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

# Structure of the N-terminal domain of human CEACAM1: binding target of the opacity proteins during invasion of *Neisseria meningitidis* and *N. gonorrhoeae*

**Alena Fedarovich,[a] Joshua Tomberg,[b] Robert A. Nicholas[b]* and Christopher Davies[a]***

[a]Department of Biochemistry and Molecular Biology, Medical University of South Carolina, Charleston, SC 29425, USA, and [b]Department of Pharmacology, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, USA

Correspondence e-mail: nicholas@med.unc.edu, davies@musc.edu

Received 15 May 2006
Accepted 1 June 2006

**PDB Reference:** N-terminal domain of human CEACAM1, 2gk2, r2gk2sf.

CEACAM1 is a cellular adhesion molecule whose protein expression is down-regulated in several carcinomas and which also contributes to the pathogenicity of *Neisseria* by acting as a receptor for Opa proteins. The crystal structure of the N-terminal (D1) domain of human CEACAM1 has been determined at 2.2 Å resolution. The structure shows several differences compared with a lower resolution model of the same domain from mouse solved previously, especially in the functional regions. Mapping of the sites of mutations that lower or abolish the binding of CEACAM1 to Opa proteins shows a distinct clustering of residues on the *GFCC′C″* face of the molecule. Prominent amongst these are residues in the *C*, *C′* and *F* strands and the *CC′* loop. A similar analysis shows that the region responsible for homophilic or heterophilic interactions of CEACAM1 is also on the *GFCC′C″* face and overlaps partially with the Opa-binding region. This higher resolution structure of CEACAM1 will facilitate a more precise dissection of its functional regions in the context of neisserial pathogenesis, cellular adhesion and immune evasion.

## 1. Introduction

Carcinoembryonic antigen-related cell-adhesion molecules (CEACAMs) belong to the CEA subgroup of the carcinoembryonic antigen family of immunoglobulins (Ig). There are seven members of the CEACAM family in humans (CEACAM1 and CEACAMs3–8), but only CEACAM1, CEACAM5, CEACAM6 and CEACAM8 appear to function as adhesion molecules (Rojas *et al.*, 1990; Benchimol *et al.*, 1989; Oikawa *et al.*, 1989, 1991). CEACAMs are differentially expressed on the surface of various leukocytic, epithelial and endothelial cells, where they are involved in a wide variety of normal and pathological processes, including cancer, bacterial and viral infection and inflammation (for a review, see Hammarstrom, 1999).

Each CEACAM family member consists of an N-terminal domain of 108–110 amino-acid residues that exhibits the fold of an immunoglobulin variable (V set) domain, followed by a differing number of Ig constant-like domains. CEACAMs are anchored to the cell surface either by transmembrane/cytoplasmic domains (CEACAM1–4) or by glycophosphatidylinositol (GPI) moieties (CEACAM5–8).

CEACAM1 (also known as CD66a or BGP) is the most broadly distributed member of the CEA family. This variant of CEACAM1 contains six domains: the N-terminal domain (N), three constant Ig-like domains (A1, B and A2), a transmembrane domain (TM) and a so-called longer cytoplasmic domain (L). There are also other splice variants of CEACAM1 with differing arrangements of the constant domains or with a short cytoplasmic tail (Beauchemin *et al.*,

1999). The primary function of CEACAM1 is as an adhesion molecule, whereby it interacts with itself or with other members of the CEA family. A number of studies have shown that this adhesion activity resides in the N-terminal domain of the protein (Teixeira *et al.*, 1994; Watt *et al.*, 2001; Markel *et al.*, 2004; Stern *et al.*, 2005), which is highly conserved within the CEACAM family. Homophilic interactions of human CEACAM1, as well as its heterophilic binding with CEACAM5, inhibit the killing activity of NK cells, thereby modulating immune response (Markel, Lieberman *et al.*, 2002; Markel, Wolf *et al.*, 2002; Stern *et al.*, 2005).

CEACAM1 appears to play a role as a tumor suppressor as its expression is down-regulated in several carcinomas, including those of prostate (Kleinerman *et al.*, 1995; Pu *et al.*, 1999), colon (Nollau, Prall *et al.*, 1997; Nollau, Scheller *et al.*, 1997), breast (Riethdorf *et al.*, 1997; Huang *et al.*, 1998) and endometrium (Bamberger *et al.*, 1998). The cellular mechanism by which CEACAM1 acts as a tumor suppressor depends on the source of the carcinoma. In the case of colon and breast cancers, the absence of CEACAM1 leads to reduced levels of apoptosis (Nittka *et al.*, 2004; Kirshner *et al.*, 2003), whereas in prostate cancer the presence of CEACAM1 appears to reduce angiogenesis (Volpert *et al.*, 2002). Interestingly, the protein is associated with increased vascularization of mouse endothelial cells (Horst *et al.*, 2006).

Finally, along with some other CEACAMs, CEACAM1 is a receptor for *Neisseria gonorrhoeae* and *N. meningitidis* opacity proteins (Opa) during invasion and infection (Virji, Makepeace *et al.*, 1996; Virji, Watt *et al.*, 1996; Bos *et al.*, 1997; Gray-Owen *et al.*, 1997). Opa proteins are integral outer membrane proteins that play an important role in neisserial pathogenesis. Although they are not essential for initial host colonization, there is strong selection pressure for expression of Opa proteins during infection *in vivo* (Jerse *et al.*, 1994). Over 95% of gonococcal and meningococcal isolates are able to bind to CEACAM1 (Virji, Watt *et al.*, 1996). The binding site for opacity-associated (Opa) proteins is in the N-terminal domain of the protein (Bos *et al.*, 1998; Virji *et al.*, 1999; de Jonge *et al.*, 2003). Infection by *N. gonorrhoeae* suppresses antibody production through CEACAM1–Opa interactions by killing B cells (Pantelic *et al.*, 2005) and by lowering the activation and proliferation of $CD4^+$ T cells (Boulton & Gray-Owen, 2002; Normark *et al.*, 2002). Such immune suppression may explain in part the lack of immunological memory that leads to repeated gonococcal infections. Mutagenesis experiments on CEACAM1 suggest that Opa proteins bind to the non-glycosylated GFC face of the protein and for all Opa variants tested requires residues Tyr34 and Ile91 (Virji *et al.*, 1999). Other residues in the N-terminal domain, such as Ser32, Phe29 and Gln44, appear only to be required for interaction with specific Opa proteins (Virji *et al.*, 1999; Popp *et al.*, 1999; Bos *et al.*, 1999).

Although there are several predicted structures of human CEACAM1 (Bates *et al.*, 1992; Teixeira *et al.*, 1994; Watt *et al.*, 2001), only one crystal structure, of domains 1 and 4 of murine CEACAM1 solved at the relatively low resolution of 3.32 Å, is available (Tan *et al.*, 2002). Here, we report the crystal structure of the N-terminal domain of human CEACAM1 determined by X-ray crystallography at 2.2 Å resolution. This reveals the CEACAM1 structure in greater detail, shows some interesting differences from its murine homologue and suggests how this domain interacts with Opa proteins and other host receptors.

