Acta Crystallographica Section D
Biological Crystallography
ISSN 1399-0047

Structure of a three-dimensional domain-swapped dimer of the *Helicobacter pylori* type IV secretion system pilus protein CagL

Stephan Barden,[a] Benjamin Schomburg,[a] Jens Conradi,[b] Steffen Backert,[c,d] Norbert Sewald[b] and Hartmut H. Niemann[a]*

[a]Structural Biochemistry, Department of Chemistry, Bielefeld University, Universitätsstrasse 25, 33615 Bielefeld, Germany, [b]Organic and Bioorganic Chemistry, Department of Chemistry, Bielefeld University, Universitätsstrasse 25, 33615 Bielefeld, Germany, [c]Institute of Medical Microbiology, OvG University Magdeburg, Leipziger Strasse 44, 39120 Magdeburg, Germany, and [d]Department of Biology, Chair of Microbiology, FAU University Erlangen-Nuremberg, 91058 Erlangen, Germany

Correspondence e-mail: hartmut.niemann@uni-bielefeld.de

Received 20 December 2013
Accepted 11 February 2014

PDB reference: CagL, 4cii

A new crystal form of the *Helicobacter pylori* type IV secretion system (T4SS) pilus protein CagL is described here. In contrast to two previously reported monomeric structures, CagL forms a three-dimensional domain-swapped dimer. CagL dimers can arise during refolding from inclusion bodies or can form spontaneously from purified monomeric CagL in the crystallization conditions. Monomeric CagL forms a three-helix bundle, with which the N-terminal helix is only loosely associated. In the new crystal form, the N-terminal helix is missing. The domain swap is owing to exchange of the C-terminal helix between the two protomers of a dimer. A loop-to-helix transition results in a long helix of 108 amino acids comprising the penultimate and the last helix of the monomer. The RGD motif of dimeric CagL adopts an α-helical conformation. In contrast to the previously reported structures, the conserved and functionally important C-terminal hexapeptide is resolved. It extends beyond the three-helix bundle as an exposed helical appendage. This new crystal form contributes to the molecular understanding of CagL by highlighting rigid and flexible regions in the protein and by providing the first view of the C-terminus. Based on the structural features, a previously unrecognized homology between CagL and CagI is discussed.

1. Introduction

The term three-dimensional domain swapping, first introduced by Eisenberg and coworkers (Bennett *et al.*, 1995), is used to describe an exchange of identical structure elements (domains, secondary-structure elements or even smaller elements) between two or more otherwise identical proteins to form dimers or higher oligomers (open form). Within this arrangement, the swapped structure elements replace one another and thus reconstitute the functional unit, resembling the monomeric (closed) form of the protein (Liu & Eisenberg, 2002). To date, the structures of about 60 domain-swapped proteins are known, but general requirements for or mechanisms of domain swapping are elusive (Rousseau *et al.*, 2012). A hinge loop connecting the swapped structure element to its residual protein part is the only region that needs to change its conformation upon domain swapping. Domain-swapped oligomers are stabilized by (i) closed interfaces also found in the monomer, (ii) open interfaces newly formed between the protomers and (iii) the structural reorientation of

the hinge (Bennett *et al.*, 1995; Rousseau *et al.*, 2001). Loss of translational and rotational entropy energetically disfavours the formation of dimers and oligomers (Schlunegger *et al.*, 1997).

Depending on the protein concentration, domain-swapped structures may form coincidentally (Rousseau *et al.*, 2003). However, a great energy barrier must be overcome to disrupt the interactions between the swapping structure element and the residual protein upon formation of the open monomer (Bennett *et al.*, 1994). Strain in the flexible hinge region may contribute to opening the folded monomer under slightly unfolding conditions (Rousseau *et al.*, 2001). Three-dimensional domain swapping not only occurs from stably folded monomers but also upon (re)folding or reconstitution oligomerization (Crestfield *et al.*, 1962; Bennett *et al.*, 1994; Carey *et al.*, 2007; Rousseau *et al.*, 2001, 2003; Griebenow & Klibanov, 1995).

CagL is a pathogenicity-associated factor from *Helicobacter pylori* that mediates contact with surface receptors of the human gastric epithelium (Backert *et al.*, 2011). It is part of a type IV secretion system (T4SS) that enables *H. pylori* to modulate host-cell signalling (Smolka & Backert, 2012; Tegtmeyer *et al.*, 2011). CagL is important for primary T4SS responses such as effector-protein (CagA) translocation and interleukin-8 secretion (Fischer *et al.*, 2001), and it binds to host-cell integrins (Kwok *et al.*, 2007). CagL can be co-purified with CagI and CagH (Shaffer *et al.*, 2011; Pham *et al.*, 2012), two proteins which are encoded within the cytotoxin-associated gene pathogenicity island (*cag*PAI) by genes contiguous with the *cagL* gene (Censini *et al.*, 1996; Shaffer *et al.*, 2011). All three proteins share a conserved C-terminal S/T-K-I/V-I-V-K hexapeptide, and deleting any of these hexapeptides inhibited functional pilus assembly (Shaffer *et al.*, 2011). Additional sequence homology has been described between CagL and CagH, which share 32% identical and 47% similar amino acids in their C-terminal parts (185 amino acids of CagL and 183 amino acids of CagH; Shaffer *et al.*, 2011).

We recently reported two crystal structures of CagL variants, $CagL^{KKQEK}$ and $CagL^{meth}$, both of which showed a monomeric protein composed of four long α-helices ($\alpha1$, $\alpha2$, $\alpha5$ and $\alpha6$) with a small two-helix appendage ($\alpha3$ and $\alpha4$) perpendicular to the long molecule axis (Barden *et al.*, 2013). A structurally invariant hydrophobic core is mainly formed by aromatic residues from $\alpha2$, $\alpha5$ and $\alpha6$. The ends of the rod-like structure exhibit substantial flexibility. The N-terminal helix $\alpha1$ packs into a groove formed by $\alpha2$ and $\alpha6$, but its precise location differs between the two structures owing to a shift of one helix turn along the helical axis. CagL exposes an Arg-Gly-Asp (RGD) motif mediating cell adhesion (Kwok *et al.*, 2007; Tegtmeyer *et al.*, 2011; Barden *et al.*, 2013). The RGD motif is located in the middle of $\alpha2$ and forms a hinge region allowing the N-terminus of $\alpha2$ to move towards $\alpha5$. Reducing the flexibility around the RGD motif by the introduction of artificial disulfides stabilized CagL but abrogated cell adhesion (Barden *et al.*, 2013).

Here, we present the crystal structure of a three-dimensional domain-swapped CagL dimer with C_2 symmetry.

2. Materials and methods

2.1. Cloning of expression constructs, protein expression and purification

Expression, refolding and purification of the $CagL^{C\text{-}His}$ used for crystallization were performed as described elsewhere (Conradi, Tegtmeyer *et al.*, 2012). The protein was further dialyzed against crystallization buffer (10 m*M* Tris adjusted with HCl to pH 7.5, 20 m*M* NaCl), concentrated by ultrafiltration (Vivaspin 20, 5000 MWCO, Sartorius) and used without freezing. $CagL^{wt}$ and its variants were produced as described previously (Barden *et al.*, 2013).

2.2. Crystallization, crystal harvesting and data collection

Crystals of $CagL^{C\text{-}His}$ grew at 2.5 mg ml^{-1} in two conditions consisting of a mixture of 420 or 400 µl of 40% 2-methyl-2,4-pentanediol (MPD), 100 m*M* phosphate/citrate buffer pH 4.2 with 80 or 100 µl 30% *tert*-butanol premixed in the reservoir with a protein:reservoir ratio of 2 µl:1 µl at 4°C by hanging-drop vapour diffusion. Single crystals were mounted in a Cryo-Loop (Hampton Research) and flash-cooled in liquid nitrogen. Two data sets from $CagL^{C\text{-}His}$ were collected at 100 K and a wavelength of 0.97950 Å on beamline ID14-4 at the ESRF, Grenoble, France using an ADSC Quantum Q315r CCD detector with an oscillation range of 0.5°.

