattis *et al*., 2007). In particular, Couturier and colleagues have shown that CagF localizes at the bacterial cytoplasmic membrane and have demonstrated that it may function as a secretion chaperone-like molecule that recruits CagA to the TFSS (Couturier *et al*., 2006). These data were also supported by Pattis and colleagues, confirming the CagF membrane localisation and identifying a specific region of CagF-CagA interacting domain. They have shown that it is localised close to the C-terminal

translocation signal of CagA (Pattis *et al*., 2007), whereas Couturier and colleaugues demonstrated that the interaction occurs at least within the amino and central 100 kDa region of CagA (Couturier *et al*., 2006). All these results suggest that the CagF binding might be a quality control mechanism that ensure the integrity or the correct folding of the CagA substrate prior to the entry into the translocation chanel. To validate and elucidate the molecular basis of these interactions, we performed some biochemically and structurally studies on some of them.

6.1.3 Human proteins interacting with CagA

Many signal transduction cascades involve post-translational modifications of proteins by phosphorylation (Blume-Jensen *et al*., 2001; Schlessinger *et al*., 2003; Pawson, 2004). Specific kinases can attach a phosphate group to the side chains of histidine, serine, threonine and tyrosine; this chemical reaction changes the biochemical properties of a protein. Although histidine, serine and threonine phosphorylation occurs in all species from bacteria to mammals, tyrosine phosphorylation is thought to be restricted largely to multicellular organisms. Thus, it might be surprising that more and more effector proteins of bacterial pathogens are found to become phosphorylated on tyrosines in host cells. During the last twenty years, it was reported that bacterial proteins translocated by bacterial type III and type IV secretion systems can also be tyrosine-phosphorylated. In particular, the enteropathogenic *E. coli* (EPEC) was found to use a type III system (TTSS) to translocate the Tir protein into the epithelial cell, where it is tyrosine phosphorylated (Tir^{PY}) (Kenny *et al*.,1997; Deibel *et al*., 1998). Two years later, it was shown that *H. pylori* uses a type IV secretion system (TFSS) to deliver CagA, which becomes tyrosine-phosphorylated ($CagA^{PY}$) in gastric epithelial cells (Segal *et al*., 1999; Stein *et al*., 2000; Asahi *et al*., 2000; Odenbreit *et al*., 2000; Backert *et al*., 2000). Recently, it was also found that effector proteins Tarp and BepD–F of *Chlamydia trachomatis* and *Bartonella henselae,* respectively, were injected into host cells and phosphorylated on tyrosine residues (Clifton *et al*., 2004; Schulein *et al*., 2005) These results indicate that the translocation of tyrosine kinase substrates could be a common strategy used by bacterial pathogens to disturb host-cell signal transduction and thereby to provoke cellular dysfunction that eventually leads to cell transformation.

A comparison with the tyrosine phosphorylation sites of bacterial effectors suggest that all of them are especially similar to the consensus substrates of **Src family**

kinases (SFKs). The SFKs regulate signal transduction by a diverse set of cell surface receptors in many different cell types (Parsons *et al*., 2004). The Src family has several members with similar structural features and significant sequence homology (Thomas *et al*., 1997). The N-terminus of SFKs (sometimes referred to as the SH4 domain) is lipid modified, which helps to attach the proteins to the cytoplasmic face of cellular membranes. The SH3 domain is involved in protein–protein interactions and can associate with specific PxxP motifs in interaction partners, whereas the SH2 domain binds specifically to phosphorylated tyrosines. The kinase (or SH1) domain is instead responsible for the tyrosine kinase activity. Enhanced SFK activity can induce cell transformation and is involved in carcinogenesis. SFK activity is tightly regulated by a molecular switch: autophosphorylation at Y418 is required for maximal catalytic activity. When Y527 is phosphorylated by Csk, it can interact with the SH2 domain. This results in a closed and catalytically inactive conformation of Src. The SH3 domain is also involved in its swithch. The *H. pylori* CagA protein was the first bacterial effector injected into the human cell via a TFSS for which the cellular tyrosine kinase was identified. *In vitro* and *in vivo* studies of CagA cellular activities in epithelial cells revealed that upon membrane localization and subsequent tyrosine phosphorylation in the C-terminal EPIYA motifs by several members of **SFKs** (**c-Src, Lyn, Fyn** and **Yes**) (Stein *et al*., 2002; Selbach *et al*., 2002) and the **non-receptor tyrosine kinase c-Abl** (Poppe *et al*., 2006), $CagA^{PY}$ seems to function as a scaffolding adaptor and interacts with a large number of host proteins involved in a variety of signal transduction pathways that regulate cell cycle, cell growth, cell motility and cell polarity in both CagA phosphorylation-dependent and phosphorylation-independent manners.

Phosphorylation-dependent biological activities of CagA

It has been hypothesised that phosphorylated CagA might function as an adapter protein to recruit eukaryotic proteins in a multi-protein and –enzyme complex, which includes a large number eukaryotic partners (Higashi *et al*., 2002; Tsutsumi *et al*., 2003; Churin *et al*., 2003; Mimuro *et al*., 2002, Suzuki *et al*., 2005; Saadat *et al*., 2007). Eukaryotic signal transduction depends on specific domains that mediate protein–protein interactions (Pawson *et al*., 2003). Two different types of protein modules, termed Src homology 2 (SH2) and phosphotyrosine binding (PTB) domains, can bind to tyrosines in a phosphorylation specific manner (Schlessinger *et al*., 2003). SH2 domains, which bind with high affinity to tyrosine-phosphorylated proteins, were first identified in Src family kinases; hence the name Src homology 2. Tyrosine

phosphorylation of a pathogenic bacteria effector protein can therefore stimulate the recruitment of SH2-domain-containing proteins, which then initiate further signalling processes. Phosphotyrosine binding (PTB) domains are also found in a growing number of cellular proteins with diverse functions. They are structurally unrelated to SH2 domains, but they can also interact specifically with tyrosine-phosphorylated targets. However, in contrast to SH2 domains, the majority of PTB domains appear to bind NPxY motifs in a phosphorylation-independent manner. So far, PTB domains that interact with tyrosine-phosphorylated bacterial effectors have not been identified yet. Similarly to kinases that prefer tyrosines within a specific sequence motif, SH2 and PTB domains bind tyrosine only within the context of specific flanking amino acid residues. The specificity of several SH2 domains was identified systematically using a peptide-based approach (Songyang et al., 1993; Songyang et al., 1994). In particular, the CagA phosphorylated protein ($CagA^{PY}$) was shown to interact with several host signalling molecules containing a SH2 domain, including the SHP-2 (Higashi *et al*., 2002), Csk (Tsutsumi *et al*., 2003) and Crk adapter protein (Brandt *et al*., 2007). The sequence of the CagA phosphorylation site is moreover similar to the consensus high-affinity binding sequences pYIXI and pYAKI for interacting with the SH2 domains of SHP-2 and Csk, respectively (pY-[V/T/A/I/S]-X-[L/I/V]-X-[F/W]).

The interaction between $CagA^{PY}$ and **Src-homology 2 (SH2) domain containing tyrosine phosphatase, (SHP-2)**, a cytoplasmic protein tyrosine phosphatase transducer of receptor tyrosine kinases (RTKs), is the best characterised (Higashi *et al*., 2002). The complex formation between tyrosine phosphorylated CagA and SHP-2 was reported *in vitro* in cells transfected with CagA, but also *in vivo* by co-immunoprecipitation in biopsy specimens of gastric mucosa in CagA positive *H. pylori*-infected patients, indicating the importance of the complex formation in the pathogenesis of *H. pylori*. During *H. pylori* infection, upon CagA tyrosine phosphorylation by SFK, SHP-2 is recruited to the cell membrane and is activated by interactions with $CagA^{PY}$ via its two tandem-repeated SH2 domains, localised on the N-terminal half (N-SH2) and on the C-terminal half (C-SH2) protein. This indicates that two of the tyrosine-phosphorylated EPIYA sites, either in *cis* or *trans*, are actively involved in the stable complex formation between CagA and SHP-2. The crystal structure of SHP-2 suggests that the N-SH2 domain occludes the catalytic cleft of the PTP domain, blocking substrate access. This closed (inactive) structure keeps the basal phosphatase activity level of SHP-2 low. Binding of tyrosine-phosphorylated CagA to the SH2 domains induces a conformational change in SHP-2 that relieves inhibition of the PTP domain by the N-SH2 domain, resulting in the activation of SHP-2

phosphatase activity (Higashi *et al*., 2002a and 2002b; Hatakeyama *et al*., 2004; Hof *et al*., 1998). In principle, SHP-2 activation could explain how $CagA^{PY}$ induces dephosphorylation of host-cell proteins. Physiologically, SHP-2 is known to activate the extracellular signal-regulated (Erk-MAP) kinase by both Ras-dependent and -independent mechanisms. Consistent with this, the interaction between $CagA^{PY}$ and SHP-2 appears to activate the Erk-MAPK pathway via Rap1 and B-Raf that is important for cell-morphological transformation, termed the hummingbird phenotype, which is characterized by elongated cell-shape with dramatic cytoskeletal rearrangements (Higashi *et al*., 2004). Inhibition of CagA tyrosine phosphorylation or disruption of the CagA–SHP-2 complex inhibits induction of the hummingbird phenotype (Higuchi *et al*., 2004). Thus, CagA-deregulated SHP-2 plays a crucial role in induction of the hummingbird phenotype by CagA in gastric epithelial cells. Recent studies have shown that gain-of-function mutations in *PTPN11*, the gene encoding SHP2, are associated with various human malignancies, such as child leukemia and some solid tumors, indicating that SHP2 is an oncoprotein that is involved in human malignancies (Loh *et al*., 2004).

However, $CagA^{PY}$ mediated cortactin dephosphorylation does not require SHP-2, but is induced by Src inactivation (Selbach *et al*., 2003). Src inactivation is probably a result of the interaction between $CagA^{PY}$ and the SH2 domain of the **C-terminal Src kinase (Csk)** (Tsutsumi *et al*. 2003). Csk is a tyrosine kinase that negatively regulates SFKs. It was speculated that through the interaction, CagA stimulates the kinase activity of Csk, which in turn phosphorylates SFK at the C-terminal inhibitory tyrosine residue. Thus SFK are kinases that phosphorylate CagA but at the same time CagA seems to inhibt the SFK activity through Csk interaction and its activation; this suggests the presence of a feedback mechanism that attenuates phosphorylation-dependent activities of CagA that could be not completely unregulated, limiting acute and serious damages to the gastric epithelial cells. (Tsutsumi *et al*., 2003). The negative feedback control seems to ensure a long-term equilibrium between *cagA*-positive *H. pylori* and the human host without excessive mucosal damage to the host.

The inactivation of Src kinases by phosphorylated CagA protein is important not only in limiting the activity of the same CagA, but it also results in the dephosphorylation by CagA of other AGS cell actin binding proteins, namely cortactin and ezrin (Selbach *et al*., 2003; Selbach *et al*., 2004). They may promote cell motility, leading to cell scattering and host cell elongation, a morphology which was called the "hummingbird" phenotype. However, how $CagA^{PY}$ stimulates rearrangement of the actin cytoskeleton is still under investigation. What we know from systematic

mutagenesis studies is that the CagA phosphorylation at a single tyrosine site (Y^{972}) is required for the elongation phenotype (Backert *et al*., 2001). It was proposed that the infection causes a redistribution of cortactin to actin-rich cellular protrusions, a hallmark of the elongation phenotype. Both ezrin and cortactin modulate actin organization and were shown to regulate actin dynamics during cell contact with pathogens, including enteropathogenic *E. coli* and *Salmonella* (Daly, 2004). **Cortactin** binds **F-actin** and **actin-related protein 2/3 (Arp2/3)** and in its non-tyrosine-phosphorylated form increases actin cross-linking capability, which may be important for stabilizing actin bundles in CagA-induced cellular elongation. **Ezrin** is a member of the ERM (ezrinradixin-moesin) protein family and is found in specialized membrane areas, such as microvilli, where it links F-actin to membrane proteins (Bretscher *et al*., 2002). The downstream effects for the dephosphorylation events of cortactin and ezrin after Src inactivation is still poorly understood (Bourzac *et al*., 2005).

Recently, it was also reported that CagA interferes with cell–cell adhesion (Amieva *et al.*, 2003) by another Src substrates which was found to be tyrosine dephosphorylated upon CagA phosphorylation, namely **vinculin**, an important regulator of cell adhesion (Moese *et al*., 2007). Moese and collegues found that $CagA^{PY}$ reduces the adhesion interfering with the Focal adhesions (FAs), dynamic structures which are costantly assembled and disassembled. FA assembly is often regulated by the **Rho GTPases Cdc42** and **Rac1**, two key mediators of actin dynamics and polymerization that control cytoskeleton-based movement via signaling to the Arp2/3 complex through their downstream nucleation-promoting factors **N-WASP** and **WAVE**, respectively. The Cdc42 and Rac1 signaling cascades are hijacked by several important human pathogens, both viral and bacterial, as part of their infectious life cycles (Gouin *et al*., 2005, Finlay, 2005; Bourzac *et al*., 2007). In the case of vinculin, the reduced phosphorylation of tyrosines Y^{100} and Y^{1065} after CagA translocation disrupts the interaction between vinculin and p34Arc, a subunit of the Arp2/3 complex (Moese *et al*., 2007). Another substrate that was recently reported to be dephosphorylated by the CagA-SHP-2 complex is the **focal adhesion kinase (FAK)**. CagA-activated SHP-2 dephosphorylates at the activating tyrosine phosphorylation sites of FAK in gastric epithelial cells, leading to an inhibition of FAK kinase activity (Tsutsumi *et al*., 2006). Recently, other CagA-interacting proteins were determined by immunoprecipitation using anti-CagA antibody and proteomic analysis. Baek and collegues found that **alpha-Pix (PAK-interactive exchange factor)**, a protein constitutively expressed in AGS cells, interacts with $CagA^{PY}$. They have shown that the interaction of alpha-Pix

with CagA was increased by *H. pylori* infection in AGS cells and the phosphorylation of CagA induces the dephosphorylation of alpha-Pix in AGS cells (Baek *et al.*, 2007). alpha-Pix is a family of PAK-binding proteins that strongly activates **PAK (p21-activated tyrosine kinase),** which in turn regulates changes in gene expression and mediates actin cytoskeletal and cell morphological changes (Baek *et al.*, 2007).

$CagA^{PY}$ was also shown to activate the **mitogen-activated protein kinases (MAPKs)**, including **extracellular signal regulated kinase (ERK)**, **p38**, and **JNK**, leading to the phosphorylation of **ATF-2** and **CREB1**. It also increases the nuclear expression of **c-Fos** and **c-Jun** transcription factors, resulting in activation of the **CRE** and **activating protein 1 (AP1)** promoter elements on the human cyclin D1 promoter, with initiation **of cyclin D1** expression. Cyclin D1 overexpression, leading to G1-S phase progression, and host cell survival, favours continued *H. pylori* adherence to host tissues, providing both positive selection for $cagA^{+}$ genotypes and enhancing the risk for malignant transformation in the host (Chang *et al*., 2006).

It remained unknown why CagA is constantly phosphorylated, even when Src kinases are inactivated, suggesting the presence of another host tyrosine kinase to ensure constant CagA phosphorylation in sustained *H. pylori* infections. Recently, two different research groups identified the **non-receptor tyrosine kinase c-Abl** as a new binding partner of CagA in epithelial cells (Poppe *et al*., 2007; Tammer *et al*., 2007). The tyrosine kinase c-Abl localizes to F-actin-associated cellular structures and focal adhesions, and binds directly to F-actin. This could be important for actin-generated structures involved in membrane ruffling, cell spreading and cell migration in response to growth factors and extracellular matrix signals. Poppe and collegues demonstrated that, upon *H. pylori* infection, c-Abl directly interacts with CagA, affects the phosphorylation of CagA and localizes in focal adhesion complexes and membrane ruffles, which are highly dynamic cytoskeletal structures necessary for cell motility (Poppe *et al*., 2007). Tammer and collegues performed a detailed functional characterization of Abl tyrosine kinase in signaling during *H. pylori* infections using the AGS gastric epithelial cell model. They showed that sustained activity of c-Abl is required to maintain CagA in a phosphorylated state, and that phosphorylated CagA forms a physical complex with c-Abl and activated **CrkII** *in vivo*. It was recently published that $CagA^{PY}$ interacts with Crk adaptor proteins and that this interaction is crucial in inducing *H. pylori*-dependent cell migration (Suzuki *et al*., 2005). These pathways possibly involves Crk>Sos1>H-Ras>Raf1 or Crk>Dock180>Rac1>WAVE signalling, leading to actin reorganization (Suzuki *et al*., 2005; Brandt *et al*., 2007).

In conclusion, c-Abl and SFKs seem to play together an important and differential role in CagA phosphorylation and subsequent actin-cytoskeletal rearrangements leading to cell scattering and elongation (Tammer *et al*., 2007).

Phosphorylation-independent biological activities of CagA

In addition to the functions of CagA as a phosphorylation-dependent scaffolding adaptor, recent studies revealed phosphorylation-independent partners of CagA, which were reported to lead to cell growth and cell motility by changes in actin dynamics (Churin *et al*., 2003).

Non-phosphorylated CagA binds to **growth factor receptor bound protein 2 (Grb2)**, an adaptor molecule that activates Ras via son of sevenless (Sos) (Mimuro *et al*., 2002). The CagA–Grb2 complex promotes proliferation of gastric epithelial cells through activation of the Ras-dependent **extracellular signal-regulated kinases 1 and 2 (Erk1/2)** (Ras>Raf>Mek>Erk), that in turn activate transcription factors like **serum responsive factor (SRF)** and the **nuclear factor kappa B (NF-κB)** (Hirata *et al*., 2002; Brandt *et al*., 2005; Kim *et al*., 2006). In particular, the evidence that IL-8 production is induced by a CagA fragment (1-812 aa) which does not contain EPIYA domains suggests other signalling patways, indipendent of CagA tyrosine phosphorylation, that may converge on MAP kinase and NF-κB (Kim *et al*., 2006). Moreover, a recent study reported also an injection of the *H. pylori*-derived proteoglycan via its TFSS to the **intracellular receptor Nod1**, which then leads to NF-κB activation and subsequent induction of **interleukin-8 (IL-8)** expression (Viala *et al*., 2004). Other previous studies have shown that IL-8 release could be achieved by activation of others tyrosine kinase receptors, such as the **epidermal growth factor receptor (EGFR)**, and **Her2-Neu (ErbB-2)** present in gastric epithelial cells, by a yet unknown TFSS effector and activation of Ras (Peek *et al*., 2000; Keates *et al*., 2001; Wallasch *et al*., 2001; Churin *et al*., 2003).

All these results suggest that *H. pylori* activates NF-κB through multiple distinct pathways. One of the less explored cell phenotypes induced by *H. pylori* is cellular invasion. Little is known about the mechanisms involved in this process. *H. pylori* was shown to activate tyrosine kinase receptors frequently involved in invasion-related pathways, such as EGFR, ErbB-2 and c-Met (Peek *et al*., 2000; Keates *et al*., 2001; Wallasch *et al*., 2001; Churin *et al*., 2003). In particular, nonphosphorylated CagA interacts intracellularly with the **scatter factor receptor c-Met**, enhancing the cell scattering (motogenic response) and inducing cell invasion (Churin *et al*., 2003; Oliveira *et al*., 2006). c-Met is a receptor tyrosine kinase with an important and well

documented participation in cell invasion (Trusolino *et al*., 2001) and upon binding with CagA, c-Met seems to undergo conformational changes leading to phosphorylation of specific tyrosine residues at the receptor intracellular domains, which act as docking sites for downstream signaling molecules. This results in the phosphorylation and binding of adaptor proteins and activation of signal transducers such as PI3K, eventually leading to cell invasion (Trusolino *et al*., 2001). Interaction of CagA with the **phospholipase C-gamma (PLC-γ)** was also reported, although the functional consequence of the interaction remains to be elucidated (Churin *et al*., 2003).

Another group of molecules associated with cancer cell invasion and influenced by *H. pylori* are **matrix metalloproteinases (MMPs)**. It was shown that *H. pylori* up-regulates the expression and activity of several MMPs, both in gastric epithelial cell lines and in the gastric mucosa (Bebb *et al*., 2003; Crawford *et al*., 2003). Increased MMP levels may facilitate the process of invasion by degradation of matrix components and cleavage of cell surface receptors that act as signal transducers in invasion pathways.

Previous reports evidenced an increase in MMP-2, MMP-3, MMP-7 and MMP-9 activity in the supernatants of cells cultured with *H. pylori*. (Gooz *et al*., 2001; Bebb *et al*., 2003; Crawford *et al*., 2003; Danese *et al*., 2004; Kundu *et al*., 2006; Oliveira *et al*., 2006). Their overexpression was demonstrated to be *H. pylori cag*-PAI dependent (Oliveira *et al*., 2006), whereas a recent study has shown that the **MMP-1** is instead directly stimulated by CagA (Pillinger *et al*., 2007). Up-regulated MMPs expression was associated with gastric ulceration and neoplasia in humans and in animal models, suggesting that MMPs participate in gastric tissue erosion and/or tumor invasion (Egeblad *et al*., 2002).

Recently, Murata-Kamiya and collegues found that CagA physically interacts with **E-cadherin** independently of CagA tyrosine phosphorylation. The membrane-associated CagA associated with E-cadherin impairs the complex formation between E-cadherin and β-catenin, causing cytoplasmic and nuclear accumulation of β-catenin by inhibiting its membrane association. **β-Catenin** is localized at cell–cell junctions by interacting with the cytoplasmic tail of the transmembrane protein E-cadherin (Franco *et al*., 2005). The E-cadherin/β-catenin complex plays a crucial role in epithelial cell–cell interaction and in the maintenance of the normal architecture of epithelial tissues. CagA-deregulated β-catenin then transactivates β-catenin-dependent genes encoding transcription factors that play a crucial role in the regulation of cell proliferation and differentiation. These results indicate that perturbation of the E-cadherin/β-catenin

complex by *H. pylori* CagA plays an important role in the development of gastric carcinoma (Franco *et al*., 2005; Murata-Kamiya *et al*., 2007). The adaptive advantage of CagA translocation for *H. pylori* is still unclear. One hypothesis is that *H. pylori* increases the permeability of the gastric epithelium, which then releases the food in the gastric lumen (Montecucco *et al*., 2001). Interestingly, CagA was found to interfere with the epithelial barrier function by disrupting tight junctions in Madin-Darby Canine Kidney (MDCK) polarized epithelial cells, a non-human and non-gastric cell model, and causing loss of apical basolateral polarity (Amieva *et al.*, 2003; Bagnoli *et al*., 2005). The tight junction controls paracellular permeability across the epithelial cell monolayer. It also has a fundamental role in the establishment and maintenance of epithelial cell polarity by delimiting the apical and the basolateral membrane domains (Shin *et al*., 2006). When delivered into the host cell, CagA localizes to the cell membrane under the point of bacterial attachment (Segal *et al*., 1999; Tanaka *et al*., 2003) and colocalizes with the **scaffolding proteins zonula occludens-1 (ZO-1)** and **tight junctional adhesion protein (JAM)**, normally found at the cell junctions, but it is not clear if this involves direct binding (Amieva *et al*., 2003; Bagnoli *et al*., 2005). The ability to mislocalize the tight junction marker ZO-1 is also consistent with the results of a study showing that the tight junction structure was irregular and fragmented in *H. pylori*-associated gastritis mucosa (Noach *et al*., 1994). An important aspect is that the redistribution of junction proteins and the alteration of the function of the apical-junctional complex occurs independently of CagA phosphorylation and promotes leakiness of the tight junctions (Amieva *et al*., 2003), while other cellular changes, previously described, including cell elongation, are mediated by the phosphorylated form of CagA.

More recently it was also demonstrated that CagA interacts with **PAR1** through the CagA multimerization sequence, which is required for CagA multimerization and subsequent stable association of CagA with SHP2 (Saadat *et al.*, 2007).

To investigate the mechanism by which CagA causes junctional and polarity defects, Saadat and collegues explored host cell proteins that bind to the native or mutants CagA. They have shown that CagA specifically interacts with PAR1, which has an essential role in the establishment and maintenance of an organized epithelial architecture. PAR1 was initially identified as a MARK kinase that phosphorylates microtubule-associated proteins.

Association of CagA inhibits PAR1 kinase activity and prevents atypical protein kinase C (aPKC)-mediated PAR1 phosphorylation, which dissociates PAR1 from the membrane, collectively causing junctional and polarity defects (Saadat *et al*., 2007).

Owing to the multimeric nature of PAR1, PAR1 also promotes CagA multimerization, which stabilizes the CagA–SHP-2 interaction (Ren *et al*., 2006). Furthermore, induction of the hummingbird phenotype by CagA-activated SHP2 requires simultaneous inhibition of PAR1 kinase activity by CagA. Thus, the CagA–PAR1 interaction not only elicits the junctional and polarity defects, but also promotes the morphogenetic activity of CagA. PAR1 is a key target of CagA in the disorganization of gastric epithelial architecture, underlying mucosal damage, inflammation and carcinogenesis.

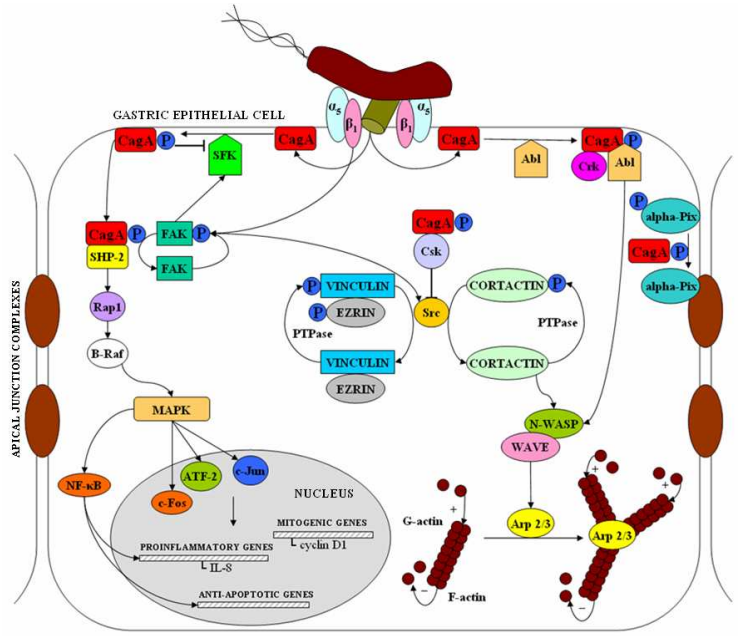
Recently, CagA was found to activate also the **nuclear factor of activated T cells (NFAT)** which is translocated from the cytoplasm to the nucleus where it transactivates NFAT-dependent genes (Yokoyama *et al*., 2005). The CagA activity toward NFAT is again independent of CagA phosphorylation. Surprisingly, one of the NFAT-dependent genes activated by CagA in gastric epithelial cells is **p21/Cip1 cyclin-dependent kinase inhibitor** that causes cell-cycle arrest (Yokoyama *et al*., 2005). Thus, whereas CagA activates a growth-promoting signal via SHP-2 deregulation, it simultaneously inhibits progression of the cell cycle through NFAT activation. As a result, CagA may cause proliferation, apoptosis or differentiation, depending on the cellular pathways.

CagA also affects **ROS** production in gastric epithelial cells. Reactive oxygen species (ROS) has been shown to play an important role in carcinogenesis by inducing DNA damage. It was found that ROS production in gastric epithelial cells was significantly enhanced by the infection of *cagA*-positive *H. pylori* species with an extensive accumulation of neutrophils. By transfecting *cagA* gene in gastric epithelial cells, Handa and collegues found that fraction of expressed CagA protein localizes to mitochondria and produces significant amount of ROS in the cells. In addition, increased ROS production might be involved in acceleration of cell cycle and subsequent cell proliferation. However, the mechanism by which CagA induces ROS production in gastric epithelial cells is still not known (Davies *et al*., 1994; Handa *et al*., 2006 and 2007).

The complexity and the number of signalling cascades induced by CagA indicates that this protein is a multifunctional effector, acting on both phosphorylation-dependent and –indipendent pathways (Backert *et al*., 2006). This protein probably consists only of protein-interaction domains and lacks any catalytic activity. By simultaneousely binding to different proteins it might form essential connections in signal transduction pathways. The knowledge of its structure could help us to understand the signal transduction in *H. pylori*-infected gastric epithelial cells and the molecular mechanism

of *H. pylori*-induced carcinogenesis. This knowledge in turn could be important for designing strategies against eradication therapy-resistant *H. pylori* and for developing new strategies against gastric cancer.

A) Phosphorylation-dependent biological activities of CagA



B) Phosphorylation-independent biological activities of CagA

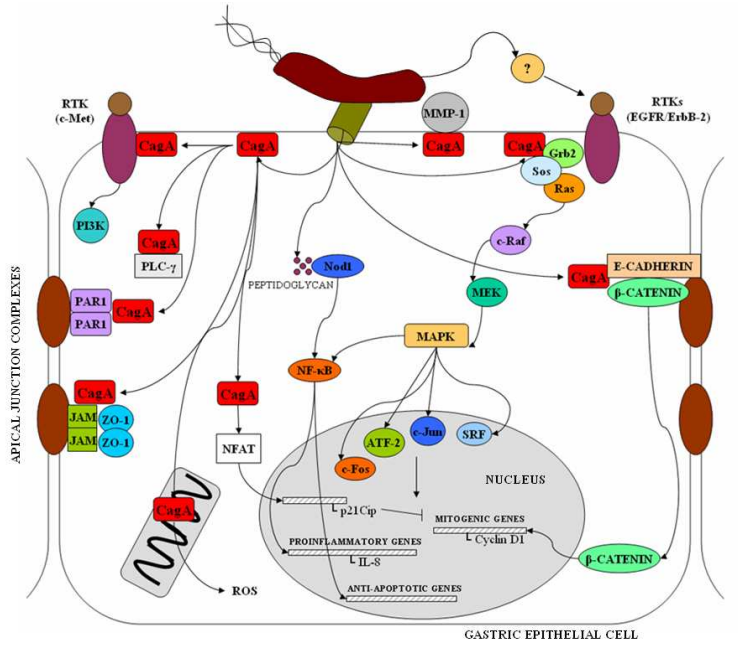


Figure 6.5. Model for the role of CagA effector molecules in host cell interaction and signal transduction.
The *H. pylori* CagA protein (shown in red) is inserted into the membrane of infected gastric epithelial cells by a TFSS and, interacting with many host signaling molecules, modulates various signaling cascades in both tyrosine phosphorylarion-dependent (A) and –independent (B) manner and thus is associated with cell differentiation, proliferation, motility, actin cytoskeletal rearrangements, disruption of tight junctions and proinflammatory responses, which might promote early stage of gastric carcinigenesis. Receptor tyrosine kinases and small Rho GTPases Rac1 and Cdc42 can also be activated, possibly by an as yet unknown T4SS effector. Additionally, peptidoglycan appears to activate the intracellular receptor Nod1, followed by NF-kB stimulation in a T4SS-dependent manner.

6.2 Soluble Domains of CagA identified by progressive deletion of its N-terminus and high-throughput screening.

Structural and biophysical studies of full-length CagA (HP0547) were till now limited by the inability to produce large amounts of recombinant protein. There are many reasons that make some proteins hard to express in bacterial expression systems. For instance, the protein can be toxic for the bacterium or it can be improperly folded, thus being more unstable and degraded.

Here we describe a high-throughput procedure to map the C-terminal regions of *H. pylori*'s CagA protein that is overexpressed in a soluble form in *Escherichia coli* using a two tag-fusion system. Dissection of a whole protein into domains is in fact an useful technique for functional analysis of large proteins comprising two or more folding domains. Therefore, we decided to examine bacterial expression of N-terminus deletion variants of CagA. We prepared a library of expression plasmids carrying *cagA* (*hp0547*) gene, unidirectionally deleted from 5' end by *Exonuclease* III.

The transformed bacteria were plated on the agar plate and, after induction of protein expression with IPTG from inducible promoter, the screening was performed using a fluorescent secondary antibody anti-mouse which recognises the anti-His antibody and a fluorescent streptavidin which recongnises the biotin-peptide complex. The screening revealed not only transformants with the correct reading frame of CagA, but also indicated expression levels of individual variants. Colonies giving the strongest signals were then selected for large culture expression. The method revealed several soluble C-terminal domains. These results provide a tool for future structural and biophysical studies of CagA.

6.2.1 Material and Methods

CagA (HP0547) library vector construction

The *cagA* (*hp0547*) gene from *Helicobacter pylori* strain 26695 genomic DNA was amplified by PCR with flanking primers *hp0547*for1 (5'-CACCATGACTAACGAAACTATTGATC-3') and *hp0547*rev1 (5'-TTAAGATTTTTGGAAACCACCTTTTG-3') and inserted into the expression pET151/D-TOPO directional vector (Invitrogen) by following the manufacturer procedure, obtaining the pET151-*cagA* construct. The *cagA* gene was then mutated to remove silently internal *Nsi*I site (ATGCAT to ATGCGT) using QuikChange® Site-directed Mutagenesis Kit (Stratagene). A *cagA** mutant was obtained by PCR reaction using a high fidelity thermostable DNA polymerase (*PfuTurbo* DNA polymerase; Stratagene).

The plasmid pET151-*cagA** was used as template and the following mutagenic primers complementary to opposite strands were synthesized in order to introduce the mutation (*hp0547Nsi*IXfor: 5'-GCACTTTTTAGCCAACGCATTTTCTACAGG-3'; *hp0547Nsi*IXrev: 5'-TCCTGTAGAAAATGCGTTGGCTAAAAAGTG-3'). The PCR product was treated with *Dpn*I (Stratagene) to digest the parental DNA template. This procedure allowed us to select the newly synthesized and potentially mutated plasmids. Single clones were then sequenced to confirm the occurrence of the desired mutation.

The mutated gene *cagA** was successively amplified by PCR with flanking primers *hp0547*for2 (5'-GATCCTAGGGCGCGCCACTAACGAAACTATTGATCAAACAA GAACACCAG-3') and *hp0547*rev2 (5'-CTAGGATCATGCATTAGATTTTTGGAA ACCACCTTTTGTATTAACATT-3') introducing *Asc*I and *Nsi*I restriction sites at the 5' and 3' ends, respectively. The amplified mutated *cagA** gene was then digested with *Asc*I and *Nsi*I enzymes and ligated into similarly digested pET modified vector to generate pHAR3011 encoding an in frame fusion of the CagA ORF with the DNA encoding a hexahistidine tag followed by a TEV protease cleavage site at the 5' end of the target gene and a DNA encoding the biotin acceptor peptide (BAP) at the 3' end of the target gene. If the protein is expressed and correctly folded also the biotin acceptor peptide upstream the protein will produced in frame folded and will be able to bind the biotin *in vivo* and following binding the streptavidin producing a signal.

N-terminal Truncation library construction

The truncation protocol was performed according to the ITCHY method (Ostermeier *et al.*, 2003; Tarendeau *et al*., 2007). To obtain large amount of plasmid DNA, 200 ml of LB medium, containing Kanamycin at 50 μg ml^{-1} were inoculated with a single colony of *E. cloni* 10G cells (Lucigen) harbouring the pHAR3011 plasmid and left to grow overnight at 37 °C under vigorous shaking. In order to obtain high quality and unnicked plasmid DNA, the pellet cells were firstly subjected to an alkaline lysis treatment and then to a Phenol:Chloroform:Isoamyl alcohol (25:24:1; Biosolve) extraction and Isopropanol precipitation. Successively, to remove salt and RNA, it was performed a polish using QIAprep kit (Qiagen). The pure plasmid DNA obtain was stored at 4°C and its concentration determined by biophotometer.