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

Values in parentheses are for the last resolution shell of data.

| | |
|---|---|
| Space group | $P6_3$ |
| Unit-cell parameters (Å) | $a = b = 88.2$, $c = 62.0$ |
| Resolution (Å) | 43.0–2.2 (2.28–2.20) |
| No. of observations | 140829 |
| No. of unique observations | 13366 |
| Redundancy | 10.5 (6.7) |
| Completeness (%) | 99.6 (96.9) |
| $I/\sigma(I)$ | 38.8 (2.0) |
| $R_{\rm merge}$† (%) | 6.7 (73.6) |
| No. of protein atoms | 1714 |
| No. of water molecules | 75 |
| $R_{\rm cryst}$ (%) | 21.3 |
| $R_{\rm free}$ (%) | 25.8 |
| R.m.s.d. bond lengths (Å) | 0.013 |
| R.m.s.d. bond angles (°) | 1.54 |
| Ramachandran statistics (%) | |
| Most favorable region | 86.4 |
| Additional allowed region | 12.5 |
| Generously allowed region | 1.1 |
| Disallowed region | 0.0 |
| Mean $B$ factor, all atoms (Å$^2$) | 41.5 |
| R.m.s. deviation in $B$ factors (main-chain atoms) (Å$^2$) | 0.37 |
| R.m.s. deviation in $B$ factors (side-chain atoms) (Å$^2$) | 1.12 |

† $R_{\rm merge} = \sum_{hkl}\sum_i |I_i(hkl) - \langle I(hkl)\rangle| / \sum_{hkl}\sum_i I_i(hkl)$.

## 2. Experimental

### 2.1. Protein expression and purification

A plasmid encoding the N-terminal D1 domain of human CEACAM1 [amino acids 1–107 of the mature protein; hCEACAM1(D1)] was obtained from Dr Martine Bos (Utrecht University; Bos *et al.*, 1998). The gene was amplified from this plasmid by PCR and cloned into the expression vector pGEX-2V. This plasmid is a derivative of pGEX-2T (GE Healthcare, Piscataway, NJ, USA) in which the thrombin-cleavage site linking GST and the fusion protein was replaced with the cleavage site (ENLYFQ↓G) for tobacco etch virus protease (TEV). In addition, the spacer amino acids GSGGA were included immediately after the TEV site and before amino acid 2 of CEACAM. The plasmid was maintained in *Escherichia coli* MC1061 cells for protein production.

Cells harboring the expression plasmid were grown in 6 l Luria–Bertani broth containing 50 µg ml$^{-1}$ ampicillin at 310 K until the $A_{600\,\rm nm}$ reached 0.6–0.7. Protein expression was induced with 0.5 m*M* isopropyl $\beta$-D-thiogalactopyranoside followed by overnight shaking at 298 K. The cells were harvested by centrifugation at 4500*g* for 15 min and resuspended in 20 m*M* Tris–HCl pH 8.0, 0.15 *M* NaCl, 2 m*M* EDTA and 10% glycerol pH 8.0. After addition of PMSF (0.06 mg ml$^{-1}$), the cells were lysed by sonication or emulsiflex

disruption, membranes and cell debris were removed by centrifugation at 100 000*g* for 1 h and proteins were precipitated from the supernatant by the addition of ammonium sulfate to 55% saturation followed by stirring for 1 h at 277 K. After centrifugation at 12 000*g* for 25 min, the pellet [containing GST-hCEACAM1(D1)] was resuspended in 30 ml 20 m*M* Tris–HCl pH 8.0, 0.15 *M* NaCl and 10% glycerol, centrifuged for 20 min at 12 000*g* and filtered through a 0.45 µm filter.

Purification of the fusion protein was performed using a GS-Trap HP column (GE Healthcare) equilibrated in 20 m*M* Tris pH 7.3, 0.15 *M* NaCl, 2 m*M* dithiothreitol (DTT) and 10% glycerol at 277 K. After sample loading, the column was washed with ten column volumes of starting buffer and the protein was eluted with 10 m*M* reduced glutathione in the same buffer. The purified fusion protein was then digested with TEV protease during overnight dialysis against 20 m*M* Tris pH 7.3, 0.15 *M* NaCl and 10% glycerol. hCEACAM1(D1) was purified from GST and TEV protease by passing the digestion mixture over a HR Sephacryl S-200 gel-filtration column (26/60 mm; GE Healthcare) equilibrated in 20 m*M* Tris–HCl pH 8.0, 0.5 *M* NaCl and 10% glycerol. Fractions containing pure hCEACAM1(D1) were pooled, concentrated to 11 mg ml$^{-1}$ (as measured by the Bradford assay) and stored at 193 K.