2.3. Data reduction, structure determination and refinement

The data sets were processed with *XDS* (Kabsch, 2010) and scaled with *SCALA* (Evans, 2006) from the *CCP*4 suite (Winn *et al.*, 2011). The structure of $CagL^{C\text{-}His}$ was solved by molecular replacement (MR) using *Phaser* (McCoy *et al.*, 2007). The model was manually modified in *Coot* (Emsley *et al.*, 2010) and refined with *PHENIX* (Adams *et al.*, 2010) using simulated annealing in the first cycle. Initial *B* values were taken from the starting model. Translation/libration/screw (TLS) groups were identified with the *TLSMD* server (Painter & Merritt, 2006) and included in refinement. Figures were prepared with *PyMOL* (v.0.99rc6; Schrödinger) including the *APBS* plug-in (Baker *et al.*, 2001).

2.4. Dimerization assay and analytical size-exclusion chromatography

50 µl N-terminally His_6-tagged or TEV-cleaved $CagL^{wt}$ at 10 mg ml^{-1} in Tris-buffered saline (TBS) were mixed with 40 µl MPD and 100 µl citrate–phosphate buffer at pH 4.0 and pH 8.0, respectively, and incubated at 4°C for 20 d. 50 µl of the samples were analyzed on a Superdex 75 10/300 GL (GE Healthcare) size-exclusion chromatography (SEC) run with phosphate-buffered saline (PBS). The SEC column was calibrated using the Gel Filtration Calibration Kit LMW (GE Healthcare).

2.5. Limited proteolysis and immunoblotting

60 µl of CagL variants at 1.33 mg ml^{-1} in TBS were mixed with 20 µl trypsin at 0.04 mg ml^{-1} (final protease:protein ratio of 1:100) and incubated on ice. 10 µl samples were taken after

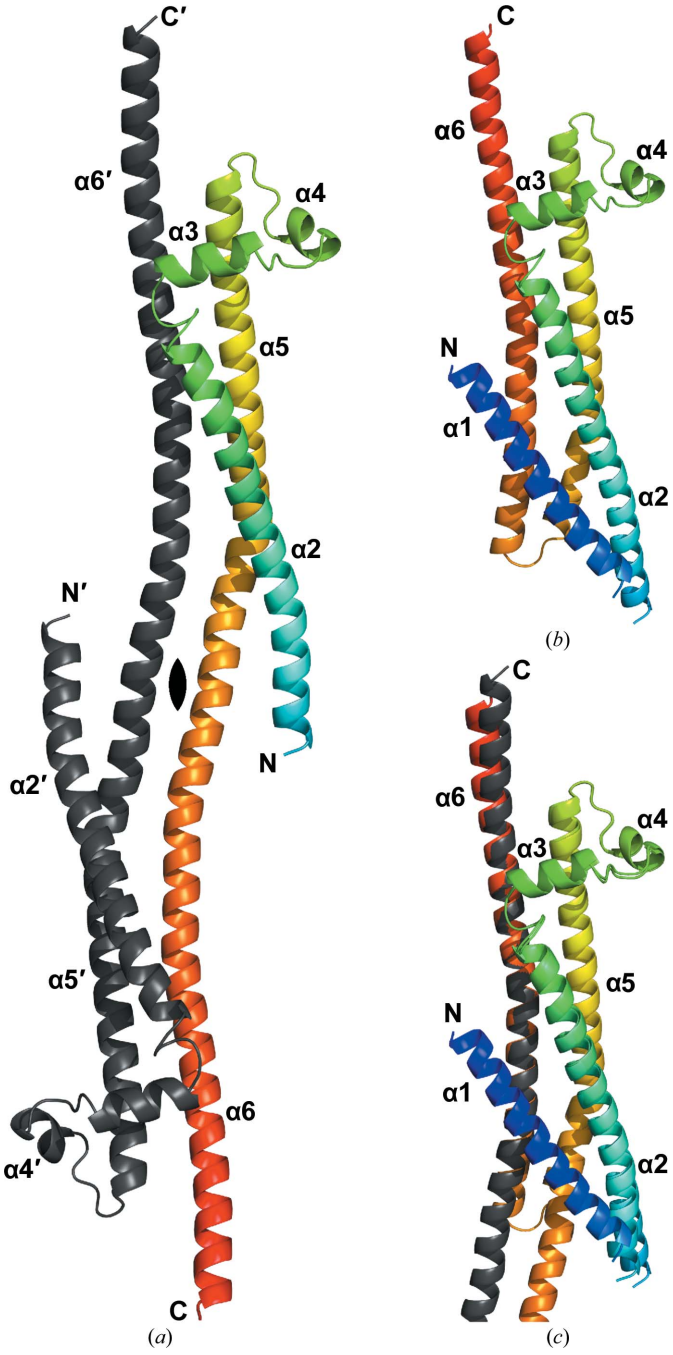

Figure 1
Domain-swapped dimer of CagL$^{C\text{-}His}$. (*a*) Two molecules of CagL$^{C\text{-}His}$ related by a crystallographic twofold axis form an elongated domain-swapped dimer. One molecule is coloured blue to red from the N-terminus to the C-terminus. The second molecule is shown in grey. Helix $\alpha1$ is missing. (*b*) The structure of CagL^{meth} chain *A* is shown as representative of monomeric CagL. Colouring is identical to that in (*a*). (*c*) Overlay of monomeric CagL^{meth} and the upper part of the domain-swapped dimer. Small structural deviations are apparent only at the N-terminus of $\alpha2$, around $\alpha4$ and at the C-terminus of $\alpha6$. The central core region of CagL formed by the three-helix bundle $\alpha2$, $\alpha5$ and $\alpha6$ is virtually identical in the monomeric and the domain-swapped structures.

Table 1
Data-collection and refinement statistics.

Values in parentheses are for the highest resolution shell.

Data collection	
Space group	$P6_522$
Unit-cell parameters (Å)	$a = b = 61.58$, $c = 244.53$
Resolution (Å)	44.75–2.15 (2.27–2.15)
No. of observed reflections	144575 (14229)
No. of unique reflections	16080 (2252)
Multiplicity	9.0 (6.3)
Completeness (%)	99.8 (99.5)
R_{merge}	0.085 (0.503)
R_{meas}	0.090 (0.545)
$R_{p.i.m.}$	0.026 (0.197)
$\langle I\rangle/\langle\sigma(I)\rangle$	15.7 (3.2)
$CC_{1/2}$	0.999 (0.877)
Wilson *B* factor (Å^2)	26.2
Refinement	
Resolution (Å)	40.28–2.15 (2.28–2.15)
R_{cryst} (%)	19.85 (23.65)
R_{free} (%)	23.74 (26.21)
No. of reflections	
Work set	15193 (2443)
Test set	784 (130)
No. of molecules/atoms	
Protein	1/1443
Ligands	1/8
Water	46
Solvent content† (%)	63
R.m.s.d	
Bond lengths (Å)	0.009
Angles (°)	0.968
Ramachandran plot	
Favoured (%)	96.7
Allowed (%)	3.30
Mean *B* factor (Å^2)	51.67

† The solvent content was calculated with *MATTHEWS_COEF* (Kantardjieff & Rupp, 2003) using the final PDB file.

the given periods and mixed with 1 µl phenylmethanesulfonylfluoride (PMSF) at 1.7 mg ml^{-1} in ethanol on ice. Following separation by 16% Tricine–SDS–PAGE (Schägger, 2006), the proteins were blotted onto Immobilon-P (Millipore) PVDF membrane. The membrane was blocked with 100 ml 5%(*w*/*v*) nonfat dry milk powder in PBS at room temperature for 4 h, incubated with 5 ml 1:2000 anti-polyhistidine–peroxidase (Sigma–Aldrich) antibody in TBS supplemented with 0.05%(*v*/*v*) Tween-20 (TBST) overnight at 4°C, washed three times with TBST and analyzed with a Luminescent Image Analyzer LAS-3000 (Fujifilm) apparatus using Roti-Lumin (Carl Roth) solution.