About 10 μg of plasmid pHAR3011 were digested with *Aat*II (New England Biolabs), which makes ends resistant to *Exonuclease* III, and *Asc*I (New England Biolabs), which gives ends sensitive to *Exonuclease* III. To obtain deletion variants differing by only a few nucleotides, four micrograms of purified, linearized vector were diluted in 1 X Buffer 1 (New England Biolabs), 30 mM NaCl (in addition to the buffer to slow down the reaction) to a final volume of 130 μl. To the 130 μl at 22 ºC, 400 units (100 U per μg DNA) of *Exonuclease* III (New England Biolabs) were added and mixed. At 2 min intervals, 2 μl ($1/60^{th}$) of the enzyme-DNA reaction were removed and added to a single "quenching tube" comprising 200 μl of 3 M NaCl on ice. This was continued for a total of 2 h until 120 μl of the reaction mix were transferred. When finished, the reaction was heated up to 70 °C for 20 min to denature the *Exonuclease* III and cleaned up using Nucleospin Extract II spin columns (Macherey-Nagel) and successively eluted in 40 μl of NE buffer.

In order to remove the single stranded overhang left after the exonuclease digest, the 40 μl of library mix in NE buffer was diluted in 1 X Mung Bean Nuclease buffer (New England Biolabs) and 5 units of *Mung Bean Nuclease* enzyme (New England Biolabs) added in a final volume of approximately 50 μl. The reaction was then incubated at 30 ºC for 30 min and subsequently cleaned up using Nucleospin Extract II spin columns (Macherey-Nagel) and eluted in 40 μl of NE. To polish the ends of the vector prior to ligation, 40 ul of the library DNA in NE were diluted in 1 X Pfu polymerase Native buffer (Stratagene) with 2.5 mM dNTPs and 1 unit of *Pfu* DNA polymerase (Stratagene) in a final volume of 50 μl. The reaction was incubated at 72 ºC for 20 min and then stored at 4 ºC.

The reaction mix was loaded onto a 0.5 % (w/v) TBE agarose gel. After electrophoresis in the cold room, the DNA in the size range of interest (larger than

linearized vector with no insert) was excised from the gel, fractionated in 3 sublibraries (insert size 0-1 kb, 1-2 kb and 2-3.5 kb), purified using the QIAexII gel extraction kit (Qiagen) and eluted in 60 µl of EB. The size-selected DNA corresponding to linearized vector containing truncated gene fragments was recircularized by ligation with T4 DNA ligase. This procedure was achieved by incubating 8 µl of DNA solution from the QIAexII purification with reagents from the Rapid Ligation Kit (Roche) according to the manufacturer's instructions. The ligation mix was desalted using a Montage cleanup spin column (Millipore) and 2 µl were used to transform One Shot OmniMAX 2 T1 Phage-Resistant *E. coli* ($5x10^9$ cfu ug^{-1}) competent cells (Invitrogen) by electroporation (1.8 kv) and grown at 37 °C for 1 h. After recovery of the transformation mix in SOC media, the library was plated out on 22 cm square Qtray LB-agar plates (Genetix) with Kanamycin (50 µg ml^{-1}). After overnight growth at 37 ºC, approximately 18,000 colonies were scraped from the agar, resuspended in 1 X PBS buffer and plasmid DNA prepared from a small aliquot of cells using a Midiprep kit (Qiagen). This was used to transform the protein expression strain of *E. coli* BL21 CodonPlus-RIL (Stratagene) by electroporation (1.8 kv). The size distribution of truncations was confirmed by a colony PCR screen with flanking primers (T7-promoter and T7-terminator) and agarose gel electrophoresis.

Robotical Colony picking

The library was plated out on new 22 cm Qtray square LB agar plates (Kanamycin 50 µg ml^{-1}; Chloramphenicol 50 µg ml^{-1}) at a density of approximately 2,000 colonies per plate grown at 30 ºC. About 18,000 colonies were robotically picked, using a biosystems picker-gridder robot, into 384 well plates filled with 70 µl TB-HMFM medium per well (supplemented with Kanamycin and Chloramphenicol). Liquid cultures were grown overnight with shaking (300 rpm) at 30 ºC in a HiGro shaker incubator (Genomic Solutions) [Fig 6.2.1].

Figure 6.2.1. HiGro shaker incubator

To prepare a colony array filter, squares of nitrocellulose membrane (GE Healthcare) were cut and laid on the top of 22 cm LB agar plates (supplemented with Kanamycin and Chloramphenicol). Using a gridding pin tool and the arraying robot [Fig. 6.2.2], the cultures were printed on to the membranes at high density. Plates were then incubated overnight at 25 ºC until colonies were visible.

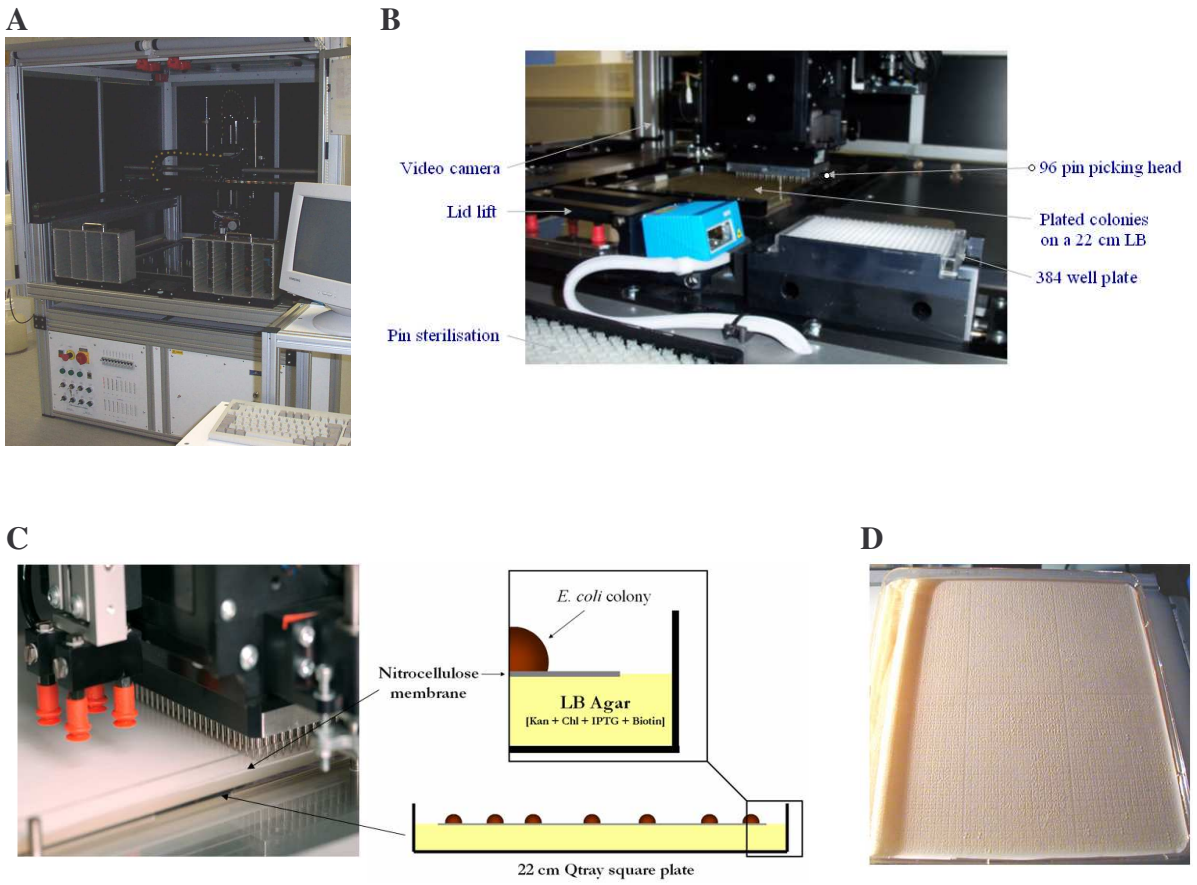


Figure 6.2.2. Robotic library picking and gridding. A. Picture of the Picker-gridder robot B. Zooming picture A, the video camera, the 96 pin picking head and the 384 well plate are visible; C, instantaneous picture of the robot printing clones on membranes; D, Colony arrays for protein expression 18,000 genetic constructs.

The membranes were then lifted from the agar and laid over fresh Qtray LB agar plates (supplemented with Kanamicyn, Chloramphenicol, 1 mM IPTG and 50 µM biotin) to induce recombinant protein expression within the colonies [Fig. 6.2.2 C and D]. Membranes were incubated for 4-5 h at 30 ºC, lifted from the inducing agar and placed at -80 ºC. Prior to analysis, the membranes were warmed to room temperature and laid over filter paper soaked into a Denaturation solution (0.5 M NaOH, 1.5 M NaCl) for 10 min at room temperature. The membranes were successively neutralized in a Neutralisation solution (1 M Tris-HCl, pH 7.5; 1.5 M NaCl) for 10 min and then for 15 min in 2 X SSC buffer. Then, with a glass spreader, the colonies were gently scraped from the membrane and the cell debris removed. Finally, the membranes were incubated with 1 X PBS buffer plus 0.1 % Tween-20 (PBS-T) and subsequently blocked overnight using a Superblock solution (Pierce) in a cold room on a cylinder rotator.

Detection of His_6 and Biotin tags

Excess blocking reagent was removed by washing in PBS-T for 10 min. In initial experiments, the membrane was firstly incubated with mouse monoclonal anti-His antibodies (Roche) diluited 1:3,000 in PBS-T buffer for 1 h at 4 °C and then washed several times in PBS-T buffer. The detection of biotinylated and His_6-tag- containing proteins were achieved using fluorescent Streptavidin labelled with Alexa 488 dye (Invitrogen) and a fluorescent secondary antibody anti-mouse labelled with Alexa 532 dye (Invitrogen), diluited 1:5,000 and 1:1,000, respectively. Both were diluted in a volume of 50 ml of PBS-T and added to the membrane for 1-2 h on a stir cylinder rotor in a cold dark room. The membrane was then washed three times with PBS-T buffer for 10 min each and subsequently with two changes of H_2O milliQ. The membrane was visualized with a Typhoon 9400 fluorescence imager (GE Helathcare) [Fig. 6.2.3] with blue laser (excitation wavelength 488 nm) and emission filter of 519 nm for fluorescent Streptavidin labelled with Alexa 488 dye and with yellow laser (excitation wavelength 532 nm) and emission filter of 553 nm for fluorescent secondary antibody anti-mouse labelled with Alexa 532 dye.

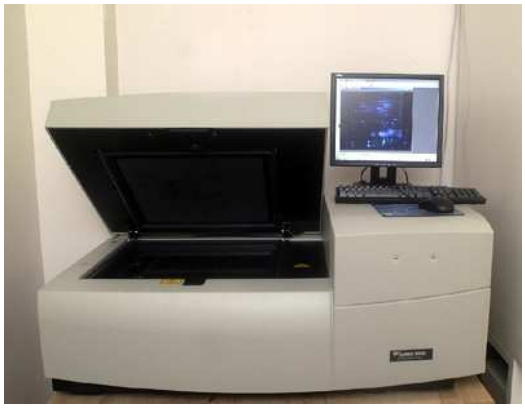


Figure 6.2.3. Typhoon Fluorescence Imager.

Signals from the array were analysed and quantified from digitized images using Visual Grid software (GPC Biotech) and clones ranked for expression level in MS Excel. Of the approximately 18.000 clones analyzed, 96 were selected and rearrayed robotically into a single plate for further analysis since the data indicated significant levels of protein expression above the background level of endogenous cellular biotinylation. The colonies giving the strongest signal were selected for further experimenting.

Colony PCR screen and analysis of selected clones

Insert size was screened rapidly using PCR screen of 96 clones with vector-specific flanking primers and agarose gel-electrophoresis. Colony PCR was performed in a 96-well format. Less than 1 µl sample containing recombinant *E. coli* from a 96 overnight culture plate well was used as template in 20 µl reaction mix. The universal T7

promoter primer and T7 terminator primer were used in this setup. PCR reactions were run and analyzed on 1 % TAE agarose gels using standard procedures.

Protein expression and purification in 24 plates using TECAN

For small scale expression tests, 40 µl a saturated overnight growth of 96 selected colonies were picked and transferred to a fresh 4 ml TB medium containing 50 µg ml^{-1} Kanamycin plus 50 µg ml^{-1} Chloramphenicol in 4x24 ml deep well plates. They were then incubated with shaking (300 rpm) in a HiGro shaker incubator (Genomic Solutions) at 37 °C to an OD_{600} of 0.6 before induction with 1 mM IPTG and reduction of the temperature to 25 °C. Cells were then harvested by centrifugation 16 h after induction (3700 rpm 10 min 4 °C) and the cell pellet obtained was subjected to lysis directly in the 24 well plates. The harvested cells were resuspended in 4 ml cold Spheroplasts buffer (20 mM Tris pH 8.0, 250 mM NaCl, 20 % Sucrose and 1 mg ml^{-1} lysozyme) and then frozen at -80°C to burst cells by one cycle of freeze and thaw. After 1 h, the plates were centrifuged again and the supernatant discarded. The pellet was following resuspended in 700 µl Lysis buffer (10 mM Tris pH 7.5, Brij 0.5%, Benzonase and a cocktail of Inhibitors without EDTA) per well using a thermomixer (500 rpm) at 4°C. Finally, the plates were incubated for 20' on a rocking platform in cold room. For purification on Ni-NTA agarose beads (Qiagen) using TECAN robot, the supernatant after lysis was loaded into a 96 well filter plates previously filled with 60 µl of Ni-NTA resin (Sigma) and equilibrated with Buffer A (50 mM Na_2HP0_4/NaH_2PO_4 pH 7.0, 300 mM NaCl, 5 mM Imidazolo). After extensive washing with Buffer A, the proteins were eluted with 80 µl of Buffer B (50 mM Na_2HP0_4/NaH_2PO_4 pH 7.0, 300 mM NaCl, 300 mM Imidazolo). For solubility and protein purification assays, samples from soluble, pellet

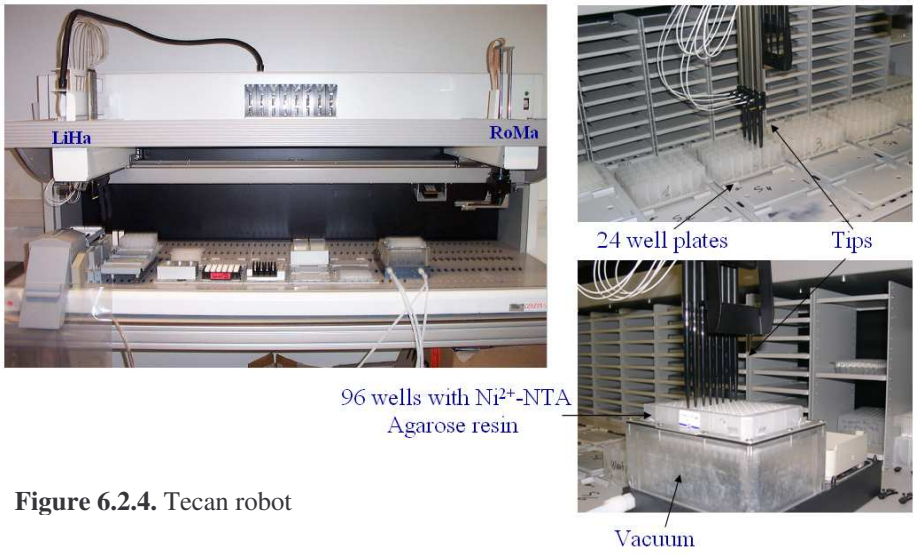


Figure 6.2.4. Tecan robot

and purified fractions from small-scale expression and purification were separated on 12-15 % SDS–PAGE gels and compared to the predicted insert size from the PCR screen. Plasmids from colonies that showed purifible proteins were sequenced to check the integrity of open reading frame, the presence of His_6-tag on the N-terminal and the Biotin acceptor peptide at the C-terminal of expressed protein fragments. The sequence analyses were carried out using programs available on EXPASY server and the deduced amino acid sequence was aligned with the HP0547 sequence available in GenBank.

6.2.2 Results and Discussion

The large size of the CagA protein and the weak homology sequence with other proteins of known structure and function make it difficult to predict which regions of the protein constitute well-folded domains. Previous studies, in fact, demonstrated that in infected gastric epithelial cells (AGS), human cell line THP-1 and several phagocytic cell lines (U937, Josk-M and J774A.1), CagA was partially found to be processed into two major fragments (Xiang *et al*., 1995; Backert *et al*., 2001; Moese *et al*., 2001; Odenbreit *et al*., 2001) [Fig. 6.2.5 A, C and D], whereas in *E.coli*, CagA is completely degraded in several fragments (Couturier *et al*., 2006; Pattis *et al*., 2007) [Fig. 6.2.5 B].

2-DE and immunoblotting studies of infected cell lines revealed the presence of two major CagA subfragments, a 100–105 kDa fragment ($p100^{CagA}$) and a 30–40 kDa fragment ($p35^{p\text{-}Tyr}$), in addition to full-length CagA ($p135^{p\text{-}Tyr}$) (Moese *et al*., 2001). They were further confirmed by MALDI-MS analysis. Both the pI of the CagA fragments (Backert *et al*., 2001) and the distribution of potential tyrosine phosphorylation motifs (particularly in the variable C-terminal part of CagA variants) provided further evidence that $p35^{p\text{-}Tyr}$ represents the C-terminal part of CagA and $p100^{CagA}$ the remaining N-terminal fragment. This suggested that CagA full-lenght is a very fragile protein and it undergoes proteolysis or autoproteolysis, breaking itself into subunits at defined positions. Although both $p100^{CagA}$ and $p35^{p\text{-}Tyr}$ were detectable in gastric epithelial cells (AGS) (Backert *et al*., 2001) and phagocytes (Odenbreit *et al*., 2001; Moese *et al*., 2001) [Fig. 6.2.5 A, C and D], they differed in the tyrosine phosphorylation pattern. In phagocytes, $p35^{p\text{-}Tyr}$ was predominantly phosphorylated, whereas in epithelial cells the major phosphorylated fragment was $p100^{CagA}$ (Backert *et al*., 2001). This suggested that differential phosphorylation of certain tyrosine residues in CagA occurs, perhaps due to diverse profiles of protein kinase expression pattern in both cell types.

In order to circumvent these problems, a high-throughput screening of all possible fragments of the CagA protein by progressive deletion of its N-terminus was performed using two enzymes, *Exonuclease* III and *Mung Beam nuclease*, and a modified pET vector carrying a biotin-fusion solubility reporter at the 3' end and a His_6-tag at the 5' end in order to identify and produce in *E. coli* stable, soluble and purifible C-terminal fragments with a yield high enough for structural studies. At the beginning, we decided to focalise our attention on the C-terminal domain, because it is

considered the most important domain of CagA, since it contains the phosphorylation and multimerisation motifs and it is directly involved into the interaction with eukaryotic proteins like kinases, phosphatases and receptors. Very little is instead known about the function of the CagA central and N-terminal domains.

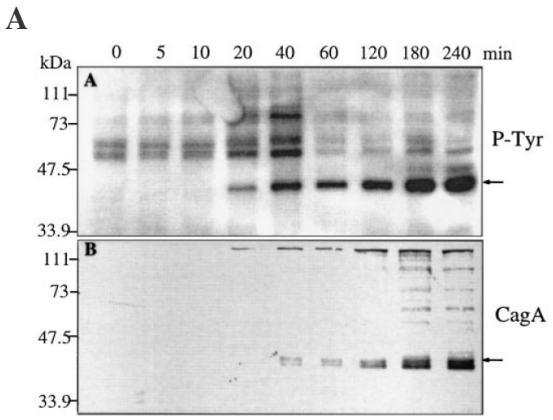


Figure 6.2.5 A. Kinetics of tyrosine phosphorylation and processing of CagA. Professional phagocytes J774A.1 cells were incubated with *H. pylori* P12 strain and the infection was terminated at different time points. Lysates were separated by 10 % PAGE, blotted to nitrocellulose and reacted with anti-phosphotyrosine-specific monoclonal antibody PY99 (A) or anti-CagA antiserum AK257 (B). Arrows indicate the position of p35-45 (Adapted from Odenbreit *et al*., 2001)

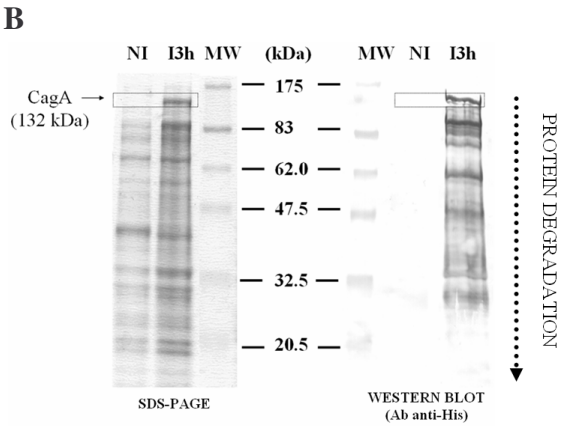


Figure 6.2.5 B. SDS-PAGE (left) and Western-blot (right) respectively of the full-lenght recombinant His_6-CagA expressed in *E. coli*. The protein is improperly folded, being more unstable and completely degraded when produced in *E. coli*.

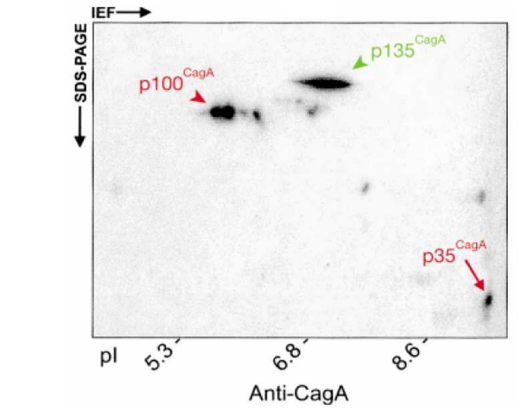


Figure 6.2.5 C. Identification of CagA subfragments in *H. pylori* and in infected phagocytic cell lines U937 by immunoblot analysis of 2-DE gels. The 2-DE blots were probed with a polyclonal anti-CagA antibody (Anti-CagA). Three major CagA protein species were identified: (i) native 130–140 kDa CagA protein ($p135^{p\text{-}Tyr}$), green; (ii) a 100–105 kDa fragment ($p100^{CagA}$) and (iii) a 30–40 kDa fragment ($p35^{p\text{-}Tyr}$), red (Adapted from Moese *et al*., 2001).

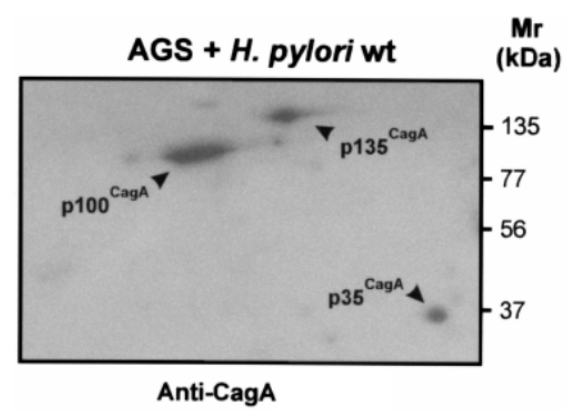


Figure 6.2.5 D. 2-DE immunoblotting of *H. pylori* with AGS cell contact. Three CagA protein species: (i) full-length 135 kDa CagA protein ($p135^{CagA}$), (ii) a 100 kDa fragment ($p100^{CagA}$), and (iii) a 35 kDa fragment ($p35^{CagA}$) were identified using anti-CagA antibody (Adapted from Backert *et al*., 2001).

Cloning of *cagA* full length gene into pET modified vector

To construct the library-plasmid, the amplified *cagA* gene was firstly cloned into the pET151/D-TOPO vector using the TOPO cloning kit (Invitrogen), which permits a quick directional cloning into the expression system obtaining pET151-*cagA*. Because the nucleotide sequence of *cagA* contained an internal *Nsi*I site, a mutant was produced in order to remove it. The mutated plasmid pET151-*cagA** was then selected and newly amplified using specific primers introducing *Asc*I and *Nsi*I sites at the 5' and 3' ends, respectively. The PCR product was then digested with the two restriction enzymes and inserted into a pET modified expression vector which encodes a N-terminal His_6-tag and a C-terminal biotin acceptor peptide. The vector obtained, termed pHAR3011, was used to produce the library [Fig. 6.2.6].

Construction of N-terminal deletion variants of CagA

In order to obtain CagA variants sustainable in bacteria, we utilized progressive *Exonuclease* III digestion from 5' end of *cagA* gene. The constructs pHAR3011 was digested with enzymes *Asc*I and *Aat*II yielding linearized plasmids. In particular, the *Asc*I enzyme leaves 5' overhang, located between the His_6-tag and the beginning of the *cagA* gene. The latter is an *Exonuclease* III sensitive site, whereas the *Aat*II enzyme leaves a 3' overhang that is an *Exonuclease* III reisistant site. The *Exonuclease* III enzyme then recognises the 5' overhang and starts to digests one strand of the dsDNA progressively from the sensitive in the 3' to 5' direction. The *Mung Beam nuclease* blunts this DNA with long single-stranded overhangs left by *Exonuclease* III, as well as the short overhangs left by the second restriction enzyme digestion. To obtain deletion variants differing by only a few nucleotides, the *Exonuclease* III digestion was performed at 30 mM NaCl concentration, to slow down the reaction, at room temperature; samples were taken at frequent intervals, every 2 min. All samples were then mixed together in one tube and loaded on TBE agarose gel electrophoresis to check the progress of the reaction (such as in Fig. 6.2.6). A continuum of deletion variants were obtained, ranging from 8.0 kb (original length) to approximately 5.0 kb fragments. These fragments were then excised from the gel, divided in three tubes and purified. The plasmids obtained by recircularisation of the purified products were then used to transform the bacteria. Obtained colonies were blotted on a nitrocellulose membrane and after IPTG induction several colonies giving strong red signal, from the fluorescent secondary antibody, and the green signal, from the strepatavidin (Fig. 6.2.7 A, B, C and D). In this way, we obtained a set o recombinants molecules wich codes for proteins differing only in few amino acids on the N-terminus of the expressed

protein. Protein aggregation and inclusion body formation are frequently encountered when recombinant proteins are overexpressed in *E. coli*. By expressing biotin acceptor as a C-terminal fusion to the protein of interest, the solubility of the fusion protein *in vivo*, indicative of productive protein folding, can be monitored by fluorescence streptavidin. If the protein of interest is prone to aggregation when expressed alone, it would likely also do so in the fusion product. It appears that the aggregation of the upstream protein domain causes misfolding of the downstream domain. Without a correctly folded structure, the biotin acceptor peptide cannot bind the biotin and be recognised by Streptavidin (Fig. 3.12 B).

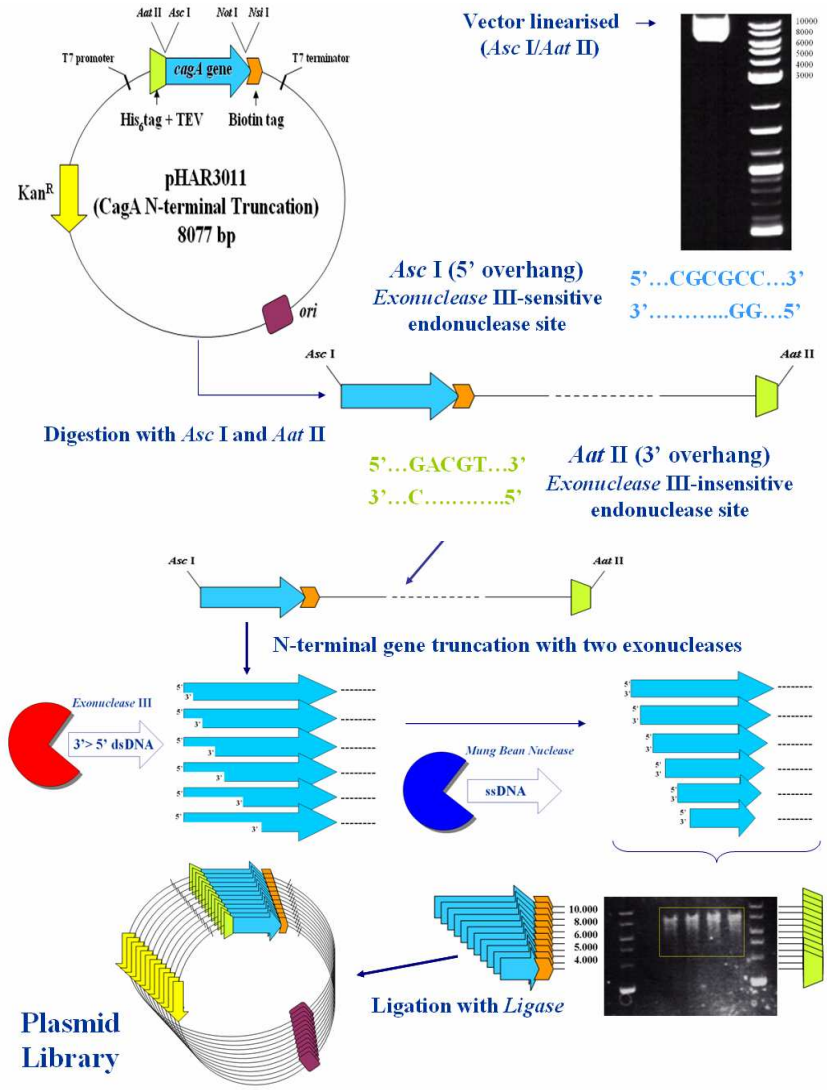


Figure 6.2.6. Progressive deletion of pHAR3011 vector carrying *cagA* gene previously linearized with *Asc*I and *Aat*II restriction enzymes and subjected to *Exonuclease* III/ *Mung Beam Nuclease* digestion at room temperature, with 30 mM NaCl.

As described earlier, the screening was performed by robotic picking and gridding, followed by detection of protein expression by hybridisation of membranes. The colony arrays were probed with fluorescent streptavidin (labelled with Alexa 488 dye), which binds the biotin-peptide complex, and with a fluorescent secondary antibody anti-mouse (labelled with Alexa 532 dye) which recognises the anti-His antibody. Detection was then performed using Typhoon 9400 fluorescent imager (GE Helathcare). The colonies yielding a signal (green) from the streptavidin hybridisation

indicate that the protein is soluble, whereas the presence of a N-terminal hexahistidine tag signal (red) indicates that the protein is full-length. In this way, false positives resulting from internal initiation of translation within the gene, or premature termination of translation, may be simply screened out, since the presence of both end signals of the protein indicates that it is a full-length (Fig. 6.2.7 A). Additionaly, the presence of a hexahistidine tag allows rapid verification of solubility via small-scale metal affinity chromatography column and, secondly, it allows direct scale-up and purification of the truncated proteins.

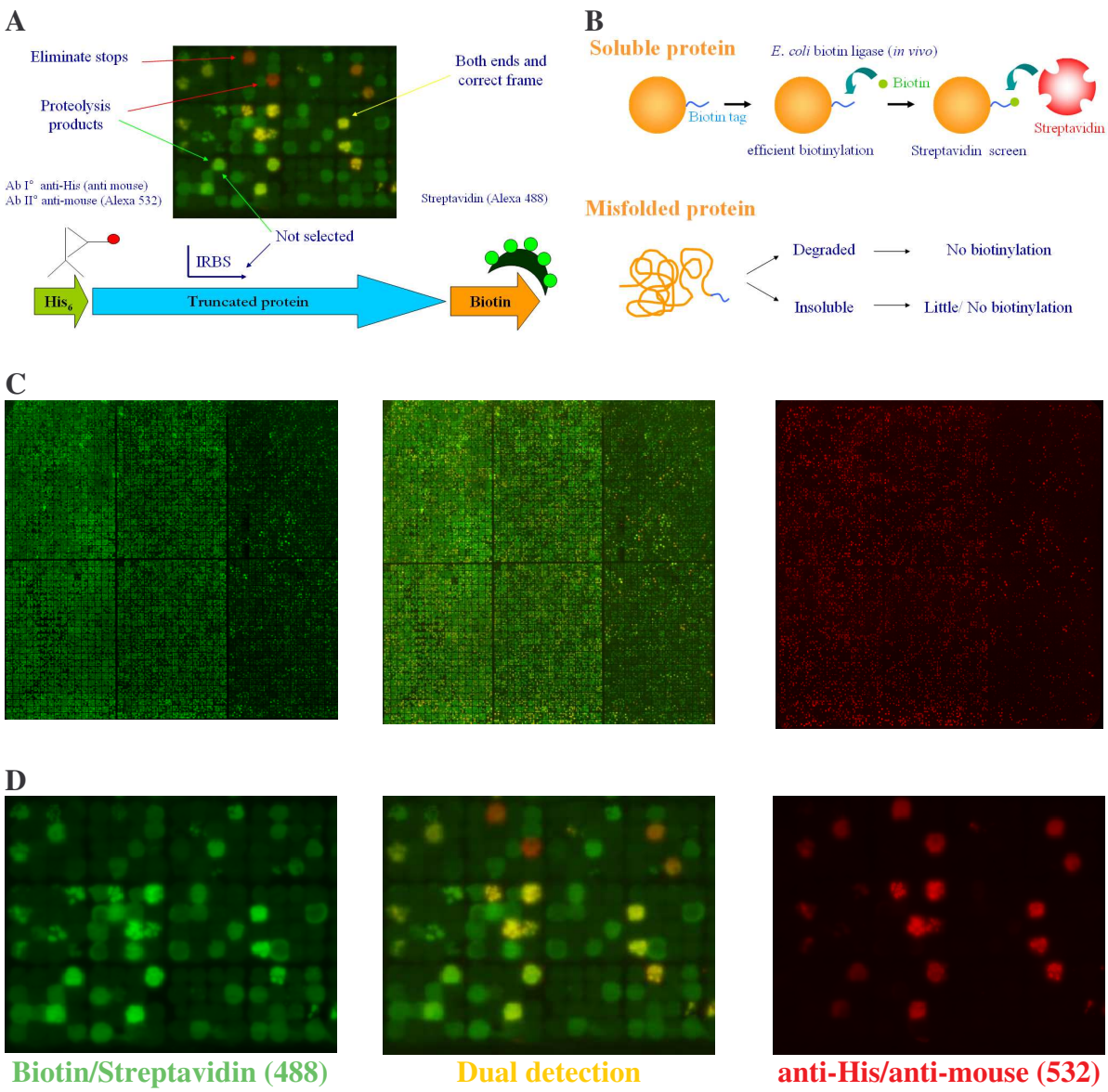


Figure 6.2.7. Dual detection of clones with high level expression of CagA deletion variants after colony-blotting using anti-His/anti-mouse hybridisation at the N-terminal (red signal) and Biotin/Strepatavidin hybridisation at the C-terminal (green signal).

Figure 6.2.7 shows an example of colony array expressing the His_6-CagA-Biotin fusion library probed with fluorescent streptavidin and anti-mouse antibody after irradiation with UV light (Fig. 6.2.7 C and D).