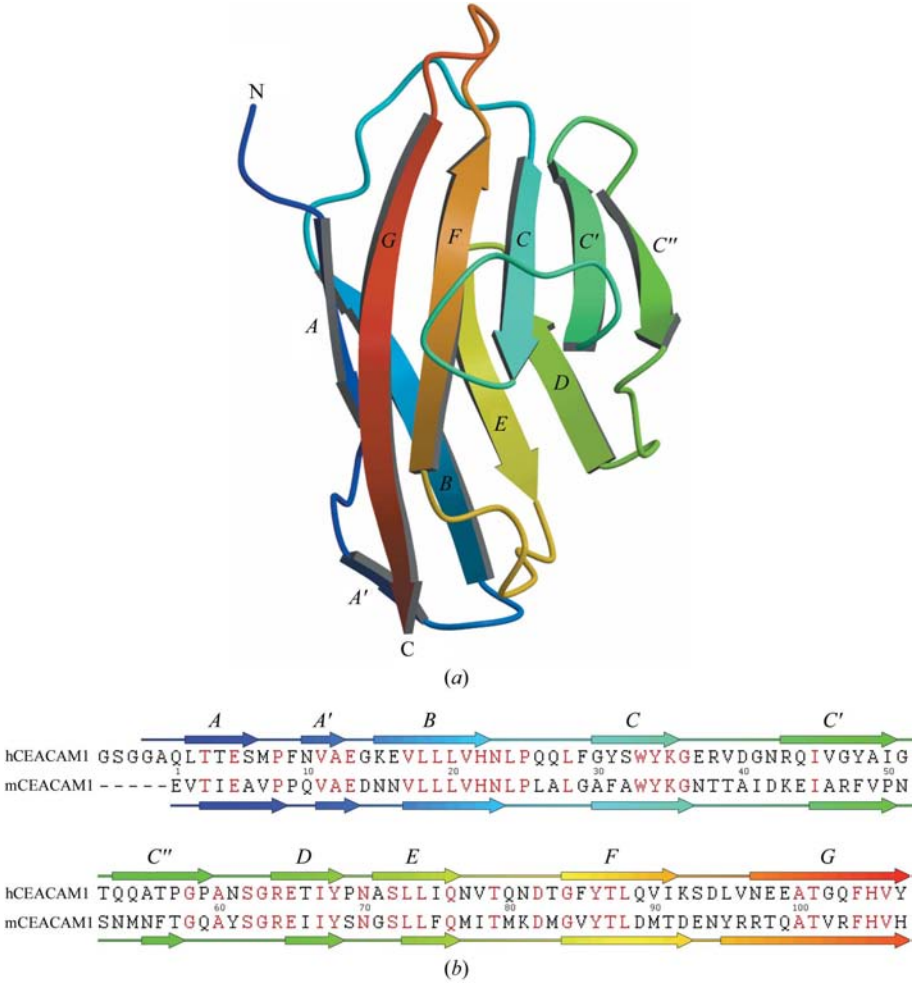


**Figure 1**
The structure of the D1 domain of human CEACAM1. (*a*) The structure is displayed in ribbon format showing the secondary structure and is color-ramped from blue to red in the N-terminal to C-terminal direction. The elements of secondary structure are labeled according to the convention of Chothia & Jones (1997). (*b*) An alignment of the sequences of the D1 domains of human CEACAM1 and mouse CEACAM1a, including the secondary-structure assignments for each molecule based on their respective structures.

### 2.2. Crystallization and X-ray data collection

hCEACAM1(D1) was subjected to a search for crystallization conditions, initially beginning with Crystal Screens 1 and 2 (Hampton Research, Aliso Viejo, CA, USA). Crystal trays were set up using the hanging-drop technique in which 2 µl protein was mixed with 2 µl well solution. After optimization, crystals were obtained over wells containing 16–18% polyethylene glycol monomethyl ether (PEG MME) 2000, 0.1 *M* Tris–HCl pH 8.5 and 5 m*M* nickel(II) chloride. Prior to freezing at 100 K, crystals were cryoprotected by passage through a solution containing 35% PEG MME 2000, 0.1 *M* Tris–HCl pH 8.5 and 10% glycerol. Preliminary diffraction analysis showed that the crystals belonged to point group *P*6, with unit-cell parameters $a = b = 88.2$, $c = 62.0$ Å. At this stage, however, the exact space group was not apparent from examination of the systematic absences. Estimation of the solvent content suggested that there were two molecules of hCEACAM1(D1) in the asymmetric unit (Matthews, 1968).

An initial data set used for molecular-replacement calculations was collected on an R-AXIS IV$^{++}$ detector mounted on an RU-H3R rotating-anode generator (Rigaku) operating at 50 kV and 100 mA. For these data, the crystal-to-detector distance was 150 mm and the exposure time was 3 min per 0.5° oscillation. 130° of data were collected and processed using *d*TREK* (Pflugrath, 1999). Later, X-ray diffraction data of higher resolution were collected on a MAR 300 CCD at the SER-CAT 22-ID beamline at the Advanced Photon Source, Argonne National Laboratory, Chicago, IL, USA. In this experiment, the crystal-to-detector distance was 260 mm, the exposure time was 4 s per frame, the wavelength was 1.00 Å and 225° of data were collected in increments of 1.0°. The data were integrated and scaled with *HKL*-2000 (Otwinowski & Minor, 1997). These data were used for refinement of the model.

### 2.3. Structure determination and refinement

The structure was determined by molecular replacement using *MOLREP* (Vagin & Teplyakov, 1997) with the N-terminal domain (D1) of mouse CEACAM1 as a search model (mCEACAM1a; Tan *et al.*, 2002; PDB code 1l6z), which is 41% identical in sequence. This calculation was performed at 3.0 Å using the initial data set collected on the 'home source' and all possible space groups in the $P6_x$ system were tested. A possible solution arose in space group $P6_3$, with two molecules in the asymmetric unit, an *R* factor of 53.3% and a correlation coefficient of 0.362. This peak was only marginally better than other peaks, but showed good packing of molecules in the crystal lattice with no clashes. The initial structure was refined using *REFMAC*5 (Murshudov *et al.*, 1997), after which the correct sequence was inserted into the model using the program *O* (Jones *et al.*, 1991). This model was refined using *CNS* (Brünger *et al.*, 1998) and the model was then improved with alternating rounds of manual revision using *O* and refinement by *REFMAC*5. The higher resolution synchrotron data were introduced in later rounds of refinement. 5% of the data was set aside throughout for calculation of the free *R* factor (Brünger, 1992) and these assignments were retained when switching data sets. The final model is refined to 2.2 Å resolution with an *R* factor of 21.3% (the free *R* factor is 25.8%; Table 1).

In the protein construct used, there are five non-native residues at the N-terminus (numbered Gly−4, Ser−3, Gly−2, Gly−1 and Ala0). Several of these are visible in the electron density: Gly−2, Gly−1 and Ala0 in molecule *A* and Ser−3, Gly−2, Gly−1 and Ala0 in molecule *B*. For both molecules, the model ends at the last amino acid of the construct (Tyr107) and its carboxylate is clearly visible in both molecules. Interestingly, a nickel ion derived from the crystallization solution is observed in the structure on a threefold axis of symmetry coordinated to three His105 residues derived from molecule *A* and three water molecules. Thus, a coordination shell is formed around the nickel that is comprised of three symmetry-related histidines and water molecules and accordingly the nickel and water molecules were each modeled at 33% occupancy during refinement. Since diffracting crystals could only be obtained in the presence of nickel, this interaction appears to be critical for organizing the protein molecules in the lattice. In comparison to the same histidine residue in mouse CEACAM1 (Tan *et al.*, 2002), the side chain has rotated approximately 120° to contact the nickel but otherwise this region is unchanged in structure. There is also additional positive electron density near His105 of molecule *B* that was also modeled as a nickel ion. Finally, a large region of positive electron density is also visible in a region of molecule *B* bordered by Ala0, Asn23, Pro25 and Glu98. Attempts to model Tris or glycerol molecules into this density were unsuccessful, leaving the identity of this molecule unknown.

The common main chains of the two molecules in the asymmetric unit of the crystal could be superimposed with a root-mean-square (r.m.s.) deviation of 0.39 Å, indicating that the molecules are essentially identical in structure. Only minor differences occur at the termini of the molecules.

## 3. Results and discussion

### 3.1. Structure description

hCEACAM1(D1) has the topology of a V-set fold of the immunoglobulin superfamily (IgSF; Wang & Springer, 1998). It comprises two $\beta$-sheets, which, following convention (Chothia & Jones, 1997), are labeled *ABDE* in one sheet and $A'GFCC'C''$ (*GFC* for short) in the other (Fig. 1*a*). The two $\beta$-sheets are linked by the *BC*, *EF*, $C''D$ and $AA'$ loops, which all cross over between the sheets. Both sheets are antiparallel, with the exception of the $\beta$-strands $A'$ and *G*. As found in other V-type IgSF folds, there is a salt bridge between Arg64 and Asp82 in hCEACAM1(D1) that tethers the two $\beta$-sheets by linking the loops preceding the $\beta$-strands *D* and *F*. In contrast, hCEACAM1(D1) lacks a disulfide bond connecting the $\beta$-strands *B* and *F* that occurs in many proteins of the Ig superfamily.