2.6. Software used for structure and sequence analysis

Superposition of the structures was performed with *LSQKAB* (Kabsch, 1976), secondary-structure assignment with *DSSP* (Joosten *et al.*, 2010) and calculation of the buried surface area with the *PDBePISA* server (Krissinel & Henrick, 2007). The alignment was generated with *ClustalW* (Larkin *et al.*, 2007); secondary-structure predictions were performed with *Jpred* (Cole *et al.*, 2008), disulfide predictions with *DISULFIND* (Ceroni *et al.*, 2006), signal peptide predictions with *SignalP* (Bendtsen *et al.*, 2004) and transmembrane predictions with *TMHMM* (Krogh *et al.*, 2001).

3. Results

3.1. CagL$^{C\text{-}His}$ forms a three-dimensional domain-swapped dimer

The variant CagL$^{C\text{-}His}$ comprising amino acids 21–237 (*H. pylori* strain 26695) with an artificial C-terminal non-cleavable LEH$_6$ tag crystallized non-reproducibly 4–5 months after crystallization setup as hexagonal needles with one molecule per asymmetric unit. The structure of CagL$^{C\text{-}His}$ was solved by molecular replacement with CagL^{KKQEK} (PDB entry 3zci; Barden *et al.*, 2013) as a search model and refined to a resolution of 2.15 Å. Data-collection and refinement statistics are provided in Table 1.

CagL$^{C\text{-}His}$ is resolved from Gly61 to His240 in the artificial His$_6$ tag with no electron density for α1. In contrast to the monomeric structures CagL^{KKQEK} and CagL^{meth} (Barden *et al.*, 2013), CagL$^{C\text{-}His}$ reveals an elongated domain-swapped crystal dimer with C_2 symmetry (Fig. 1*a*). Residues Ala176–Thr179, which connect α5 and α6 in a loop conformation in monomeric CagL, adopt a α-helical conformation in CagL$^{C\text{-}His}$ to form one long helix of about 108 amino acids (Lys133–His240) covering α5 and α6. Thus, in CagL$^{C\text{-}His}$ the invariant three-helix bundle previously identified by comparison of CagL^{KKQEK} and CagL^{meth} (Barden *et al.*, 2013) is formed by α2, α5 and a symmetry-related α6′. The core region of CagL$^{C\text{-}His}$ comprising amino acids 80–114 (α2 or α3), 144–166 (α5) and 187–219 (α6′) aligns almost perfectly with the same regions of CagL^{KKQEK} (backbone r.m.s.d. of 0.502 Å) and CagL^{meth} (backbone r.m.s.d. of 0.424 Å) (Figs. 1*b* and 1*c*). However, the domain swap induces structural rearrangements in the lower, more flexible part of the three-helix bundle, mostly in the C-terminal part of α5 (Fig. 2). In monomeric CagL, a short stretch around the double glycine Gly168-Gly169 adopts a 3$_{10}$-helical conformation, whereas it is α-helical in the domain-swapped CagL$^{C\text{-}His}$. This induces a shift in the helical register of one amino acid from Ala171/Ser172 to Ala176/Ser177 (Figs. 2*a* and 2*c*). Concomitantly, Thr170-Ala171 and Ile174-Thr175 face towards α6′ and form a hydrophobic patch with Phe189′-Ile190′ and Ile193′. A hydrophilic interface between the protomers involves Gln178-Thr179, Glu182 and Lys185-Asn186 around a twofold symmetry axis between the side chains of Glu182 and Glu182′ (Fig. 2*d*). The overall inter-domain interface between CagL$^{C\text{-}His}$ and its symmetry mate comprises ∼3087 Å^2, corresponding to ∼1540 Å^2 per monomer, which is similar to the interface area formed by α6 with α2 and α5 in CagL^{KKQEK} (∼1500 Å^2) and in CagL^{meth}

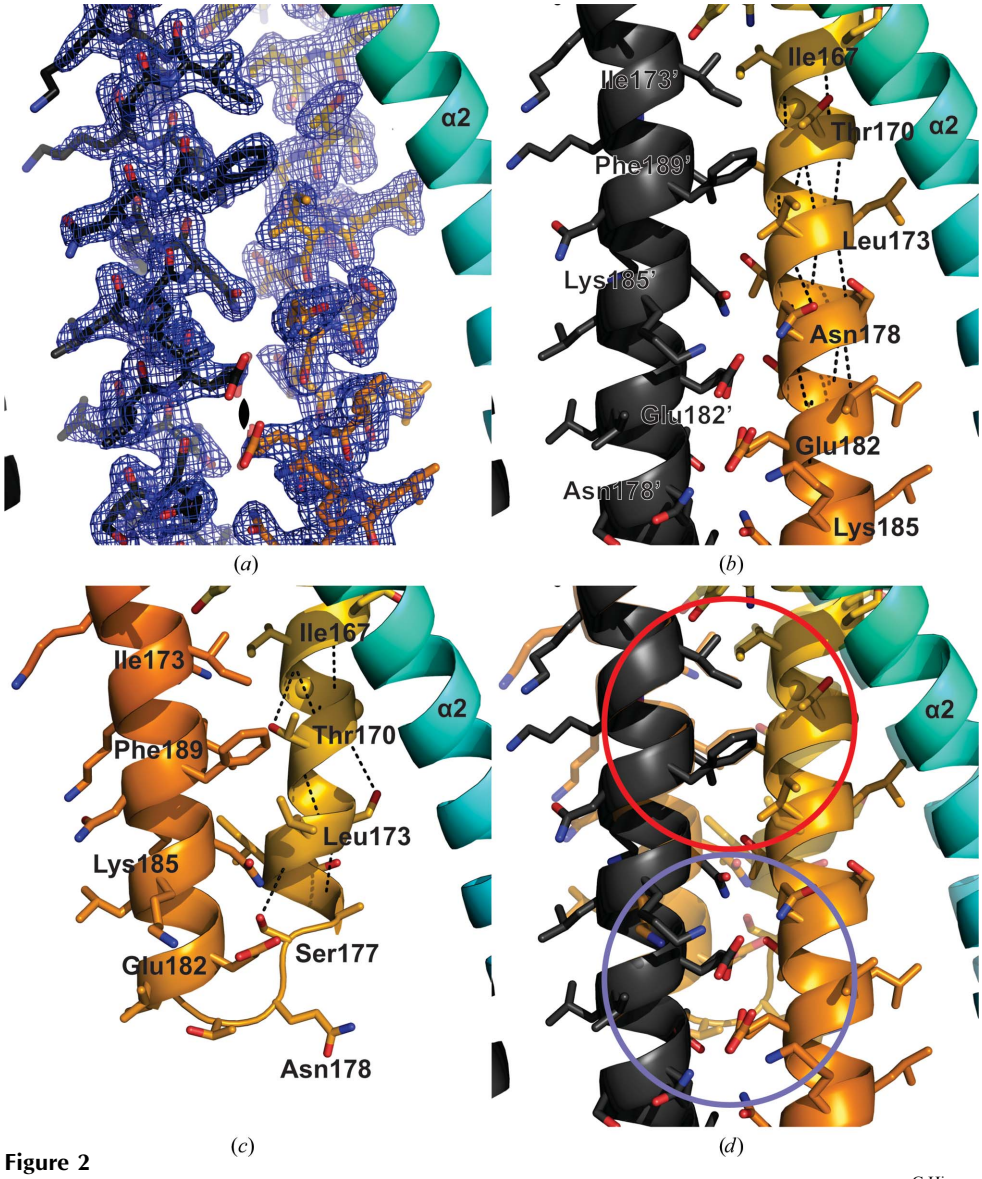

Figure 2
Comparison of the hinge loop in the domain-swapped and monomeric forms of CagL. (*a*) In CagL$^{C\text{-}His}$ α5 and α6 form a continuous α-helix. The helical hinge regions from both monomers contact each other. Colouring is as in Fig. 1. (*b*) The electron density around the hinge region in the domain-swapped CagL contoured at 1σ unambiguously shows the dimeric assembly. (*c*) The same region and orientation as in (*a*) is shown for CagL^{meth}. A short region around Gly168 and Gly169 (C$^{\alpha}$ atoms shown as spheres) adopts a 3$_{10}$-helical rather than α-helical conformation, and α5 and α6 are connected by a short loop. Helix α1 is omitted for clarity. (*d*) Overlay of (*b*) and (*c*). The α-helical conformation around Gly168 and Gly169 results in a register shift of the dimer relative to the monomer for residues Ala171–Thr175. The newly formed open interface is hydrophilic and marked with a blue circle. The hydrophobic region of the interface (circled in red) is also present in the closed, monomeric form.