High-throughput screening of deletion products of the CagA protein

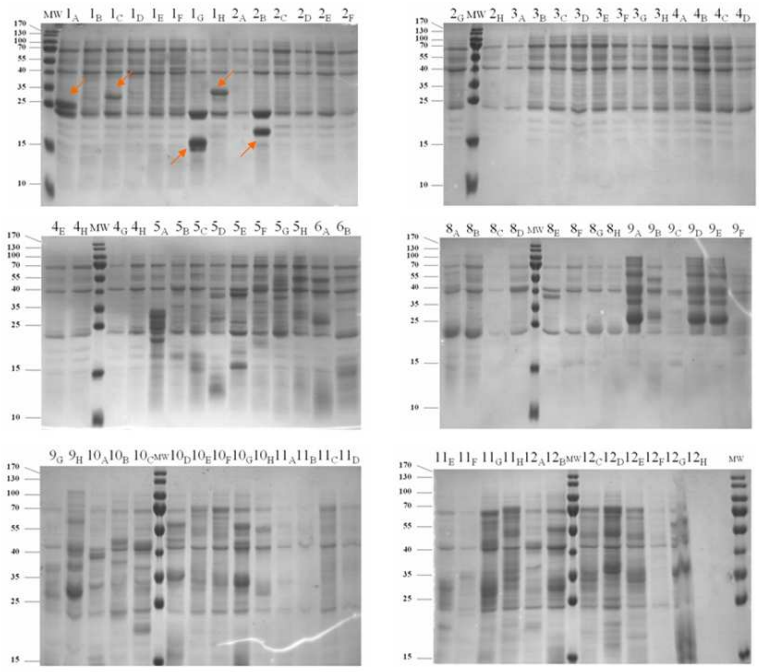


Figure 6.2.8 A. Size distribution of 96 selected C-terminal CagA protein domains from small scale expression and purification experiments.

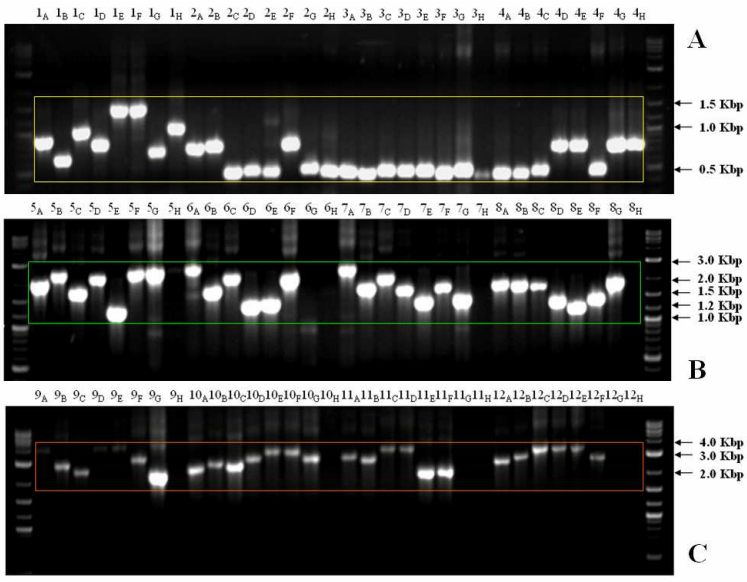


Figure 6.2.8 B. Size distribution of 96 DNA fragment from PCR screen.

Approximately 18.000 colonies showed a strong fluorescence signal. From this library, 96 colonies displaying a strong fluorescent yellow signal were selected for expression tests in 24-well tissue culture plates in 4 mL cultures. Addition of IPTG to the suspension cultures resulted in the overexpression of fusion proteins for most fluorescent colonies, as visualized by Coomassie stained SDS-PAGE (Fig. 6.2.8 A). To confirm that the 96 positive signals on the dot blot indicating soluble fragments correspond to the CagA clones visualised into the SDS-PAGE, each clone were subjected

to colony PCR. Using the universal T7 promoter primer and T7 terminator primer (they both bind to the vector region of pHAR3011), they were amplified from *E. coli* directly from each well of the microtiter plate as described under Materials and Methods. After separating the PCR-amplified products on a 2 % TAE agarose gel, the size of the insert clones has been compared to the size of protein domain identified by SDS-PAGE (Fig. 6.2.8 B). We therefore focused on these C-terminal fragments of CagA in order to validate the hypothesis that the high-throughput assay was indeed identifying regions of the protein that could be expressed with a N-terminal histidine tag in a soluble form and purified by Ni^{+2} affinity chromatography. Plasmid DNA was isolated and sequenced from about seventy out of 96 colonies overexpressing fusion proteins larger than 20 kDa in size. Small C-terminal polypeptide fragments in fact are usually highly soluble and yield bright colonies, but they are usually meaningless.

Many of selected colonies did not result in a clear overexpression of a fusion protein, perhaps due to post-translational degradation of the fusion protein or to the exposure of an internal cryptic ribosome-binding site within the CagA sequence. It is conceivable that the fragments screen might produce protein constructs that are soluble, but that are made up of multiple structured domains or that are not folded.

These results are partially in agreement with previus studies: in fact, our purified protein samples show the same degradation pattern that was previously reported in several *in vivo* and *in vitro* experiments (Xiang *et al*., 1995; Backert *et al*., 2001; Moese *et al*., 2001; Odenbreit *et al*., 2001; Couturier *et al*., 2006; Pattis *et al*., 2007) [Fig. 6.2.5 A, B, C and D]

However, we were able to identify some potential stable fragments of about 30 – 20 kDa, indicated by orange arrows in figure 6.2.8 A, that are now used for future large scale expression and purification experiments. We believe that the production and purification of these domains separately would provide a key tool for studying its structure and function.

A similar approach could be used in the future to delete gradually sequences from both N-terminus and C-terminus of the recombinant protein in order to obtain more stable internal soluble protein domains.

6.3 Expression, purification, characterisation of CagA central domain (392-733) and preliminary crystallization and interaction studies with CagF and *Hp*YbgC proteins.

The Cag TFSS contains 18 gene products which are considered essential for CagA translocation. However, only 14 codified proteins are thought to encode components of the secretion apparatus, since besides their involvement in CagA translocation they are also necessary for induction of IL-8 secretion from epithelial cells (Fischer *et al*., 2001). The four additional proteins that are not considered component of the TFSS but were identified to be necessary for the CagA translocation include: two proteins of unknown function, CagI (HP0540) and CagZ (HP0526), the coupling protein homologue Cagβ (HP0524) and CagF (HP0543). The latter was recently shown to interact with CagA in the bacterial cell (Couturier *et al*., 2006; Pattis *et al*., 2007). The specific region of CagF-CagA interacting domain was shown to be localized close to the C-terminal translocation signal of CagA (Pattis *et al*., 2007), whereas Couturier and colleaugues suggested that the interaction occurs at least within the amino and central 100 kDa region of CagA (Couturier *et al*., 2006).
Another protein-protein interaction was reported between CagA and *Hp*YbgC (HP0496), a thioesterase protein. The interaction is domain specific: in fact, the minimum fragment of CagA which was identified to interact with *Hp*YbgC full-length protein is the central domain between residues from 392 to 733 (Rain *et al*., 2001; Terradot *et al*., 2004). *Hp*YbgC does not belong to the *cag*-PAI and shows homology (30 % identity) to the thioesterase protein, especially to the YbgC-like protein, a cytoplasmic component of the Tol-Pal system. Very recently, we solved the crystal structure of *Hp*YbgC protein and characterized its enzymatic activity (Angelini *et al*., Chapter 7). To validate and elucidate the molecular basis of these interactions we performed some biochemically and structurally studies.

6.3.1 Material and Methods

CagA$^{392-733}$ Protein Expression and Purification

Cloning of the CagA (HP0547) Central Domain (CagA$^{392-733}$) encoding gene into the pPROEX HTb expression vector (Invitrogen) was previously described in Ref.

(Terradot *et al.*, 2004). *E. coli* BL21 (DE3) Star cells (Invitrogen) were transformed with the recombinant plasmid and grown at 37 °C in LB medium containing 100 μg ml^{-1} of Ampicillin. The expression was induced at OD_{600} = 0.7-0.8 with 1 mM IPTG for 5 at 25 °C under shaking.

The cells harvested by centrifugation at 11.000 g for 20 min were suspended with Lysis Buffer (30 mM Tris pH 8.0, 300 mM NaCl, 5 % Glycerol, 0.2 % Triton-X100) and stored at -80 °C. The frozen suspended bacteria were supplemented with *Dnase*I, Lysozyme and a mini "complete" EDTA-free protease inhibitor tablet (Roche) and disrupted using a French press. The soluble lysate obtained by centrifugation for 25 min at 40.000 g was applied onto a 5 ml His-Trap column (GE Healthcare) equilibrated with Buffer A (30 mM Tris pH 8.0, 300 mM NaCl, 5 % Glycerol) at 2 ml min^{-1}. After a first washing in the same buffer and a second washing with 10 % Buffer B (30 mM Tris-HCl, pH 8.0, 300 mM NaCl, 5 % Glycerol, 500 mM Imidazole), the protein was eluted in a single peak at 250 mM Imidazole by applying a linear gradient (10 %-100 %) of Buffer B using an AKTA primer system.

The fractions containing the pure His_6-$CagA^{392-733}$ were pooled, concentrated by ultrafiltration (Amicon Ultra 10,000 MWCO, Millipore), diluted 20 times into Buffer A supplemented with 0.5 mM EDTA and 1 mM DTT and incubated over-night at 4 °C with rTEV protease to cleave off the His-tag. Further purification was obtained loading the protein sample after cleavage again onto a 5 ml His-Trap column (GE Healthcare) equilibrated with Buffer A. The flow through containing the pure and homogeneous $CagA^{392-733}$ without His_6-tag was collected and concentrated by ultrafiltration (Amicon Ultra 10,000 MWCO, Millipore). All these purification procedures were performed at 4°C to avoid/limited any protein degradation process. The concentrated sample was injected on gel filtration column (Superdex 75, Prep grade HR 16/60) equilibrated with Buffer C (30 mM Tris pH 8.0, 300 mM NaCl) in order to dialyse the protein in a new buffer and to obtain a pure and homogeneous sample (as confirmed by SDS-PGE). Pure protein fractions were concentrated by ultrafiltration (Amicon Ultra 10,000 MWCO, Millipore) to 20 mg ml^{-1} and stored at -80 °C.

CagF and *Hp*YbgC protein Expression and Purification

The *cagF (hp0543)* gene, from the *H. pylori* G27 strain, was amplified and cloned in a pGEX-4T-3 vector (GE Healthcare) as described in Ref. (Seydel *et al.*, 2002). *E. coli* BL21 (DE3) cells (Invitrogen) harvesting the plasmid pGEX-*cag*F were grown at 37 °C in 2 l LB medium containing 100 μg/ml of ampicillin. The expression was induced

at 0.6-0.7 OD_{600} by adding 1 mM isopropyl-β-D-thiogalactopyranoside (IPTG) and prolonged for 3 h at 30 °C, under vigorous shaking. The cells were harvested by centrifugation at 11.000 g for 20 min, resuspended in Buffer D (30 mM Tris pH 8.0, 500 mM NaCl, 1 mM LDAO) supplemented with protease inhibitor cocktail (Roche, Complete-mini EDTA-free) and disrupted by sonication. The soluble fraction, obtained after 30 min centrifugation at 40.000 g, was applied onto a 3 ml Glutathione Sepharose affinity resin (Glutathione Sepharose 4B Fast flow, GE Healthcare) previously equilibrated with Buffer D at about 1-2 ml min^{-1}. After many extensive washes with Buffer D, the recombinant protein GST-CagF was eluted by applying Buffer E (30 mM Tris pH 8.0, 500 mM NaCl, 1 mM LDAO and 10 mM Glutathione). The fractions containing the pure GST-CagF were pooled, concentrated by ultrafiltration (Amicon Ultra 5,000 MWCO, Millipore), diluted into Buffer D and incubated over-night at 4 °C with Thrombin protease to remove the N-terminal GST-tag. Further purification, to separate the GST, was obtained loading the protein sample after cleavage onto 3 ml Glutathione Sepharose affinity resin (Glutathione Sepharose 4B Fast flow, GE Healthcare) previously equilibrated with Buffer D. The flow through containing the pure and homogeneous CagF without GST-tag was collected and concentrated by ultrafiltration. The cleavage product was further purified by gel filtration (Superdex 200 10/30, GE Healthcare), equilibrated with buffer D.
The *Hp*YbgC protein was expressed and purified as described in the Chapter 7 (Angelini *et al.*, Accepted).

Protein Characterization

The purified samples were further characterised by UV/VIS spectroscopy (240-360 nm, b = 1 cm, DU640 UV/VIS spectrophotometer, Beckman Coulter Inc.) using the theoretical absorption coefficient of 0.27 ml mg^{-1} cm^{-1}, 0.70 ml mg^{-1} cm^{-1} and 0.67 ml mg^{-1} cm^{-1} for $CagA^{392-733}$, CagF and *Hp*YbgC, respectively. CD measurement of $CagA^{392-733}$ (2 mg ml^{-1}) and CagF (3 mg ml^{-1}) were performed using a JASCO-J720 Spectropolarimeter, in a 0.02 cm path length cell. The spectra in the far-UV (190-260 nm) were recorded at a scanning speed of 20 nm min^{-1}. Ten spectra were accumulated in each case and averaged, followed by baseline correction by subtraction of the buffer soluttion absorption. Mean residue weight ellipticities were calculated and expressed in units of degree cm^2 $dmol^{-1}$. The circular dichroism spectrum has been deconvoluted by software program CDNN (Version 2.1) to estimate the content of secondary structure of both $CagA^{392-733}$ and CagF. The CD spectra of *Hp*YbgC is not reported since its 3D structure has been determined (Chapter 7).

Analytical Gel Filtration experiments

The oligomeric state of purified proteins (CagA$^{392-733}$, CagF and *Hp*YbgC) was then examined by analytical size exclusion chromatography experiments using a Superdex 200 10/300 column (GE Healthcare). Flow rates used were 0.5 ml min^{-1} and the experiments were performed at room temperature. The column was previously calibrated using the following weight markers: Ribonuclease A (13.7 kDa), Chymotrypsinogen A (25 kDa), Bovine serum albumin (67 kDa), Catalase (232 kDa), Ferritin (440 kDa) and Blue dextran 2000 (2 MDa).

In particular, to characterize the behaviour of CagA$^{392-733}$ in solution several analytical gel filtrations were performed using different salt concentration buffer solution. In particular, a first gel filtration was performed loading 100 ul volume [25 ul of CagA$^{392-733}$ (20 mg ml^{-1}) diuited adding 75 ul Buffer I] on the column previously equilibrated with Buffer I (30 mM Tris pH 8.0, 300 mM NaCl) and the eluition profile was monitored at 280 nm [Fig. 6.3.1B]. A second gel filtration was then performed loading 100 ul volume [25 ul of CagA$^{392-733}$ (20 mg ml^{-1}) diluited adding 75 ul Buffer II] on the column previously equilibrated with Buffer II (30 mM Tris pH 8.0, 100 mM NaCl) [Fig. 6.3.1A]. Finally, a third gel filtration was performed loading 100 ul volume [25 ul of CagA$^{392-733}$ (20 mg ml^{-1}) diluited adding 75 ul Buffer III) on the column previously equilibrated with Buffer II (30 mM Tris pH 8.0, 500 mM NaCl) [Fig. 6.3.1C].

The oligomeric state of purified *Hp*YbgC and CagF proteins was also examined by size exclusion chromatography using a Superdex 200 10/30 column previously equilibrated with buffer III (30 mM Tris pH 8.0, 500 mM NaCl) and D (30 mM Tris pH 8.0, 500 mM NaCl and 1 mM LDAO), respetively [Fig 6.3.1 E and F].

For the complex analysis the stock CagF, *Hp*YbgC and CagA$^{392-733}$ solutions (kept frozen at -80 °C) were incubated for at least 1 h before use and then injected on the same column previously equilibrated with buffers I, II and D. The eluition profiles were monitored at 280 nm [Fig. 6.3.4 A and B; Fig. 6.3.5 A and B].

In particular, for the CagA$^{392-733}$-*Hp*YbgC complex formation at 100 mM NaCl, 100 ul volume [25 ul of *Hp*YbgC (10 mg ml^{-1}) mixed with 25 ul CagA$^{392-733}$ (20 mg ml^{-1}) plus 50 ul Buffer I] were incubated for 1 h at room temperature and then loaded on the column previously equilibrated with Buffer I [Fig. 6.3.5 A]. An identical procedure has been adopted to analyse the CagA$^{392-733}$-*Hp*YbgC complex formation at 300 mM NaCl, using buffer II [Fig. 6.3.5 B].

To characterized the the CagA$^{392-733}$-CagF complex formation, two different ratio of the proteins were used. For the 1:1 CagA$^{392-733}$-CagF complex formation, 100 ul

volume [50 ul of *Hp*YbgC (10 mg ml^{-1}) mixed with 25 ul $CagA^{392-733}$ (20 mg ml^{-1}) plus 25 ul Buffer D] were incubated for 1 h at room temperature and then loaded on the column previously equilibrated with Buffer D [Fig. 6.3.4 A]. For the 2:1 $CagA^{392-733}$-CagF complex formation instead, 100 ul volume [25 ul of *Hp*YbgC (10 mg ml^{-1}) mixed with 25 ul $CagA^{392-733}$ (20 mg ml^{-1}) plus 50 ul Buffer D] were incubated for 1 h at room temperature and then loaded on the column previously equilibrated with Buffer D [Fig. 6.3.4 B].

Far Western Blotting Assay

Several amounts (0 ng, 50 ng, 100 ng, 250 ng, 500 ng) of the purified proteins CagA (20 mg ml^{-1}), CagF (10 mg ml^{-1}) and *Hp*YbgC (10 mg ml^{-1}) were separately spotted onto nitrocellulose membranes (GE Healthcare) and allowed to dry at room temperature. The membranes were then blocked in phosphate-buffered saline solution containing 0.2 % Tween 20 (PBS-T) and 5 % BSA-free for 1 h at 4 °C. The membranes were then washed with PBS-T buffer and then incubated with the potential interacting partner His_6-$CagA^{392-733}$ for CagF and *Hp*YbgC proteins, His_6-*Hp*YbgC for $CagA^{392-733}$ for 1 h at 4 °C. The membranes were then washed several times with PBS-T and incubated over-night at 4 °C with penta anti-His antibody, BSA-free, Mouse Monoclonal IgG, diluted 1:1.000 in PBS-T buffer. The day after, the blots were washed with PBS-T and incubated for 2 h at room temperature with rabbit anti-mouse antibody Alkaline Phosphatase (AP) conjugate diluted 1:10.000 in the same buffer.
The membranes were then washed and incubated for 30 min in AP buffer (100 mM Tris, pH 9.5, 100 mM NaCl and 5 mM $MgCl_2$) and further processed with NBT/BCIP system (Sigma). The same procedure was performed without adding the $CagA^{392-733}$, *Hp*YbgC and CagF interacting partners as negative control.

6.3.2 Results and Discussion

Single protein expression, purification and characterization

In this work, the three recombinant proteins His_6-$CagA^{392-733}$, His_6-*Hp*YbgC and GST-CagF expressed separately in *E. coli* were obtained in good yield and at high degree of purity using a two-step purification protocol. The proteins were purified to homogeneity and shown to be monodisperse and folded by analytical gel filtration and CD spectroscopy, respectively [Fig. 6.3.1 and Fig. 6.3.2].

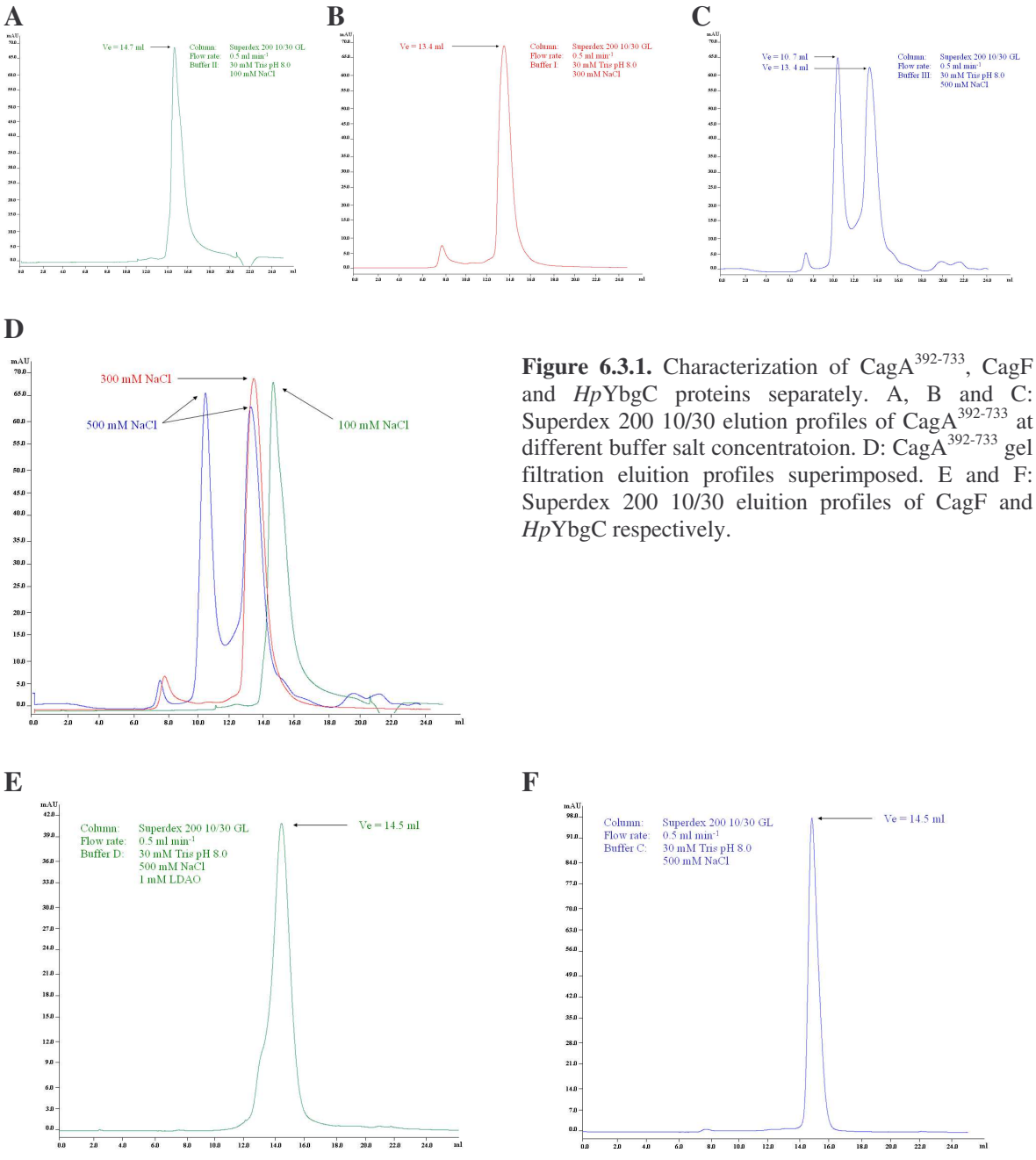
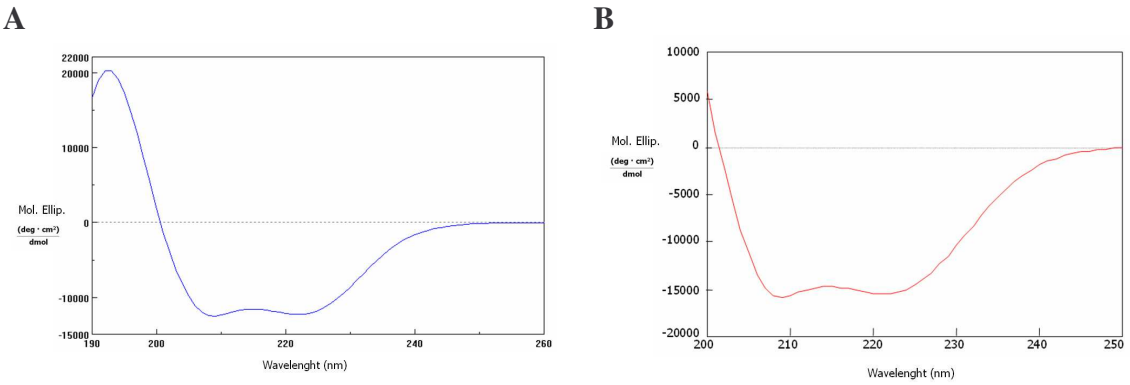


Figure 6.3.1. Characterization of $CagA^{392-733}$, CagF and *Hp*YbgC proteins separately. A, B and C: Superdex 200 10/30 elution profiles of $CagA^{392-733}$ at different buffer salt concentratoion. D: $CagA^{392-733}$ gel filtration eluition profiles superimposed. E and F: Superdex 200 10/30 eluition profiles of CagF and *Hp*YbgC respectively.

From the gel filtration, the eluition volume of CagA$^{392-733}$ domain indicates that the protein probably behaves in solution as a trimer or a tetramer. This could be due to the fact that probably CagA$^{392-733}$ is not strictly spherical, but could form in solution a long nonglobular structure that may result in a relatively slower passage through the gel material and consequently lead to understimation of the molecular weight.
Moreover, the following analytical gel filtration experiments on CagA$^{392-733}$ protein show the presence of an affect of the ionic strength on its oligomerization. At low salt concentration (100 mM NaCl), CagA$^{392-733}$ was eluted as a single peak with a retention volume (Ve = 14,7 ml) corresponding to a monomer or dimer [Fig. 6.3.1A]. At 300 mM NaCl concentration the protein behaves in solution (Ve = 13,4 ml) as a trimer or a tetramer [Fig. 6.3.1B], whereas, increasing the ionic strength (500 mM NaCl), it is prone to become a decamer or dodecamer (Ve = 10,7 ml) [Fig. 6.3.1C]. To determine its precise hydrodynamic radius and oligomerisation state, light scattering measurement have been performed and similar results were obtained (data not shown), suggesting that the protein oligomerization strictly depends upon salt concentration.



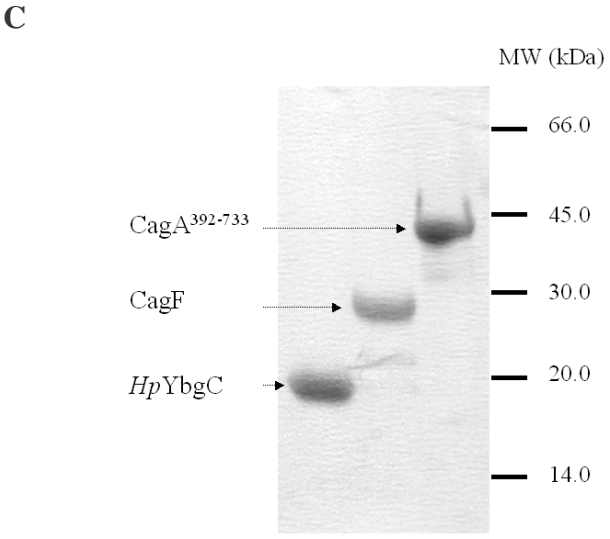


Figure 6.3.2 A, B. Characterization of CagA$^{392-733}$ (blu line) and CagF (red line) proteins separately. A and B: far-UV CD spectra of CagA$^{392-733}$ (40 % α-helices, 30 % β-sheets) and CagF (49 % α-helices, 12.5 % β-sheets), respectively.

Figure 6.3.2 C. SDS-PAGE of purified CagA$^{392-733}$, CagF and HpYbgC proteins.

The effect of the ionic strength on the CagA domain oligomerization and behavior in solution can be explained by the presence of several positive charged residues (pI 9.3) that characterize the CagA$^{392-733}$ domain sequence. Since the positive charged residues on the surface of a protein are sometime important for the interaction with the acidic phospholipids present in the membrane and the CagA protein is known to interact with the membrane, future experiments will be performed using seval acidic phospholipids, as weel as anionic detergents, in order to obtain a more homogenus and stable form of CagA in solution.

About CagF and *Hp*YbgC, as indicated by size exclusion chromatography experiments, the two proteins behave in solution as a dimer (70 kDa) and as a tetramer (65 kDa), respectively. The *Hp*YbgC structure has been determined (Chapter 7), whereas far-UV CD spectra of CagA$^{392-733}$ and CagF indicated that all these proteins exhibited a high proportion of secondary structure [Fig. 6.3.2]. The major contribution is that of the α-helix structure, about 49 % α-helices, compared with the β-sheets contribution of 12.5 % for CagF and of about 40 % compared with the β-sheets contribution of 30 % for CagA$^{392-733}$ domain.

Characterization of the CagA$^{392-733}$-CagF interaction

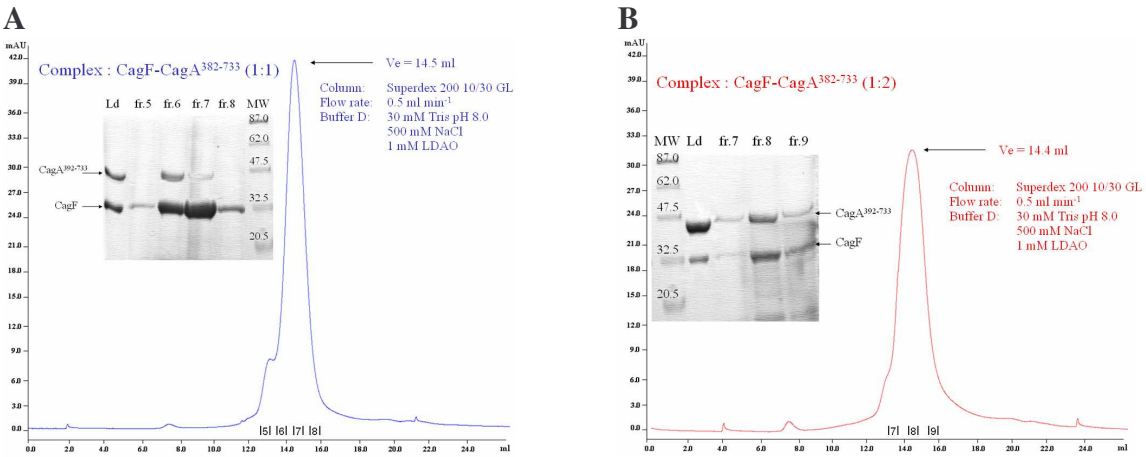
Since two different studies (Couturier *et al*., 2006; Pattis *et al*., 2007) reported that CagA$^{392-733}$ seems to bind specifically CagF, to validate this protein–protein interaction we performed far-western blotting assay as well as analytical gel filtration experiments.

About far western blotting experiments, increasing amounts of CagA$^{392-733}$ domain and CagF protein were separately spotted on a nitrocellulose membrane, followed by washing and incubation with the CagA$^{392-733}$ domain as a His-tagged protein [Fig. 6.3.6 B]. After extensive washing, the bound proteins were probed with an anti His-tag antibody. Moreover, several proteins like GST, Lysozyme, Myoglobin, Anhydrase and Ribonuclease, which were not expected to interact with CagA$^{392-733}$, were included as a negative control for non-specific binding [Fig. 6.3.6 C].

His_6-CagA$^{392-733}$ and CagF co-precipitate showing that CagA domain interact, even if weakly, with CagF under these conditions. A different observation was made in the case of His_6-CagA$^{392-733}$ and CagA$^{392-733}$, where the CagA domain seems to bind strongly to itself [Fig. 6.3.6 B]. Since these preliminary results supported a direct but week interaction between the CagA domain and CagF, mixtures of the two purified proteins were co-incubated to allow complex formation. The samples were then subjected to gel filtration chromatography over a Superdex 200 10/30 column and the

eluted fractions were analized by SDS-PAGE. In order to distinguish between co-elution due to similar molecular masses and co-elution as an effect of an interaction, we compared the elution volume of the complex with the elution volume of each single protein alone [Fig. 6.3.4]. CagA$^{392-733}$ was eluted as a trimer or tetramer (Ve = 13.4 ml) from the gel filtration whereas CagF (Ve = 14.5 ml) behaves in solution as a homodimer.

Figure 6.3.4. Characterization of the CagA$^{392-733}$-CagF complex by anlitycal gel filtration experiments



(A) Superdex 200 10/300 eluition profile of a roughly 1:1 mixture of CagA$^{392-733}$ and CagF. Coomassie Blue-stained 15 % SDS-PAGE analysis of the peak fractions. Purified proteins (10 mg ml-1 CagA$^{392-733}$ and CagF) were mixed 1:1 and incubated for 1 h at 4 °C prior to injection on a Superdex 200 10/300 analitycal gel filtration equilibtaed with Buffer D. The flow rate was 0.5 ml min-1 and proteins were detected by UV absorbance at 280 nm.

(B) Superdex 200 10/300 eluition profile of a roughly 2:1 mixture of CagA$^{392-733}$ and CagF. Coomassie Blue-stained 15 % SDS-PAGE analysis of the peak fractions. Purified proteins (10 mg ml-1 CagA$^{392-733}$ and CagF) were mixed 2:1 and incubated for 1 h at 4 °C prior to injection on a Superdex 200 10/300 analitycal gel filtration equilibtaed with Buffer D. The flow rate was 0.5 ml min-1 and proteins were detected by UV absorbance at 280 nm.

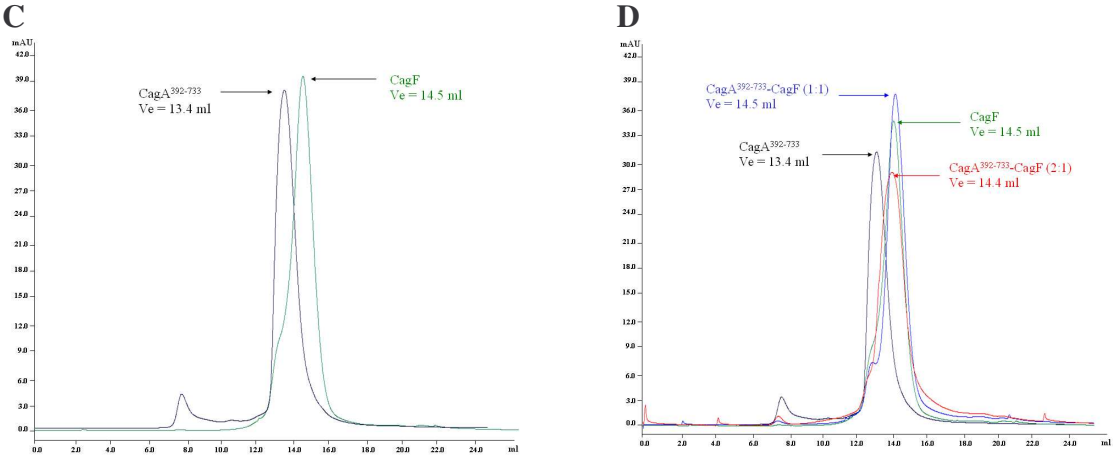


Figure **C** represents the eluition profiles of CagA$^{392-733}$ and CagF alone whereas the figure **D** represents the eluition profile of CagA$^{392-733}$ and CagF superimposed to the eluition profile of complexes A and B.

Generally, complex formation is assumed to lead to shifts of the elution volumes, but in our case we did not observe any increase of the eluition volume due to the complex. However, the complex was detected at a Ve of 14.5 ml, representing a complex that has a lower molecular weight of $CagA^{392-733}$ alone but the same molecular weight of CagF alone. These results, that are difficult to interpret, could suggest two hypotheses. The first one assumes that CagF interacts with $CagA^{392-733}$ forming a 1:1 stoichiometric complex. This finding is consistent with the formation of a complex containing one molecule of $CagA^{392-733}$ (38.5 kDa) and one molecule of CagF (35 kDa) would be of about 70 kDa, which corresponds quite well with the eluition volume (Ve = 14.5) of the peak fraction I in Fig. 6.3.4 A and B, as well as with the eluition volume of CagF heterodimer alone [Fig. 6.3.1 E and 6.3.4 C].