For the most part, the structure of hCEACAM1(D1) complies closely with the V and I set of IgSF folds, except that the $A'$ strand is significantly shorter than other representatives. An important feature of the structure with respect to the

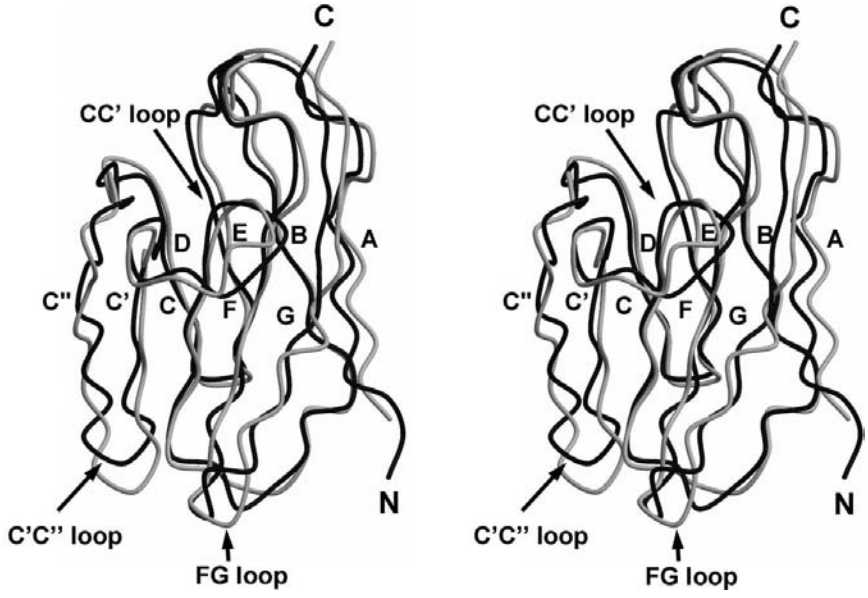


**Figure 2**
A comparison of the D1 domains of human CEACAM1 and mouse CEACAM1a. The two structures were superimposed and are shown in backbone form with hCEACAM1 in black and mCEACAM1a in gray. Regions of the two structures that differ between the two molecules are noted with arrows and are labeled.

function of hCEACAM1(D1) is the *CC′* loop. Rather than forming a conventional β-ribbon at the base of the *GFC* face (as viewed in Fig. 1*a*), as has been predicted in molecular models of CEACAM1 (Bates *et al.*, 1992; Virji *et al.*, 1999; Popp *et al.*, 1999; Watt *et al.*, 2001), the *CC′* loop folds back to lie across the face of the β-sheet, where it forms a distinct protrusion containing residues involved in binding interactions with Opa and other CEACAMs (see below). A very similar position for this loop was also observed in the structure of mCEACAM1a(D1,4) (Tan *et al.*, 2002).

### 3.2. Comparison with mCEACAM1a(D1)

The D1 domains of hCEACAM1 and mCEACAM1a share 41% sequence identity and are very similar in structure (Fig. 1*b*). Their common main-chain atoms (residues 1–107) can be superimposed with an r.m.s. deviation of 1.6 Å (Fig. 2). The major sequence diversity between the D1 domains of human and mouse CEACAM1 domains lies on the *GFC* face of the molecule, principally in the *CC′*, *C′C″* and *FG* loops and the *C′* and *C″* strands. Accordingly, these regions also vary most in structure between the two molecules. The crossover *BC* loop is also slightly different at positions 27–29.

**3.2.1. The *CC′* loop**. The biggest difference between the two structures occurs in a region stretching from residues 37 to 65 encompassing the *C′* and *C″* strands and connecting loops. In particular, the *CC′* loop, which projects from the *GFC* face of the molecule, is significantly different in structure (Fig. 3*a*). In hCEACAM1(D1) the *CC′* loop is markedly more hydrophilic in character owing to the substitution of Glu37, Arg38 and Asp40 for Asn37, Thr38 and Ala40, respectively, in mCEACAM1a(D1,4). In the superimposition of the two structures, the positions of equivalent $C^{\alpha}$ atoms in this region are very divergent. For example, the $C^{\alpha}-C^{\alpha}$ distance between Arg38 (human) and Thr38 (mouse) is 3.4 Å. In the structure of hCEACAM1(D1), a water molecule forms hydrogen bonds with the carbonyls of Lys35 and Arg38 within the bend of the *CC′* loop, which appears to be important for the maintaining the conformation of this loop.

The *CC′* loop is important for function in both human CEACAM1 and mouse CEACAM1a. In mouse CEACAM1a, it is involved in binding to the spike glycoprotein of mouse hepatitis virus (MHV; Rao *et al.*, 1997; Wessner *et al.*, 1998) and to a monoclonal antibody directed against murine CEACAM1a that blocks the binding of the virus to the receptor (Dveksler *et al.*, 1991). In hCEACAM1, this region binds *Neisseria* Opa proteins (see below). Hence, the structural differences of this loop may be attributed to these disparate activities. In particular, the conformation of the *CC′* loop in mCEACAM1a(D1,4) is probably influenced by the glycosylation of Asn37, whereas the equivalent residue in hCEACAM1 is Glu37 and would not be glycosylated *in vivo*.

**3.2.2. *C″D* loop**. The *C′* and *C″* strands adopt similar hairpin structures in both molecules but are shifted with respect to one another, such that the *C′C″* loop lies closer to the main body of the protein in hCEACAM1(D1) (Fig. 2). Almost no residues are conserved between the two proteins in this region. By contrast, the *C″D* loop is mostly conserved in sequence and yet is significantly different in structure compared with the same loop in mCEACAM1(D1) (Fig. 3*b*). Ala60, Ser62 and Gly63 are conserved between human and mouse CEACAM1 but in the superimposition their respective $C^{\alpha}$ atoms are separated by 2.3, 3.1 and 2.0 Å. Most strikingly, the side chains of Asn61 and Ser62 (human) and of Tyr61 and Ser62 (mouse) point in opposite directions in the two structures. In mCEACAM1(D1,4) Tyr61 points towards the outside of the molecule, whereas in hCEACAM1(D1) Asn61 points inwards within the curve made by the *C″D* loop, where it is hydrogen bonded to the carbonyl of Ile45 and the amide N atom of Gly63. Likewise, Ser62 in hCEACAM1(D1) forms contacts with Arg64 and Glu65, whereas in mCEACAM1a(D1,4), Ser62 points into solvent and forms no contacts.

A key residue in this network in hCEACAM1(D1) is Glu65, which is mostly buried and inaccessible to solvent. One side of Glu65 is bounded by the hydrophobic residues Ile45, Leu73 and Ile75 and the other side by residues 61–64 of the *C″D* loop. The side chain forms a hydrogen bond with the side chain of Asn61 and with both the main chain and side chain of Ser62. Along with the conserved salt bridge between Arg64 and Asp82, all these contacts appear to be important for maintaining a specific conformation of the *C″D* loop.