(∼1655 Å^2). Helix α4, which is resolved in only some of the seven previously reported crystallographically independent CagL chains, is resolved in CagL$^{C\text{-}His}$ but is flexible, as indicated by high *B* factors. The conformation of α3 in CagL$^{C\text{-}His}$ resembles that in CagL^{meth}, while it adopts a double conformation in CagL^{KKQEK}, most likely owing to a destabilizing effect of the mutations that were introduced into CagL^{KKQEK} to promote crystallization (Barden *et al.*, 2013).

3.2. The N-terminal helix is not resolved

We previously identified α1 (amino acids 21–52 of CagL^{meth}; PDB entry 3zcj) to be only loosely associated with the rest of CagL (Barden *et al.*, 2013). In CagL^{meth} and CagL^{KKQEK} α1 packs into a hydrophobic groove formed by α2 and α6, but it is shifted by about one helix turn in its axial direction when comparing the two structures. In CagL$^{C\text{-}His}$ α1 is not visible. Given the crystal packing of CagL$^{C\text{-}His}$, α1 would clash with α3 and α4 of a symmetry-related molecule if it retained the same position as found in either CagL^{meth} or CagL^{KKQEK}. Presumably, α1 is proteolytically cleaved off during the long time required for crystal growth. The linker connecting α1 and α2 was not resolved in any of the previous CagL crystal structures, indicating that it is flexible.

3.3. The RGD motif is *α*-helical in CagL$^{C\text{-}His}$

Helix α2 embedding the potential integrin-binding RGD (Arg76-Gly77-Asp78) motif is resolved from Gly61 and is involved in several crystal contacts with three symmetry mates. It is helical from Glu62 to Thr102, including the RGD motif. Comparison of CagL$^{C\text{-}His}$ with CagL^{meth} and CagL^{KKQEK} divides α2 into an N-terminal and a C-terminal part: amino acids C-terminal to the RGD motif align almost perfectly, whereas the N-terminus of α2 in CagL$^{C\text{-}His}$ adopts an intermediate position between CagL^{meth} and CagL^{KKQEK}, slightly kinked sideways (Fig. 3). A hinge region for the N-terminal movement can be assigned to the RGD motif, which was similarly found to be flexible in our previous study (Barden *et al.*, 2013).

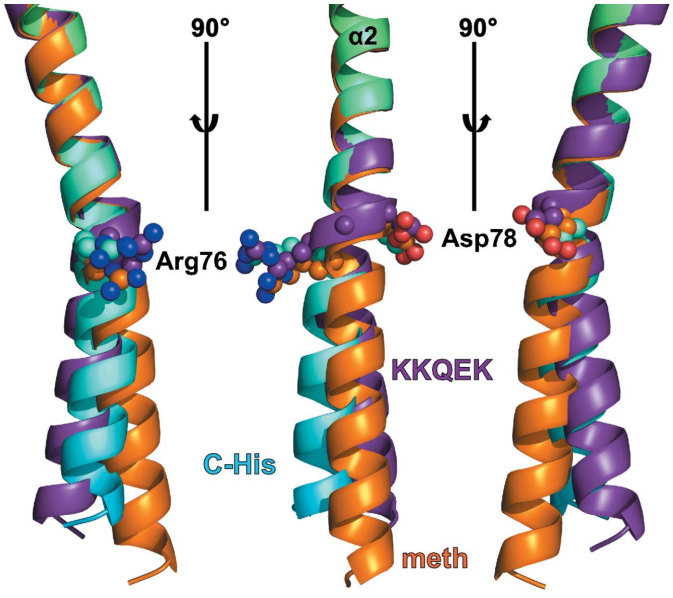

Figure 3
Flexibility of the α2 N-terminal region. Structures of the three CagL variants CagL$^{C\text{-}His}$ (light blue and light green), CagL^{meth} (orange) and CagL^{KKQEK} (purple) were aligned on the core region (amino acids 80–114, 144–166 and 187–219) to highlight structural differences in the N-terminal region of α2. The RGD motif acts as hinge. Side chains of Arg76, Gly77 and Asp78 from the RGD motif are shown as spheres.

3.4. CagL dimerizes upon refolding and under crystallization conditions

After refolding, CagL eluted from size-exclusion chromatography (SEC) in two peaks with apparent molecular masses of ∼50 and ∼75 kDa compared with a reference curve of globular proteins (Fig. 4*a*). Given the nonglobular, elongated shape of CagL, we assumed that these peaks represent monomeric and dimeric species (the theoretical molecular mass of the monomer is ∼25 kDa). The dimer peak contained 10–40% of the total refolded protein based on integration of the absorption peak area. Roughly half of the dimers were mediated by intermolecular disulfide bridges as judged by nonreducing SDS–PAGE, whereas the others ran identically to the presumed monomer. Following incubation with 10 m*M* dithiothreitol (DTT), the latter species were separated by SEC and snap-frozen. These dimers were not stable towards snap-freezing and partially converted to monomeric CagL (Fig. 4*b*). In contrast, we did not observe the conversion of purified monomeric CagL to dimeric protein upon freezing.

Based on the crystal structure of CagL$^{C\text{-}His}$, we wondered whether monomeric CagL converts to dimeric protein under the nonphysiological conditions within the crystallization drop. To this end, N-terminally His$_6$-tagged CagL^{wt} and *Tobacco etch virus* (TEV) protease-cleaved CagL^{wt} (lacking the N-terminal His$_6$ tag) were incubated for three weeks in ∼40% 2-methyl-2,4-pentanediol (MPD) at pH 4.5 and pH 8 and analyzed by SEC. No dimerization was observed for His$_6$-tagged CagL^{wt}, whereas TEV protease-cleaved CagL^{wt} dimerized to about 40% at pH 4.5 and in minor amounts at pH 8. Nonreducing SDS–PAGE confirmed the absence of disulfide-linked dimers. These results suggest that CagL$^{C\text{-}His}$ might have undergone dimerization to the domain-swapped dimer in the crystallization drop prior to crystallization. Interestingly, dimerization of CagL occurred without cleavage of α1.