On the other hand, we could hypothesize that the affinities are not high enough to form a stable complex during gel filtration and what we see is just a co-elution due to similar molecular masses of the two proteins. However, $CagA^{392-733}$ is eluted with a lower Ve compared to that of the domain alone (Ve = 13, 4 ml) [Fig. 6.3.4 C and D]. This can be attributed to the presence of CagF: the latter, acting as a chaperone, could disaggregate and stabilize CagA, inducing its dimerization even in the presence of high salt. This means that the two proteins do not interact strongly and probably the interaction is weaker and transient or may need additional interaction partners. Secretion chaperones and chaperone-like molecues are in fact involeved in different types of protein scretion, but were extensively characterized only in type II and III secretion systems. In these systems, their funtion is to maintain the stability and secretion competence of effector proteins, for istance by keeping them in unfolded state, to prevent premature effector protein interactions and/or to recruit the effector proteins to the secretion apparatus (Parsot *et al*., 2003; Pattis *et al*., 2007). Like many well characterised type III secretion system chaperones, CagF has an acidic isoletric point (pI 4.5) and a high content of α-helix [Fig. 6.3.1 D]. However, with a size of 35 kDa, CagF is considerably much larger than a typical secretion chaperone, which usually has a molecular mass of about 20 kDa. This unusual size may be related to the membrane-association properties of CagF. Moreover, as most secretion chaperone in type III secretion systems, CagF forms homodimers and our gel filtrationd data are consistent with it [Fig. 6.3.1 E].

Alternative methods will be employ in the future to characterize this complex. In particular, to confirm that CagF interacts with the CagA domain we are planning several experiments, like pull-down assays as well as the co-expression of the two proteins.

Characterization of the CagA$^{392-733}$-*Hp*YbgC interaction

As for CagF, to verify that CagA$^{392-733}$ interacts with *Hp*YbgC, we performed far-western blotting assay as well as analytical gel filtration experiments.

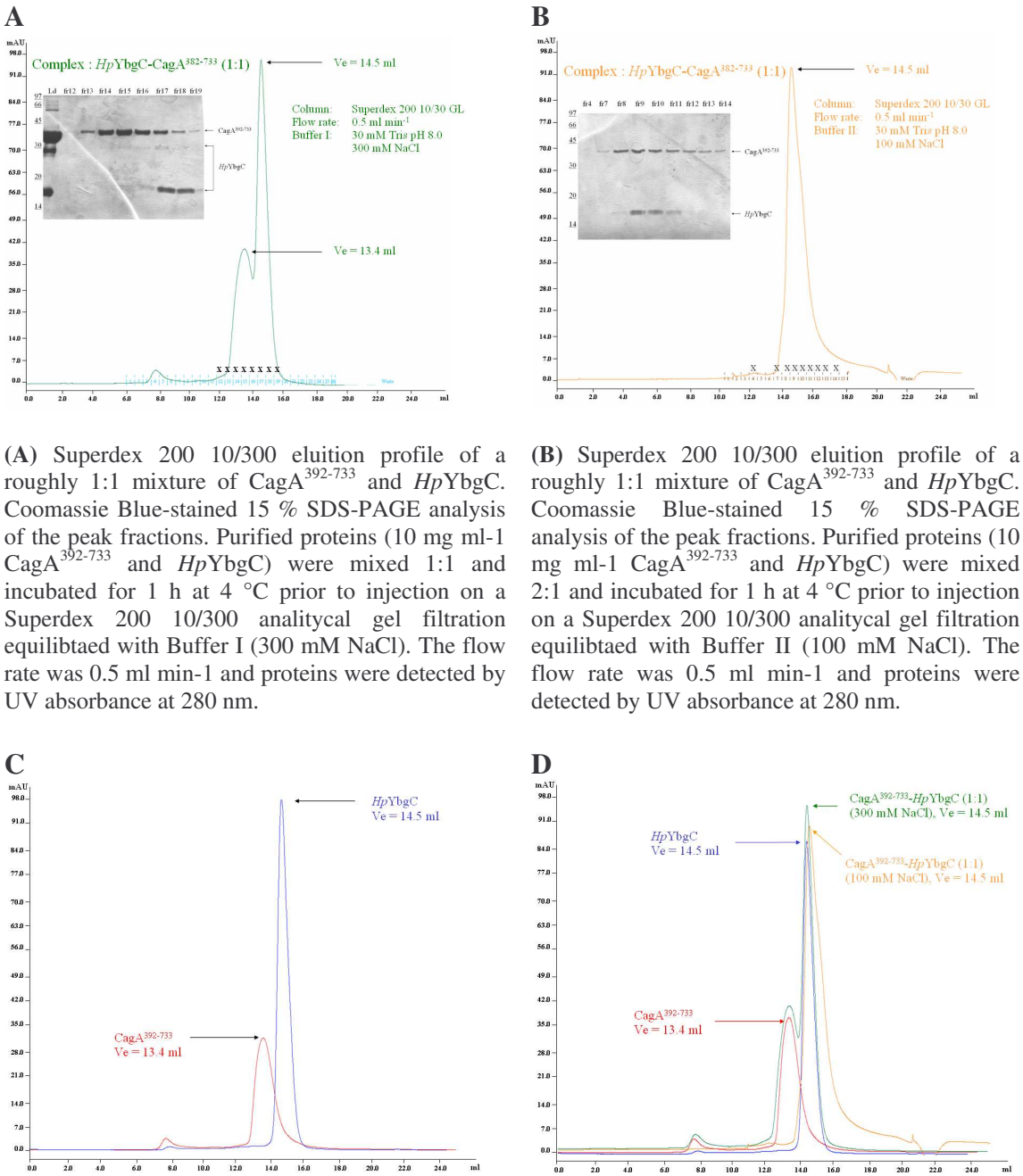
About far western blotting experiments, increasing amounts of the CagA$^{392-733}$ domain and *Hp*YbgC protein were separately spotted on a nitrocellulose membrane followed by washing and incubation with *Hp*YbgC and the CagA$^{392-733}$ domain, respectively [Fig. 6.3.6 A] After extensive washing, the bound proteins were probed with an anti His-tag antibody. Also in this case, several proteins, which were not expected to interact with CagA$^{392-733}$, were included as a negative control for non-specific binding [Fig. 6.3.6 C].

His_6-CagA$^{392-733}$ and *Hp*YbgC co-precipitate, showing that CagA domain interacts, strongly, with *Hp*YbgC under these conditions [Fig. 6.3.6 A]. These preliminary results suggest that the CagA domain seems to bind *Hp*YbgC stronger than CagF. To characterize this complex in solution, mixtures of the two purified proteins were co-incubated to allow complex formation. The samples were then subjected to gel filtration chromatography over a Superdex 200 10/30 column and the eluted fractions were analized by SDS-PAGE.

As mentioned before, complex formation is assumed to lead to shifts of the elution volumes, but, as for CagF, also in this case we did not observe any increase of the eluition volume due to the complex formation. Differently by CagF, to allow complex formation, the two proteins were incubated with 1:1 ratio using two different buffers. In particular, using a buffer containing 300 mM NaCl we could verify that the two proteins do not interact [Fig. 6.3.5 A]. In fact, they were eluted with identical eluition volumes of CagA$^{392-733}$ (Ve = 13.4 ml) and *Hp*YbgC (Ve = 14.5 ml) alone, respectively, as confirmed also by gel electrophoresis. In a different way, using a buffer containing 100 mM NaCl [Fig. 6.3.5 B] we could assess that the proteins were detected with a Ve of 14.5 ml. As in case of CagA$^{392-733}$-CagF complex, the hypothetical CagA$^{392-733}$-*Hp*YbgC complex has a lower molecular weight of CagA$^{392-733}$ alone, but the same molecular weight of *Hp*YbgC alone. However, in this case it is difficult to hypothesize that the two proteins interact, since we know that *Hp*YbgC is a stable tetramer of about 75 kDa that is eluted with a Ve of 14.5 ml. In an optimistic way we could assume that *Hp*YbgC (16 kDa) interacts with CagA$^{392-733}$ (38,5 kDa) forming a 2:1 complex of about 70 kDa, which could correspond to the eluition volume detected. However, since we know that CagA$^{392-733}$ alone at 100 mM NaCl [Fig. 6.3.5 B] is eluted with a Ve of 14.7 ml, we believed that it is most realistic to

think that also in this case, where we have used a buffer containing 100 mM NaCl, CagA$^{392\text{-}733}$ domain (Ve = 14.7 ml) and *Hp*YbgC (14.5 ml) are coeluted separately. However, as for CagF, alternative methods will be employ in the future to characterized this complex like pull-down assays and co-expression as well.

Figure 6.3.5. Characterization of the CagA$^{392\text{-}733}$-*Hp*YbgC complex by anlitycal gel filtration experiments



(A) Superdex 200 10/300 eluition profile of a roughly 1:1 mixture of CagA$^{392\text{-}733}$ and *Hp*YbgC. Coomassie Blue-stained 15 % SDS-PAGE analysis of the peak fractions. Purified proteins (10 mg ml-1 CagA$^{392\text{-}733}$ and *Hp*YbgC) were mixed 1:1 and incubated for 1 h at 4 °C prior to injection on a Superdex 200 10/300 analitycal gel filtration equilibtaed with Buffer I (300 mM NaCl). The flow rate was 0.5 ml min-1 and proteins were detected by UV absorbance at 280 nm.

(B) Superdex 200 10/300 eluition profile of a roughly 1:1 mixture of CagA$^{392\text{-}733}$ and *Hp*YbgC. Coomassie Blue-stained 15 % SDS-PAGE analysis of the peak fractions. Purified proteins (10 mg ml-1 CagA$^{392\text{-}733}$ and *Hp*YbgC) were mixed 2:1 and incubated for 1 h at 4 °C prior to injection on a Superdex 200 10/300 analitycal gel filtration equilibtaed with Buffer II (100 mM NaCl). The flow rate was 0.5 ml min-1 and proteins were detected by UV absorbance at 280 nm.

Figure **(C)** represents the eluition profiles of CagA$^{392\text{-}733}$ and *Hp*YbgC alone whereas the figure **(D)** represents the eluition profile of CagA$^{392\text{-}733}$ and *Hp*YbgC superimposed to the eluition profile of complexes A and B.

A

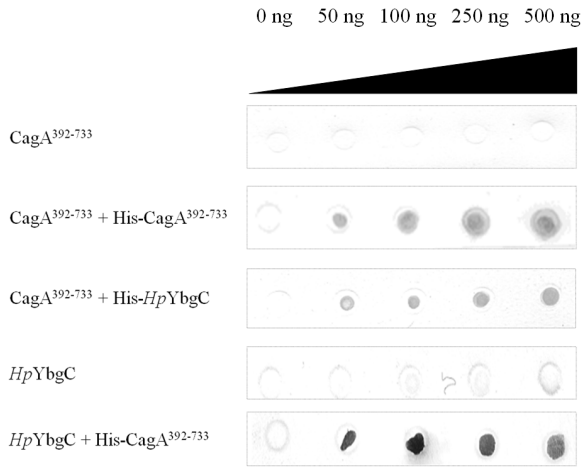


Fig. 6.3.6 A. Dot-lot/Far western blot assay between His-CagA$^{392\text{-}733}$ and *Hp*YbgC. From the top, the corresponding lanes contain the following: the purified CagA$^{392\text{-}733}$ protein alone, the purified CagA$^{392\text{-}733}$ incubated with the purified His-CagA$^{392\text{-}733}$; the purified CagA$^{392\text{-}733}$ incubated with the purified His-*Hp*YbgC, the purified *Hp*YbgC protein and the purified *Hp*YbgC protein incubated with the His-CagA$^{392\text{-}733}$ proteins. The presence of a dark signal that is indicative of an interaction between the two proteins.

B

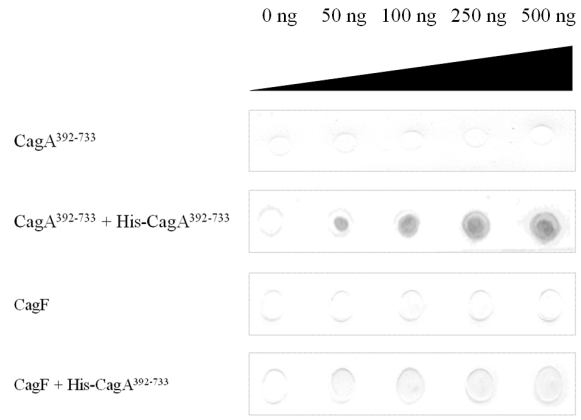


Fig. 6.3.6 B. Dot-lot/Far western blot assay between His- CagA$^{392\text{-}733}$ and CagF. Increasing amounts of CagA$^{392\text{-}733}$ domain and CagF were separately spotted on a nitrocellulose membrane. From the top, the corresponding lanes contain the following: the prified CagA$^{392\text{-}733}$ protein alone, the purified CagA$^{392\text{-}733}$ incubated with the purified His-CagA$^{392\text{-}733}$; the purified CagF alone and the purified CagF protein incubated with the His-CagA$^{392\text{-}733}$ protein. The presence of a dark signal that is indicative of an interaction between the two proteins.

C

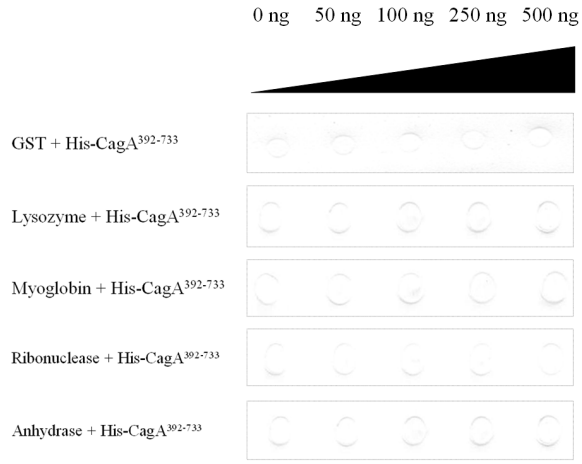


Fig. 6.3.6 C. Dot-lot/Far western blot assay between His-CagA$^{392\text{-}733}$ and several proteins which were used as negative controls. From the top, the corresponding lanes contain the following: the purified GST protein incubated with the purified His-CagA$^{392\text{-}733}$; the pure Lysozyme incubated with the purified His-CagA$^{392\text{-}733}$; the pure Myoglobin incubated with the purified His-CagA$^{392\text{-}733}$; the pure Ribonuclease incubated with the purified His-CagA$^{392\text{-}733}$ and the pure Anhydrase incubated with the purified His-CagA$^{392\text{-}733}$. The lack of dark signals suggest no detectable binding of the CagA$^{392\text{-}733}$ domain to any of the control proteins tested.

Figure 6.3.6. Characterisation of the CagA$^{392\text{-}733}$-CagF (A) and CagA$^{392\text{-}733}$-HpYbgC (B) complexes by the far western blot experiments. Clear evidence for binding was observed for the HpYbgC wheraes the binding signal with CagF is quite low.

Crystallization and proteolysis trials

Many sparse matrix crystallization tests were performed with the pure and concentrated CagA$^{392\text{-}733}$ domain, using the vapor diffusion technique. Several standard crystallization screens were used (Hampton Research, Qiagen and Molecular Dimension Ltd.) at different CagA$^{392\text{-}733}$ protein concentrations and salt concentrations. Crystals have not been obtained, yet.

To check if this was due to the presence of flexible regions in the protein, we performed proteolyis experiment on CagA$^{392\text{-}733}$, in order to identify a more stable and crystallizable domain. The limited proteolysis is successfully used in combination with mass spectrometry to improve crystal growth of proteins by cutting off amino- or carboxyl-terminal flexible end. In fact, limited proteolysis of a whole protein is a classical approach to probe domain structure, because well structured domains are in general more resistant to proteolysis.

For this purpose, the enzymatic reactions were performed using several proteases (Trypsin, α-Chimotrypsin, Thrombin, Thermolysin, Proteinase k and Subtilisin) at different concentration and temperatures (0 – 25 °C). Following digestion, products were analyzed by 15 % SDS-PAGE (data not shown) and we observed the complete protein degradation. This is probably due to the presence of several charged positive residues within the domain as well as flexible regions, that are recognised by the proteases and are susceptible to cleavage. On the other hand, we could assess that CagA$^{392\text{-}733}$ is not aggregated, since in general aggregated forms of the proteins are protease-resistant. Moreover, since the flexible domains are in general directly involved into the interaction with other partener, several crystallization trials of CagA domain with CagF and *Hp*YbgC were performed in order to obtain stable and crystallizable complex.

Chapter 7

Structural and enzymatic characterization of HP0496, a YbgC thioesterase from *Helicobacter pylori*

Contents

This chapter has been adapted from:
Alessandro Angelini [1,2], Laura Cendron [2], Susana Goncalves [1], Giuseppe Zanotti [2] and Laurent Terradot [1]; (2007). Structural and enzymatic characterisation of Hp0496 a YbgC thioesterase from *Helicobacter pylori*. *Proteins*. Accepted.

[1] Macromolecular Crystallography Group, European Synchrotron Radiation Facility, B.P. 220, 6 rue Jules Horowitz, F-38043 Grenoble Cedex, France.
[2] Department of Chemistry, University of Padua, and ICB-CNR, Section of Padua, Via Marzolo 1, 35131 Padua, and Venetian Institute of Molecular Medicine, Via Orus 2, 35127 Padua, Italy

Atomic coordinates and structure factors have been deposited at the Protein Data Bank (http://www.rcsb.org) for immediate release with PDB ID code 2PZH.

7.1 Introduction

Thioesterases are a large protein superfamily, widespread in bacteria, archea and eukaryotes involved in a wide range of cellular processes, such as thioester hydrolysis, phenylacetic acid degradation and fatty acid biosynthesis (Dillon *et al*., 2004). These enzymes catalyze the hydrolysis of acyl-CoA thioesters to free fatty acids and coenzyme A (CoA-SH). Acyl-CoA thioesterases adopt a so-called "hot-dog" fold, characterized by a central helix wrapped by a six-stranded antiparallel β-sheet (Dillon *et al*., 2004; Leesong *et al*., 1996). The hot-dog fold was also recently shown to be a component domain of proteins other than thioesterases. For instance in Gram-positive bacteria, the fatty acid and phospholipid biosynthesis regulator (FapR) utilises a hot-dog fold to sense malonyl-CoA and activate the transcription of genes involved in lipid biogenesis (Schujman *et al*., 2006).

Many crystal structures of proteins belonging to the hot-dog superfamily are available, including the 4-hydroxybenzoyl-CoA thioesterase from *Pseudomonas sp.* strain CBS-3 (*Ps*HTE; Benning *et al*., 1998) and from *Arthrobacter sp.* Strain SU (*Ar*HTE; Thoden *et al*., 2003), the thioesterase II from *E. coli* (*Ec*TEII; Li *et al*., 2000), the human thioesterase superfamily member 2 (hTHEM2; Cheng *et al*., 2006), the putative Acyl-CoA thioesterase from *Xanthomonas campestris* (XC229; Chin *et al*., 2006), the phenylacetate degradation protein PaaI from *Thermus thermophilus* (*Tt*PaaI; Kunishima *et al*., 2005), the YbgC protein from *E. coli* (*Ec*YbgC; PDB ID code 1S5U) and two hypothetical proteins YbaW and YdiI from *E. coli* (*Ec*YbaW; *Ec*YdiI; Badger *et al*., 2005). These enzymes are all defined by the arrangement of two hot-dog folds in γ dimers (Song *et al*., 2006) but these dimeric subunits form tetramers by two different modes. In *Ec*TEII, hTHEM2, *Ar*HTE, *Tt*PaaaI and *Ec*YdiI, the dimers associate on the β sheet side (the so called βγ mode), while in *Ps*HTE, *Ec*YbgC, *Ec*YbaW and XC229 the dimers associate by their helical sides (the so called εγ mode). Despite the number of structures available, the function of many proteins containing hot-dog folds are still poorly understood.

As much as 1357 proteins which are structurally related to the "hot-dog" domain-containing proteins have been classified into 17 subfamilies (Dillon *et al*., 2004). Among them, the YbgC-like subfamily includes a large number of poorly characterised acyl-CoA thioesterases, which show significant identity to the *Haemophilus influenzae* YbgC protein (*Hi*YbgC) and whose active site was suggested to be very similar to the 4-hydroxybenzoyl-CoA thioesterase *Ps*HTE (Dillon *et al*., 2004; Zhuang *et al*., 2002). *Hi*YbgC catalyzes the hydrolysis of the short chain

aliphatic acyl-CoA thioesters and is a member of the bacterial Tol-Pal system, a system important for transport of macromolecules between the bacterial membranes and for the maintenance of the membrane integrity in bacteria (Lazzaroni *et al*. 2002; Cascales *et al*., 2002; Lloubes *et al*., 2001). Recently, Tol-Pal system components were shown to accumulate at constriction sites and to be involved in cell division that drives the connections near the septal rings (Gerding *et al*., 2007). A recent study, using a modified version of the tandem affinity purification (TAP) system, showed that *Ec*YbgC interacts with acyl carrier protein (ACP), PlsX, PlsB and PssA, all these four proteins belonging to the phospholipids synthesis pathway (Gully *et al*., 2006). Together with *Hi*YbgC, YbgCs from *Shigella flexneri* and *E. coli* are representative members of a Tol/Pal cluster, and share the consensus sequence [DTD-$X_{(2)}$-GVV-X-H-$X_{(2)}$-Y] that defines the active site core (Dillon *et al*., 2004). Until now, the only YbgC member whose structure has been solved is the *Ec*YbgC protein (PDB ID code 1S5U, Kim *et al.*, unpublished). However little is known about YbgC and YbgC-like proteins functions in bacteria other than *E. coli*.

H. pylori is a gram negative bacterium that colonizes the human stomach. Over half of the human population is infected with *H. pylori*, with the highest rates in developing countries (Rothenbacher and Brenner, 2003). In 10 % of the cases, the persistence of the bacterium results in gastric pathologies, like peptic ulcer disease, atrophic gastritis and sometimes gastric cancer. *Helicobacter pylori* genome encodes for a protein of unknown function, HP0496, which displays 30 % sequence identity with *Ec*YbgC (Terradot *et al*., 2004). A high throughput yeast two hybrid system study has shown that HP0496 interacts with the cytotoxin associated gene protein CagA (Rain *et al*., 2001). This interaction was further confirmed by in an *in vitro* pull down assay after over expression of the two proteins in *E. coli* (Terradot *et al*., 2004), but the biological significance of the interaction between HP0496 and CagA has yet to be established.

To gain insight into HP0496 function, we have determined its crystal structure and characterized its enzymatic activity. We establish that HP0496 is a member of the hot-dog family of proteins, with a εγ tetrameric arrangement. As suggested from the sequence, we confirm that its closest structural homologue is *Ec*YbgC. We then show that HP0496 is an acyl-CoA thioesterase with a preference for long-chain acyl-CoA derivatives. Our results and analysis, together with the available studies on YbgCs, point toward a species-specific role of these proteins in the cell that could be linked to the phospholipids biogenesis.

7.2 Materials and methods

Cloning, expression and purification of HP0496

HP0496 open reading frame was PCR amplified from the genomic DNA of *H. pylori* strain 26695. The PCR product was inserted into the expression pET151/D-TOPO directional vector (Invitrogen) according to the manufacturer procedure. The resulting expression vector was used to transform *E.coli* BL21 Star (DE3) pLysS (Invitrogen). Cells harboring pET151-*hp0496* were grown at 37 °C in LB medium containing 100 μg ml^{-1} ampicillin and 40 μg ml^{-1} chloramphenicol up to an OD_{600nm} value of 0.6-0.8. The protein expression was induced by adding 1 mM iso-propyl β-thiogalactopyranoside (IPTG) for 16 h at 20 °C. Cells were harvested by centrifugation at 11.000 g for 20 min and suspended in 30 mM Tris-HCl, pH 7.5, 500 mM NaCl, 5 % Glycerol, 0.2 % Triton-X100 and 1 mM DTT. Resuspended cells were disrupted using a French press in the presence of DnaseI, Lysozyme and a mini "complete" EDTA-free protease inhibitor tablet (Roche). After centrifugation (25 min at 40.000 g), the soluble fraction was loaded onto a 5 ml His-Trap column (GE Healthcare) equilibrated with Buffer A (30 mM Tris-HCl, pH 7.5, 500 mM NaCl, 5% Glycerol). After a first wash in the same buffer and a second with 10 % Buffer B (Buffer A plus 500 mM Imidazole), the protein was eluted in a single peak at 400 mM Imidazole by applying a linear gradient (10 %-100 %) of Buffer B. The His-tag was cleaved by enzymatic digestion with Tobacco Etch Virus (TEV) protease. The protein was purified from the cleavage solution by loading it onto a 5 ml His-Trap column (GE Healthcare) and collecting the flow through. The protein was finally concentrated to 10 mg ml^{-1}. The oligomeric state of purified HP0496 was examined by size exclusion chromatography with a Superdex 200 10/30 column (GE Healthcare) equilibrated with Buffer A.

Crystallization and structure determination

The protein was mixed with an equal volume of reservoir solution containing 100 mM Tris pH 8.5, 20 % Ethanol (Crystal Screen 2 formulation, Hampton) and left to equilibrate at 20 °C. The hanging drop method was used and the crystals were obtained in 1 day. Diffraction data were collected at the beamline ID14EH3 of ESRF synchrotron (Grenoble). Crystals were brought to cryogenic temperature (100 K) after soaking for a few seconds into a cryo-protectant solution (reservoir solution with 15% glycerol). Data integration and scaling were performed with the programs MOSLFM and SCALA (CCP4 suite, 1994; Leslie, 1992). The crystals belong to the

orthorhombic space group $P2_12_12_1$ and contain one tetramer per asymmetric unit. The crystal structure was solved by molecular replacement using software MOLREP (Vagin and Teplyakov, 1997) starting from a model of the monomer generated through the Swiss-Model automatic modeling server using YbgC from *E. coli* (PDB ID code 1S5U) as a template (Schwede *et al.*, 2003). After a rigid body minimization, the structure was refined using programs CNS (Brunger *et al.*, 1998) and REFMAC (Murshudov *et al.*, 1999) to a final R factor of 0.21 (R_{free} 0.25). Model visualization and rebuilding were performed with the Coot program (Emsley and Cowtan, 2004). The quality of the structures was checked with the PROCHECK software . Statistics on data processing and refinement are reported in Table 7.1. Figures were generated using PYMOL (DeLano, 2002).

X-ray data [a]	
Space group	P $2_12_12_1$
Wavelength [Å]	0.931
Cell parameters, [Å, °]	**a**=49.860, **b**=99.205, **c**=107.288
Resolution (Å)	45-1.70 (1.79-1.70)
Independent reflections	53355 (6452)
Multiplicity	4.8 (4.2)
Completeness (%)	90.3 (76.5)
<I/σ(I)>	22.0 (2.7)
Rmerge	0.050 (0.50)
Refinement	
Total number of protein atoms	4440
Solvent molecules	295
$R_{cryst.}$ /R_{free} (%)	20.9/25.1
Ramachandran plot [%]	
Most favored	91.5
Additionally allowed	8.0
Generously allowed	0.4
Overall G-factor	0.1
R.m.s.d. on ideals	
bonds length [Å]	0.009
angles (°)	1.18

Table 7.1. Statistics on data collection and refinement. [a)] Numbers in parentheses refer to the last resolution shell.

Assay of thioesterase activity of HP0496

Substrates were purchased from Sigma-Aldrich. Hydrolysis reactions of the acyl-CoA thioester substrates were monitored at 25 °C by measuring the absorbance of 5-thio-2-nitrobenzoate at 412 nm, which was formed by reacting DTNB (5,5′-dithio-bis(2-nitrobenzoic acid)) with the CoA-SH released after the acyl-CoA substrate hydrolysis (Cheng *et al*., 2006). Each assay reaction was carried out in a quartz 10 mm light path cuvette at 303 K and monitored for 20 minutes. The 600 µl reaction mixture contained HP0496 protein (4.3 – 13 µM), 40-500 µM acyl-CoA substrate, 5 mM DTNB, 50 mM HEPES, pH 7.4, 300 mM KCl. The eleven different substrates tested were all purchased from Sigma-Aldrich (see Table 7.2). Kinetic data were collected by a Shimadzu UV-250 1PC recording spectrophotometer and processed by the software GraphPad Prism (version 5.0). V_{max}, K_m and K_{cat} were extrapolated using a non-linear regression fitting, constraining the enzyme concentration to the value estimated by UV absorption at 280 nm (extinction coefficient 10555 M^{-1} cm^{-1}) (Cheng *et al*., 2006).

Acyl-CoA Substrate	**K_m (µM)**	**K_{cat} (s^{-1})**	**K_{cat} / K_m ($s^{-1} \times M^{-1}$)**
acetyl-CoA (lithium salt) [C1:0]	-	$< 10^{-4}$	-
n-Propionyl-CoA (lithium salt) [C3:0]	-	$< 10^{-4}$	-
Butyryl-CoA (dilithium salt hydrate) [C4:0]	$2.7 \pm 0.8 \times 10$	$8 \pm 1 \times 10^{-4}$	3×10
B-Methylcrotonyl-CoA (lithium salt) [C5:1]	-	$< 10^{-4}$	-
Octanoyl-CoA (lithium salt monohydrate) [C8:0]	$4 \pm 1 \times 10$	$1.3 \pm 0.1 \times 10^{-3}$	3.3×10
Lauroyl-CoA (lithium salt) [C12:0]	$1.6 \pm 0.5 \times 10^{2}$	$1.5 \pm 0.4 \times 10^{-2}$	9.2×10^{2}
Myristoyl-CoA (lithium salt) [C14:0]	$2.4 \pm 0.8 \times 10^{2}$	$2.5 \pm 0.5 \times 10^{-2}$	1.1×10^{2}
Palmitoyl-CoA (lithium salt) [C16:0]	$5.7 \pm 1.5 \times 10^{2}$	$4.5 \pm 0.8 \times 10^{-2}$	7.9×10
Stearoyl-CoA (lithium salt) [C18:0]	$2.1 \pm 0.5 \times 10^{2}$	$3.4 \pm 0.4 \times 10^{-2}$	1.6×10^{2}
Benzoyl-CoA (lithium salt)	$2.0 \pm 0.8 \times 10^{2}$	$2.8 \pm 0.9 \times 10^{-2}$	1.4×10^{2}

Table 7.2. Kinetic constants measured for wild-type HP0496-catalyzed hydrolyses of acyl-CoA thioesters at pH 7.4 and 30 °C. The kinetic parameters of K_m and K_{cat} were determined from initial velocity data, assuming a Michaelis-Menten model.

7.3 Results and discussion

Analysis of the *Hp*YbgC encoding gene (*hp0496*)

A search for putative homologues of the HP0496 protein suggests that it is an homologue of the YbgC proteins from *E. coli* (33 % sequence identity and 58 % similarity, Fig. 7.2 A), *Haemophilus influenzae* (31 % sequence identity and 59 % similarity) and *Shigella flexneri* (33 % sequence identity and 58 % similarity), all located next to the Tol-Pal cluster of genes (Sturgis, 2001) (Fig. 7.1 A).

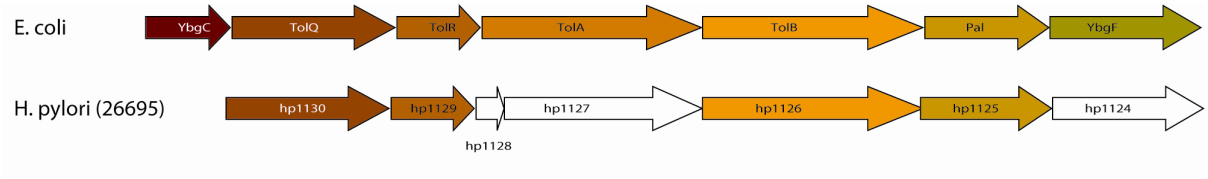


Figure 7.1 A. Comparison of the Tol/Pal clusters from *Escherichia coli* and *Helicobacter pylori.* Diagram showing (top) the genes from the Tol/Pal cluster in *E. coli* with and (bottom) the corresponding genes in *H. pylori* colored as in *E. coli* except for the open reading frames *hp1128*, *hp1127* and *hp1124* (in white) for which no significant sequence homology was found with *E. coli* genes.

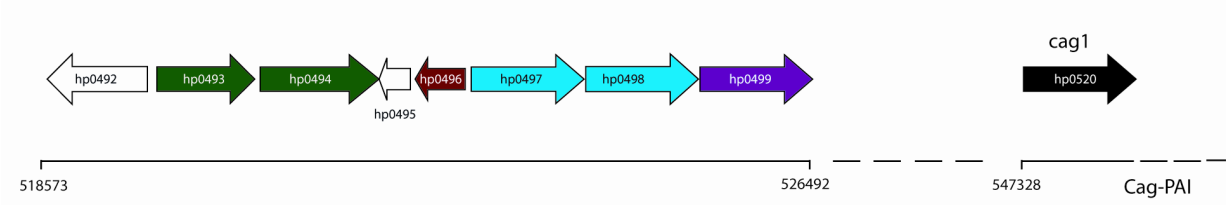


Figure 7.1 B. Location of *hp0496* ORF (colored in red) in *H. pylori* genome. *hp0496* is located before the Cag-PAI island (*hp0520* to *hp0547*) and is surrounded by genes involved in lipid biogenesis (magenta), peptidoglycan biosynthesis (green), membrane transporters (light blue) or of unknown function (white).

Based on this result we name *Hp*YbgC the protein encoded by *hp0496* throughout the rest of the text. The *hp*0496 gene is nevertheless not located within or close to the Tol-Pal cluster of genes in *H. pylori.* Homologues of the Tol/Pal proteins were found in the genome of the strain 26695: TolQ (*hp*1113, TolR (*hp*1129), TolB (*hp*1126), Pal (*hp*1125), and possibly YbgF (*hp*1124), although in this latter case the sequence has no clear homology with *E. coli* YbgF (Sturgis, 2001; Tomb *et al.*, 1997). *hp*0496 gene is surrounded by homologues of phospho-N-acetylmuramoyl-pentapeptide-transferase (*hp*0493) and UDP-N-acetylmuramoyl-L-alanyl-D-glutamate synthetase (*hp*0494), both involved in peptidoglycan biosynthesis in *E. coli* (Tomb *et al.*, 1997; Fig. 7.1 B). It is also close to a gene coding for a protein of unknown function (*hp*0495) which structure has been recently determined (Seo *et al.*, 2007). Downstream of the *hp*0496

genes are found membrane sodium- and chloride-dependent transporters (*hp*0497 and *hp*0498), followed by a phospholipase A1 precursor *PldA* (*hp*0499). Interestingly, the latter gene encodes for PldA, a protein with lecithinase, PLA_2, and haemolytic activities that was shown to have a role in colonization of the gastric mucosa and possibly tissue damage after colonization (Istivan *et al*., 2006; Dorrell *et al*., 1999). Therefore, despite its sequence homology with YbgC proteins, *hp*0496 gene does not associate with the Tol/Pal cluster of genes in *H. pylori.* In contrast with other YbgCs, the genome location of *hp*0496 suggests that it associates with phospholipid or LPS biogenesis.

*Hp*YbgC structure is a member of the hot-dog fold family of proteins

The crystal structure of *Hp*YbgC was determined at 1.7 Å resolution with four monomers per asymmetric unit. No residue shows dihedral angles outside the allowed regions of the Ramachandran plot and all the other geometric parameters are as expected or better for this resolution. The monomer adopts the so-called "hot-dog" fold, with a long α-helix (α2) surrounded by a four-stranded antiparallel β-sheet (strands β2 to β5, Figure 2B). Other two short helices and a short β-strand (β1) complete the structure. Two monomers of HP0496 assemble to form a dimer, mainly through a net of hydrogen bonds between β-strands β2 of each monomer: main chain atoms of strand β2 of one monomer (A or C), form six hydrogen bonds with strand β2 of the adjacent monomer (B or D, respectively), giving rise to a continuous β-sheet of eight strands (Fig. 7.2 B). Other three H-bonds stabilize the interaction, whilst the two central helices are in contact mainly through hydrophobic interactions.