**3.2.3. *FG* loop**. Finally, another region that exhibits large differences between mouse and human CEACAM1 is the *FG* loop and *G* strand (Fig. 3*c*). Immediately following the *F* strand, the hairpin turn between *F* and *G* diverges in structure, commencing with residue 90, and equivalent $C^{\alpha}-C^{\alpha}$ distances thereafter vary from 1.5 to 3.7 Å. The pinnacle of the β-turn is quite different in the two structures, with the side chains of Ser93 in human CEACAM1 and Glu93 in mouse CEACAM1a pointing in opposite directions. Most notably, the register between the two structures alters in this region such that residues 96–99 of human CEACAM1 overlap with residues 95–98 of mouse CEACAM1a. This structural 'shunt' causes the *FG* loop to be one residue longer and the *G* strand to be one residue shorter in hCEACAM1(D1) when compared with mCEACAM1a(D1,4).

**3.2.4. Functional *versus* modeling differences**. It is important to establish whether the structural differences between mouse and human CEACAM1 could be genuine and relate, for instance, to the lack of glycosylation in the recombinant human protein or are more simply modeling differences arising from the lower resolution (3.32 Å) of mCEACAM1a(D1,4). Indeed, some of the crystallographic statistics of the mouse structure do suggest problems with the model. For instance, the *R* factor is relatively high (29.5%), only 68.5% of the residues lie in the most favored region of the Ramachandran plot (Tan *et al.*, 2002) and the r.m.s. deviation in *B* factors for main-chain atoms is 3.8 $Å^2$. 26 water molecules were modeled but, when we examined these, the electron density was either absent or unlikely to be of a water molecule and very few had appropriate hydrogen-bonding partners. Most pertinently, as would be expected at this resolution, the electron density for many of the side chains in this structure

was relatively weak. Of the three regions highlighted above, the *FG* loop is most likely to have been modeled incorrectly in mCEACAM1a(D1,4). This region exhibits particularly high *B* factors (80–100 Å$^2$) and its density is such that the human model in this region could fit equally well as the mouse. The conformation of this loop is biologically interesting because in some members of the pregnancy-specific group (PSG) subgroup of the CEA family it contains an integrin-binding Arg-Gly-Asp (RGD) motif (Zhou & Hammarstrom, 2001). Any ambiguities in this region will only be resolved with a higher resolution structure of mCEACAM1a(D1,4).

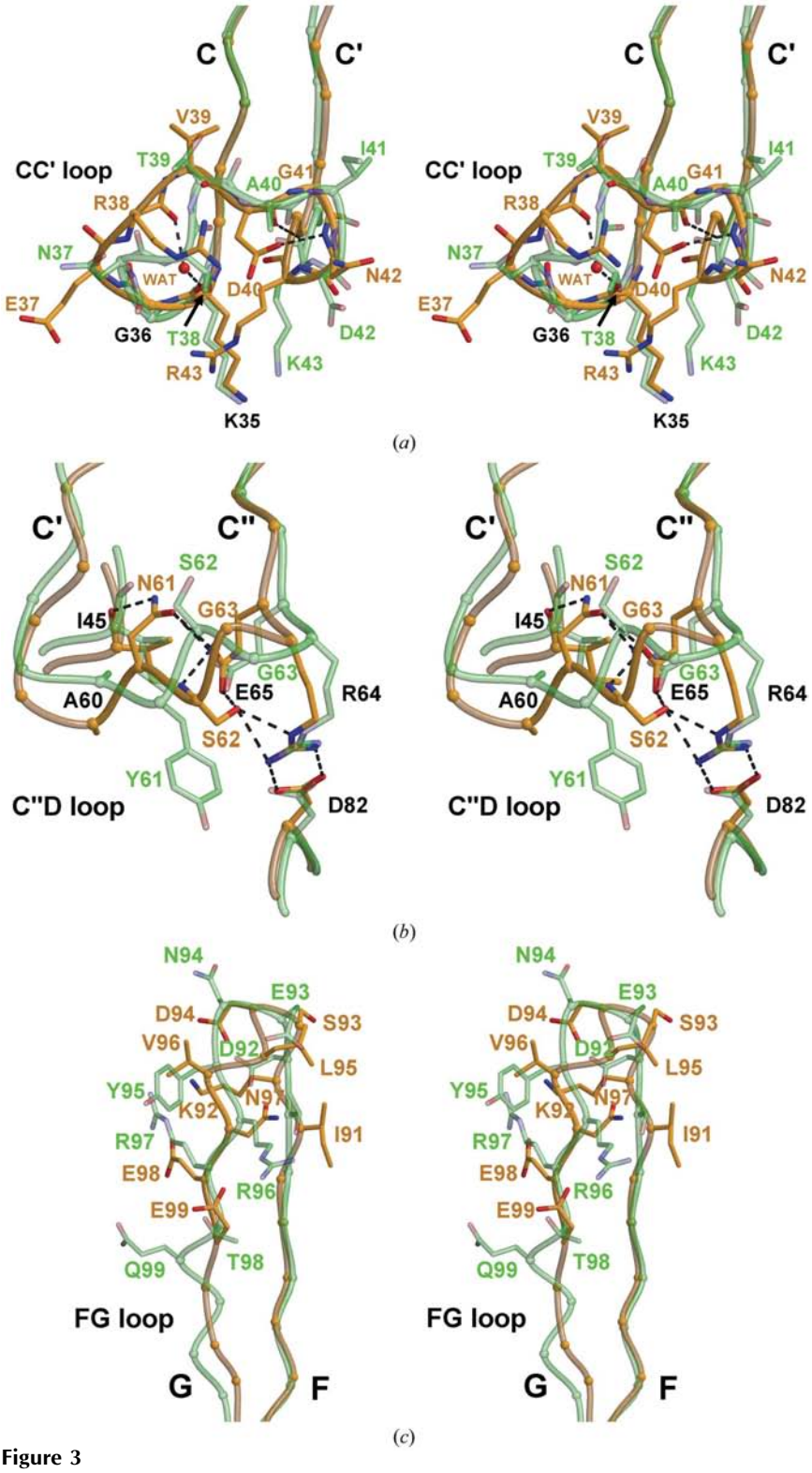


**Figure 3**
Regions that differ significantly in structure between the D1 domains of human CEACAM1 and mouse CEACAM1a. In each stereoview, human CEACAM1 is colored orange and mouse CEACAM1a is colored green, with corresponding labels, except for those residues that are the same in each, which are colored black. Hydrogen-bonding interactions in human CEACAM are shown as dashed lines. (*a*) The *CC*' loop (*b*), the *C*′*D* loop and (*c*) the *FG* loop.