3.5. The C-terminus of CagL$^{C\text{-}His}$ is helical and accessible

The C-terminus of CagL harbouring a conserved S/T-K-I/V-I-V-K hexapeptide has been shown to be important for pilus assembly (Shaffer *et al.*, 2011). In the crystal structures of CagL^{KKQEK} and CagL^{meth} there was no electron density for this hexapeptide (Barden *et al.*, 2013). In contrast, the C-terminus of CagL$^{C\text{-}His}$ is resolved up to the first histidine (His240) of the artificial C-terminal LEH$_6$ tag (Fig. 5*a*). The C-terminus is α-helical and protrudes beyond the helical bundle. Lys233 and Lys237 point into one direction and form a positively charged patch on the surface of CagL together with

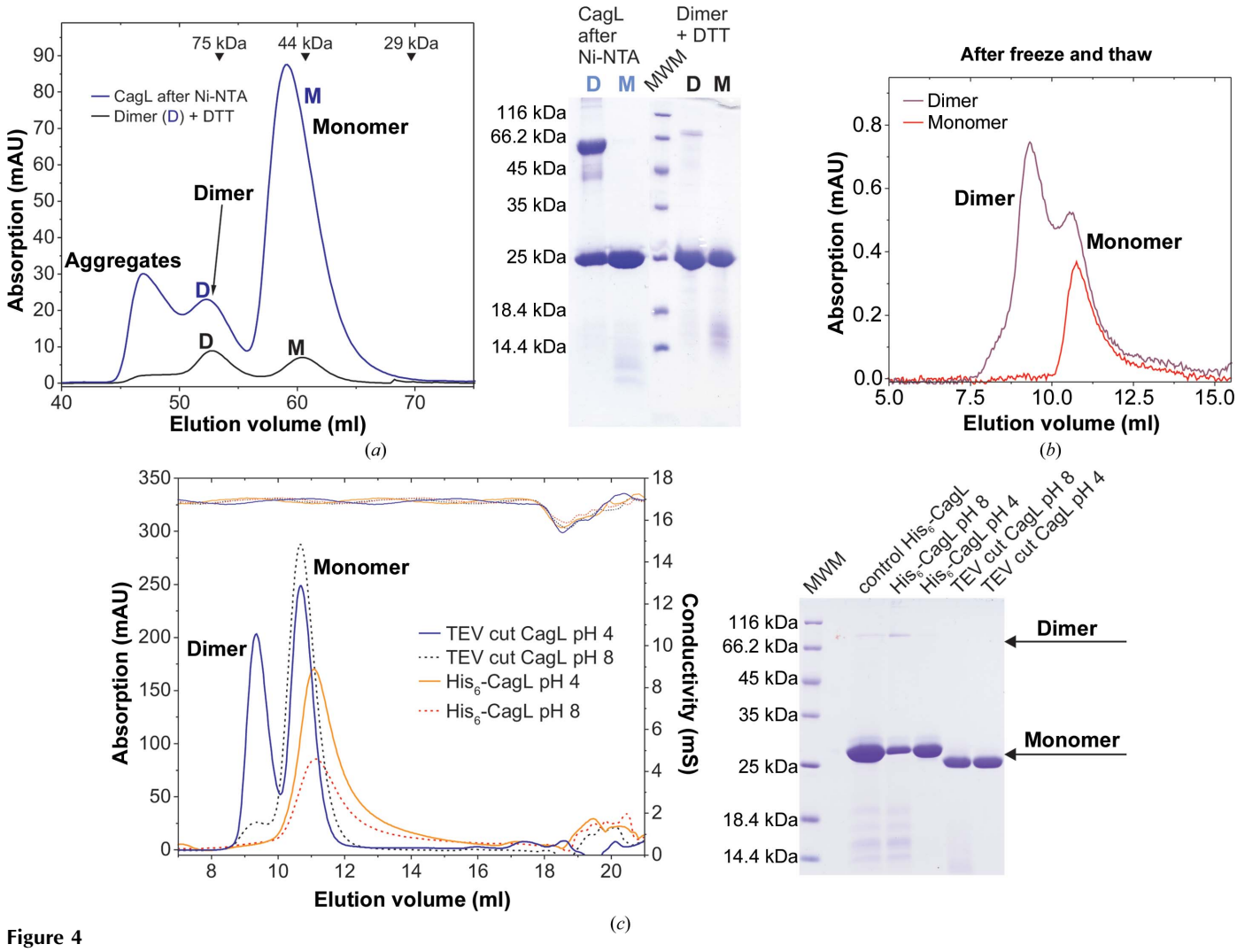

Figure 4
CagL forms dimers upon refolding and in the crystallization cocktail. (*a*) Preparative size-exclusion chromatography (SEC) of CagL revealed monomers, dimers and aggregates after refolding of CagL by dialysis and Ni–NTA affinity chromatography (blue curve). Fractions containing the dimer (D) were incubated with 10 m*M* DTT and rerun on the same column (black curve). Nonreducing SDS–PAGE revealed that no disulfide-linked dimers remained after the addition of DTT. (*b*) Analytical SEC of CagL monomer and noncovalent CagL dimer showed that the dimer dissociated partly after a single freeze–thaw cycle. (*c*) ∼5 mg ml^{-1} monomeric N-His$_6$-CagL^{wt} and monomeric TEV-cleaved CagL^{wt} were incubated for 20 d in 40%(*v*/*v*) MPD, 100 m*M* citrate–phosphate buffer pH 4.5 and pH 8 at 4°C and analyzed by analytical SEC. N-His$_6$-CagL^{wt} remained monomeric, while TEV-cleaved CagL^{wt} showed a small dimer fraction at pH 8 and about 50% dimer at pH 4.5. Nonreducing SDS–PAGE revealed no disulfide-linked dimers.

Lys222, Arg223 and Arg229 (Fig. 5*b*). The helical structure might not represent the native conformation of the hexapeptide, as the C-terminus is involved in crystal contacts with two symmetry mates. In solution, the C-terminus of CagL is susceptible to proteolysis, indicating an extended conformation. Upon limited proteolysis with trypsin, the C-terminal His$_6$ tag was cleaved within 5 min as revealed by anti-penta-histidine immunoblotting. In contrast, residual amounts of full-length CagL^{wt} with an N-terminal His$_6$ tag of 26 amino acids were observed even after 2 h of incubation (Fig. 5*c*).

4. Discussion

4.1. The CagL RGD motif acts as a hinge at the border of a rigid structural core

Including the structure described here, three crystal structures of CagL have been described: (i) CagL$^{C\text{-}His}$, native CagL with a C-terminal His$_6$ tag; (ii) CagL^{meth}, a variant of CagL with chemically methylated primary amines that contained six molecules per asymmetric unit; and (iii) CagL^{KKQEK}, a surface-entropy reduction variant. The latter two variants were generated to facilitate crystallization (Barden *et al.*, 2013). Hence, there are eight crystallographically independent molecules. The domain swap represents the main structural difference between CagL$^{C\text{-}His}$ and the two CagL variants. The structural core region previously identified by comparing the seven crystallographically independent molecules of CagL^{meth} and CagL^{KKQEK} fits strikingly well to the hydrophobic core of CagL$^{C\text{-}His}$ formed by α2, α3, α5 and a symmetry-related α6′. The structure of CagL$^{C\text{-}His}$ confirms the previously described flexibility around the RGD motif. Acting as a hinge, the RGD motif allows the N-terminus of α2 to approach α5 more closely. Within the two extremes represented by a helical RGD motif embedded in a straight α2 (chain *E* of CagL^{meth})

and a nonhelical RGD motif within a kinked α2 (CagL^{KKQEK}), CagL$^{C\text{-}His}$ adopts an intermediate conformation with a helical RGD motif and a kinked α2. The structure of CagL$^{C\text{-}His}$ thus confirms our previous findings but now for native protein without any surface modifications that could potentially alter the structure. We conclude that the structural differences between the three CagL variants are owing to crystal-packing forces that induce rearrangements in flexible regions of CagL rather than to the exchange or modification of individual surface-exposed side chains.