*Hp*YbgC is a tetramer in solution gel filtration (data not shown) and in the crystal, formed by the assembly of four identical subunits, characterized by 222 symmetry (Fig. 7.2C). Interactions in the tetramer are mainly between loops α1-α2 and α2 itself. There are 11 hydrogen bonds and 4 ion pair interactions between monomers A and C and 4 hydrogen bonds between monomers A and D. The surface buried by each monomer in the tetramer formation is about 1800 $Å^3$, for a total contact area of 7080 $Å^3$. The tetramer formation produces a small cavity in the centre of the molecule, in correspondence with the intersection of the two-fold axes. This cavity is heavily charged (Lys23, Glu8, Asn20) and filled with solvent molecules, but it is not accessible to the solvent.

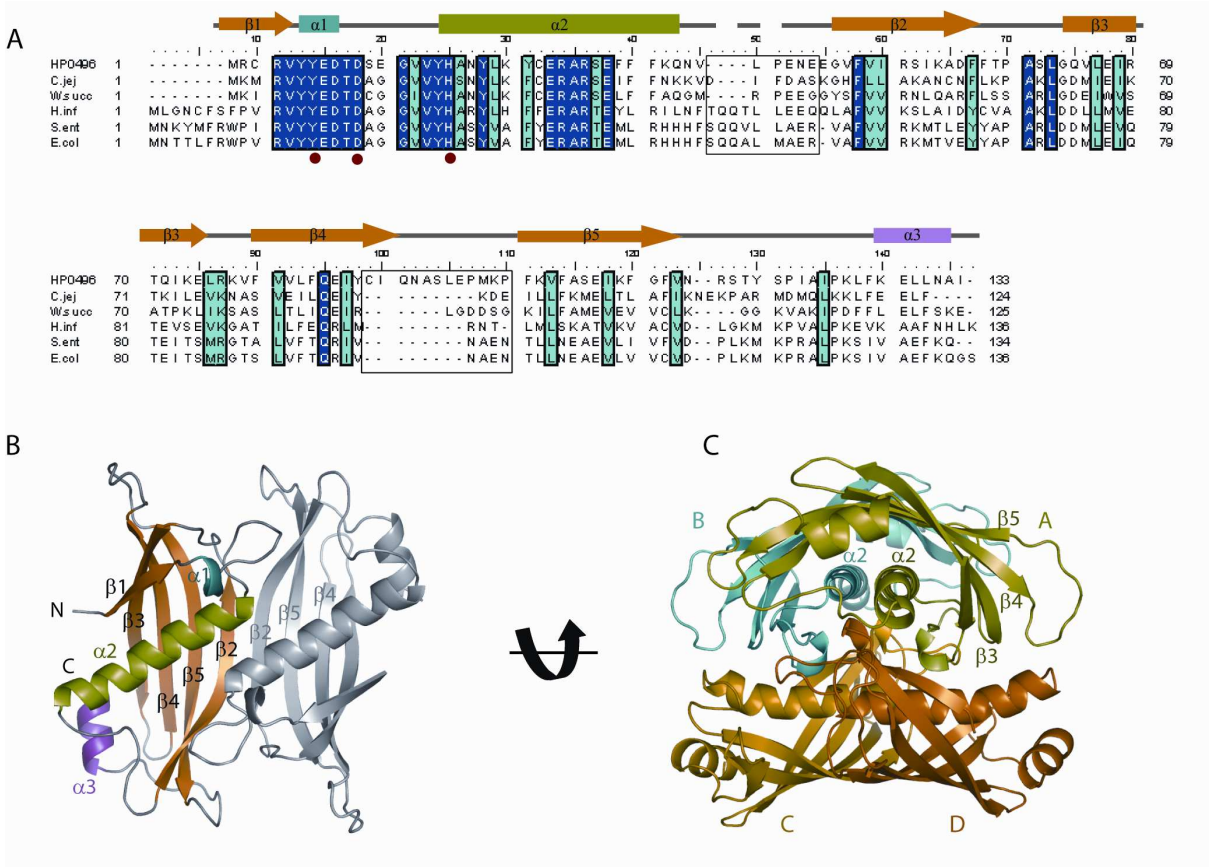


Figure 7.2. Crystal structure of *Hp*YbgC. (A) Sequence alignment of YbgC proteins and secondary structure assignment. Amino acids of five representative homologues of *Hp*YbgC (HP0496; gi|15645123) were aligned, from *Campylobacter jejuni* (C. jej; gi|6968402), *Wolinella succinogenes* (W. suc; gi|34483342), *Salmonella enterica* (S. ent; gi|16759683) *Haemophilus influenzae* (H. inf; gi|42629844) and *E. coli* YbgC (E. col ; gi|49258898). Strictly conserved and strongly conserved residues are marked dark and light blue, respectively. The model of HpYbgC is shown as a grey line above the sequences for non-regular structure, blue, green and pink boxes for α-helices α1, α2 and α3, respectively and orange arrows for β-strands. Black boxes indicate the two most divergent regions between *Hp*YbgC and *Ec*YbgC. Red dots indicate the residues belonging to the putative active site. B) Structure of the monomer and dimer of *Hp*YbgC. Secondary structure representation and labels are as in A. C) Tetrameric arrangement of *Hp*YbgC with subunit A colored in green, B in blue, C in yellow and D in orange.

HP0496 is a YbgC protein with a εγ tetrameric arrangement

A DALI search performed with *Hp*YbgC identified 19 structures with Z-scores higher than 6 sharing a highly similar hot-dog fold and the dimer structure of *Hp*YbgC is also well conserved. *Hp*YbgC is most similar to the *Ec*YbgC (r.m.s.d. 1.2 Å, PDB ID code 1S5U, Z-score = 18) and also adopts the same arrangement. *Hp*YbgC dimers pack with the helices buried in the interior of the tetramer and the β-sheets exposed to the solvent, an arrangement defined as the "εγ mode" (Kunishima et al., 2005) in *Ps*HTE, *Ec*YbgC, *Ec*YbaW and XC229 structures (Fig. 7.3 A). The two members of the YbgC family therefore adopts the εγ mode of tetramerisation, suggesting that this is likely to

be a general characteristic of YbgCs. Conserved residues among YbgC proteins (Fig. 7.2 A) were mapped onto *Hp*YbgC protein. The strictly conserved residues clearly point towards the centre of the tetramer and indicate only buried residues (Arg4, Glu8, Asp9, Arg26 and Glu30), with only two conserved residues being solvent accessible (Leu74, Arg76) (Fig. 7.3 B). Compared with *Ec*YbgC (Fig. 7.3 B and C), *Hp*YbgC displays a particular loop connecting strands β4 and β5 which is about 10 residues (90 to 100) longer than in *Ec*YbgC and the other YbgC proteins (Fig. 7.3 C). While Gln90 and Asn91 extend the length of the β sheet, the main chain of Ala92 and Ser93 (of chain A) hydrogen bond with the side chain of Asn37 and main chain of Tyr34 from the adjacent subunit (chain D; Fig. 7.3 D).

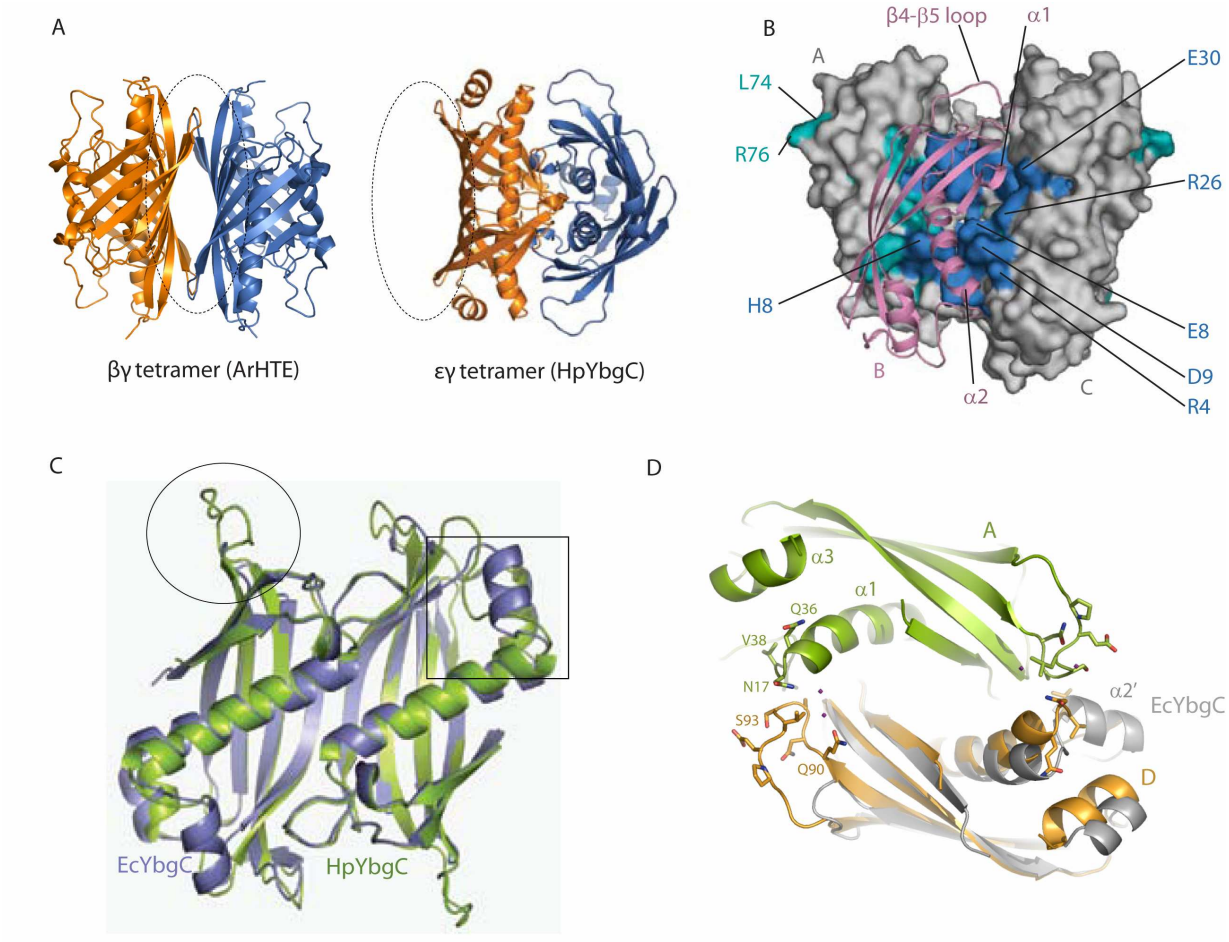


Figure 7.3. A) Ribbon diagram comparing the tetramers assembled according to the βγ mode of (*Ar*HTE) with the εγ mode (*Hp*YbgC). B) Conserved residues from the YbgC family of proteins. Surface representation of the tetramer of *Hp*YbgC for chain A, C and D with strictly conserved residues colored in dark blue and conserved residues in light blue. Chain B is represented by a pink ribbon.
C) Structural superimposition of the dimers (Chains A and B) of *Hp*YbgC (green) and *Ec*YbgC (blue). The fold is very well conserved but two areas are divergent with an extended loop in *Hp*YbgC (circle) and an additional helix, α2' in *Ec*YbgC (box). Detailed view of the residues involved in specific interactions between *Hp*YbgC chain A (green) and D (orange) and comparison with the superimposed chain D of *Ec*YbgC.

Water-mediated hydrogen bonds strengthen the packing of these subunits and thus of the tetramer. Interestingly, *Ec*YbgC displays an additional helix, that we name α2', between α2 and β2 compared with *Hp*YbgC (Fig. 7.3 C and D). This additional helix doest not make any contact with residues from chain D in the structure of *Ec*YbgC, but provides additional hydrogen bonds that could help stabilising the subunits interactions.

*Hp*YbgC is a thioesterase that favors long-chain acyl-CoA subtrates

YbgC-like proteins are similar to 4-hydroxybenzoyl-CoA thioesterase, an enzyme that catalyses the hydrolysis of the thioester moiety in 4-hydroxybenzoyl-CoA to yield 4-hydroxybenzoate and Coenzyme A (Benning *et al.*, 1998; Sturgis *et al.*, 2001). *Hi*YbgC was shown to hydrolyze acyl-CoA chains with a preference for short-chain substrates. To determine whether *Hp*YbgC has a similar activity, we analysed its thioesterase activity by monitoring the hydrolysis of several substrates at 303 K and pH 7.4 (Table 7.2).

We tested some short and long chain thioesters, along with benzoyl-CoA thioester. As indicated in Table 7.2, *Hp*YbgC enzyme has a preferred activity for long chain Acyl-CoA substrates and benzoyl-CoA, with weakest activity on short and medium chain substrates (<C12:0). The k_{cat} values observed for the five most active substrates tested, palmitoyl-CoA (C16:0), stearoyl-CoA (C18:0), myristoyl-CoA (C14:0), lauroyl-CoA (C12:0) and benzoyl-CoA, range from 0.045 s^{-1} to 0.015 s^{-1}, while the Km values range from 570 to 160 μM. *Hi*YbgC, the only YbgC family member whose activity has been partially characterized by now, showed k_{cat} values ten times higher and Km around 10 mM for short chain acyl-CoA and favoured propionyl-CoA as a substrate; no activity was detected for long chains acyl-CoA substrates. *Hp*YbgC shows relatively modest k_{cat} /Km values, suggesting that the identified substrates are similar but not identical to the actual physiological ones, as also suggested for *Hi*YbgC. Interestingly, *H. pylori* was found to have an unusual cellular fatty acid profile, with several palmitoyl-based lipids in its membrane composition (Moran, 2007). The fatty acid composition of *H. pylori* have been analyzed by gas-liquid chromatography, identifying as major components myristic acid ($C_{14:0}$), palmitic acid ($C_{16:0}$), stearic acid ($C_{18:0}$), oleic acid ($C_{18:1}$), linoleic acid ($C_{18:2}$), 19-carbon cyclopropane fatty acid ($C_{19:0}$cyc), β-hydroxy-palmitic acid (3-OH-$C_{16:0}$), and β-hydroxystearic acid (3-OH-$C_{18:0}$). Myristic acid (41 to 55 %) and 19-carbon cyclopropane fatty acid (22 to 30 %) were the major phospholipid fatty acids detected. Interestingly, it has been shown that the cellular fatty acids 3-OH-$C_{16:0}$ and 3-OH-$C_{18:0}$ are unique to this organism,

suggesting that specific enzymes are necessary to catalyze their synthesis (Moran, 2007). It is therefore worth noting that *Hp*YbgC substrate specificity correlates with the unique lipid composition of *H. pylori.*

Putative site and substrate specificity

Substrate-bound crystal structures are available for several hot-dog fold thioesterases including *Ps*HTE, *Ar*HTE and PaaI. Among these substrate/enzyme complexes, *Ps*HTE is the closest homologue of *Hp*YbgC (Z score = 15.9). *Ps*HTE has also been crystallized with the inhibitor 4-hydroxyphenacyl-CoA (PDB ID code 1LO7) and an inactive mutant (Asp17Asn) was crystallized with the substrate 4-hydroxybenzoyl-CoA (PDB ID code 1LO9) allowing a detailed description of its active site (Thoden *et al*., 2002). The active site of the enzyme are located at the interface between two subunits A and B and between C and D (Thoden *et al*., 2002). Whereas the nucleotide moiety is positioned in a depression located at the surface of one of the subunit, the rest of the substrate is in a deep crevice formed at the subunit-subunit interface (Fig. 7. 4 A). Here the aromatic moiety of the substrate is positioned by interactions with the side chain of Trp47 and main chain of Thr59 from one subunit and of Tyr24 from the other subunit (Fig. 7.4 A). The Asp17 is located at 3.2 Å from the carbonyl group of the substrate where it can act as a nucleophile as derived from mutagenesis (Thoden *et al*., 2002). We generated a model of the *Hp*YbgC/4-hydroxybenzoyl-CoA and *Ec*YbgC/4-hydroxybenzoyl-CoA complex using the 4-hydroxybenzoyl-CoA coordinates from *Ps*HTE mutant complex (PDB ID code 1LO9) and superimposed it to *Ps*HTE (Fig. 7.4). The two YbgCs present very similar organization of the active site. The thioester carbonyl group of the substrate is positioned at the end of helix α2, where it can form a hydrogen bond with the main chain of His18 in *Hp*YbgC (His25 in *Ec*YbgC), which corresponds to Tyr24 of *Ps*HTE. Tyr17, which superimposes well with Trp47, could play a similar role by interacting with the hydroxyl group of the ligand (Fig. 7.4). At the same time the side chain of Asp11 points towards the carbonyl carbon of the substrate and it could function as the nucleophile, as suggested for the corresponding Asp17 in *Ps*HTE or Asp18 in *Hi*YbgC (or *Ec*YbgC).

Interestingly, in *Hp*YbgC, Glu13 also points towards the carbonyl group and could act as a nucleophile. However, the mutagenesis carried out on *Hi*YbgC suggests that Asp17 is likely involved in the catalysis in *Hi*YbgC (Zhuang *et al*., 2002) and therefore Asp11 could play this role in *Hp*YbgC.

Most of the other residues in the active site are conserved including or are substituted by residues with comparable properties. The major difference between the three active

sites is represented by loop connecting α2 to strand β2, which is shorter in *H. pylori* compared with *Ec*YbgC and *Ps*HTE where this connection consists in two helical turns.

In this study we show that *Hp*YbgC thioesterase activity is higher with substrates that have long acyl chains, such as palmitoyl-, stearoyl- and myristoyl-CoA, or aromatic chains, like benzoyl-CoA, whilst *Hi*YbgC, is active only on short-chains acyl-CoA's.

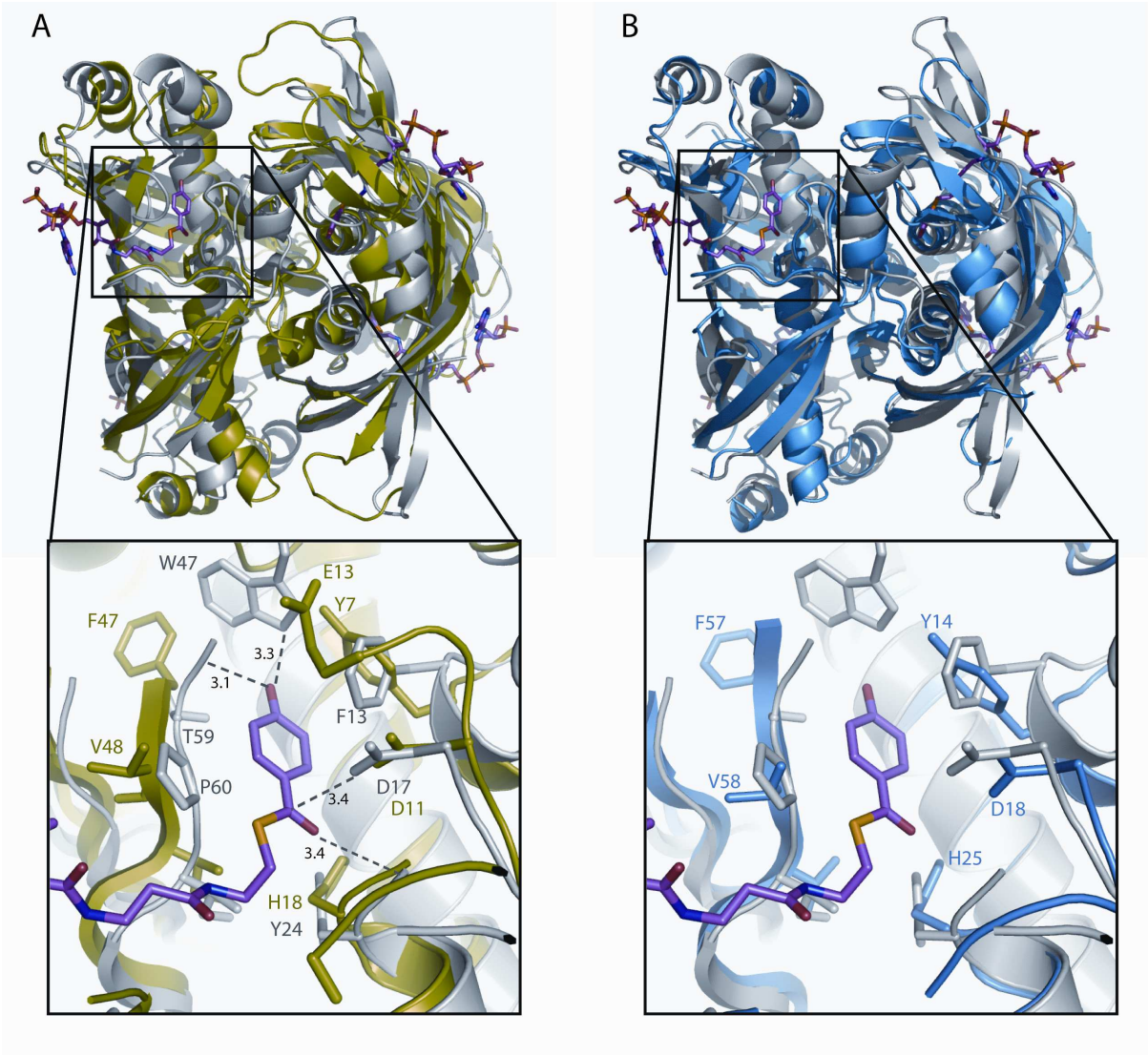


Figure 7.4. Structural comparison of *Ps*HTE active site (pdb code 1LO7) with (A) *Hp*YbgC and (B) *Ec*YbgC. Upper panels: ribbon representations of the tetramers of *Ps*HTE (grey) superimposed to *Hp*YbgC (green, left) or *Ec*YbgC (blue, right). The substrate 4-hydroxybenzoyl-CoA superimposed from the structure of *Ps*HTE (pdb code 1LO9) is represented by sticks and colored by atoms (carbon, magenta; yellow, sulfur; nitrogen, blue; oxygen, red). Lower panels: close-up view of the active site of *Ps*HTE and the putative active site of *Hp*YbgC (left) and *Ec*YbgC (right). The side chains of the residues belonging to the active site are represented by sticks. Distances separating the substrate from the residues of *Ps*HTE are indicated (in Å). The substrate is represented and colored as in the upper panels.

This suggests that *Hp*YbgC proteins, while having similar catalytic residues might have differences in their structures explaining their substrate specificities. We noticed major differences in the substrate binding region. *Hp*YbgC presents in fact a long tunnel associated with the region binding the acyl moiety of the substrate, whereas this tunnel is absent in *Ps*HTE and in *Ec*YbgC enzymes (Fig. 7.5).

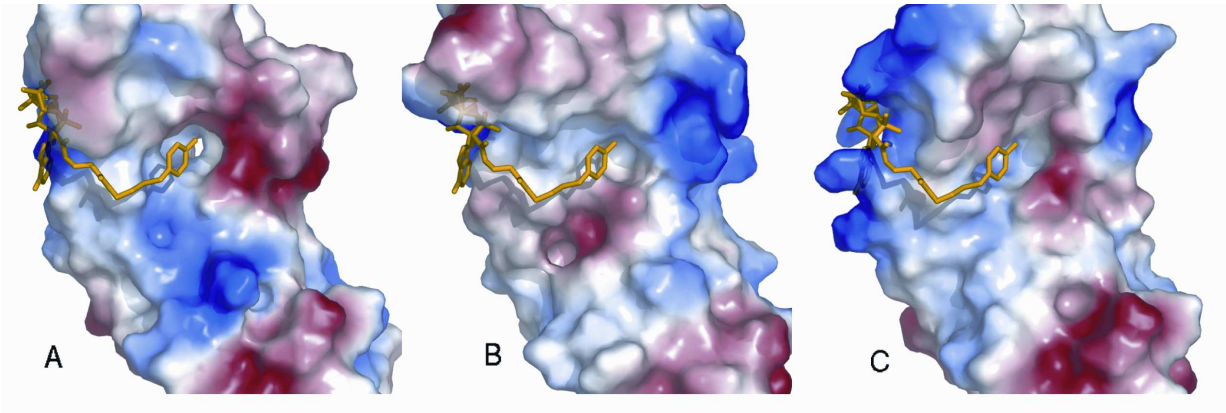


Figure 7.5. Electrostatic potential surface of the dimer of (A) *Hp*YbgC, (B) *Ps*HTE, (C) *Ec*YbgC in the region around the active site. The substrate 4-hydroxybenzoyl-CoA, as it was observed in *Ps*HTE structure (pdb code 1LO9) is superimposed onto the three structures. A tunnel in the active site of *Hp*YbgC is clearly visible, whilst it is absent in the other two, possibly preventing the binding of a substrate with a long chain.

A superposition of hydroxybenzoyl-CoA substrate, as identified in the *Ps*HTE complex structure, to the *Hp*YbgC and *Ec*YbgC structures shows that the benzoyl group points towards the active site tunnel in *Hp*YbgC (Fig. 7.5 A), where there is space for a longer chain, whilst in the other two enzymes the entrance of the tunnel is blocked, hindering the binding of a long-chain acyl-CoA to the active site (Fig. 7.5 B and C). This difference in *Hp*YbgC active site is due to the flexible loop between residues Val38 and Glu44 , which does not display the α-helical arrangement present in *Ec*YbgC, leaving an extra space in the active site. This fragment corresponds to one of the two gaps in the sequence alignment with other YbgC proteins (Fig. 7.2 A). Interestingly, the distribution of charges in the active site varies between each structure, which could also play a role in the substrate specificity. In contrast, the region of the active site that binds the CoA moiety is well preserved, both in shape and charge distribution in the three structures.

7.4 Conclusion

In this study we have determined the structure of a new member of the YbgC family of proteins that we named *Hp*YbgC (HP0496) and identified its enzymatic activity. The crystal structure shows that *Hp*YbgC and *Ec*YbgC share the same thioesterase-characteristic hot-dog fold. Together with the structure of *Ec*YbgC and the strong similarity between protein sequences within this family, our results suggest that this is a general feature of the YbgC family. The crystal structure of *Hp*YbgC also shows that the εγ tetrameric assembly mode is conserved among the YbgC proteins. This particular mode of tetramerisation leaves an important surface, generated by the 12-stranded β sheet, available for additional protein-protein interactions, such as in βγ mode found in other hot-dog fold proteins. These areas are more species-specific than the highly conserved active site and could be involved in specific interaction such as *Hp*YbgC interaction with CagA or *Ec*YbgC with ACP.

*Hp*YbgC is an acyl-CoA thioesterase and, in contrast with the *Hi*YbgC, has a maximum activity with long-chain acyl-CoA, such as palmitoyl-CoA and aromatic chain like benzoyl-CoA. Based on the activity and on the comparison of *Hp*YbgC structure with *Ec*YbgC and related enzymes, we suggest that the catalytic core of the YbgC is conserved. YbgCs are therefore likely involved in hydrolyzing acyl-CoA derivatives in the cell and process via a mechanism similar to the one used by other thioesterase such as *Ps*THE. YbgC proteins are widely distributed in gram negative bacteria and are part of the tol/pal system, a machinery important in cell membrane integrity and cell division. The role of YbgC in this system is unclear. No phenotype could be attributed to deletion mutants of YbgC in *E. coli,* suggesting that YbgC might not have a direct role in the Tol/Pal function. This seems also to be the case of *Hp*YbgC, since the gene is not part of the Tol/Pal system of this organism. A recent study showed that *Ec*YbgC (but not its homologue YbaW) specifically binds to the ACP and is part of a protein network involved in the phospholipids and LPS biogenesis at the cell membrane. *Ec*YbgC activity could therefore be coupled to the Tol/Pal system while not being necessary to its function. Gully and Bouveret hypothesized that *Ec*YbgC could be required for the synthesis, termination and release from ACP of certain specific lipids to be transferred to phospholipids synthetic enzymes. In *H. pylori,* the gene coding for *Hp*YbgC does not belong to the Tol/Pal cluster of gene and could also have a distinct function. Moreover, it is located nearby genes involved in phospholipids synthesis. Our study therefore supports the idea that YbgCs might be involved in the biogenesis of species-specific phospholipids.

Appendix

Appendix A: Abbreviations and Symbols

Å	Angstrom
aa	amino acid
Abs	Absorption
ADP	Adenosine Di-Phosphate
ATP	Adenosine Tri-Phosphate
cag	cytotoxin associated gene (gene)
Cag	Cytotoxin associated gene (associated protein)
cag-PAI	cytotoxin associated genes - PAthogenicity Island (genes)
Cag-PAI	Cytotoxin associated gene- PAthogenicity Island (associated proteins)
CCD	Charge-Coupled Device
CHAPSO	3-[(3-CHolamidopropyl) dimethylAmmonio]-2-hydroxyPropaneSulfonic acid
CMC	Critical micelle concentration
$C_{12}E_8$	Octaethyleneglycol Mono-n-dodecyl Ether
Da	Dalton
DTT	DiThioThreitol
EDTA	Ethylene Diamino Tetracetic Acid
ESRF	European Synchrotron Radiation Facility
F(hkl)	Structure factor amplitude
Fobs, Fcalc	Observed and calculated structure factor amplitudes
FOM	Figure Of Merit
FPLC	Fast Protein Liquid Chromatography
Hepes	N-[2-Hydroxyethyl] piperazine-N'-[2-ethanesolfonic] acid
I	Measured Intensity of the diffraction spots
IPTG	IsoPropyl-β-D-ThioGalactopyranoside
IS	Insertion Sequences
LB	Luria Bertani liquid medium
LDAO	LaurylDimethylAmine Oxide (detergent)
MAD	Multiple Anomalous Dispersion
mAu	milli Absorption unit
MES	2-(N-Morpholin) ethansulfonate
MW	Molecular Weight
O.N.	Over Night
OD	Optical Dispersion
ORF	Open Reading Frames
PAI	PAathogenicity Island
PCR	Polymerase Chain Reaction
PDB	Protein Data Bank
PEG	PolyEthylene Glycol

pI	Isoelectric point
r.m.s.d.	Root-mean-square deviation
SAD	Single Anomalous Dispersion
SDS	Sodium Dodecyl Sulfate
SDS-PAGE	SDS-PolyAcrylamide Gel Electrophoresis
SEM	Scanning Electron Microscopy
Tris	2-Amino-2-(hydroxymethyl)-1,3-propanediol
Triton	Octylphenoxypolyethoxyethanol Polyethylene Glycol-p-isooctylphenyl Ether
σ(I)	Standard deviation of the measured Intensities (I)

Amino acids

Ala	A	Alanine
Arg	R	Arginine
Asp	D	Aspartic acid
Asn	N	Asparagine
Cys	C	Cysteine
Gly	G	Glycine
Gln	Q	Glutamine
Glu	E	Glutamic acid
His	H	Histidine
Ile	I	Isoleucine
Lys	K	Lysine
Leu	L	Leucine
Met	M	Methionine
Phe	F	Phenylalanine
Pro	P	Proline
Ser	S	Serine
Thr	T	Threonine
Tyr	Y	Tyrosine
Trp	W	Tryptophan
Val	V	Valine

Appendix B: Methods of cloning, protein expression, solubility, purification, crystallization trials and structure solving

Cloning for protein expression

The first step toward protein expression is the cloning of the gene of interest in a proper expression vector. Almost without exception, we cloned the selected genes with an affinity tag, which makes the following purification steps easier and can increase solubility, yield, and stability of the target protein (Esposito *et al*., 2006). For each protein various expression construct may need to be tested for optimal expression, solubility and crystallization. The selected genes were amplified by PCR of chromosomal/genomic DNA extracted from the *H. pylori* pathogenic type I strain CCUG 17874 (Genebank numbers: AF282853 and AF282852) using specific primers (forward and reverse), designed individually for each open reading frame and able to introduce restriction sites at the 5' and 3' ends of the gene. The PCR products were then agarose gel purified and firstly cloned into the pBSK (Stratagene) intermediate cloning vector before subcloning into an expression vector using standard techniques (Sambrook *et al*., 1989; Hartley *et al*., 2006). In most cases, full-length *cag* genes were cloned. For those genes that are predicted to encode a secretion signal sequence, only the portion of the gene encoding the mature protein was cloned, in order to avoid that the hydrophobic signal sequence made the product more prone to aggregation. In the cases where *E. coli* did not express properly large or multi-domain proteins we decided to select and clone a smaller fragment of the target protein to improve its expression levels. Moreover, the solubility of difficult targets , was also improved in some cases, by selecting a more hydrophilic and likely stable domain for expression and purification trials.

The expression vectors used are predominantly pET vectors (Novagen), although we also tested pGEX vectors (GE Healthcare) and pMal vectors (New England Biolabs) carrying multiple different fusion tags (hexahistidine, glutathione-S-transferase and maltose-binding-protein respectively). The benefit of one tag over another often resulted protein specific. The vectors used in this project generally contain a low copy origin of replication (pBR322) and are under the control of an inducible promoter (T7 RNA polymerase promoter) with tight control for high expression as well as plasmid stability.

To limit as much as possible the poorly reproducible and time-consuming procedures of traditional cloning methods, we recently decided to adopt with success a new strategy using the TOPO-cloning kit (Invitrogen). By this approach, the amplified and purified gene is inserted directly into the expression vector by a specific Topoisomerase, in a single step, and the resulting protein is expressed with an N-terminal hexahistidine tag and a tobacco etch virus (TEV) cleavage site, between the tag and the sequence of interest.

Prior to proceed with protein expression studies, all the constructs were confirmed by DNA sequencing.

Expression of protein in *Escherichia coli*

Once each gene was cloned into the selected vector, systematic tests to control and optimize the level of protein expression were performed. *Escherichia coli* remains the preferred host for heterologous expression of recombinant proteins for structural studies. It is the most characterized host for protein expression and has numerous advantages over other hosts (Makrides, 1996; Sørensen *et al*., 2005a and 2005b). These advantages include ease of handling, relatively inexpensive culture medium, availability of vectors and fast growth and expression. Furthermore, *E. coli* is the host of choice for metabolic labeling with selenomethionine for MAD phasing in X-ray crystallography (Goulding *et al*., 2003). Finally, in our case *E. coli*, like *H. pylori* is a gram-negative bacterium and it can be consider the best choice to express *H. pylori* proteins (Terpe, 2006). The ligated sequenced products were directly transformed into *E. coli* BL21(DE3) cells (Invitrogen), which were used for preliminary small-scale expression tests.

Expression studies were first carried out by growing cultures inoculated with single colonies in small scale using 10-100 ml of different media (LB or TB) and taking aliquots of the cultures prior to and following isopropyl-β-D-thiogalactoside (IPTG) induced growth. Protein production was induced with 0.1-1 mM IPTG and continued for 3–5 h at 30 °C and 37 °C or overnight at 20 °C. Whole-cell lysates were prepared from cultured cells and analyzed by Coomassie-stained denaturing polyacrylamide gel electrophoresis (SDS-PAGE) and Western blot.

If the protein expression was too low or not detected, we tested other *E. coli* strains like BL21 (DE3) STAR (Invitrogen), that encodes a truncated RNase E (*rne*131) enzyme that lacks the ability to degrade mRNA, reulting in an increase in mRNA stability, and the CodonPlus (DE) RP/RIL (Stratagene) or BL21 (DE3) Rosetta (Novagen), engineered to contain plasmid that encodes for *tRNA* gene that recognise rare codons. A variety of additional cell lines have also been used. The BL21 Origami cells (Novagen) contain mutations in thioredoxin reductase (*trxB*) and glutathione reductase (*gor*) gene, to enhance disulfide bond formations. If the protein is toxic the use of BL21 (DE3) pLysS cell (Invitrogen) is suggested. This strain has been engineered to express T7 lisozyme in order to inhibit T7 RNA polymerase activity prior to induction thereby preventing premature expression of the protein guaranteeing/assuring a more tight control. The *E. coli* C41 and C43 strain (Avidis) were instead tested/used to produce membrane protein and the *E. coli* B834 cells (Invitrogen), that are methionine-auxotroph cells engineered, are used for Selenomethionine incorporation and subsequent crystallographic studies (Protein Expression and Purification Facility, EMBL: http://www.embl-heidelberg.de).