### 3.3. Opa binding region of CEACAM1

Human CEACAMs are receptors for several bacterial pathogens, including *Haemophilus influenzae*, *N. meningitidis* and *N. gonorrhoeae*, and several studies have implicated the non-glycosylated *GFC* face of the N-terminal domain as the site of molecular recognition for opacity-associated (Opa) proteins of *Neisseria* spp. (Virji, Watt *et al.*, 1996; Bos *et al.*, 1998, 1999; Virji *et al.*, 1999; Popp *et al.*, 1999). Using a series of CEACAM6-derived chimaeras, Popp and coworkers demonstrated that Ser32 was important for binding to $Opa_{52}$ of *N. gonorrhoeae* and that residues 27–29 confer specificity of binding to certain Opa molecules (Popp *et al.*, 1999). Residues 27–29 were also implicated in CEACAM1 binding to three Opa variants tested (Bos *et al.*, 1999). In CEACAM5, mutation of Phe29 or Ser32 to Ala abrogated binding to a panel of three Opa variants, whereas mutation of Gly41 or Gln44 to Ala blocked the interaction with some but not all of the tested Opa variants (Popp *et al.*, 1999). These residues are all conserved in CEACAM1. Site-directed mutagenesis of residues on the *GFC* face of CEACAM1 identified Tyr34 and Ile91 as residues involved in binding to all nine strains of *N. gonorrhoeae* or *N. meningitidis* tested that expressed various Opa molecules, whilst Ser32, Val39, Gln44 and Gln89 were shown to be involved in binding a smaller set of Opa molecules (Virji *et al.*, 1999).

These Opa-binding residues were mapped onto the crystal structure of hCEACAM1(D1), where they form a distinct patch on the *GFC* face of the molecule comprised of residues projecting from the *BC* loop, the *C* strand, the *CC*′ loop, the *C*′ strand and the *F* strand (Figs. 4

and 5). As revealed by the electrostatic plot of this surface (Fig. 5*a*), this region is markedly hydrophobic in character, as expected for a protein–protein interaction domain. The surface is also concave in shape, largely owing to the *CC′* loop, which folds back to lie across the surface of the *GFCC′C″* $\beta$-sheet and creates the lower part of the binding surface. In the centre of this patch are residues Tyr34 and Ile91 (colored red in Fig. 5*b*), which are the most important for Opa binding (Virji *et al.*, 1999), and these residues are surrounded by others that are involved in binding some but not all Opa molecules (Val39, Gly41, Ser43, Gln44, Gln89 and Phe29; colored orange and purple in Fig. 5*b*). Because most of its side chain packs against the *CC′* loop, only the hydroxyl group of Tyr34 is present on the surface and it can be predicted that this residue will interact with Opa proteins *via* hydrogen-bonding interactions. The same is also true for the hydroxyl group of Ser32, which lies adjacent to Tyr34 on the *C* $\beta$-strand. The three glutamines in this region (27, 44 and 89) may also form hydrogen bonds with Opa proteins, whereas Ile91, Val39 and Phe29 probably interact *via* hydrophobic interactions.

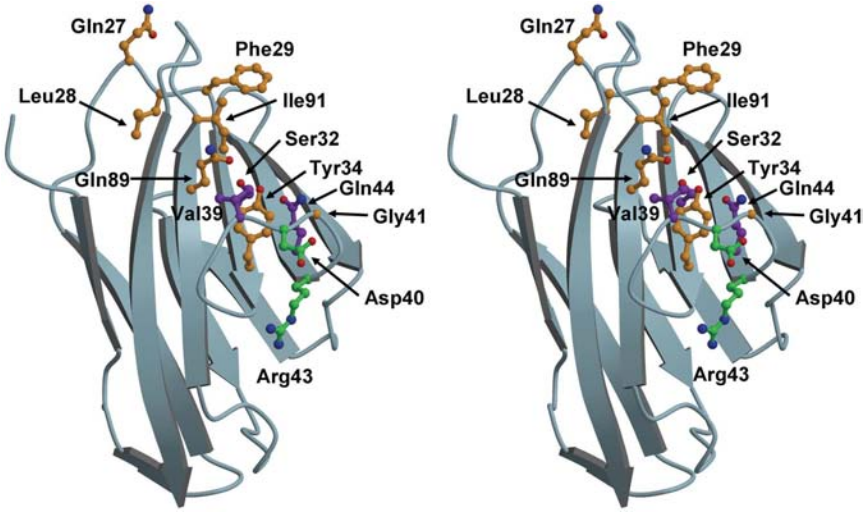


**Figure 4**
The residues of hCEACAM1 that bind Opa proteins and mediate homophilic and heterophilic interactions. In this stereoview, the structure is shown as a blue ribbon and important residues are shown in ball-and-stick format. The coloring scheme for these residues is as follows: red, residues that are required for interactions with all Opa proteins tested; orange, residues that are required for interactions with some Opa proteins; purple, residues that are involved in interactions with Opa proteins and in homophilic or heterophilic interactions with CEACAMs; green, residues involved in homophilic or heterophilic interactions (see text for details).

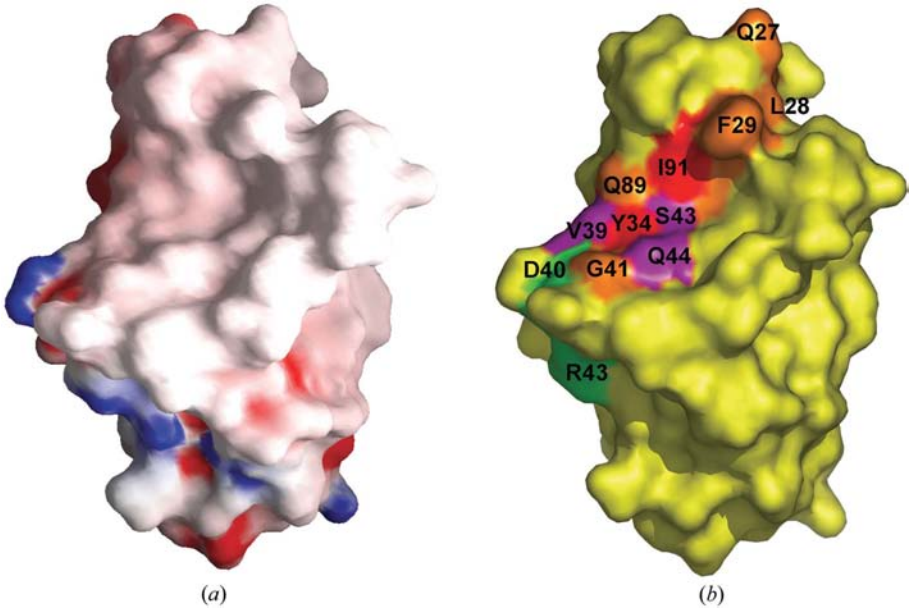


**Figure 5**
The binding surfaces of the D1 domain of human CEACAM1. (*a*) An electrostatic plot, calculated using *GRASP* (Nicholls *et al.*, 1991), in which electropositive regions of the molecule are colored blue, electronegative regions are red and hydrophobic regions are white. (*b*) The regions of hCEACAM1 that bind Opa proteins and mediate homophilic and heterophilic interactions. The surface of the protein is rendered in yellow and the coloring scheme for residues important for binding is the same as in Fig. 4. Compared with the view in Fig. 4, the molecule has been rotated slightly to the left about the vertical axis.