4.2. Energetic basis and biological significance of CagL dimerization

For the crystallization of CagL^{meth} and CagL^{KKQEK}, which both crystallized as monomers, we used only protein from the presumed monomer peak of preparative SEC (Barden *et al.*, 2013). In contrast, we had not purified CagL$^{C\text{-}His}$ by gel filtration prior to crystallization. CagL$^{C\text{-}His}$ was the first construct of CagL that we had worked with, and we were unaware of heterogeneous oligomerization at this point in time. However, dimers of CagL$^{C\text{-}His}$ may not only have formed upon refolding. As we have shown here, CagL dimers may also have formed from monomeric protein under the conditions employed for crystallization. This would require the opening of the hydrophobic core. Interestingly, CagL is stable to only about 42°C under physiological conditions, but its stability increases at lower pH (Choudhari *et al.*, 2013; Barden *et al.*, 2013). This suggests that the pH and the temperature applied for crystallization of CagL$^{C\text{-}His}$ may not be sufficient to induce dimerization. More likely, opening of the CagL hydrophobic core could be driven by the high concentration of MPD. At low concentration, MPD stabilizes proteins by binding to hydrophobic surfaces, releasing the trapped water (Anand *et al.*, 2002). In contrast, high concentrations of MPD can destabilize proteins under certain pH conditions (Arakawa *et al.*, 1990). A pH dependency was indeed found for CagL dimerization. Partially unfolded CagL might be stabilized by MPD shielding hydrophobic amino acids from water. Finally, dimerization of CagL would release the MPD molecules and result in the three-dimensional domain-swapped low-energy state. The loose association of the N-terminal helix may additionally favour partial unfolding of CagL because it provides access to the hydrophobic core upon dissociation (Barden *et al.*, 2013).

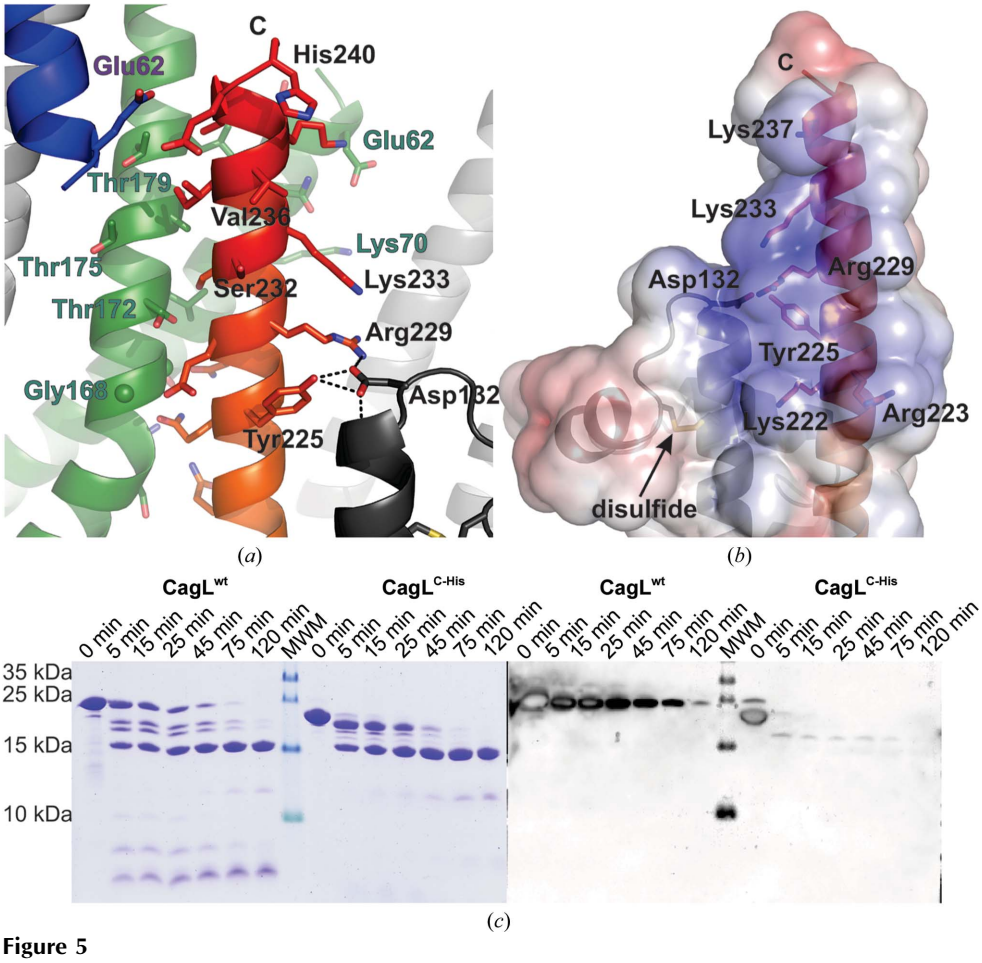

Figure 5
Helical structure of the conserved CagL C-terminus. (*a*) The C-terminus of CagL that extends beyond α5 (coloured red as in Fig. 1*a*) is helical and contacts two symmetry-related molecules (shown in green and blue). (*b*) Electrostatic surface potential of the CagL C-terminus. The view is rotated 160° relative to (*a*). Amino acids Lys222, Arg223, Arg229, Lys233 and Lys237 form a basic patch directly adjacent to the disulfide bridge that is conserved between CagL, CagI and CagH. (*c*) Coomassie-stained SDS–PAGE and anti-His$_6$ Western blot of a limited proteolysis of CagL^{wt} (N-terminal His$_6$ tag) or CagL$^{C\text{-}His}$ (C-terminal His$_6$ tag) with trypsin. For both constructs, a protease-resistant fragment of the same size is formed. The N-terminal His$_6$ tag is only slowly degraded over a time course of 2 h. The C-terminal His$_6$ tag can no longer be detected in the Western blot after only 5 min.

Both the formation of new interactions between the protomers of a dimer and the reorientation of the hinge region contribute to the energy stabilizing the domain-swapped dimer. However, for almost any interaction between the two protomers in dimeric CagL there is an equivalent interaction in the monomeric form. Only a few amino acids form new interprotomer interactions, *e.g.* Gln178 and Thr179, which participate in the loop-to-helix transition of the hinge loop. The energy gain from formation of the open interface might thus be quite small. In contrast, the energy gain from reorientation of the hinge region could be larger. Upon the rearrangement of the α5 C-terminus and the loop connecting

α5 and α6 into an α-helical segment, several new intra-main-chain hydrogen bonds are formed. Comparison of the main-chain hydrogen-bond network of CagL$^{\text{C-His}}$ with that of the equally well resolved CagL$^{\text{KKQEK}}$ structure reveals at least six new main-chain hydrogen bonds. Concomitantly, water molecules satisfying the hydrogen-bonding potential of these main-chain atoms in monomeric CagL are released. The low concentration of water in 40%(*v*/*v*) MPD may thus shift the equilibrium to the dimeric form.

The biological relevance of three-dimensional domain swapping is diverse (Bennett *et al.*, 2006; Liu & Eisenberg, 2002; Rousseau *et al.*, 2012). The pilus protein CagL is a T4SS adhesin targeting host-cell integrins (Conradi, Huber *et al.*, 2012; Kwok *et al.*, 2007; Tegtmeyer *et al.*, 2011; Wiedemann *et al.*, 2012). It has been shown that CagL is capable of mediating cell adhesion in its monomeric form (Barden *et al.*, 2013). Owing to the low stability of the domain-swapped dimer, we were not able to generate reliable functional data for dimeric CagL. Dimerization by domain swapping might increase the binding affinity of CagL for integrins owing to an avidity effect. A similar mechanism is found in the adhesion-mediating extracellular matrix molecule fibronectin, which forms elongated heterodimers (Pankov & Yamada, 2002). However, binding of two rather large integrin molecules (~220 kDa) to a CagL dimer may be unlikely as the distance between the two RGD motifs is only 43 Å (the distance between Gly77 C$^{\alpha}$ atoms).