Solubility studies

Following the determination of expression conditions, an evaluation of the conditions which could guarantee the highest yield of soluble protein was carried out. In our experience, the protein solubility represents a major bottleneck for production of protein for X-ray crystallography. Our first step toward defining a suitable lysis buffer includes a strategy to reduce the additives in the solubilising buffer to the minimum necessary to keep each protein

folded and soluble (Georgiou and Valax, 1996). If the protein is poorly soluble we tested its solubility using several screening buffers changing the pH value, the salt concentration or adding detergents, reducing agents and cofactors. The cells are lysed in the presence of these buffers and centrifuged. The supernatant and pellet from the centrifugation are analysed by SDS-PAGE and Western Blot to compare solubility of the protein in the various buffers.
If inclusion bodies are present, conditions to decrease the rate of protein translation or enhance protein folding or denaturation and refolding protocols have been employed (Lilie *et al*., 1998; Middelberg, 2002; Swietnicki, 2006). On the other hand, if soluble protein partitioning is the issue, lysate buffer conditions should be explored. Improving the solubility of recombinant proteins in *E. coli* commonly involves changing some of the expression conditions (Baneyx and Mujacic, 2004). Factors such as temperature, changes in the *E. coli* expression strain, different parameters or induction conditions have all been examined. In particular, to decrease the amount of protein found in inclusion bodies we routinely lower the induction temperature at 20-25 °C. However, the improvements can be minimal and many proteins will remain insoluble or behave poorly even under these condition. When good expression levels of insoluble protein were obtained, refolding was attempted Following large-scale production with induction at 37°C for 4–5 h, the protein was purified from inclusion bodies using solubilising buffers supplemented with 8 M urea or 6 M Guanidine-HCl (Carrió and A. Villaverde 2002; Tsumoto *et al*., 2002; Jungbauer *et al*., 2004; Li *et al*., 2004; Oganesyan *et al*., 2005). We have tried purifying proteins under denaturing conditions, followed by renaturation, with limited success (see Chapter 5).

Protein Purification

Since every protein behaves in a unique manner, every purification usually requires constant assessment and modification based upon these peculiar features. Purification is generally approached once a relatively high yield of soluble protein has been achieved. Purification is begun in the lysis buffer found to solubilize the highest fraction of protein.
The cells are lysed by French press, sonication, or using lysis reagents. To increase the lysis efficiency and reduce the solution viscosity we prefer to add to lysis buffer Lysozime and DNase respectively. All lysis buffers also contain a protease inhibitor cocktail to suppress protein degradation. After the lysis cell debris is removed by centrifugation at 40.000 g for 30 min. Protein solubility was checked on SDS–PAGE and Western-blot by loading both the cell extract and the soluble fraction after centrifugation. The tags might help to isolate the recombinant fusion protein by affinity chromatography and discard the majority of the other *E. coli* protein. As the majority of our constructs are hexahistidine tagged, the first chromatographic step of purification is standard immobilized metal affinity chromatography (IMAC). After extensive washing the proteins are eluted applying an Imidazole gradient (0-0.5 M Imidazole).
Other than hexahistidine tag, other affinity tags used include the Maltose binding protein (MBP) and the Glutathione S-transferase (GST). The proteins that are expressed as MBP or GST fusions are purified first by amylose or glutathione affinity chromatography respectively. The MBP fusion is consider on of the most well-studied solubility factor, in fact it has been

shown to not only increase expression yield but also passively promote the folding of its fused partner, thereby significantly facilitating the production of soluble protein (Kapust *et al*., 1999). The ability of solubility fusion tags to produce soluble and pure proteins is only the first step in the pathway towards protein production. The affinity tag, in fact, despite its positive features, sometimes may interfere with folding, oligomerization, protein localization, especially with membrane protein, and protein crystallization. As many of these tags are large proteins it is often necessary to remove the tags. For these reasons, the vectors we generally used contain all affinity tags with either a protease site between the tag and the construct of interest. The best one is the TEV protease from the Tobacco Etch Virus which shows high specificity, is relatively easy to make in large amounts/quantities and cleaves in most case leaving a native N-terminus. In particular, we use a recombinant hexahistidine-tagged Tev protease, which aids in the purification of the protease as well as the removal of the protease a new affinity chromatography after digestion. We have also found Thrombin to be a robust protease, in most cases, although all reactions have required some optimization for high yields of cleaved product. We have found less than optimal results with Factor Xa and Enterokinase due to an apparent narrow range of conditions for activity and some unspecific proteolysis. Sometimes, removal of solubility partner causes them insolubility and precipitation. Possible solutions to this problem include the use of additives, such as detergents r molecular chaperones. The use of MBP or GST tags also require a next step after purification of the fusion proteins to remove the MBP or GST after cleavage. To separate our protein from these tag a further purification by gel filtration or Ion-exchange chromatography is included/necessary. A problem we frequently met producing recombinant proteins as MBP or GST fusion is the their aggregation or decreased solubility after the tag cleavage. If the protein is not pure enough after affinity chromatography, the protein is concentrated by ultrafiltration and loaded onto a gel filtration columns, depending on the protein size. The third step of purification is generally an ion exchange chromatography (anion or cation) or a hydrophobic interaction chromatography (Ahrer et al., 2006).

Generally, we proceed with characterization if the protein is >90% pure. Proteins are then concentrated using to 5-20 mg ml^{-1} as a starting concentration for crystallization trials. The crystallization results are then used as a guide to suggest optimal concentrations.

Protein Characterization

All purified proteins are characterized by UV/VIS spectroscopy (UV/VIS), circular dichroism (CD) (Kelly *et al*., 2005), analytical gel filtration (HPLC), dynamic light scattering (DLS) and mass spectroscopy (MS) before crystallization trials are attempted. The CD (Jasco 720 Spectrophotometer, Jasco Inc.) has been/used is used to ensure that the protein product is folded estimating the content of secondary structures. (software program CDNN version 2.1; Bohm *et al*., 1992). The mass spectroscopy (MS) data are instead used to determine the precise protein molecular weightand and to determine the presence of degraded protein. The MS can also confirm the presence of Se-methionine substitution.

Analytical gel filtration (HPLC) and DLS are instead used to evaluate the homogeneity and eventual oligomeric form of the samples. Other characterization studies that are often

employed include: UV/VIS spectroscopy, Fluorescence Spectroscopy, Western Blot, native gel electrophoresis and, if an enzymatic activity is present, functional assays are performed.

Protein Crystallization

If the protein sample presents high degree of purity and homogeneity, it is submitted to crystallization trials. After the solubility, the crystallization of proteins is considered one of the most serious bottlenecks for structure determination by X-ray crystallography (Dale *et al*., 2003). Optimal conditions for crystallization are very difficult to predict. An effective way to search through a large number of variables (pH, precipitant, salt, protein concentration, additives, buffer, temperature, detergents, organics, etc.) that may affect crystal growth is to conduct a sparse matrix search of specific crystallization conditions (DeLucas *et al*., 2003). Generally, our first approach begins with crystallization trials using commercial crystal screening kits (Structure Screen I and II, Clear Strategy Screen I and II from Molecular Dimension; the PACT suite, MbClass I and II from Qiagen; Crystal Screen I and II, Membfac and PEG/Ion Screen from Hampton).

We are currently using both hanging drop and sitting drop vapour-diffusion methods, which rely on the exchange of either water and/or some volatile agents between a micro-drop of mother liquor, containing the protein, and much larger reservoir solution. Protein crystals are very often thin and small (microns to millimetres, in a few lucky events) and rich in solvent content, between 30 to 70 %. Therefore, it is often a great advantage to have access to high brilliance synchrotron sources to collect X-ray data with good quality and high resolution. Often, the initial screens are as much a test of protein concentration as of crystallization conditions. If approximately half of the conditions yield a precipitate, we consider the protein concentration to be appropriate for further trials. If not, the concentration of the protein is adjusted accordingly or the buffer composition modified.

The crystallization experiments are performed manually and also using an Oryx8 crystallization robot (Douglas Instruments). In a typical experiment, 0.1 µl screening solution was added to 0.1 µl protein solution on 96-well crystal plates (Douglas Instruments); the reservoir wells contained 100 µl screening solution. The use of nanodroplet high-throughput crystallization allows for numerous conditions to be sampled while keeping the requirements of protein quantities to a minimum. Conditions that yielded crystals were optimized with manually prepared condition using 24-wells Linbro plates and glass cover slides (Hampton Research). Drops containing equal volumes of protein and reservoir solution (1-4 ul of each) were allowed to equilibrate against 400-500 µl reservoir solution. Manual inspection of crystallisation experiment is carried under a microscope.

Data Collection and Structure Determination

The synchrotron radiation source is available (ESRF, Grenoble and ELETTRA, Trieste), and it allows the use of any suitable wavelength in the spectral range selected (usually between 0.9 and 2 Å with proteins). This property is particularly relevant in multiple wavelength anomalous dispersion (MAD) technique and when wavelengths shorter than 1 Å are necessary. Since all the proteins studied in this work do not present structural homologues,

two strategies were used to determine their 3D-structures: MAD (Multiple Anomalous Dispersion) and MIR/MIRAS, SIR/SIRAS (Single and Multiple Isomorphous Replecement).
In the last years, crystallographers are performing more and more MAD experiments, especially around the K-edge of Selenium (0.98 Å), using crystals of Se-Methionine derivatives of the protein of interest. The 3D structure of the proteins CagD (Chapter 4), CagS (Cendron *et al*., 2007; Chapter 3) and CagZ have been solved with this technique. Another method for solving the phase problem is the single or multiple isomorphous replacement (SIR and MIR). This require the soaking of crystals of the native protein into the crystallization mother liquor containing heavy atoms, which might eventually become coordinated to the proteins inside the crystal, thus affecting the diffracted intensities. This kind of method has been used with one of the proteins studied in this thesis, CagV (Chapter 5).

References

Ahrer K., Jungbauer A. (2006). Chromatographic and electrophoretic characterization of protein variants. J. *Chromatogr. B Analyt Technol. Biomed. Life Sci.* **841**, 110-22

Alm R.A., Ling L.-S.L., Moir D.T., King B.L., Brown E.D., Doig P.C., Smith D.R., Noonan B., Guild B.C., deJonge B.L., Carmel G., Tummino P.J., Caruso A., Uria-Nickelsen M., Mills D.M., Ives C., Gibson R., Merberg D., Mills S.D., Jiang Q., Taylor D.E., Vovis G.F., Trust T.J. (1999). Genomic-sequence comparison of two unrelated isolates of the human gastric pathogen *Helicobacter pylori*. *Nature* **397**, 176-80.

Altschul S.F., Madden T.L., Schaffer A.A., Zhang J., Zhang Z., Miller W., Lipman DJ. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–402.

Akopyants N., Clifton S. W., Kersulyte D., Crabtree J. E., Youree B. E., Reece C. A., Bukanov N. O., Drazek E. S., Roe B. A., Berg D. E. (1998). Analyses of the *cag* pathogenicity island of *Helicobacter pylori*. *Mol. Microbiol.* **28**, 37-53.

Amieva M.R., Vogelmann R., Covacci A., Tompkins L.S., Nelson W.J., Falkow S. (2003). Disruption of the epithelial apical-junctional complex by *Helicobacter pylori* CagA. *Science* **300**, 1430–4

Andrzejewska J., Kyung Lee S., Olbermann P., Lotzing N., Katzowitsch E., Linz B., Achtman M., Kado C. I., Suerbaum S., Josenhans C. (2006). Characterization of the Pilin ortholog of the *Helicobacter pylori* type IV *cag* pathogenicity apparatus, a surface-associated protein expressed during infection. *J. Bateriol.* **188**, 5865-77.

Aras R.A., Fischer W., Perez-Perez G.I., Crosatti M., Ando T., Haas R., Blaser M.J. (2003). Plasticity of repetitive DNA sequences within a bacterial type IV secretion system component. *J. Exp. Med.* **198**, 1349–60.

Asahi M., Azuma T., Ito S., Ito Y., Suto H., Nagai Y., Tsubokawa M., Tohyama Y., Maeda S., Omata M., Suzuki T., Sasakawa C. (2000). *Helicobacter pylori* CagA protein can be tyrosine phosphorylated in gastric epithelial cells. *J. Exp. Med.* **191**, 593–602.

Aspholm M., Olfat F.O., Nordén J., Sondén B., Lundberg C., Sjöström R., Altraja S., Odenbreit S., Haas R., Wadström T., Engstrand L., Semino-Mora C., Liu H., Dubois A., Teneberg S., Arnqvist A., Borén T. (2006). SabA is the *H. pylori* hemagglutinin and is polymorphic in binding to sialylated glycans. *PLoS Pathog*. **2**, 989-1001.

Aspholm-Hurtig M., Dailide G., Lahmann M., Kalia A., Ilver D., Roche N., Vikström S., Sjöström R., Lindén S., Bäckström A., Lundberg C., Arnqvist A., Mahdavi J., Nilsson U.J., Velapatiño B., Gilman R.H., Gerhard M., Alarcon T., López-Brea M., Nakazawa T., Fox J.G., Correa P., Dominguez-Bello M.G., Perez-Perez G.I., Blaser M.J., Normark S., Carlstedt I., Oscarson S., Teneberg S., Berg D.E., Borén T. (2004). Functional adaptation of BabA, the H. pylori ABO blood group antigen binding adhesin. *Science* **305**, 519–22.

Atherton J.C., Cao P., Peek R.M. Jr, Tummuru M.K., Blaser M.J., Cover T.L. (1995). Mosaicism in vacuolating cytotoxin alleles of *Helicobacter pylori*. Association of specific *vacA* types with cytotoxin production and peptic ulceration. *J. Biol. Chem.* **270**, 17771–7.

Backert S., Ziska E., Brinkmann V., Zimny-Arndt U., Fauconnier A., Jungblut P.R., Naumann M., Meyer T.F. (2000). Translocation of the *Helicobacter pylori* CagA protein in gastric epithelial cells by a type IV secretion apparatus. *Cell. Microbiol.* **2**, 155–64.

Backert S., Moese S., Selbach M., Brinkmann V., Meyer T.F. (2001). Phosphorylation of tyrosine 972 of the *Helicobacter pylori* CagA protein is essential for induction of a scattering phenotype in gastric epithelial cells. *Mol. Microbiol.* **42**, 631–44.

Backert S., Muller E.C., Jungblut P.R., Meyer T.F. (2001) Tyrosine phosphorylation patterns and size modification of the *Helicobacter pylori* CagA protein after translocation into gastric epithelial cells. *Proteomics* **1**, 608-17.

Backert S., Schwarz T., Miehlke S., Kirsch C., Sommer C., Kwok T., Gerhard M., Goebel U.B., Lehn N., Koenig W., Meyer T.F. (2004). Functional analysis of the *cag* pathogenicity island in *Helicobacter pylori* isolates from patients with gastritis, peptic ulcer, and gastric cancer. *Infect. Immunol*. **72**, 1043-56.

Backert S., Kwok T., Schmid M., Selbach M., Moese S., Peek R.M. Jr, Konig W., Meyer T.F., Jungblut P. R. (2005). Subproteomes of soluble and structure-bound *Helicobacter pylori* proteins analyzed by two-dimensional gel electrophoresis and mass spectrometry. *Proteomics* **5**,1331-45.

Backert S., Selbach M. (2005). Tyrosine-phosphorylated bacterial effector proteins: the enemies within. *Trends Microbiol*. **10**, 476-84.

Backert S., Meyer T.F. (2006). Type IV secretion systems and their effectors in bacterial pathogenesis. *Curr. Opin. Microbiol.* **9**, 207-17.

Backstrom A., Lundberg C., Kersulyte D., Berg D.E., Boren T., Arnqvist A. (2004). Metastability of *Helicobacter pylori* bab adhesin genes and dynamics in Lewis b antigen binding. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 16923-28.

Badger J., Sauder J.M., Adams J.M., Antonysamy S., Bain K., Bergseid M.G., Buchanan S.G., Buchanan M.D., Batiyenko Y., Christopher J.A., Emtage S., Eroshkina A., Feil I., Furlong E.B., Gajiwala K.S., Gao X., He D., Hendle J., Huber A., Hoda K., Kearins P., Kissinger C., Laubert B., Lewis H.A., Lin J., Loomis K., Lorimer D., Louie G., Maletic M., Marsh C.D., Miller I., Molinari J., Muller-Dieckmann H.J., Newman J.M., Noland B.W., Pagarigan B., Park F., Peat T.S., Post K.W., Radojicic S., Ramos A., Romero R., Rutter M.E., Sanderson W.E., Schwinn K.D., Tresser J., Winhoven J., Wright T.A., Wu L., Xu J., Harris T.J. (2005) Structural analysis of a set of proteins resulting from a bacterial genomics project. *Proteins* **60**, 787-96.

Baek H.Y., Lim J.W., Kim H. (2007). Interaction between the *Helicobacter pylori* CagA and alpha-Pix in gastric epithelial AGS cells. *Ann. N. Y. Acad. Sci.* **1096**, 18-23.

Bagnoli F., Buti L., Tompkins L., Covacci A., Amieva M. R. (2005). *Helicobacter pylori* CagA induces a transition from polarized to invasive phenotypes in MDCK cells. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 16339-44.

Bailey S., Ward D., Middleton R., Grossmann J.G., Zambryski P.C. (2006). *Agrobacterium tumefaciens* VirB8 structure reveals potential protein-protein interaction sites. *Proc. Natl. Acad. Sci. U.S.A.* **21**, 2582-7.

Baneyx F., Mujacic M. (2004). Recombinant protein folding and misfolding in *Escherichia coli*. *Nat. Biotechnol.* **22**, 1399-408.

Baron C., Llosa M., Zhou S., Zambryski P. C. (1997). VirB1, a component of the T-complex transfer machinery of *Agrobacterium tumefaciens*, is processed to a C-terminal secreted product, VirB1*. *J. Bacteriol.* **179**, 1203–10.

Baron C. (2006). VirB8: a conserved type IV secretion system assembly factor and drug target. *Biochem. Cell Biol.* **84**, 890-9.

Bayliss R., Leung S.W., Baker R.P., Quimby B.B., Corbett A.H., Stewart M. (2002). Structural basis for the interaction between NTF2 and nucleoporin FxFG repeats. *EMBO J.* **21**,2843-53.

Bayliss R., Harris R., Coutte L., Monier A., Fronzes R., Christie P.J., Driscoll P.C., Waksman G. (2007). NMR structure of a complex between the VirB9/VirB7 interaction domains of the pKM101 type IV secretion system. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 1673-8

Bebb J.R., Letley D.P., Thomas R.J., Aviles F., Collins H.M., Watson S.A., Hand N.M., Zaitoun A., Atherton J.C. (2003). *Helicobacter pylori* upregulates matrilysin (MMP-7) in epithelial cells *in vivo* and *in vitro* in a Cag dependent manner. *Gut.* **52**,:1408-13.

Bendtsen J. D., Nielsen H., von Heijne G., Brunak S. (2004). Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* **340**, 783-95.

Benning M.M., Wesenberg G., Liu R., Taylor K.L., Dunaway-Mariano D., Holden H.M. (1998). The three-dimensional structure of 4-hydroxybenzoyl-CoA thioesterase from *Pseudomonas sp*. Strain CBS-3. *J. Biol. Chem.* **273**, 33572-79.

Bhasin M., Garg A., Raghava G.P. (2005). PSLpred: prediction of subcellular localization of bacterial proteins. *Bioinformatics* **21**, 2522-4.

Bijlsma J.J., Waidner B., Vliet A.H., Hughes N.J., Hag S., Bereswill S., Kelly D.J., Vandenbroucke-Grauls C.M., Kist M., Kusters J.G. (2002). The *Helicobacter pylori* homologue of the ferric uptake regulator is involved in acid resistance. *Infect. Immunol.* **70**, 606-11.

Bizzozero G. (1893). Uber die schlauchformingen Drusen des Magendarmkanals und die Beziehung ihres Epithels zu dem Oberflachenepithel de Schleimhaut. *Arch. Mikr. Anat.* **42**, 82-152.

Bjorkholm B., Lundin A., Sillén A., Guillemin K., Salama N., Rubio C., Gordon J.I., Falk P., Engstrand L. (2001). Comparison of genetic divergence and fitness between two subclones of *Helicobacter pylori*. *Infect. Immun.* **69**, 7832-38.

Blaser M. J. (1998). *Helicobacter pylori* and gastric diseases. *Br. Med. J.* **316**, 1507-10.

Blaser, M. J. (2001). Heterogeneity of *Helicobacter pilori* cag genotypes in experimentally infected mice. *FEMS Microbiol. Lett.* **203**, 109-14.

Blaser M. J., Atherton J. C. (2004). *Helicobacter pylori* persistence: biology and disease. *J. Clin. Invest.* **113**, 321-33.

Blume-Jensen P., Hunter T. (2001). Oncogenic kinase signalling. *Nature* **411**, 355-65.

Boncristiano M., Paccani S.R., Barone S., Ulivieri C., Patrussi L., Ilver D., Amedei A., D'Elios M.M., Telford J.L., Baldari C.T. (2003). The *Helicobacter pylori* vacuolating toxin inhibits T cell activation by two independent mechanisms. *J. Exp. Med.* **198**, 1887-97.

Boren T., Falk P., Roth K.A., Larson G., Normark S. (1993). Attachment of *Helicobacter pylori* to human gastric epithelium mediated by blood group antigens. *Science* **262**, 1892-5.

Bourhis L.L., Werts C. (2007). Role of Nods in bacterial infection. *Microbes Infect.* **9**, 629-36.

Bourzac K. M., Botham C.M., Guillemin K. (2007). *Helicobacter pylori* CagA induces AGS cell elongation through a cell retraction defect that is independent of Cdc42, Rac1, and Arp2/3. *Infection and Immunity* **75**, 1203-13.

Bourzac K. M., Guillemin K. (2005). *Helicobacter pylori*-host cell interactions mediated by type IV secretion. *Cell. Microbiol.* **7**, 911-9.

Bourzac K.M., Satkamp L.A., Guillemin K. (2006). The *Helicobacter pylori* cag pathogenicity island protein CagN is a bacterial membrane-associated protein that is processed at its C-terminus. *Infection and Immunity* **74**, 2537-43.

Brandt S., Kwok T., Hartig R., König W., Backert S. (2005). NF-kB activation and potentiation of proinflammatory responses by the *Helicobacter pylori* CagA protein. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 9300-5.

Brandt S., Shafikhani S., Balachandran P., Jin S., Hartig R., Konig W., Engel J., Backert S. (2007). Use of a novel coinfection system reveals a role for Rac1, H-Ras, and CrkII phosphorylation in *Helicobacter pylori*-induced host cell actin cytoskeletal rearrangements. *FEMS Immunol. Med. Microbiol.* **50**, 190-205.

Bretscher A., Edwards K., Fehon R.G. (2002). ERM proteins and merlin: integrators at the cell cortex. *Nat. Rev. Mol. Cell. Biol.* **3**, 586-99.

Brünger A.T., Adams P.D., Clore G.M., DeLano W.L., Gros P., Grosse Kunstleve R.W., Jiang J.S., Kuszewski J., Nilges M., Pannu N.S., Read R.J., Rice L.M., Simonson T., Warren G.L. (1998). Crystallography and NMR system: A new software suite for macromolecular structure determination. *Acta Crystallogr*. **D54**, 905-21.

Buhrdorf R., Forster C., Haas R., Fischer W. (2003). Topological analysis of a putative *virB8* homologue essential for the *cag* type IV secretion system in *Helicobacter pylori*. *Int. J. Med. Microbiol*. **293**, 213-7.

Bury-Mone S., Thiberge J.M., Contreras M., Maitournam A., Labigne A., De Reuse H. (2004). Responsiveness to acidity via metal ion regulators mediates virulence in the gastric pathogen *Helicobacter pylori*. *Mol. Microbiol*. **53**, 623-38.

Busler V.J., Torres V.J., McClain M.S., Tirado O., Friedman D.B., Cover T.L. (2006). Protein-protein interactions among *Helicobacter pylori* cag proteins. *J. Bacteriol.* **188**, 4787-800.

Cao T.B., Saier M.H. Jr. (2001). Conjugal type IV macromolecular transfer systems of Gram-negative bacteria: organismal distribution, structural constraints and evolutionary conclusions. *Microbiology* **147**, 3201-14.

Carrió M.M., Villaverde A. (2002). Construction and deconstruction of bacterial inclusion bodies. *J. Biotechnol.* **96**, 3-12.

Cascales E., Bernadac A., Gavioli M., Lazzaroni J.C., Lloubes R. (2002) Pal lipoprotein of *Escherichia coli* plays a major role in outer membrane integrity. *J. Bacteriol.* **184**, 754-9.

Cascales E., Christie P.J. (2003). The versatile bacterial type IV secretion systems. *Nat. Rev. Microbiol.* **1**, 137-49.

Cendron L., Seydel A., Angelini A., Battistutta R., Zanotti G. (2004). Crystal structure of CagZ, a protein from the *Helicobacter pylori* pathogenicity island that encodes for a type IV secretion system. *J. Mol. Biol.* **340**, 881-9.

Cendron L., Tasca E., Seraglio T., Seydel A., Angelini A., Battistutta R., Montecucco C., Zanotti G. (2007). The crystal structure of CagS from the *Helicobacter pylori* pathogenicity island. *Proteins* **69**, 440-3.

Censini S., Lange C., Xiang Z., Crabtree J. E., Ghiara P., Borodovsky M., Rappuoli R., Covacci A. (1996). *cag*, a pathogenicity island of *Helicobacter pylori*, encodes type I-specific and disease-associated virulence factors. *Proc. Natl. Acad. Sci. U.S.A.* **93**, 14648-53.

Chang Y.J., Wu M.S., Lin J.T., Pestell R.G., Blaser M.J., Chen C.C. (2006). Mechanisms for *Helicobacter pylori* CagA-induced cyclin D1 expression that affect cell cycle. *Cell. Microbiol.* **8**, 1740-52.

Cheng Z., Song F., Shan X., Wei Z., Wang Y., Dunaway-Mariano D., Gong W. (2006). Crystal structure of human thioesterase superfamily member 2. *Biochem. Biophys. Res. Commun.* **349**, 172-7.

Chin K.H., Chou C.C., Wang A.H., Chou S.H. (2006). Crystal structure of a putative acyl-CoA thioesterase from *Xanthomonas campestris* (XC229) adopts a tetrameric hotdog fold of epsilongamma mode. *Proteins* **64**, 823-6.

Christie P.J., Atmakuri K., Krishnamoorthy V., Jakubowski S., Cascales E. (2005). Biogenesis, architecture, and function of bacterial type IV secretion systems. *Annu. Rev. Microbiol.* **59**, 451-85.

Churin Y., Al-Ghoul L., Kepp O., Meyer T.F., Birchmeier W., Naumann M. (2003). *Helicobacter pylori* CagA protein targets the c-Met receptor and enhances themotogenic response. *J. Cell Biol.* **161**, 249-55

Clifton D.R., Fields K.A., Grieshaber S.S., Dooley C.A., Fischer E.R., Mead D.J., Carabeo R.A., Hackstadt T. (2004). A chlamydial type III translocated protein is tyrosine-phosphorylated at the site of entry and associated with recruitment of actin. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 10166-71.

Clyne M., Labigne A., Drumm B. (1995). *Helicobacter pylori* requires an acidic environment to survive in the presence of urea. *Infect. Immunol*. **63**, 1669-73.

Clyne M., Dolan B., Reeves E.P. (2007). Bacterial factors thatmediate colonization of the stomach and virulence of *Helicobacter pylori*. *FEMS Microbiol. Lett.* **268**, 135-43.

Collaborative Computing Project 4. (1994). The CCP4 suite: programs for protein crystallography. *Acta Crystallogr.* **D50**, 760-3.

Contreras M., Thiberge J.M., Mandrand-Berthelot M.A., Labigne A. (2003) Characterization of the roles of NikR, a nickel-responsive pleiotropic autoregulator of *Helicobacter pylori*. *Mol. Microbiol*. **49**, 947-63.

Couturier M.R., Tasca E., Montecucco C., Stein M. (2006). Interaction with CagF is required for translocation of CagA into the host via the *Helicobacter pylori* type IV secretion system. *Infect. Immun.* 74, 273-81.

Covacci A., Censini S., Bugnoli M., Petracca R., Burroni D., Macchia G., Massone A., Papini E., Xiang Z., Figura N. (1993). Molecular characterization of the 128-kDa immunodominant antigen of *Helicobacter pylori* associated with cytotoxicity and duodenal ulcer. *Proc. Nat. Acad. Sci. U.S.A.* **90**, 5791-5.

Covacci A., Falkow S., Berg D.E., Rappuoli R. (1997). Did the inheritance of a pathogenicity island modify the virulence of *Helicobacter pylori* ? *Trends Microbiol*. **5**, 205-8.

Covacci A., Rappuoli R. (1998). *Helicobacter pylori*: molecular evolution of a bacterial quasi-species. *Curr. Opin. Microbiol.* **1**, 96-102.

Covacci A., Telford J.L., Del Giudice G., Parsonnet J., Rappuoli R. (1999). *Helicobacter pylori* virulence and genetic geography. *Science* **284**, 1328-33.

Covacci A., Rappuoli R. (2000). Tyrosine-phosphorylated bacterial proteins: Trojan horses for the host cell. *J. Exp. Med.* **191**, 587-92.

Cover T.L., Blaser M.J. (1992). Purification and characterization of the vacuolating toxin from *Helicobacter pylori*. *J. Biol. Chem.* **267**, 10570-5.

Cover T.L., Hanson P.I., Heuser J.E. (1997). Acid-induced dissociation of VacA, the *Helicobacter pylori* vacuolating cytotoxin, reveals its pattern of assembly. *J. Cell Biol.* **138**, 759-69.

Cover T.L., Blanke S.R. (2005). *Helicobacter pylori* VacA, a paradigm for toxin multifunctionality. *Nat. Rev. Microbiol.* **3**, 320-2.

Crabtree J.E., Xiang Z., Lindley I.J., Tompkins D.S., Rappuoli R., Covacci A. (1995). Induction of interleukin-8 secretion from gastric epithelial cells by a cagA negative isogenic mutant of *Helicobacter pylori*. *J. Clin. Pathol.* **48**, 967-9.

Crawford H.C., Krishna U.S., Israel D.A., Matrisian L.M., Washington M.K., Peek R.M. Jr. (2003). *Helicobacter pylori* strain-selective induction of matrix metalloproteinase-7 *in vitro* and within gastric mucosa. *Gastroenterology* **125**, 1125-36.

Croxen M.A., Sisson G., Melano R., Hoffman P.S. (2006). The *Helicobacter pylori* chemotaxis receptor TlpB (HP0103) is required for pH taxis and for colonization of the gastric mucosa. *J. Bacteriol.* 188, 2656-65.

Dale G.E., Oefner C., D'Arcy A. (2003). The protein as a variable in protein crystallization. *Journal of Structural Biology* **142**, 88–97.

Daly R.J. (2004) Cortactin signalling and dynamic actin networks. *Biochem. J.* **382**, 13–25.

Danese S., Papa A., Gasbarrini A., Ricci R., Maggiano N. (2004). *Helicobacter pylori* eradication down-regulates matrix metalloproteinase-9 expression in chronic gastritis and gastric ulcer. *Gastroenterology* **126**, 369-71.

Das A., Xie Y. H. (1998). Construction of transposon Tn3phoA: its application in defining the membrane topology of the *Agrobacterium tumefaciens* DNA transfer proteins. *Mol. Microbiol.* **27**, 405-14.

Das A., Xie Y. H. (2000). The *Agrobacterium* T-DNA transport pore proteins VirB8, VirB9, and VirB10 interact with one another. *J. Bacteriol.* **182**, 758-63.

Davies G.R., Simmonds N.J., Stevens T.R., Sheaff M.T., Banatvala N., Laurenson I.F., Blake D.R., Rampton D.S. (1994). *Helicobacter pylori* stimulates antral mucosal reactive oxygen metabolite production *in vivo*. *Gut*. **35**, 179-85.

De Bernard M., Burroni D., Papini E., Rappuoli R., Telford J., Montecucco C. (1998). Identification of the *Helicobacter pylori* VacA toxin domain active in the cell cytosol. *Infect. Immunol.* **66**, 6014-6.

De Lucas L.J., Bray T.L., Nagy L., McCombs D., Chernov N., Hamrick D., Cosenza L., Belgovskiy A., Stoops B., Chait A. (2003). Efficient protein crystallization. *Journal of Structural Biology* **142**, 188-206.

Del Giudice G., Covacci A., Telford J.L., Montecucco C., Rappuoli R. (2001). The design of vaccines against *Helicobacter pylori* and their development. *Annu. Rev. Immunol*. **19**, 523-63.

DeLano W.L. (2002). http://pymol.sourceforge.net.

Delany I., Spohn G., Rappuoli R., Scarlato V. (2001a). The Fur repressor controls transcription of iron-activated and -repressed genes in *Helicobacter pylori*. *Mol. Microbiol.* **42**, 1297-309.

Delany I., Pacheco A.B.F., Spohn G., Rappuoli R., Scarlato V. (2001b). Iron-dependent transcription of the *frpB* gene of *Helicobacter pylori* is controlled by the Fur repressor protein. *J. Bacteriol.* **183**, 4932-7

Deibel C., Krämer S., Chakraborty T., Ebel F. (1998). EspE, a novel secreted protein of attaching and effacing bacteria, is directly translocated into infected host cells, where it appears as a tyrosine-phosphorylated 90 kDa protein. *Mol. Microbiol.* **28**, 463-74.

Deindl S., Kadlecek T.A., Brdicka T., Cao X., Weiss A., Kuriyan J. (2007). Structural Basis for the Inhibition of Tyrosine Kinase Activity of ZAP-70. *Cell* **129**, 735-46.

Di Nitto J.P., Lambright D.G. (2006). Membrane and juxtamembrane targeting by PH and PTB domains. *Biochim. Biophys. Acta.* **1761**, 850-67.

Dian C., Schauer K., Kapp U., McSweeney S.M., Labigne A., Terradot L. (2006). Structural basis of the nickel response in *Helicobacter pylori*: crystal structures of *Hp*NikR in Apo and nickel-bound states. *J. Mol. Biol.* **361**, 715-30.

Dillon S.C., Bateman A. (2004). The Hotdog fold: wrapping up a superfamily of thioesterases and dehydratases. *BMC Bioinformatics* **5**, 109.

Dorrell N., Martino M.C., Stabler R.A., Ward S.J., Zhang Z.W., McColm A.A., Farthing M.J., Wren B.W. (1999). Characterization of *Helicobacter pylori* PldA, a phospholipase with a role in colonization of the gastric mucosa. *Gastroenterology* **117**, 1098-104.

Dosanjh N.S., Michel S.L. (2006). Microbial nickel metalloregulation: NikRs for nickel ions. *Curr. Opin. Chem. Biol.* **10**, 123-30.

Dossumbekova A., Prinz C., Mages J., Lang R., Kusters J.G., Van Vliet A.H., Reindl W., Backert S., Saur D., Schmid R.M., Rad R. (2006). *Helicobacter pylori* HopH (OipA) and bacterial pathogenicity: genetic and functional genomic analysis of *hopH* gene polymorphisms. *J. Infect. Dis.* **194**, 1346-55.