Of all the residues involved in binding Opa proteins identified by site-directed mutagenesis and chimeric proteins, only Leu28 appears poorly placed in the structure to interact directly with Opa. This residue is present on the *BC* loop and rather than pointing outside, projects toward the hydrophobic core of the molecule. Residues 27–29 were identified as potential Opa-binding residues by their ability to confer Opa-binding properties to CEACAM5 following their replacement with those from CEACAM1 (Popp *et al.*, 1999). The structure suggests that the least important of residues in this triplet for binding Opa in CEACAM1 is Leu28.

### 3.4. Adhesion surface of CEACAM1

Aside from being a receptor for bacterial invasion, CEACAM1 is also an adhesion molecule, mediating both homophilic and heterophilic interactions (Rojas *et al.*, 1990). In the crystal structure of hCEACAM1(D1) there are two molecules in the asymmetric unit and these interact *via* hydrophobic interactions across the *ABED* faces of the two molecules. Hence, this opens the question of whether such an interaction might mimic the association of CEACAM1 with itself (or with other CEACAMs) *in vivo*. Examination of the potential sites of glycosylation,

however, shows this to be very unlikely because glycosylation at Asn70, which lies on the *ABED* face immediately prior to strand *D*, would prevent such interactions. Indeed, in the crystal structure of mCEACAM1a(D1,4) (Tan *et al.*, 2002) this position is glycosylated by two units of *N*-acetylglucosamine and one mannose and, if this were the same in the human protein, the packing arrangement of molecules in crystals of hCEACAM1(D1) would be impossible.

In agreement with this, mutagenesis studies demonstrate unequivocally that the binding region of CEACAM1 for homophilic and heterophilic interactions lies on the *GFC* face of CEACAM1 and at least partially overlaps with the Opa-binding site (Watt *et al.*, 2001). Mutation of Val39 and Asp40 to Ala abolished homophilic adhesion of CEACAM1 and mutation of Ser32 to Ala lowered adhesion, whereas mutation of several other residues of the *GFCC′C″* face, including Tyr34, Ile91 and Gln44, had no effect (Watt *et al.*, 2001). In a separate study, however, Gln44, as well as Arg43, was important for homophilic interactions (Markel *et al.*, 2004). Thus, it appears that the interface for homophilic interactions lies just below the Opa-binding region and includes the *CC′* loop (Figs. 4 and 5*b*). A similar region is also used for heterophilic interactions because Arg43 and Gln44 are also required for binding of CEACAM1 to CEACAM5 and, conversely, mutation of Ser32 and Leu44 to Arg and Gln, respectively, confers CEACAM1-binding activity to CEACAM6, which otherwise does not bind CEACAM1 (Markel *et al.*, 2004).

Interestingly, gel filtration demonstrated that our construct of hCEACAM1(D1) is monomeric in solution. Thus, although residues required for homophilic interactions are located in the N-terminal domain, other domains of CEACAM absent in this construct must also contribute to binding. Consistent with this view, it has been shown previously that domains A1, B and A2 of CEACAM1 (the three other domains in the extracellular region of CEACAM1) also contribute to homophilic adhesion (Watt *et al.*, 2001).

## 4. Conclusion

Here, we present the crystal structure of the N-terminal domain of human CEACAM1 at a resolution of 2.2 Å. This is the highest resolution view of any CEACAM molecule to date and highlights interesting structural differences with the lower resolution structure from mouse. The structure also reveals deficiencies in the various homology models that have been generated of hCEACAM1, principally because the *CC′* loop is shifted significantly in comparison with the immunoglobulin folds on which these models were based.

The *GFC* face of CEACAM1 contains residues responsible for binding to *Neisseria* Opa proteins and for homophilic and heterophilic interactions. The *CC′* loop, which folds back to lie across this face of the molecule, is a prominent feature of this surface. The Opa-binding site maps to a single contiguous region containing Tyr34 and Ile91, which mutagenesis studies suggest are important for interactions with all Opa proteins, at its center. Residues that mediate homophilic and heterophilic interactions partially overlap with the Opa-binding region.

Although a similar analysis of the Opa-binding and adhesion surfaces of human CEACAM1 has been made using the structure of mouse CEACAM1(D1,D4) as a model (Tan *et al.*, 2002), the higher resolution of the current structure and the greater certainty in the placement of side chains provides a better framework for a more precise dissection of the functional regions of human CEACAM1 in the context of *Neisseria* pathogenesis, cellular adhesion and immune evasion. It is also a prelude to the structural determination of the complex between CEACAM1 and members of the Opa family of proteins.

This work was supported by the National Institutes of Health grants GM66861 to CD and AI36901 to RAN. Use of the Advanced Photon Source was supported by the US Department of Energy, Office of Science, Office of Basic Energy Sciences under Contract No. W-31-109-ENG-38. Data were collected at Southeast Regional Collaborative Access Team (SER-CAT) 22-ID beamline at the Advanced Photon Source, Argonne National Laboratory. Supporting institutions may be found at http://www.ser-cat.org/members.html. We thank Dr Martine Bos for providing the plasmid encoding CEACAM1 and Dr Janne Cannon and Nan Fulcher, UNC-Chapel Hill, for many helpful discussions.

## References

Bamberger, A. M., Riethdorf, L., Nollau, P., Naumann, M., Erdmann, I., Gotze, J., Brummer, J., Schulte, H. M., Wagener, C. & Loning, T. (1998). *Am. J. Pathol.* **152**, 1401–1406.
Bates, P. A., Luo, J. & Sternberg, M. J. (1992). *FEBS Lett.* **301**, 207–214.
Beauchemin, N. *et al.* (1999). *Exp. Cell Res.* **252**, 243–249.
Benchimol, S., Fuks, A., Jothy, S., Beauchemin, N., Shirota, K. & Stanners, C. P. (1989). *Cell*, **57**, 327–334.
Bos, M. P., Grunert, F. & Belland, R. J. (1997). *Infect. Immun.* **65**, 2353–2361.
Bos, M. P., Hogan, D. & Belland, R. J. (1999). *J. Exp. Med.* **190**, 331–340.
Bos, M. P., Kuroki, M., Krop-Watorek, A., Hogan, D. & Belland, R. J. (1998). *Proc. Natl Acad. Sci. USA*, **95**, 9584–9589.
Boulton, I. C. & Gray-Owen, S. D. (2002). *Nature Immunol.* **3**, 229–236.
Brünger, A. T. (1992). *Nature (London)*, **355**, 472–474.
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). *Acta Cryst.* D**54**, 905–921.
Chothia, C. & Jones, E. Y. (1997). *Annu. Rev. Biochem.* **66**, 823–862.
Dveksler, G. S., Pensiero, M. N., Cardellichio, C. B., Williams, R. K., Jiang, G. S., Holmes, K. V. & Dieffenbach, C. W. (1991). *J. Virol.* **65**, 6881–6891.
Gray-Owen, S. D., Dehio, C., Haude, A., Grunert, F. & Meyer, T. F. (1997). *EMBO J.* **16**, 3435–3445.
Hammarstrom, S. (1999). *Semin. Cancer Biol.* **9**, 67–81.
Horst, A. K., Ito, W. D., Dabelstein, J., Schumacher, U., Sander, H., Turbide, C., Brummer, J., Meinertz, T., Beauchemin, N. & Wagener, C. (2006). *J. Clin. Invest.* **116**, 1596–1605.
Huang, J., Simpson, J. F., Glackin, C., Riethorf, L., Wagener, C. & Shively, J. E. (1998). *Anticancer Res.* **18**, 3203–3212.