A physiological relevance of CagL dimerization appears to be questionable for several reasons: (i) CagL dimers arise under nonphysiological conditions whereas (ii) no spontaneous dimerization was found in physiological solutions; (iii) the majority of CagL forms monomers upon refolding and only a small fraction of domain-swapped dimers is found which are (iv) not stable upon freezing in aqueous solutions. Unfortunately, analyzing the dimerization/oligomerization of CagL *in vivo* is hardly possible as the amount of CagL naturally expressed in *H. pylori* is low. CagL was not detected in two-dimensional gel electrophoresis of *H. pylori* cell lysate (Busler *et al.*, 2006; Backert *et al.*, 2005) and immuno-affinity enrichment of CagL with polyclonal antisera co-purifies CagI and CagH (among others; Shaffer *et al.*, 2011; Pham *et al.*, 2012; Kutter *et al.*, 2008). To the best of our knowledge, to date

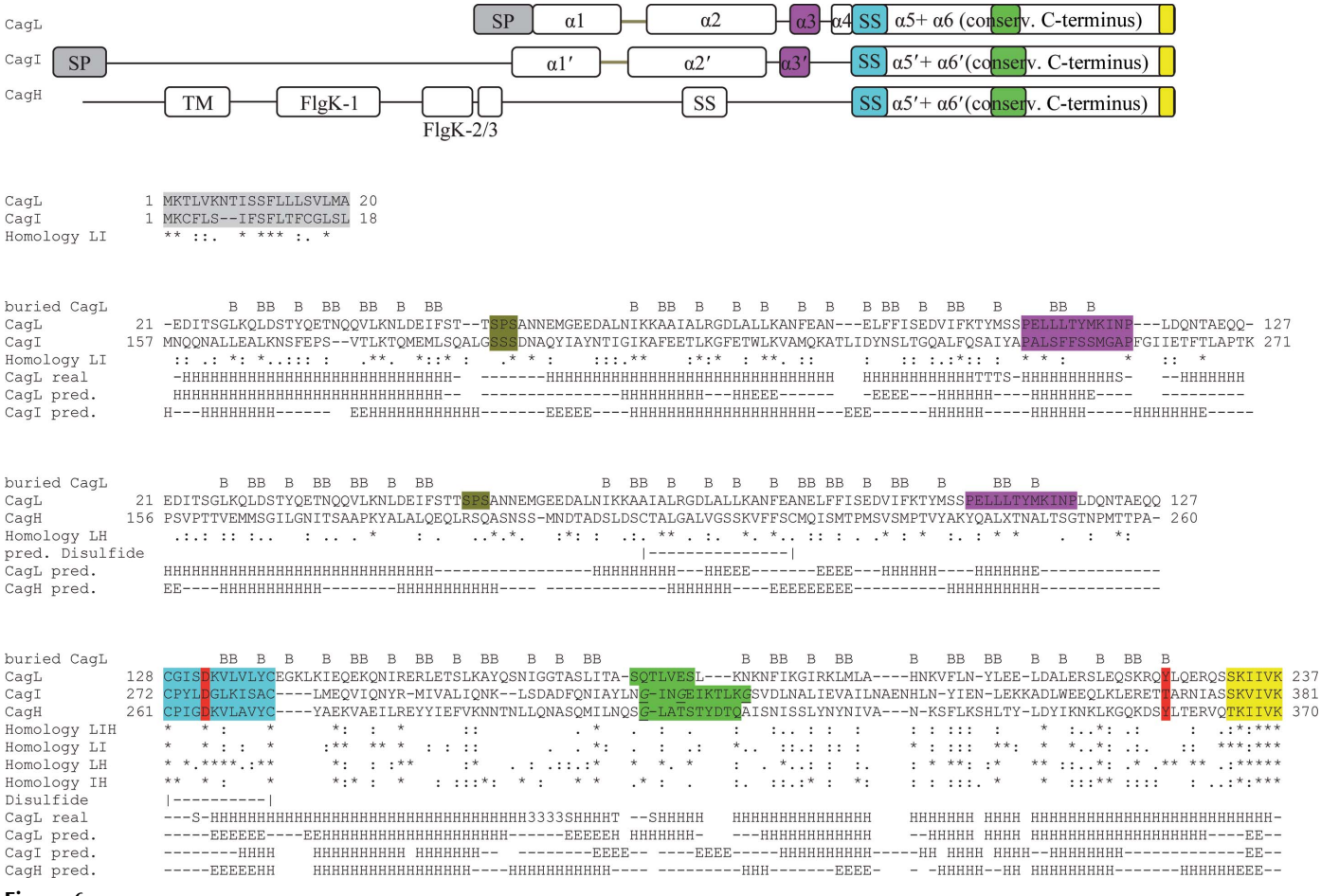

Figure 6
Comparison and sequence alignment of CagL, CagI and CagH. Upper panel: sequence features, secondary structure of CagL and secondary-structure prediction of CagI and CagH [SP, signal peptide; SS, (predicted) disulfide bond; TM, transmembrane helix; FlgK1/2/3, flagellar hook-associated protein K domain (Shaffer *et al.*, 2011); α, α-helix]. Lower panel: sequence alignment of CagL, CagI and CagH. For clarity, the sequence alignment was split into three parts: the predicted signal peptide of CagL and CagI; the middle part of CagL, aligned separately with CagI and CagH; and the C-terminus of the three proteins. The alignment was generated as described in §2. Buried amino acids of CagL are indicated with 'B'. Structurally important parts are highlighted, *e.g.* a probable hinge region between α5 and α6 is highlighted in light green. For a detailed description, see §3.

purification of CagL from *H. pylori* has not been reported in a purity and quantity that would allow its native oligomeric state in the context of the T4SS to be addressed.

4.3. CagL may serve as a structural template for CagI and CagH

Based on sequence alignment (Fig. 6), we hypothesize that CagL, CagI and CagH share a structurally similar C-terminus beginning at the disulfide (Cys128 and Cys139 of CagL). This disulfide is highly conserved in CagI and CagH, with mostly conservative mutations in between, and exactly 109 amino acids from the C-terminal Lys. A conserved aspartate (Asp132 of CagL) at position Cys1+4 forms a hydrogen bond to a tyrosine at the C-terminus of CagL (Tyr225), which is also found in CagH. Thus, the disulfide is directly connected to the C-terminus, which exhibits the highly conserved S/T-K-I/V-I-V-K hexapeptide (Shaffer *et al.*, 2011). Structural conservation often indicates potential binding sites, and thus the arrangement supports the hypothesis of a C-terminal translocation signal. In *H. pylori*, a strictly conserved C-terminal translocation signal has not been identified, but an accumulation of positively charged amino acids arranged as R/K-X_3-R/K or R/K-X_4-R/K is found in several T4SS proteins (Hohlfeld *et al.*, 2006; Vergunst *et al.*, 2005). Deletion of the C-terminal residues leads to a dysfunctional assembly of the secretion needle or mislocalization of the proteins (Shaffer *et al.*, 2011). However, it does not affect the adhesion of MKN-45 cells to CagL, as shown previously (Barden *et al.*, 2013).

The amino-acid sequences of CagH, CagI and CagL exhibit a high degree of conservation of fold-relevant residues and are predicted to be helical from the disulfide bridge to the C-termini. It is likely that the C-termini of CagI and CagH adopt an arrangement of two helices connected by a short linker, similar to CagL. Looking for further homologies, we also found a high degree of conservation of fold-relevant residues N-terminal to the conserved disulfide bond. For example, the two prolines (Pro107 and Pro118) confining helix α3 in CagL are conserved in CagI and a loop region connecting helices α1 and α2 in CagL can also be assigned. In line with this, the secondary-structure prediction for CagI is mostly helical with exception of the predicted loop region. Interestingly, the N-terminal signal peptides of CagL and CagI also possess a high degree of conservation. However, these homologies are not found between CagL and CagH. CagH does not exhibit the N-terminal signal peptide or the two prolines. Instead, a second disulfide bond between Cys206 and Cys222 and a transmembrane helix from Val29 to Gly51 are predicted (Kutter *et al.*, 2008). The proteins CagL, CagI and CagH may thus have emerged from one common ancestor by gene duplication and then diversified to fulfill specialized functions in type IV secretion of *H. pylori*. The observed three-dimensional domain swap in CagL and our suggestion that helices α5 and α6 may also be present in CagH and CagI raises the tantalizing question of whether an exchange of α6 might be involved in the formation of heterodimers or higher oligomers of CagL, CagH and CagI, which have been observed in the context of the intact type IV secretion pilus by co-immunoprecipitation and mass spectrometry (Shaffer *et al.*, 2011).