Dundon W.G., De Bernard M., Montecucco C. (2001). Virulence factors of *Helicobacter pylori*. *Int. J. Med. Microbiol.* **290**, 647-58.

Dunn B.E., Phadnis S.H. (1998). Structure, function and localization of *Helicobacter pylori* urease. *J. Biol. Med.* **71**, 63-73.

Eaton K.A., Brooks C.L., Morgan D.R., Krakowka S. (1991). Essential role of urease in pathogenesis of gastritis induced by *Helicobacter pylori* in gnotobiotic piglets. *Infect. Immun.* **59**, 2470-5.

Eaton K.A., Morgan D.R., Krakowka S. (1992). Motility as a factor in the colonisation of gnotobiotic piglets by *Helicobacter pylori*. *J. Med. Microbiol.* **37**, 123-7.

Eaton K. A., Suerbaum S., Josenhans C., Krakowka S. (1996). Colonization of gnotobiotic piglets by *Helicobacter pylori* deficient in two flagellin genes. *Infect. Immun.* **64**, 2445-8.

Eckhart W., Hutchinson M.A., Hunter T. (1979). An activity phosphorylating tyrosine in polyoma T antigen immunoprecipitates. *Cell* **18**, 925-33.

Egeblad M., Werb Z. (2002). New functions for the matrix metalloproteinases in cancer progression. *Nat. Rev. Cancer* **2**, 161-74.

Emsley P., Cowtan K. (2004). Coot: model-building tools for molecular graphics. *Acta D Biol. Crystallogr.* **60**, 2126-32.

Esposito D., Chatterjee D.K. (2006). Enhancement of soluble protein expression through the use of fusion tags. *Current Opinion in Biotechnology* **17**, 1-6.

Evans D.J. Jr, Evans D.G., Takemura T., Nakano H., Lampert H.C., Graham D.Y., Granger D.N., Kvietys P.R. (1995). Characterization of a *Helicobacter pylori* neutrophil-activating protein. *Infect. Immun.* **63**, 2213-20.

Evans D.J. Jr, Queiroz D.M.M., Mendes E.N., Evans D.G. (1998). Diversity in the variable region of *Helicobacter pylori cagA* gene involves more than simple repetition of a 102-nucleotide sequence. *Biochem. Biophys. Res. Commun.* **245**, 780-4.

Evans D.J., Evans D.G. (2001). *Helicobacter pylori* CagA: Analysis of sequence diversity in relation to phosphorylation motifs and implications for the role of CagA as a virulence factor. *Helicobacter* **6**, 187-98.

Figura N., Vindigni C., Covacci A., Presenti L., Burroni D., Vernillo R., Banducci T., Roviello F., Marrelli D., Biscontri M., Kristodhullu S., Gennari C., Vaira D. (1998). *cagA* positive and negative *Helicobacter pylori* strains are simultaneously present in the stomach of most patients with nonulcer dyspepsia: relevance to histological damage. *Gut*. **42**, 772-8.

Finlay B.B. (2005). Bacterial virulence strategies that utilize Rho GTPases. *Curr. Top. Microbiol. Immunol.* **291**, 1-10.

Fischer W., Puls J., Buhrdorf R., Gebert B., Odenbreit S., Haas R. (2001). Systematic mutagenesis of the *Helicobacter pylori* cag pathogenicity island: essential genes for CagA translocation in host cells and induction of interleukin-8. *Mol. Microbiol.* **42**, 1337-48.

Forsyth M.H., Cao P., Garcia P.P., Hall J.D., Cover T.L. (2002). Genome-wide transcriptional profiling in a histidine kinase mutant of *Helicobacter pylori* identifies members of a regulon. *J. Bacteriol.* 184, 4630-5.

Franco A.T., Israel D.A., Washington M.K., Krishna U., Fox J.G., Rogers A.B., Neish A.S., Collier-Hyams L., Perez-Perez G.I., Hatakeyama M., Whitehead R., Gaus K., O'Brien D.P., Romero-Gallo

J., Peek R.M. Jr. (2005). Activation of β-catenin by carcinogenic *Helicobacter pylori*. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 10646-51.

Fritz J.H., Ferrero R.L., Philpott D.J., Girardin S.E. (2006). Nod-like proteins in immunity, inflammation and disease. *Nat. Immunol.* **7**, 1250-7.

Galmiche A., Rassow J., Doye A., Cagnol S., Chambard J.C., Contamin S., de Thillot V., Just I., Ricci V., Solcia E., Van Obberghen E., Boquet P. (2000). The N-terminal 34 kDa fragment of *Helicobacter pylori* vacuolating cytotoxin targets mitochondria and induces cytochrome c release. *EMBO J.* **19**, 6361-70.

Gangwer K.A., Mushrush D.J., Stauff D.L., Spiller B., McClain M.S., Cover T.L., Lacy D.B. (2007). Crystal structure of the *Helicobacter pylori* vacuolating toxin p55 domain. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 16293-8.

Gardy J.L., Laird M.R., Chen F., Rey S., Walsh C.J., Ester M., Brinkman F.S. (2005). PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics* **21**, 617-23.

Gebauer D., Li J., Jogl G., Shen Y., Myszka D.G., Tong L. (2004). Crystal structure of the PH-BEACH domains of human LRBA/BGL. *Biochemistry* 43, 14873-80.

Gebert B., Fischer W., Weiss E., Hoffmann R., Haas R. (2003). *Helicobacter pylori* vacuolating cytotoxin inhibits T lymphocyte activation. *Science* **301**, 1099-102.

Geis G., Leying H., Suerbaum S., Mai U., Opferkuch W. (1989). Ultrastructure and chemical analysis of *Campylobacter pylori* flagella. *J. Clin. Microbiol.* **27**, 436-41.

Georgiou G., Valax P. (1996). Expression of correctly folded proteins in *Escherichia coli*. *Curr. Opin. Biotechnol.* **7**, 190-7.

Gerding M.A., Ogata Y., Pecora N.D., Niki H., de Boer P.A. (2007). The trans-envelope Tol-Pal complex is part of the cell division machinery and required for proper outer-membrane invagination during cell constriction in *E. coli*. *Mol. Microbiol.* **63,** 1008-25.

Gerhard M., Lehn N., Neumayer N., Boren T., Rad R., Schepp W., Miehlke S., Classen M., Prinz C. (1999). Clinical relevance of the *Helicobacter pylori* gene for blood-group antigen-binding adhesin. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 12778-83.

Christie P.J. (2004). Type IV secretion: the *Agrobacterium* VirB/D4 and related conjugation systems. *Biochim. Biophys. Acta* **1694**, 219-34.

Girardin S.E., Boneca I.G., Carneiro L.A., Antignac A., Jehanno M., Viala J., Tedin K., Taha M.K., Labigne A., Zahringer U., Coyle A.J., DiStefano P.S., Bertin J., Sansonetti P.J., Philpott D.J. (2003) Nod1 detects a unique muropeptide from gram-negative bacterial peptidoglycan. *Science* **300**, 1584-7.

Gomis-Ruth F.X., Moncalian G., Perez-Luque R., Gonzalez A., Cabezon E., de la Cruz F., Coll M. (2001). The bacterial conjugation protein TrwB resembles ring helicases and F1-ATPase. *Nature* **409**, 637-41.

Göõz M., Göõz P., Smolka A.J. (2001). Epithelial and bacterial metalloproteinases and their inhibitors in *H. pylori* infection of human gastric cells. *Am. J. Physiol. Gastrointest. Liver Physiol.* **281**, 823-32.

Gouin E., Welch M. D., Cossart P. (2005). Actin-based motility of intracellular pathogens. *Curr. Opin. Microbiol.* **8**, 35-45.

Goulding C.W., Perry L.J. (2003). Protein production in *Escherichia coli* for structural studies by X-ray crystallography. *Journal of Structural Biology* **142**, 133-43.

Gressmann H., Linz B., Ghai R., Pleissner K.P., Schlapbach R., Yamaoka Y., Kraft C., Suerbaum S., Meyer T.F., Achtman M. (2005). Gain and loss of multiple genes during the evolution of *Helicobacter pylori*. *PLOS Genet*. **1**, 419-28.

Gully D., Bouveret E. (2006). A protein network for phospholipid synthesis uncovered by a variant of the tandem affinity purification method in *Escherichia coli*. *Proteomics* **6**, 282-93.

Guillemin K., Salama N.R., Tompkins L.S., Falkow S. (2002). Cag pathogenicity island-specific responses of gastric epithelial cells to *Helicobacter pylori* infection. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 15136-41.

Ha N.C., Oh S.T., Sung J.Y., Cha K.A., Lee M.H., Oh B.H. (2001). Supramolecular assembly and acid resistance of *Helicobacter pylori* urease. *Nat. Struct. Biol.* **8**, 505-9.

Haapalainen A.M., van Aalten D.M., Merilainen G., Jalonen J.E., Pirila P., Wierenga R.K., Hiltunen J.K., Glumoff T. (2001). Crystal structure of the liganded SCP-2-like domain of human peroxisomal multifunctional enzyme type 2 at 1.75 A resolution. *J. Mol. Biol.* **313**, 1127-38.

Haas R., Kahrs A. F., Facius D., Allmeier H., Schmitt R., Meyer T. F. (1993b). TnMax - a versatile mini-transposon for the analysis of cloned genes and shuttle mutagenesis. *Gene* **130**, 23-31.

Haas R., Meyer T.F., van Putten J.P.M. (1993). Aflagellated mutants of *Helicobacter pylori* generated by genetic transformation of naturally competent strains using shuttle mutagenesis. *Mol. Microbiol.* **8**, 753-60.

Hacker J., Kaper J.B. (2000). Pathogenicity islands and the evolution of microbes. *Annu. Rev. Microbiol.* **54**, 641-79.

Hart D.J., Tarendeau F. (2006). Combinatorial library approaches for improving soluble protein expression in *Escherichia coli*. *Acta Crystallogr. D Biol. Crystallogr*. **62**, 19-26.

Handa O., Naito Y., Yoshikawa T. (2007). CagA protein of *Helicobacter pylori*: a hijacker of gastric epithelial cell signaling. *Biochemical pharmacology* **73**, 1697-702.

Hartley J. L. (2006). Cloning technologies for protein expression and purification. *Current Opinion in Biotechnology* **17**, 1-8.

Hatakeyama M. (2003). *Helicobacter pylori* CagA, a potential bacterial oncoprotein that functionally mimics the mammalian Gab family of adaptor proteins. *Microbes Infect.* **5**, 143-50.

Hatakeyama M. (2004). Oncogenic mechanisms of the *Helicobacter pylori* CagA protein. *Nature Rev. Cancer* **4**, 688-94.

Hatakeyama M., Higashi H. (2005). *Helicobacter pylori* CagA: a new paradigm for bacterial carcinogenesis. *Cancer Sci.* **96**, 835-43.

Hatakeyama M. (2006). The Role of *Helicobacter pylori* CagA in Gastric Carcinogenesis. *Int. J. Hematol.* **84**, 301-8.

Hatakeyama M., Brzozowski T. (2006). Pathogenesis of *Helicobacter pylori* infection. *Helicobacter* **1**, 14-20.

Hegde S.S., Shrader T.E. (2001). FemABX family members are novel nonribosomal peptidyltransferases and important pathogen-specific drug targets. *J. Biol. Chem.* **276**, 6998-7003.

Hennig E.E., Mernaugh R., Edl J., Cao P., Cover T.L. (2004). Heterogeneity among *Helicobacter pylori* strains in expression of the outer membrane protein BabA. *Infect. Immunol.* **72**, 3429-35.

Hessey S.J., Spencer J., Wyatt J.I., Sobala G., Rathbone B.J., Axon A.T., Dixon M.F. (1990). Bacterial adhesion and disease activity in *Helicobacter* associated chronic gastritis. *Gut*. **31**, 134-8.

Higashi H., Tsutsumi R., Muto S., Sugiyama T., Azuma T., Asaka M., Hatakeyama M. (2002). SHP-2 tyrosine phosphatase as an intracellular target of *Helicobacter pylori* CagA protein. *Science* **295**, 683-6.

Higashi H., Tsutsumi R., Fujita A., Yamazaki S., Asaka M., Azuma T., Hatakeyama M. (2002). Biological activity of the *Helicobacter pylori* virulence factor CagA is determined by variation in the tyrosine phosphorylation sites. *Proc. Natl Acad. Sci. U.S.A.* **99**, 14428-33.

Higashi H., Nakaya A., Tsutsumi R., Yokoyama K., Fujii Y., Ishikawa S., Higuchi M., Takahashi A., Kurashima Y., Teishikata Y., Tanaka S., Azuma T., Hatakeyama M. (2004). *Helicobacter pylori* CagA induces Ras-independent morphogenetic response through SHP-2 recruitment and activation. *J. Biol. Chem.* **279**, 17205-11.

Higashi H., Yokoyama K., Fujii Y., Ren S., Yuasa H., Saadat I., Murata-Kamiya N., Azuma T., Hatakeyama M. (2005). EPIYA motif is a membrane targeting signal of *Helicobacter pylori* virulence factor CagA in mammalian cells. *J. Biol. Chem.* **280**, 130-7.

Higuchi M., Tsutsumi R., Higashi H., Hatakeyama M. (2004). Conditional gene silencing utilizing the lac repressor reveals a role of SHP-2 in *cagA* positive *Helicobacter pylori* pathogenicity. *Cancer Sci.* **95**, 442-7.

Hirata Y., Maeda S., Mitsuno Y., Tateishi K., Yanai A., Akanuma M., Yoshida H., Kawabe T., Shiratori Y., Omata M. (2002). *Helicobacter pylori* CagA protein activates serum response element-driven transcription independently of tyrosine phosphorylation. *Gastroenterology* **123**, 1962-71.

Hof P., Pluskey S., Dhe-Paganon S., Ech M.J., Shoelson S.E. (1998). Crystal structure of the tyrosine phosphatase SHP-2. *Cell* **92**, 441-50.

Hohlfeld S., Pattis I., Püls J., Plano G.V., Haas R., Fischer W. (2006). A C-terminal translocation signal is necessary, but not sufficient for type IV secretion of the *Helicobacter pylori* CagA protein. *Molecular Microbiology* **59**, 1624-37.

Holm L., Sander C. (1993). Protein structure comparison by alignment of distance matrices. *J. Mol. Biol.* **233**, 123-8.

Hoppner C., Liu Z., Domke N., Binns A.N., Baron C. (2004). VirB1 orthologs from *Brucella suis* and pKM101 complement defects of the lytic transglycosylase required for efficient type IV secretion from *Agrobacterium tumefaciens*. *J. Bacteriol.* **186**, 1415-22.

Höppner C., Carle A., Sivanesan D., Hoeppner S., Baron C. (2005). The putative lytic transglycosylase VirB1 from *Brucella suis* interacts with the type IV secretion system core components VirB8, VirB9 and VirB11. *Microbiology* **151**, 3469-82.

Hsu P.I., Hwang I.R., Cittelly D., Lai K.H., El-Zimaity H.M., Gutierrez O., Kim J.G., Osato M.S., Graham D.Y., Yamaoka Y. (2002). Clinical presentation in relation to diversity within the *Helicobacter pylori cag* pathogenicity island. *Am. J. Gastroenterol.* **97**, 2231-8.

Hung C.L., Liu J.H., Chiu W.C., Huang S.W., Hwang J.K., Wang W.C. (2007). Crystal structure of *Helicobacter pylori* formamidase AmiF reveals a cysteine-glutamate-lysine catalytic triad. *J. Biol. Chem.* **282**, 12220-9

Ilver D., Arnqvist A., Ogren J., Frick I.M., Kersulyte D., Incecik E.T., Berg D.E., Covacci A., Engstrand L., Boren T. (1998). *Helicobacter pylori* adhesin binding fucosylated histo-blood group antigens revealed by retagging. *Science* **279**, 373-7.

Istivan T.S., Coloe P.J. (2006). Phospholipase A in Gram-negative bacteria and its role in pathogenesis. *Microbiology* **152**, 1263-74.

Jemth P., Gianni S. (2007). PDZ domains: folding and binding. *Biochemistry* **46**, 8701-8.

Josenhans C., Labigne A., Suerbaum S. (1995). Comparative ultrastructural and functional studies of *Helicobacter pylori* and *Helicobacter mustelae* flagellin mutants: both flagellin subunits, FlaA and FlaB, are necessary for full motility in *Helicobacter* species. *J. Bacteriol.* **177**, 3010-20.

Josenhans C., Beier D., Linz B., Meyer T.F., Suerbaum S. (2007). Pathogenomics of *Helicobacter*. *Int. J. Med. Microbiol.* **297**, 589-600.

Jungbauer A., Kaar W., Schlegl R. (2004). Folding and refolding of proteins in chromatographic beds. *Curr. Opin. Biotechnol.* **15**, 487-94.

Kapust R.B., Waugh D.S. (1999). *Escherichia coli* maltose-binding protein is uncommonly effective at promoting the solubility of polypeptides to which it is fused. *Protein Sci*. **8**, 1668-74.

Keates S., Sougioultzis S., Keates A.C., Zhao D., Peek R.M. Jr, Shaw L.M., Kelly C.P. (2001). *cag*$^+$ *Helicobacter pylori* induce transactivation of the epidermal growth factor receptor in AGS gastric epithelial cells. *J. Biol. Chem.* **276**, 48127-34.

Kelly S.M., Jess T.J., Price N.C. (2005). How to study proteins by circular dichroism. *Biochimica et Biophysica Acta* **1751**, 119-39.

Kenny B., DeVinney R., Stein M., Reinscheid D.J., Frey E.A., Finlay B.B. (1997). Enteropathogenic *E. coli* (EPEC) transfers its receptor for intimate adherence into mammalian cells. *Cell* **91**, 511-20.

Kersulyte D., Mukhopadhyay A.K., Velapatiño B., Su W., Pan Z., Garcia C., Hernandez V., Valdez Y., Mistry R.S., Gilman R.H., Yuan Y., Gao H., Alarcón T., López-Brea M., Balakrish Nair G., Chowdhury A., Datta S., Shirai M., Nakazawa T., Ally R., Segal I., Wong B.C., Lam S.K., Olfat F.O., Borén T., Engstrand L., Torres O., Schneider R., Thomas J.E., Czinn S., Berg D.E. (2000). Differences in genotypes of *Helicobacter pylori* from different human populations. *J. Bacteriol.* **182**, 3210-8.

Kersulyte D., Chalkauskas H., Berg D.E. (1999). Emergence of recombinant strains of *Helicobacter pylori* during human infection. *Mol. Microbiol.* **31**, 31-41.

Kim Y., Prestegard J.H. (1990). Refinement of the NMR structures for acyl carrier protein with scalar coupling data. *Proteins* **8**, 377-85.

Kim J.S., Chang J.H., Chung S.I., Yum J.S. (1999). Molecular cloning and characterization of the *Helicobacter pylori fliD* gene, an essential factor in flagellar structure and motility. *J. Bacteriol.* **181**, 6969-76.

Kim J., Sitaraman S., Hierro A., Beach B.M., Odorizzi G., Hurley J.H. (2005). Structural basis for endosomal targeting by the Bro1 domain. *Dev. Cell* **8**, 937-47.

Kim S.Y., Lee Y.C., Kim H.K., Blaser M.J. (2006). *Helicobacter pylori* CagA transfection of gastric epithelial cells induces interleukin-8. *Cell Microbiol.* **8**, 97-106.

Knaust R.K., Nordlund P. (2001). Screening for soluble expression of recombinant proteins in a 96-well format. *Anal. Biochem.* **297**, 79-85.

Koraimann G. (2003). Lytic transglycosylases in macromolecular transport systems of Gram-negative bacteria. *Cell Mol. Life Sci.* **60**, 2371-88.

Kostrzynska M., Betts J.D., Austin J.W., Trust T.J. (1991). Identification, characterization, and spatial localization of two flagellin species in *Helicobacter pylori* flagella. *J. Bacteriol.* **173**, 937-46.

Kraft C., Stack A., Josenhans C., Niehus E., Dietrich G., Correa P., Fox J.G., Falush D., Suerbaum S. (2006). Genomic changes during chronic *Helicobacter pylori* infection. *J. Bacteriol.* **188**, 249-54.

Kraulis P.J. (1991). MOLSCRIPT: a program to produce both detailed and schematic plot of protein structures. *J. Appl. Crystallogr.* **24**, 946-50.

Krueger S., Hundertmark T., Kuester D., Kalinski T., Peitz U., Roessner A. (2007). *Helicobacter pylori* alters the distribution of ZO-1 and p120ctn in primary human gastric epithelial cells. *Pathol. Res. Pract.* **203**, 433-44.

Kuipers E. J., Israel D.A., Kusters J.G., Gerrits M.M., Weel J., van Der Ende A., van Der Hulst R.W., Wirth H.P., Höök-Nikanne J., Thompson S.A., Blaser M.J. (2000). Quasispecies development of

Helicobacter pylori observed in paired isolates obtained years apart from the same host. *J. Infect. Dis.* **181**, 273-82.

Kumar R.B., Xie Y.H., Das A. (2000). Subcellular localization of the *Agrobacterium tumefaciens* T-DNA transport pore proteins: VirB8 is essential for the assembly of the transport pore. *Mol. Microbiol.* **36**, 608-17.

Kumar R.B., Das A. (2001). Functional analysis of the *Agrobacterium tumefaciens* T-DNA transport pore protein VirB8. *J. Bacteriol.* **183**, 3636-41.

Kundu P., Mukhopadhyay A.K., Patra R., Banerjee A., Berg D.E., Swarnakar S. (2006). Cag pathogenicity island-independent up-regulation of matrix metalloproteinases-9 and -2 secretion and expression in mice by *Helicobacter pylori* infection. *J. Biol. Chem.* **281**, 34651-62.

Kunishima N., Asada Y., Sugahara M., Ishijima J., Nodake Y., Sugahara M., Miyano M., Kuramitsu S., Yokoyama S., Sugahara M. (2005). A novel induced-fit reaction mechanism of asymmetric hot dog thioesterase PAAI. *J. Mol. Biol.* **352**, 212-8.

Kwok T., Zabler D., Urman S., Rohde M., Hartig R., Wessler S., Misselwitz R., Berger J., Sewald N., König W., Backert S. (2007). *Helicobacter* exploits integrin for type IV secretion and kinase activation. *Nature* **449**, 862-6.

Lai E.M., Eisenbrandt R., Kalkum M., Lanka E., Kado C.I. (2002). Biogenesis of T pili in *Agrobacterium tumefaciens* requires precise VirB2 propilin cleavage and cyclization. *J. Bacteriol.* **184**, 327-30.

Laskowski R.A., Watson J.D., Thornton J.M. (2005). ProFunc: a server for predicting protein function from 3D structure. *Nucl. Acids Res.* **33**, 89-93.

Lawrence M.C., Pilling P.A., Epa V.C., Berry A.M., Ogunniyi A.D., Paton J.C. (1998). The crystal structure of pneumococcal surface antigen PsaA reveals a metal-binding site and a novel structure for a putative ABC-type binding protein. *Structure* **6**, 1553-61.

Lazzaroni J.C., Dubuisson J.F., Vianney A. (2002). The Tol proteins of *Escherichia coli* and their involvement in the translocation of group A colicins. *Biochimie* **84**, 391-7.

Leesong M., Henderson B.S., Gillig J.R., Schwab J.M., Smith J.L. (1996). Structure of a dehydratase-isomerase from the bacterial pathway for biosynthesis of unsaturated fatty acids: two catalytic activities in one active site. *Structure* **4**, 253-64.

Lemmon M.A. (2007). Pleckstrin homology (PH) domains and phosphoinositides. *Biochem. Soc. Symp.* **74**, 81-93.

Leslie A.G.W. (1992). Recent changes to the MOSFLM package for processing film and image plate data. *Joint CCP4+ESF-EAMCB Newsletter on Protein Crystallography* 26.

Leying H., Suerbaum S., Geis G., Haas R. (1992). Cloning and genetic characterization of a *Helicobacter pylori* flagellin gene. *Mol. Microbiol.* **6**, 2863-74.

Ley C., Parsonnet J. (2001). *Helicobacter pylori* infection and gastric cancer. *Gastroenterology* **120**, 324-5.

Leunk R.D., Johnson P.T., David B.C., Kraft W.G., Morgan D.R. (1988). Cytotoxic activity in broth-culture filtrates of *Campylobacter pylori*. *J. Med. Microbiol.* **26**, 93-9.

Lloubes R., Cascales E., Walburger A., Bouveret E., Lazdunski C., Bernadac A., Journet L. (2001). The Tol-Pal proteins of the *Escherichia coli* cell envelope: an energized system required for outer membrane integrity ? *Res. Microbiol.* **152**, 523-9.

Li M., Su Z.G., Janson J.C. (2004). In vitro protein refolding by chromatographic procedures. *Protein Expr. Purif.* **33**, 1-10

Li J., Derewenda U., Dauter Z., Smith S., Derewenda Z.S. (2000). Crystal structure of the *Escherichia coli* thioesterase II, a homolog of the human Nef binding enzyme. *Nat. Struct. Biol.* **7**, 555-9.

Li Y., Fanning A.S., Anderson J.M., Lavie A. (2005). Structure of the Conserved Cytoplasmic C-terminal Domain of Occludin: Identification of the ZO-1 Binding Surface. *J. Mol. Biol.* **352**, 151-64.

Lilie H., Schwarz E., Rudolph R. (1998). Advances in refolding of proteins produced in *E. coli. Curr. Opin. Biotechnol.* 9, 497-501.

Linden S., Nordman H., Hedenbro J., Hurtig M., Boren T., Carlstedt I. (2002). Strain- and blood group-dependent binding of *Helicobacter pylori* to human gastric MUC5AC glycoforms. *Gastroenterology* **123**, 1923-30.

Loh J.T., Cover T.L. (2006). Requirement of histidine kinases HP0165 and HP1364 for acid resistance in *Helicobacter pylori*. *Infect. Immunol.* **74**, 3052-9.

Loh M.L., Vattikuti S., Schubbert S., Reynolds M.G., Carlson E., Lieuw K.H., Cheng J.W., Lee C.M., Stokoe D., Bonifas J.M., Curtiss N.P., Gotlib J., Meshinchi S., Le Beau M.M., Emanuel P.D., Shannon K.M. (2004). Mutations in *PTPN11* implicate the SHP-2 phosphatase in leukemogenesis. *Blood* **103**, 2325-31.

Lu H., Hsu P.I., Graham D.Y., Yamaoka Y. (2005a). Duodenal ulcer promoting gene of *Helicobacter pylori*. *Gastroenterology* **128**, 833-48.

Lu H., Yamaoka Y., Graham D.Y. (2005b). *Helicobacter pylori* virulence factors: facts and fantasies. *Curr. Opin. Gastroenterol.* **21**, 653-9.

Luo Z.Q., Isberg R.R. (2004). Multiple substrates of the *Legionella pneumophila* Dot/Icm system identified by interbacterial protein transfer. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 841-6.

Lupetti P., Heuser J.E., Manetti R., Massari P., Lanzavecchia S., Bellon P.L., Dallai R., Rappuoli R., Telford J.L. (1996). Oligomeric and subunit structure of the *Helicobacter pylori* vacuolating cytotoxin. *J. Cell. Biol.* **133**, 801-7.

Macnab R.M. (1996). Flagella and motility in *E. coli* and *Salmonella*: cellular and molecular biology. *ASM press.* 123-145.

Mahdavi J., Sonden B., Hurtig M., Olfat F.O., Forsberg L., Roche N., Angstrom J., Larsson T., Teneberg S., Karlsson K.A., Altraja S., Wadström T., Kersulyte D., Berg D.E., Dubois A., Petersson C., Magnusson K.E., Norberg T., Lindh F., Lundskog B.B., Arnqvist A., Hammarström L., Borén T. (2002). *Helicobacter pylori* SabA adhesin in persistent infection and chronic inflammation. *Science* **297**, 573-8.

Makrides S.C. (1996). Strategies for achieving high-level expression of genes in *Escherichia coli*. *Micr. Rev.* **60**, 512-38.

Marcus E.A., Scott D.R. (2001). Cell lysis is responsible for the appearance of extracellular urease in *Helicobacter pylori*. *Helicobacter* **6**, 93-9.

Marcus E.A., Moshfegh A.P., Sachs G., Scott D.R. (2005). The periplasmic alpha-carbonic anhydrase activity of *Helicobacter pylori* is essential for acid acclimation. *J. Bacteriol.* **187**, 729-38.

Marshall B.J., Warren J.R. (1984). Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. *Lancet* **1**, 1311-5.

Matthysse A.G. (1987). Characterization of nonattaching mutants of *Agrobacterium tumefaciens. J. Bacteriol.* **169**, 313-23.

McGee D.J., Radcliff F.J., Mendz G.L., Ferrero R.L., Mobley H.L. (1999). *Helicobacter pylori rocF* is required for arginase activity and acid protection *in vitro* but is not essential for colonization of mice or for urease activity. *J. Bacteriol.* **181**, 7314-22.

McGee D.J., Zabaleta J., Viator R.J., Testerman T.L., Ochoa A.C., Mendz G.L. (2004). Purification and characterization of *Helicobacter pylori* arginase, RocF: unique features among the arginase superfamily. *Eur. J. Biochem.* **271**, 1952-62.

McRee D.E. (1999). XtalView/Xfit- A versatile program for manipulating atomic coordinates and electron density. *J. Struct. Biol.* **125**, 156-65.

Merrell D.S., Goodrich M.L., Otto G., Tompkins L.S., Falkow S. (2003). pH-regulated gene expression of the gastric pathogen *Helicobacter pylori. Infect. Immunol.* **71**, 3529-39.

Merrit E.A., Bacon D.J. (1997). Raster3D photorealistic molecular graphics. *Methods. Enzymol.* **277**, 505-24.

Middelberg A. P. J. (2002) Preparative protein refolding. *Trends Biotechnol.* 20, 437-43.

Middleton R., Sjolander K., Krishnamurthy N., Foley J., Zambryski P. (2005). Predicted hexameric structure of the *Agrobacterium* VirB4 C-terminus suggests VirB4 acts as a docking site during type IV secretion. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 1685-90.

Milburn C.C., Boudeau J., Deak M., Alessi D.R., van Aalten D.M. (2004). Crystal structure of MO25 alpha in complex with the C-terminus of the pseudo kinase STE20-related adaptor. *Nat. Struct. Mol. Biol.* 11, 193-200.

Mimuro H., Suzuki T., Tanaka J., Asahi M., Haas R., Sasakawa C. (2002). Grb2 is a key mediator of *Helicobacter pylori* CagA protein activities. *Mol. Cell* **10**, 745-55.

Mizushima T., Sugiyama T., Kobayashi T., Komatsu Y., Ishizuka J., Kato M., Asaka M. (2002). Decreased adherence of *cagG*-deleted *Helicobacter pylori* to gastric epithelial cells in Japanese clinical isolates. *Helicobacter* **7**, 22-9.

Mobley H.L.T, Mendz G.L., Hazell S.L. (2001). *Helicobacter pylori* physiology and genetics. *ASM Press* **300**, 69-71, 75-76.

Moese S., Selbach M., Zimny-Arndt U., Jungblut P.R., Meyer T.F., Backert S. (2001). Identification of a tyrosine-phosphorylated 35 kDa carboxy-terminal fragment (p35CagA) of the *Helicobacter pylori* CagA protein in phagocytic cells: processing or breakage ? *Proteomics* **1**, 618-29

Moese S., Selbach M., Kwok T., Brinkmann V., König W., Meyer T.F., Backert S. (2004). *Helicobacter pylori* induces AGS cell motility and elongation via independent signaling pathways. *Infect. Immun.* **72**, 3646-9.

Moese S., Selbach M., Brinkmann V., Karlas A., Haimovich B., Backert S., Meyer T.F. (2007). The *Helicobacter pylori* CagA protein disrupts matrix adhesion of gastric epithelial cells by dephosphorylation of vinculin. *Cellular Microbiology* **9**, 1148-61.

Molinari M., Salio M., Galli C., Norais N., Rappuoli R., Lanzavecchia A., Montecucco C. (1998). Selective inhibition of Ii-dependent antigen presentation by *Helicobacter pylori* toxin VacA. *J. Exp. Med.* **187**, 135-40.

Monack D.M., Mueller A., Falkow S. (2004). Persistent bacterial infections: the interface of the pathogen and the host immune system. *Nat. Rev. Microbiol*. **2**, 747-65.

Montecucco C., Rappuoli R. (2001). Living dangerously: how *Helicobacter pylori* survives in the human stomach. *Nat. Rev. Mol. Cell Biol.* **2**, 457-66.

Moran A.P. (2007). Lipopolysaccharide in bacterial chronic infection: Insights from *Helicobacter pylori* lipopolysaccharide and lipid A. *Int. J. Med. Microbiol.* **297**, 307-19.

Moss S. F., Sood S. (2003). *Helicobacter pylori*. *Curr. Opin. Infect. Dis.* **16**, 445-51.

Mourey L., Da Re S., Pedelacq J. D., Tolstykh T., Faurie C., Guillet V., Stock J. B., Samama, J.P. (2001). Crystal structure of the CheA histidine phosphotransfer domain that mediates response regulator phosphorylation in bacterial chemotaxis. *J. Biol. Chem.* **276**, 31074-82.

Murata-Kamiya N., Kurashima Y., Teishikata Y., Yamahashi Y., Saito Y., Higashi H., Aburatani H., Akiyama T., Peek R.M. Jr, Azuma T., Hatakeyama M. (2007). *Helicobacter pylori* CagA interacts with E-cadherin and deregulates the β-catenin signal that promotes intestinal transdifferentiation in gastric epithelial cells. *Oncogene* **26**, 4617-26.

Murshudov G.N., Vagin A.A., Lebedev A., Wilson K.S., Dodson E.J. (1999). Efficient anisotropic refinement of macromolecular structures using FFT. *Acta Crystallogr. D Biol. Crystallogr.* **55**, 247-55.

Murzin A.G., Brenner S.E., Hubbard T., Chothia C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**, 536-40.

Nagai H., Roy C.R. (2001). The DotA protein from *Legionella pneumophila* is secreted by a novel process that requires the Dot/Icm transporter. *EMBO J.* **20**, 5962-70.

Noach L.A., Rolf T.M., Tytgat G.N. (1994). Electron microscopic study of association between *Helicobacter pylori* and gastric and duodenal mucosa. *J. Clin. Pathol.* **47**, 699-704.

O'Toole P.W., Lane M.C., Porwollik S. (2000). *Helicobacter pylori* motility. *Microbes Infect.* **2**, 1207-14.

O'Toole P.W., Kostrzynska M., Trust T.J. (1994). Non motile mutants of *Helicobacter pylori* and *Helicobacter mustelae* defective in flagellar hook production. *Mol. Microbiol.* **14**, 691-703.

Odenbreit S., Till M., Hofreuter D., Faller G., Haas R. (1999). Genetic and functional characterization of the *alpAB* gene locus essential for the adhesion of *Helicobacter pylori* to human gastric tissue. *Mol. Microbiol.* **31**, 1537-48.

Odenbreit S., Puls J., Sedlmaier B., Gerland E., Fischer W., Haas R. (2000). Translocation of *Helicobacter pylori* CagA into gastric epithelial cells by type IV secretion. *Science* **287**, 1497-500.

Odenbreit S., Gebert B., Püls J., Fischer W., Haas R. (2001). Interaction of *Helicobacter pylori* with professional phagocytes: role of the *cag* pathogenicity island and translocation, phosphorylation and processing of CagA. *Cell Microbiol.* **3**, 21-31.

Oganesyan N., Kim S.H., Kim R. (2005). On-column protein refolding for crystallization. *J. Struct. Funct. Genomics* **6**, 177-182.