Jerse, A. E., Cohen, M. S., Drown, P. M., Whicker, L. G., Isbey, S. F., Seifert, H. S. & Cannon, J. G. (1994). *J. Exp. Med.* **179**, 911–920.
Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. (1991). *Acta Cryst.* A**47**, 110–119.
Jonge, M. I. de, Hamstra, H. J., van Alphen, L., Dankert, J. & van der Ley, P. (2003). *Mol. Microbiol.* **50**, 1005–1015.
Kirshner, J., Chen, C. J., Liu, P., Huang, J. & Shively, J. E. (2003). *Proc. Natl Acad. Sci. USA*, **100**, 521–526.
Kleinerman, D. I., Troncoso, P., Lin, S. H., Pisters, L. L., Sherwood, E. R., Brooks, T., von Eschenbach, A. C. & Hsieh, J. T. (1995). *Cancer Res.* **55**, 1215–1220.
Markel, G., Gruda, R., Achdout, H., Katz, G., Nechama, M., Blumberg, R. S., Kammerer, R., Zimmermann, W. & Mandelboim, O. (2004). *J. Immunol.* **173**, 3732–3739.
Markel, G., Lieberman, N., Katz, G., Arnon, T. I., Lotem, M., Drize, O., Blumberg, R. S., Bar-Haim, E., Mader, R., Eisenbach, L. & Mandelboim, O. (2002). *J. Immunol.* **168**, 2803–2810.
Markel, G., Wolf, D., Hanna, J., Gazit, R., Goldman-Wohl, D., Lavy, Y., Yagel, S. & Mandelboim, O. (2002). *J. Clin. Invest.* **110**, 943–953.
Matthews, B. (1968). *J. Mol. Biol.* **33**, 491–497.
Murshudov, G. N., Vagin, A. A. & Dodson, E. J. (1997). *Acta Cryst.* D**53**, 240–255.
Nicholls, A., Sharp, K. A. & Honig, B. (1991). *Proteins*, **11**, 281–296.
Nittka, S., Gunther, J., Ebisch, C., Erbersdobler, A. & Neumaier, M. (2004). *Oncogene*, **23**, 9306–9313.
Nollau, P., Prall, F., Helmchen, U., Wagener, C. & Neumaier, M. (1997). *Am. J. Pathol.* **151**, 521–530.
Nollau, P., Scheller, H., Kona-Horstmann, M., Rohde, S., Hagenmuller, F., Wagener, C. & Neumaier, M. (1997). *Cancer Res.* **57**, 2354–2357.
Normark, S., Albiger, B. & Jonsson, A. B. (2002). *Nature Immunol.* **3**, 210–211.
Oikawa, S., Inuzuka, C., Kuroki, M., Arakawa, F., Matsuoka, Y., Kosaki, G. & Nakazato, H. (1991). *J. Biol. Chem.* **266**, 7995–8001.
Oikawa, S., Inuzuka, C., Kuroki, M., Matsuoka, Y., Kosaki, G. & Nakazato, H. (1989). *Biochem. Biophys. Res. Commun.* **164**, 39–45.
Otwinowski, Z. & Minor, W. (1997). *Methods Enzymol.* **276**, 307–326.
Pantelic, M., Kim, Y. J., Bolland, S., Chen, I., Shively, J. & Chen, T. (2005). *Infect. Immun.* **73**, 4171–4179.
Pflugrath, J. W. (1999). *Acta Cryst.* D**55**, 1718–1725.
Popp, A., Dehio, C., Grunert, F., Meyer, T. F. & Gray-Owen, S. D. (1999). *Cell. Microbiol.* **1**, 169–181.
Pu, Y. S., Luo, W., Lu, H. H., Greenberg, N. M., Lin, S. H. & Gingrich, J. R. (1999). *J. Urol.* **162**, 892–896.
Rao, P. V., Kumari, S. & Gallagher, T. M. (1997). *Virology*, **229**, 336–348.
Riethdorf, L., Lisboa, B. W., Henkel, U., Naumann, M., Wagener, C. & Loning, T. (1997). *J. Histochem. Cytochem.* **45**, 957–963.
Rojas, M., Fuks, A. & Stanners, C. P. (1990). *Cell Growth. Differ.* **1**, 527–533.
Stern, N., Markel, G., Arnon, T. I., Gruda, R., Wong, H., Gray-Owen, S. D. & Mandelboim, O. (2005). *J. Immunol.* **174**, 6692–6701.
Tan, K., Zelus, B. D., Meijers, R., Liu, J. H., Bergelson, J. M., Duke, N., Zhang, R., Joachimiak, A., Holmes, K. V. & Wang, J. H. (2002). *EMBO J.* **21**, 2076–2086.
Teixeira, A. M., Fawcett, J., Simmons, D. L. & Watt, S. M. (1994). *Blood*, **84**, 211–219.
Vagin, A. & Teplyakov, A. (1997). *J. Appl. Cryst.* **30**, 1022–1025.
Virji, M., Evans, D., Hadfield, A., Grunert, F., Teixeira, A. M. & Watt, S. M. (1999). *Mol. Microbiol.* **34**, 538–551.
Virji, M., Makepeace, K., Ferguson, D. J. & Watt, S. M. (1996). *Mol. Microbiol.* **22**, 941–950.
Virji, M., Watt, S. M., Barker, S., Makepeace, K. & Doyonnas, R. (1996). *Mol. Microbiol.* **22**, 929–939.
Volpert, O., Luo, W., Liu, T. J., Estrera, V. T., Logothetis, C. & Lin, S. H. (2002). *J. Biol. Chem.* **277**, 35696–35702.
Wang, J.-H. & Springer, T. A. (1998). *Immunol. Rev.* **163**, 197–215.
Watt, S. M., Teixeira, A. M., Zhou, G. Q., Doyonnas, R., Zhang, Y., Grunert, F., Blumberg, R. S., Kuroki, M., Skubitz, K. M. & Bates, P. A. (2001). *Blood*, **98**, 1469–1479.
Wessner, D. R., Shick, P. C., Lu, J. H., Cardellichio, C. B., Gagneten, S. E., Beauchemin, N., Holmes, K. V. & Dveksler, G. S. (1998). *J. Virol.* **72**, 1941–1948.
Zhou, G. Q. & Hammarstrom, S. (2001). *Biol. Reprod.* **64**, 90–99.