We gratefully acknowledge access to beamline ID14-4 at ESRF, Grenoble, France and the coverage of travel expenses through BAG MX-926. We thank Roman Fedorov (Hannover Medical School) for help with data collection.

References

Adams, P. D. *et al.* (2010). *Acta Cryst.* D**66**, 213–221.
Anand, K., Pal, D. & Hilgenfeld, R. (2002). *Acta Cryst.* D**58**, 1722–1728.
Arakawa, T., Bhat, R. & Timasheff, S. N. (1990). *Biochemistry*, **29**, 1924–1931.
Backert, S., Clyne, M. & Tegtmeyer, N. (2011). *Cell Commun. Signal.* **9**, 28.
Backert, S., Kwok, T., Schmid, M., Selbach, M., Moese, S., Peek, R. M., König, W., Meyer, T. F. & Jungblut, P. R. (2005). *Proteomics*, **5**, 1331–1345.
Baker, N. A., Sept, D., Joseph, S., Holst, M. J. & McCammon, J. A. (2001). *Proc. Natl Acad. Sci. USA*, **98**, 10037–10041.
Barden, S., Lange, S., Tegtmeyer, N., Conradi, J., Sewald, N., Backert, S. & Niemann, H. H. (2013). *Structure*, **21**, 1931–1941.
Bendtsen, J. D., Nielsen, H., von Heijne, G. & Brunak, S. (2004). *J. Mol. Biol.* **340**, 783–795.
Bennett, M. J., Choe, S. & Eisenberg, D. (1994). *Proc. Natl Acad. Sci. USA*, **91**, 3127–3131.
Bennett, M. J., Sawaya, M. R. & Eisenberg, D. (2006). *Structure*, **14**, 811–824.
Bennett, M. J., Schlunegger, M. P. & Eisenberg, D. (1995). *Protein Sci.* **4**, 2455–2468.
Busler, V. J., Torres, V. J., McClain, M. S., Tirado, O., Friedman, D. B. & Cover, T. L. (2006). *J. Bacteriol.* **188**, 4787–4800.
Carey, J., Lindman, S., Bauer, M. & Linse, S. (2007). *Protein Sci.* **16**, 2317–2333.
Censini, S., Lange, C., Xiang, Z., Crabtree, J. E., Ghiara, P., Borodovsky, M., Rappuoli, R. & Covacci, A. (1996). *Proc. Natl Acad. Sci. USA*, **93**, 14648–14653.
Ceroni, A., Passerini, A., Vullo, A. & Frasconi, P. (2006). *Nucleic Acids Res.* **34**, W177–W181.
Choudhari, S. P., Pendleton, K. P., Ramsey, J. D., Blanchard, T. G. & Picking, W. D. (2013). *J. Pharm. Sci.* **102**, 2508–2519.
Cole, C., Barber, J. D. & Barton, G. J. (2008). *Nucleic Acids Res.* **36**, W197–W201.
Conradi, J., Huber, S., Gaus, K., Mertink, F., Royo Gracia, S., Strijowski, U., Backert, S. & Sewald, N. (2012). *Amino Acids*, **43**, 219–232.
Conradi, J., Tegtmeyer, N., Woźna, M., Wissbrock, M., Michalek, C., Gagell, C., Cover, T. L., Frank, R., Sewald, N. & Backert, S. (2012*b*). *Front. Cell. Infect. Microbiol.* **2**, 70.
Crestfield, A. M., Stein, W. H. & Moore, S. (1962). *Arch. Biochem. Biophys.* **Suppl. 1**, 217–222.
Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. (2010). *Acta Cryst.* D**66**, 486–501.
Evans, P. (2006). *Acta Cryst.* D**62**, 72–82.
Fischer, W., Püls, J., Buhrdorf, R., Gebert, B., Odenbreit, S. & Haas, R. (2001). *Mol. Microbiol.* **42**, 1337–1348.
Griebenow, K. & Klibanov, A. M. (1995). *Proc. Natl Acad. Sci. USA*, **92**, 10969–10976.
Hohlfeld, S., Pattis, I., Püls, J., Plano, G. V., Haas, R. & Fischer, W. (2006). *Mol. Microbiol.* **59**, 1624–1637.
Joosten, R. P., te Beek, T. A. H., Krieger, E., Hekkelman, M. L., Hooft, R. W. W., Schneider, R., Sander, C. & Vriend, G. (2010). *Nucleic Acids Res.* **39**, D411–D419.
Kabsch, W. (1976). *Acta Cryst.* A**32**, 922–923.

Kabsch, W. (2010). *Acta Cryst.* D**66**, 125–132.
Kantardjieff, K. A. & Rupp, B. (2003). *Protein Sci.* **12**, 1865–1871.
Krissinel, E. & Henrick, K. (2007). *J. Mol. Biol.* **372**, 774–797.
Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. L. (2001). *J. Mol. Biol.* **305**, 567–580.
Kutter, S., Buhrdorf, R., Haas, J., Schneider-Brachert, W., Haas, R. & Fischer, W. (2008). *J. Bacteriol.* **190**, 2161–2171.
Kwok, T., Zabler, D., Urman, S., Rohde, M., Hartig, R., Wessler, S., Misselwitz, R., Berger, J., Sewald, N., König, W. & Backert, S. (2007). *Nature (London)*, **449**, 862–866.
Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J. & Higgins, D. G. (2007). *Bioinformatics*, **23**, 2947–2948.
Liu, Y. & Eisenberg, D. (2002). *Protein Sci.* **11**, 1285–1299.
McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). *J. Appl. Cryst.* **40**, 658–674.
Painter, J. & Merritt, E. A. (2006). *J. Appl. Cryst.* **39**, 109–111.
Pankov, R. & Yamada, K. M. (2002). *J. Cell Sci.* **115**, 3861–3863.
Pham, K. T., Weiss, E., Jiménez Soto, L. F., Breithaupt, U., Haas, R. & Fischer, W. (2012). *PLoS One*, **7**, e35341.
Rousseau, F., Schymkowitz, J. W. H. & Itzhaki, L. S. (2003). *Structure*, **11**, 243–251.
Rousseau, F., Schymkowitz, J. & Itzhaki, L. S. (2012). *Adv. Exp. Med. Biol.* **747**, 137–152.
Rousseau, F., Schymkowitz, J. W., Wilkinson, H. R. & Itzhaki, L. S. (2001). *Proc. Natl Acad. Sci. USA*, **98**, 5596–5601.
Schägger, H. (2006). *Nature Protoc.* **1**, 16–22.
Schlunegger, M. P., Bennett, M. J. & Eisenberg, D. (1997). *Adv. Protein Chem.* **50**, 61–122.
Shaffer, C. L., Gaddy, J. A., Loh, J. T., Johnson, E. M., Hill, S., Hennig, E. E., McClain, M. S., McDonald, W. H. & Cover, T. L. (2011). *PLoS Pathog.* **7**, e1002237.
Smolka, A. J. & Backert, S. (2012). *J. Gastroenterol.* **47**, 609–618.
Tegtmeyer, N., Wessler, S. & Backert, S. (2011). *FEBS J.* **278**, 1190–1202.
Vergunst, A. C., van Lier, M. C. M., den Dulk-Ras, A., Grosse Stüve, T. A., Ouwehand, A. & Hooykaas, P. J. J. (2005). *Proc. Natl Acad. Sci. USA*, **102**, 832–837.
Wiedemann, T., Hofbaur, S., Tegtmeyer, N., Huber, S., Sewald, N., Wessler, S., Backert, S. & Rieder, G. (2012). *Gut*, **61**, 986–996.
Winn, M. D. *et al.* (2011). *Acta Cryst.* D**67**, 235–242.