Ogura K., Maeda S., Nakao M., Watanabe T., Tada M., Kyutoku T., Yoshida H., Shiratori Y., Omata M. (2000). Virulence factors of *Helicobacter pylori* responsible for gastric diseases in Mongolian gerbil. *J. Exp. Med.* **192**, 1601-10.

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., Mardis E.R., Engstrand L.G., Gordon J.I. (2006). The complete genome sequence of a chronic atrophic gastritis *Helicobacter pylori* strain: Evolution during disease progression. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 9999-10004.

Oliveira M.J., Costa A.C., Costa A.M., Henriques L., Suriano G., Atherton J.C., Machado J.C., Carneiro F., Seruca R., Mareel M., Leroy A., Figueiredo C. (2006). *Helicobacter pylori* induces gastric epithelial cell invasion in a c-Met and type IV secretion system-dependent manner. *J. Biol. Chem*. **281**, 34888-96.

Ostermeier M., Lutz S. (2003). The creation of ITCHY hybrid protein libraries. *Methods Mol. Biol.* **231**, 129-41.

Panthel K., Dietz P., Haas R., Beier D. (2003). Two-component systems of *Helicobacter pylori* contribute to virulence in a mouse infection model. *Infect. Immunol.* **71**, 5381-5.

Pape T., Schneider T.R. (2004). HKL2MAP: a graphical user interface for phasing with SHELX programs. *J. Appl. Crystallogr.* **37**, 843-4.

Parsonnet J., Friedman G.D., Orentreich N., Vogelman H. (1997). Risk for gastric cancer in people with CagA positive or CagA negative *Helicobacter pylori* infection. *Gut*. **40**, 297-301.

Parsons S.J., Parsons J.T. (2004). Src family kinases, key regulators of signal transduction. *Oncogene* **23**, 7906-9.

Parsot C., Hamiaux C., Page A.L. (2003). The various and varying roles of specific chaperones in type III secretion systems. *Curr. Opin. Microbiol.* **6**, 7-14.

Paschos A., Patey G., Sivanesan D., Gao C., Bayliss R., Waksman G., O'callaghan D., Baron C. (2006). Dimerization and interactions of *Brucella suis* VirB8 with VirB4 and VirB10 are required for its biological activity. *Proc. Natl. Acad. Sci. U.S.A.* **103**, 7252-7.

Pattis I., Weiss E., Laugks R., Haas R., Fischer W. (2007). The *Helicobacter pylori* CagF protein is a type IV secretion chaperone-like molecule that binds close to the C-terminal secretion signal of the CagA effector protein. *Microbiology* **153**, 2896-909

Pawson T., Nash P. (2003). Assembly of cell regulatory systems through protein interaction domains. *Science* **300**, 445-52.

Pawson T. (2004). Specificity in signal transduction: from phosphotyrosine-SH2 domain interactions to complex cellular systems. *Cell* **116**, 191-203.

Peck B., Ortkamp M., Diehl K.D., Hundt E., Knapp B. (1999). Conservation, localization and expression of HopZ, a protein involved in adhesion of *Helicobacter pylori. Nucleic Acids Res.* **27**, 3325-33.

Peek R.M. Jr, Blaser M.J. (2002). *Helicobacter pylori* and gastrointestinal tract adenocarcinomas. *Nature Rev. Cancer* **2**, 28–37.

Pflock M., Dietz P., Schar J., Beier D. (2004). Genetic evidence for histidine kinase HP165 being an acid sensor of *Helicobacter pylori*. *FEMS Microbiol. Lett.* **234**, 51-61.

Pflock M., Finsterer N., Joseph B., Mollenkopf H., Meyer T.F., Beier D. (2006a). Characterization of the ArsRS regulon of *Helicobacter pylori*, involved in acid adaptation. *J. Bacteriol.* **188**, 3449-62.

Pflock M., Kennard S., Finsterer N., Beier D. (2006b). Acidresponsive gene regulation in the human pathogen *Helicobacter pylori. J. Biotechnol.* **126**, 52–60.

Phadnis S.H., Parlow M.H., Levy M., Ilver D., Caulkins CM., Connors J.B., Dunn B.E. (1996). Surface localization of *Helicobacter pylori* urease and a heat shock protein homolog requires bacterial autolysis. *Infect. Immunol.* **64**, 905-12.

Pillinger M.H., Marjanovic N., Kim S.Y., Lee Y.C., Scher J.U., Roper J., Abeles A.M., Izmirly P.I., Axelrod M., Pillinger M.Y., Tolani S., Dinsell V., Abramson S.B., Blaser M.J. (2007). *Helicobacter pylori* stimulates gastric epithelial cell MMP-1 secretion via CagA-dependent and -independent ERK activation. *J. Biol. Chem.* **282**, 18722-31.

Prilusky J., Felder C.E., Zeev-Ben-Mordehai T., Rydberg E., Man O., Beckmann J.S., Silman I., and Sussman J. L. (2005). Bioinformatics FoldIndex©: a simple tool to predict whether a given protein sequence is intrinsically unfolded. *Bioinformatics* **21**, 3435-8.

Privalov P.L., Dragan A.I. (2007). Microcalorimetry of biological macromolecules. *Biophys Chem*. **126**, 16-24.

Ponting C.P. (1997a). Evidence for PDZ domains in bacteria, yeast, and plants. *Protein Sci.* **6**, 464–8.

Ponting C.P., Phillips C., Davies K.E., Blake D.J. (1997b). PDZ domains: targeting signaling molecules to sub-membranous sites. *Bioassays* **19**, 469–79.

Poppe M., Feller S.M., Romer G., Wessler S. (2007). Phosphorylation of *Helicobacter pylori* CagA by c-Abl leads to cell motility. *Oncogene* **26**, 3462-72.

Pyndiah S., Lasserre J.P., Menard A., Claverol S., Prouzet-Mauleon V., Megraud F., Zerbib F., Bonneu M. (2007). Two-dimensional blue native/SDS gel electrophoresis of multiprotein complexes from *Helicobacter pylori. Mol. Cell Proteomics.* **6**, 193-206.

Puls J., Fischer W., Haas R. (2002). Activation of *Helicobacter pylori* CagA by tyrosine phosphorylation is essential for dephosphorylation of host cell proteins in gastric epithelial cells. *Mol. Microbiol.* **43**, 961-9.

Quezada C.M., Hamel D.J., Gradinaru C., Bilwes A.M., Dahlquist F.W., Crane B.R., Simon M.I. (2005). Structural and chemical requirements for histidine phosphorylation by the chemotaxis kinase CheA. *J. Biol. Chem.* **280**, 30581-5.

Rain J.C., Selig L., De Reuse H., Battaglia V., Reverdy C., Simon S., Lenzen G., Petel F., Wojcik J., Schachter V., Chemama Y., Labigne A., Legrain P. (2001). The protein-protein interaction map of *Helicobacter pylori. Nature* **409**, 211-5.

Ren S., Higashi H., Lu H., Azuma T., Hatakeyama M. (2006). Structural basis and functional consequence of *Helicobacter pylori* CagA multimerization in cells. *J. Biol. Chem.* **281**, 32344-52.

Reyrat J.M., Lanzavecchia S., Lupetti P., de Bernard M., Pagliaccia C., Pelicic V., Charrel M., Ulivieri C., Norais N., Ji X., Cabiaux V., Papini E., Rappuoli R., Telford J.L. (1999). 3D imaging of the 58 kDa cell binding subunit of the *Helicobacter pylori* cytotoxin. *J. Mol. Biol.* **290**, 459-70.

Rieder G., Hatz R.A., Moran A.P., Walz A., Stolte M., Enders G. (1997). Role of adherence in interleukin-8 induction in *Helicobacter pylori*-associated gastritis. *Infect. Immun.* **65**, 3622-30.

Robinson D.R., Wu Y.M., Lin S.F. (2000). The protein tyrosine kinase family of the human genome. *Oncogene* **19**, 5548-57.

Rohde M., Puls J., Buhrdorf R., Fischer W., Haas R. (2003). A novel sheathed surface organelle of the *Helicobacter pylori* cag type IV secretion system. *Mol. Microbiol.* **49**, 219-34.

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

Salama N.R., Otto G., Tompkins L., Falkow S. (2001). Vacuolating cytotoxin of *Helicobacter pylori* plays a role during colonization in a mouse model of infection. *Infect. Immun.* **69**, 730-6.

Salama N., Guillemin K., McDaniel T.K., Sherlock G., Tompkins L., Falkow S. (2000). A whole-genome microarray reveals genetic diversity among *Helicobacter pylori* strains. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 14668-73.

Sambrook J., Fritsch E.F., Maniatis T. (1989). Molecular cloning: a laboratory manual. *Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N.Y.*

Saito H., Yamaoka Y., Ishizone S., Maruta F., Sugiyama A., Graham D.Y., Yamauchi K., Ota H., Miyagawa S. (2005). Roles of *virD4* and *cagG* genes in the cag pathogenicity island of Helicobacter *pylori* using a Mongolian gerbil model. *Gut.* **54**, 584-90.

Satin B., Del Giudice G., Della Bianca V., Dusi S., Laudanna C., Tonello F., Kelleher D., Rappuoli R., Montecucco C., Rossi F. (2000). The neutrophil-activating protein (HP-NAP) of *Helicobacter pylori* is a protective antigen and a major virulence factor. *J. Exp. Med.* **191**, 1467-76.

Sauer F.G., Remaut H., Hultgren S.J., Waksman G. (2004). Fiber assembly by the chaperone-usher pathway. *Biochim. Biophys. Acta* **1694**, 259-67.

Savvides S.N., Yeo H.J., Beck M.R., Blaesing F., Lurz R., Lanka E., Buhrdorf R., Fischer W., Haas R., Waksman G. (2003). VirB11 ATPases are dynamic hexameric assemblies: new insights into bacterial type IV secretion. *EMBO J.* **22**, 1969-80.

Saxena S., Yuan P., Dhar S.K., Senga T., Takeda D., Robinson H., Kornbluth S., Swaminathan K., Dutta A. (2004). A dimerized coiled-coil domain and an adjoining part of geminin interact with two sites on Cdt1 for replication inhibition. *Mol. Cell.* **15**, 245-58.

Schar J., Sickmann A., Beier D. (2005). Phosphorylation independent activity of atypical response regulators of *Helicobacter pylori. J. Bacteriol.* **187**, 3100-9.

Schlessinger J., Lemmon M.A. (2003). SH2 and PTB domains in tyrosine kinase signaling. *Sci. STKE* **191**

Schneider T.R, Sheldrick G.M. (2002). Substructure solution with SHELXD. *Acta Crystallogr.* **D58**, 1772-9.

Schulein R., Guye P., Rhomberg T.A., Schmid M.C., Schroder G., Vergunst A.C., Carena I., Dehio C. (2005). A bipartite signal mediates the transfer of type IV secretion substrates of *Bartonella henselae* into human cells. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 856-61.

Schujman G.E., Guerin M., Buschiazzo A., Schaeffer F., Llarrull L.I., Reh G., Vila A.J., Alzari P.M., de Mendoza D. (2006). Structural basis of lipid biosynthesis regulation in Gram-positive bacteria. *Embo J.* **25**, 4074-83.

Sharma S.A., Tummuru M.K., Blaser M.J., Kerr L.D. (1998). Activation of IL-8 gene expression by *Helicobacter pylori* is regulated by transcription factor nuclear factor kappa B in gastric epithelial cells. *J. Immunol.* **160**, 2401-7.

Shin K., Fogg V.C., Margolis B. (2006). Tight junctions and cell polarity. *Annu. Rev. Cell Dev. Biol.* **22**, 207-35.

Schmees C., Gerhard M., Treptau T., Voland P., Schwendy S., Rad R., Prinz C. (2006). VacA-associated inhibition of T-cell function: reviewed and reconsidered. *Helicobacter* **11**, 144-6.

Schreiber S., Konradt M., Groll C., Scheid P., Hanauer G., Werling H.O., Josenhans C., Suerbaum S. (2004). The spatial orientation of *Helicobacter pylori* in the gastric mucus. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 5024-9.

Schreiber S., Bucker R., Groll C., Azevedo-Vethacke M., Garten D., Scheid P., Friedrich S., Gatermann S., Josenhans C., Suerbaum S. (2005). Rapid loss of motility of *Helicobacter pylori* in the gastric lumen *in vivo*. *Infect. Immunol.* **73**, 1584-9.

Schulein R., Guye P., Rhomberg T.A., Schmid M.C., Schröder G., Vergunst A.C., Carena I., Dehio C. (2005). A bipartite signal mediates the transfer of type IV secretion substrates of *Bartonella henselae* into human cells. *Proc. Natl. Acad. Sci. U.S A.* **102**, 856-61.

Schwede T., Kopp J., Guex N., Peitsch M.C. (2003). SWISS-MODEL: An automated protein homology-modeling server. *Nucleic Acids Res.* **31**, 3381-5.

Scott D.R., Weeks D., Hong C., Postius S., Melchers K., Sachs G. (1998). The role of internal urease in acid resistance of *Helicobacter pylori. Gastroenterology* **114**, 58-70.

Segal E.D., Cha J., Lo J., Falkow S., Tompkins L.S. (1999). Altered states: involvement of phosphorylated CagA in the induction of host cellular growth changes by *Helicobacter pylori*. *Proc. Natl. Acad. Sci. U.S.A.* **96**, 14559-64.

Selbach M., Moese S., Meyer T.F., Backert S. (2002a). Functional analysis of the *Helicobacter pylori* cag pathogenicity island reveals both VirD4-CagA-dependent and VirD4-CagA-independent mechanisms. *Infect. Immun.* **70**, 665-71.

Selbach M., Moese S., Hauck C.R., Meyer T.F., Backert S. (2002). Src is the kinase of the *Helicobacter pylori* CagA protein *in vitro* and *in vivo*. *J. Biol. Chem.* **277**, 6775-8.

Selbach M., Moese S., Hurwitz R., Hauck C.R., Meyer T.F., Backert S. (2003). The *Helicobacter pylori* CagA protein induces cortactin dephosphorylation and actin rearrangement by c-Src inactivation. *EMBO J.* **22**, 515-28.

Selbach M., Moese S., Backert S., Jungblut P.R., Meyer T.F. (2004). The *Helicobacter pylori* CagA protein induces tyrosine dephosphorylation of ezrin. *Proteomics* **4**, 2961-8.

Selbach M., Backert S. (2005). Cortactin: an Achille's heel of the actin cytoskeleton targeted by pathogens. *Trends Microbiol.* **13**, 181-9.

Seo M.D., Park S.J., Kim H.J., Lee B.J. (2007). Solution structure of hypothetical protein, HP0495 from *Helicobacter pylori. Proteins* **67**, 1189-92.

Skouloubris S., Labigne A., De Reuse H. (1997). Identification and characterization of an aliphatic amidase in *Helicobacter pylori. Mol. Microbiol.* **25**, 989-98.

Skouloubris S., Labigne A., De Reuse H. (2001). The AmiE aliphatic amidase and AmiF formamidase of *Helicobacter pylori*: natural evolution of two enzyme paralogues. *Mol. Microbiol.* **40**, 596-609.

Smith T.G., Lim J.M., Weinberg M.V., Wells L., Hoover T.R. (2007). Direct analysis of the extracellular proteome from two strains of *Helicobacter pylori. Proteomics* **7**, 2240-5.

Song F., Zhuang Z., Finci L., Dunaway-Mariano D., Kniewel R., Buglino J.A, Solorzano V., Wu J., Lima C.D. (2006). Structure, function, and mechanism of the phenylacetate pathway hot dog-fold thioesterase PaaI. *J. Biol. Chem.* **281**, 11028-38.

Songyang Z., Shoelson S.E., Chaudhuri M., Gish G., Pawson T., Haser W.G., King F., Roberts T., Ratnofsky S., Lechleider R.J. (1993). SH2 domains recognize specific phosphopeptide sequences. *Cell* **72**, 767-8.

Songyang Z., Shoelson S.E., McGlade J., Olivier P., Pawson T., Bustelo X.R., Barbacid M., Sabe H., Hanafusa H., Yi T. (1994). Specific motifs recognized by the SH2 domains of Csk, 3BP2, fps/fes, GRB2, HCP, SHC, Syk, and Vav. *Mol. Cell. Biol.* **14**, 2777-85.

Songyang Z., Carraway K.L., Eck M.J., Harrison S.C., Feldman R.A., Mohammadi M., Schlessinger J., Hubbard S.R., Smith D.P., Eng C. (1995). Catalytic specificity of protein-tyrosine kinases is critical for selective signalling. *Nature* **373**, 536-9.

Sørensen H.P., Mortensen K.K. (2005a). Advanced genetic strategies for recombinant protein expression in *Escherichia coli. J. Biotechnol.* **115**, 113-28.

Sørensen H.P., Mortensen K.K. (2005b). Soluble expression of recombinant proteins in the cytoplasm of *Escherichia coli*. *Microb. Cell Fact.* **4**, 1.

Stein M., Rappuoli R., Covacci A. (2000). Tyrosine phosphorylation of the *Helicobacter pylori* CagA antigen after *cag*-driven host-cell translocation. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 1263-8.

Stein M., Bagnoli F., Halenbeck R., Rappuoli R., Fantl W.J., Covacci A. (2002). c-Src/Lyn kinases activate *Helicobacter pylori* CagA through tyrosine phosphorylation of the EPIYA motifs. *Mol. Microbiol.* **43**, 971-80.

Stevens R.C. (2000). Design of high-throughput methods of protein production for structural biology. *Struct. Fold. Des.* **8**, 177-85.

Stingl K., Uhlemann E.M., Deckers-Hebestreit G., Schmid R., Bakker E.P., Altendorf K. (2001). Prolonged survival and cytoplasmic pH homeostasis of *Helicobacter pylori* at pH 1. *Infect. Immunol.* **69**, 1178-80.

Stingl K., Altendorf K., Bakker E.P. (2002). Acid survival of *Helicobacter pylori*: how does urease activity trigger cytoplasmic pH homeostasis ? *Trends Microbiol.* **10**, 70-4.

Sturgis J.N. (2001). Organisation and evolution of the *tol-pal* gene cluster. *J. Mol. Microbiol. Biotechnol.* **3**, 113-22.

Suerbaum S., Josenhans C., Labigne A. (1993). Cloning and genetic characterization of the *Helicobacter pylori* and *Helicobacter mustelae flaB* flagellin genes and construction of *H. pylori flaA*- and *flaB*-negative mutants by electroporationmediated allelic exchange. *J. Bacteriol.* **175**, 3278-88.

Suerbaum S., Josenhans C. (1999). Virulence factors of *Helicobacter pylori*: implications for vaccine development. *Mol. Med. Today* **5**, 32-9.

Suerbaum S., Michetti P. (2002). *Helicobacter pylori* infection. *N. Engl. J. Med.* **347**, 1175-86.

Suerbaum S., Josenhans C. (2007). *Helicobacter pylori* evolution and phenotypic diversification in a changing host. *Nat. Rev. Microbiol.* **5**, 441-52.

Sundrud M.S., Torres V.J., Unutmaz D., Cover T.L. (2004). Inhibition of primary human T cell proliferation by *Helicobacter pylori* vacuolating toxin (VacA) is independent of VacA effects on IL-2 secretion. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 7727-32.

Suzuki M., Mimuro H., Suzuki T., Park M., Yamamoto T., Sasakawa C. (2005). Interaction of CagA with Crk plays an important role in *Helicobacter pylori*-induced loss of gastric epithelial cell adhesion. *J. Exp. Med.* **202**, 1235-47.

Swietnicki W. (2006). Folding aggregated proteins into functionally active forms. *Current Opinion in Biotechnology* **17**, 1-6.

Szabo I., Brutsche S., Tombola F., Moschioni M., Satin B., Telford J.L., Rappuoli R., Montecucco C., Papini E., Zoratti M. (1999). Formation of anion-selective channels in the cell plasma membrane by the toxin VacA of *Helicobacter pylori* is required for its biological activity. *EMBO J.* **18**, 5517-27.

Tsutakawa S.E., Hura G.L., Frankel K.A., Cooper P.K., Tainer J.A. (2007). Structural analysis of flexible proteins in solution by small angle X-ray scattering combined with crystallography. *J. Struct. Biol.* **158**, 214-23.

Tarendeau F., Boudet J., Guilligay D., Mas P.J., Bougault C.M., Boulo S., Baudin F., Ruigrok R.W., Daigle N., Ellenberg J., Cusack S., Simorre J.P., Hart D.J. (2007). Structure and nuclear import function of the C-terminal domain of influenza virus polymerase PB2 subunit. *Nat. Struct. Mol. Biol.* **14**, 229-33.

Tammer I., Brandt S., Hartig R., König W., Backert S. (2007). Activation of Abl by *Helicobacter pylori*: A Novel Kinase for CagA and Crucial Mediator of Host Cell Scattering. *Gastroenterology* **132**, 1309-19.

Tanaka J., Suzuki T., Mimuro H., Sasakawa C. (2003). Structural definition on the surface of Helicobacter *pylori* type IV secretion apparatus. *Cell Microbiol.* **5**, 395-404.

Tang Y., Swanson K.D., Neel B.G., Eck M.J. Structural Basis for the Dimerization and Phosphoinositide Specificity of the Src Kinase-associated Phosphoproteins SKAP55 and SKAP-Hom. *To be published.*

Taylor N.S., Fox J.G., Akopyants N.S., Berg D.E., Thompson N., Shames B., Yan L., Fontham E., Janney F., Hunter F.M. (1995). Long-term colonization with single and multiple strains of *Helicobacter pylori* assessed by DNA fingerprinting. *J. Clin. Microbiol.* **33**, 918-23.

Telford J.L., Ghiara P., Dell'Orco M., Comanducci M., Burroni D., Bugnoli M., Tecce M.F., Censini S., Covacci A., Xiang Z. (1994). Gene structure of the *Helicobacter pylori* cytotoxin and evidence of its key role in gastric disease. *J. Exp. Med.* **179**, 1653-8.

Terradot L., Durnell N., Li M., Li M., Ory J., Labigne A., Legrain P., Colland F., Waksman G. (2004). Biochemical characterization of protein complexes from the *Helicobacter pylori* protein interaction map: strategies for complex formation and evidence for novel interactions within type IV secretion systems. *Mol. Cell. Proteomics* **3**, 809-19.

Terradot L., Bayliss R., Oomen C., Leonard G.A., Baron C., Waksman G. (2005). Structures of two core subunits of the bacterial type IV secretion system, VirB8 from *Brucella suis* and ComB10 from *Helicobacter pylori*. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 4596-601.

Terpe K. (2006). Overview of bacterial expression systems for heterologous protein production: from molecular and biochemical fundamentals to commercial systems. *Appl. Microbiol. Biotechnol.* **72**, 211-22.

Thoden J.B., Zhuang Z., Dunaway-Mariano D., Holden H.M. (2003). The structure of 4-hydroxybenzoyl-CoA thioesterase from *Arthrobacter sp*. strain SU. *J. Biol. Chem.* **278**, 43709-16.

Thoden J.B., Holden H.M., Zhuang Z., Dunaway-Mariano D. (2002). X-ray crystallographic analyses of inhibitor and substrate complexes of wild-type and mutant 4-hydroxybenzoyl-CoA thioesterase. *J. Biol. Chem.* **277**, 27468-76.

Thomas S.M., Brugge J.S. (1997). Cellular functions regulated by Src family kinases. *Annu. Rev. Cell Dev. Biol.* **13**, 513-609.

Togni M., Swanson K.D., Reimann S., Kliche S., Pearce A.C., Simeoni L., Reinhold D., Wienands J., Neel B.G., Schraven B., Gerber A. (2005). Regulation of *in vitro* and *in vivo* immune functions by the cytosolic adaptor protein SKAP-Hom. *Mol. Cell. Biol.* **25**, 8052-63.

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., Nelson K., Quackenbush J., Zhou L., Kirkness E.F., Peterson S., Loftus B., Richardson D., Dodson R., Khalak H.G., Glodek A., McKenney K., Fitzegerald L.M., Lee N., Adams M.D., Hickey E.K., Berg D.E., Gocayne J.D., Utterback T.R., Peterson J.D., Kelley J.M., Cotton M.D., Weidman J.M., Fujii C., Bowman C., Watthey L., Wallin E., Hayes W.S., Borodovsky M., Karp P.D., Smith H.O., Fraser C.M., Venter J.C. (1997). The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature* **388**, 539-47.

Tonello F., Dundon W.G., Satin B., Molinari M., Tognon G., Grandi G., Del Giudice G., Rappuoli R., Montecucco C. (1999). The *Helicobacter pylori* neutrophil-activating protein is an iron-binding protein with dodecameric structure. *Mol. Microbiol.* **34**, 238-46.

Trusolino L., Bertotti A., Comoglio P.M. (2001). A signaling adapter function for alpha-6 beta-4 integrin in the control of HGF-dependent invasive growth. *Cell* **107**, 643-54.

Tsumoto K., Ejima D., Kumagai I., Arakawa T. (2003). Practical considerations in refolding proteins from inclusion bodies. *Protein Expr. Purif.* **28**, 1-8.

Tsutsumi R., Higashi H., Higuchi M., Okada M., Hatakeyama M. (2003). Attenuation of *Helicobacter pylori* CagA–SHP-2 signaling by interaction between CagA and C-terminal Src kinase. *J. Biol. Chem.* **278**, 3664-70.

Tsutsumi R., Takahashi A, Azuma T, Higashi H, Hatakeyama M. (2006). Focal adhesion kinase is a substrate and downstream effector of SHP-2 complexed with *Helicobacter pylori* CagA. *Mol. Cell Biol.* 26, 261-76.

Tummuru M.K., Cover T.L., Blaser M.J. (1993). Cloning and expression of a high-molecular-mass major antigen of *Helicobacter pylori*: evidence of linkage to cytotoxin production. *Infect. Immun.* **61**, 1799-809.

Tummuru M.K., Cover T.L., Blaser M.J. (1994). Mutation of the cytotoxin-associated *cagA* gene does not affect the vacuolating cytotoxin activity of *Helicobacter pylori*. *Infect. Immun.* **62**, 2609-13.

Unemo M., Aspholm-Hurtig M., Ilver D., Bergstrom J., Boren T., Danielsson D., Teneberg S. (2005). The sialic acid binding SabA adhesin of *Helicobacter pylori* is essential for nonopsonic activation of human neutrophils. *J. Biol. Chem.* **280**, 15390-7.

Vagin A., Teplyakov A. (1997). MOLREP: an Automated Program for Molecular Replacement. *Journal of Applied Crystallography* **30**, 1022-5.

Van Doorn L.J., Quint W., Schneeberger P., Tytgat G.M. and de Boer W.A. (1997). The only good *Helicobacter pylori* is a dead *Helicobacter pylori*. *Lancet* **350**, 71-2.

Van Doorn L.J., Figueiredo C., Sanna R., Plaisier A., Schneeberger P., de Boer W., Quint W. (1998). Clinical relevance of the *cagA*, *vacA*, and *iceA* status of *Helicobacter pylori*. *Gastroenterology* **115**, 58–66.

Van Vliet A.H.M., Stoof J., Vlasblom R., Wainwright S.A., Hughes N.J., Kelly D.J., Bereswill S., Bijlsma J.J.E., Hoogenboezem T., Vandenbroucke-Grauls C.M.J.E., Kist M., Kuipers E.J., Kusters J.G. (2002). The role of the Ferric Uptake Regulator (Fur) in regulation of *Helicobacter pylori* iron uptake. *Helicobacter* **7**, 237-44.

Van Vliet A.H., Ernst F.D., Kusters J.G. (2004). NikR-mediated regulation of *Helicobacter pylori* acid adaptation. *Trends Microbiol.* **12**, 489-94.

Van Vliet A.H., Kuipers E.J., Stoof J., Poppelaars S.W., Kusters J.G. (2004). Acid-responsive gene induction of ammonia-producing enzymes in *Helicobacter pylori* is mediated via a metal-responsive repressor cascade. *Infect. Immun.* **72**, 766-73.

Vergunst A.C., van Lier M.C., den Dulk-Ras A., Stuve T.A., Ouwehand A., Hooykaas P.J. (2005). Positive charge is an important feature of the C-terminal transport signal of the VirB/D4-translocated proteins of *Agrobacterium*. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 832-7.

Vetter I.R., Parker M.W., Tucker A.D., Lakey J.H., Pattus F., Tsernoglou D. (1998). Crystal structure of a colicin N fragment suggests a model for toxicity. *Structure* **6**, 863-74.

Viala J., Chaput C., Boneca I.G., Cardona A., Girardin S.E., Moran A.P., Athman R., Memet S., Huerre M.R., Coyle A.J., DiStefano P.S., Sansonetti P.J., Labigne A., Bertin J., Philpott D.J., Ferrero R. L. (2004). Nod1 responds to peptidoglycan delivered by the *Helicobacter pylori* cag pathogenicity island. *Nat. Immunol*. **5**, 1166–74.

Wallasch C., Crabtree J.E., Bevec D., Robinson P.A., Wagner H., Ullrich A. (2002). *Helicobacter pylori*-stimulated EGF receptor transactivation requires metalloprotease cleavage of HB-EGF. *Biochem. Biophys. Res. Commun.* **295**, 695-701.

Walz A., Odenbreit S., Mahdavi J., Boren T., Ruhl S. (2005). Identification and characterization of binding properties of *Helicobacter pylori* by glycoconjugate arrays. *Glycobiology* **15**, 700-8.

Ward D.V., Draper O., Zupan J. R., Zambryski P.C. (2002). Peptide linkage mapping of the *Agrobacterium tumefaciens* virencoded type IV secretion system reveals protein subassemblies. *Proc. Natl. Acad. Sci. U.S.A.* **99**, 11493-500.

Weeks D.L., Eskandari S., Scott D.R., Sachs G. (2000). A H^+-gated urea channel: the link between *Helicobacter pylori* urease and gastric colonization. *Science* **287**, 482-5.

Weydig C., Starzinski-Powitz A., Carra G., Löwer J., Wessler S. (2007). CagA-independent disruption of adherence junction complexes involves E-cadherin shedding and implies multiple steps in *Helicobacter pylori* pathogenicity. *Exp. Cell Res.* **313**, 3459-71.

Wen Y., Marcus E.A., Matrubutham U., Gleeson M.A., Scott D.R., Sachs G. (2003). Acid-adaptive genes of *Helicobacter pylori*. *Infect. Immunol.* **71**, 5921-39.

Windle H.J., Fox A., Ni Eidhin D., Kelleher D. (2000) The thioredoxin system of *Helicobacter pylori*. *J. Biol. Chem.* **275**, 5081-9.

Wiener M., Freymann D., Ghosh P., Stroud R.M. (1997). Crystal structure of colicin Ia. *Nature* **385**, 461-4.

Witherell H.L., Hansen S., Jellum E., Orentreich N., Vogelman J.H., Parsonnet J. (1997). Risk for gastric lymphoma in persons with *cagA*$^+$ and *cagA*$^-$ *Helicobacter pylori* infection. *J. Infect. Dis.* **176**, 1641-4.

Yamada K., Nameki N., Saito K., Koshiba S., Inoue M., Kigawa T., Yokoyama S. Solution structure of the third PDZ domain of mouse harmonin. *To be Published.*

Yamaguchi H., Osaki T., Taguchi H., Hanawa T., Yamamoto T., Kamiya S. (1996). Flow cytometric analysis of the heat shock protein 60 expressed on the cell surface of *Helicobacter pylori. J. Med. Microbiol.* 45, 270-7.

Yamaguchi H., Osaki T., Taguchi H., Hanawa T., Yamamoto T., Kamiya S. (1998). Relationship between expression of HSP60, urease activity, production of vacuolating toxin, and adherence activity of *Helicobacter pylori. J. Gastroenterol.* **33**, 6-9.

Yamaoka Y., Kodama T., Kashima K., Graham D.Y., Sepulveda A. (1998). Variants of the 3'. region of the *cagA* gene in *Helicobacter pylori* isolates from patients with different *H. pylori*-associated diseases. *Infect. Immun.* **36**, 2258-63.

Yamaoka Y., Kwon D.H., Graham D.Y. (2000). A M(r) 34,000 proinflammatory outer membrane protein (OipA) of *Helicobacter pylori*. *Proc. Natl. Acad. Sci. U.S.A.* **97**, 7533-8.

Yamaoka Y., Ojo O., Fujimoto S., Odenbreit S., Haas R., Gutierrez O., El-Zimaity H.M., Reddy R., Arnqvist A., Graham D.Y. (2006). *Helicobacter pylori* outer membrane proteins and gastroduodenal disease. *Gut*. **55**, 775-81.

Yeo H.J., Savvides S.N., Herr A.B., Lanka E., Waksman G. (2000). Crystal structure of the hexameric traffic ATPase of the *Helicobacter pylori* type IV secretion system. *Mol. Cell.* **6**, 1461-72.

Yeo H.J., Yuan Q., Beck M.R., Baron C., Waksman G. (2003). Structural and functional characterization of the VirB5 protein from the type IV secretion system encoded by the conjugative plasmid pKM101. *Proc. Natl. Acad. Sci. U.S.A.* **100**, 15947-52.

Yeo H.J., Waksman G. (2004). Unveiling molecular scaffolds of the type IV secretion system. *J. Bacteriol.* **186**, 1919-26.

Yokoyama K., Higashi H., Ishikawa S., Fujii Y., Kondo S., Kato H., Azuma T., Wada A., Hirayama T., Aburatani H., Hatakeyama M. (2005). Functional antagonism between *Helicobacter pylori* CagA and vacuolating toxin VacA in control of the NFAT signaling pathway in gastric epithelial cells. *Proc. Natl. Acad. Sci. U.S.A.* **102**, 9661-6.

Yoshida N., Granger D.N., Evans D.J. Jr, Evans D.G., Graham D.Y., Anderson D.C., Wolf R.E., Kvietys P.R. (1993). Mechanisms involved in *Helicobacter pylori*-induced inflammation. *Gastroenterology* **105**, 1431-40.

Yuan Q., Carle A., Gao C., Sivanesan D., Aly K.A., Hoppner C., Krall L., Domke N., Baron C. (2005). Identification of the VirB4-VirB8-VirB5-VirB2 pilus assembly sequence of type IV secretion systems. *J. Biol. Chem.* **280**, 26349-59.

Xiang Z., Censini S., Bayeli P.F., Telford J.L., Figura N., Rappuoli R., Covacci A. (1995). Analysis of expression of CagA and VacA virulence factors in 43 strains of *Helicobacter pylori* reveals that clinical isolates can be divided into two major types and that CagA is not necessary for expression of the vacuolating cytotoxin. *Infect. Immun.* **63**, 94-8.

Zahrl D., Wagner M., Bischof K., Bayer M., Zavecz B., Beranek A., Ruckenstuhl C., Zarfel G.E., Koraimann G. (2005). Peptidoglycan degradation by specialized lytic transglycosylases associated with type III and type IV secretion systems. *Microbiology* **151**, 3455-67.

Zanotti G., Papinutto E., Dundon W., Battistutta R., Seveso M., Giudice G., Rappuoli R., Montecucco C. (2002). Structure of the neutrophil-activating protein from *Helicobacter pylori. J. Mol. Biol.* **323**, 125-30.

Zhou M.M., Meadows R.P., Logan T.M., Yoon H.S., Wade W.S., Ravichandran K.S., Burakoff S.J., Fesik S.W. (1995). Solution structure of the Shc SH2 domain complexed with a tyrosine-phosphorylated peptide from the T-cell receptor. *Proc. Natl. Acad. Sci. U.S.A.* **92**, 7784-8.

Zhuang Z., Song F., Martin B.M., Dunaway-Mariano D. (2002). The YbgC protein encoded by the *ybgC* gene of the tol-pal gene cluster of *Haemophilus influenzae* catalyzes acyl-coenzyme A thioester hydrolysis. *FEBS Lett*. **516**, 161-3.