

# Crystal structure of murine sCEACAM1a[1,4]: a coronavirus receptor in the CEA family

Kemin Tan<sup>1,2</sup>, Bruce D.Zelus<sup>3</sup>,  
Rob Meijers<sup>1,2</sup>, Jin-huan Liu<sup>1,2</sup>,  
Jeffrey M.Bergelson<sup>4</sup>, Norma Duke<sup>5</sup>,  
Rongguang Zhang<sup>5</sup>, Andrzej Joachimiak<sup>5</sup>,  
Kathryn V.Holmes<sup>3,6</sup> and Jia-huai Wang<sup>1,6,7,8</sup>

<sup>1</sup>Dana-Farber Cancer Institute, 44 Binney Street, Boston, MA 02115, Departments of <sup>2</sup>Medicine, <sup>7</sup>Pediatrics, <sup>8</sup>Biological Chemistry and Molecular Pharmacology, Harvard Medical School, <sup>3</sup>Department of Microbiology, University of Colorado Health Sciences Center, Denver, CO 80262, <sup>4</sup>The Children's Hospital of Philadelphia, University of Pennsylvania School of Medicine, Philadelphia, PA 19104 and <sup>5</sup>Biosciences Division, Argonne National Laboratory, Argonne, IL 60439, USA

<sup>6</sup>Corresponding authors

e-mail: jwang@red.dfc.harvard.edu or kathryn.holmes@uchsc.edu

K.Tan, B.D.Zelus and R.Meijers contributed equally to this work

**CEACAM1 is a member of the carcinoembryonic antigen (CEA) family. Isoforms of murine CEACAM1 serve as receptors for mouse hepatitis virus (MHV), a murine coronavirus. Here we report the crystal structure of soluble murine sCEACAM1a[1,4], which is composed of two Ig-like domains and has MHV neutralizing activity. Its N-terminal domain has a uniquely folded CC' loop that encompasses key virus-binding residues. This is the first atomic structure of any member of the CEA family, and provides a prototypic architecture for functional exploration of CEA family members. We discuss the structural basis of virus receptor activities of murine CEACAM1 proteins, binding of *Neisseria* to human CEACAM1, and other homophilic and heterophilic interactions of CEA family members.**

**Keywords:** bacterial binding/CEA family/cell adhesion/coronavirus receptor/crystal structure

## Introduction

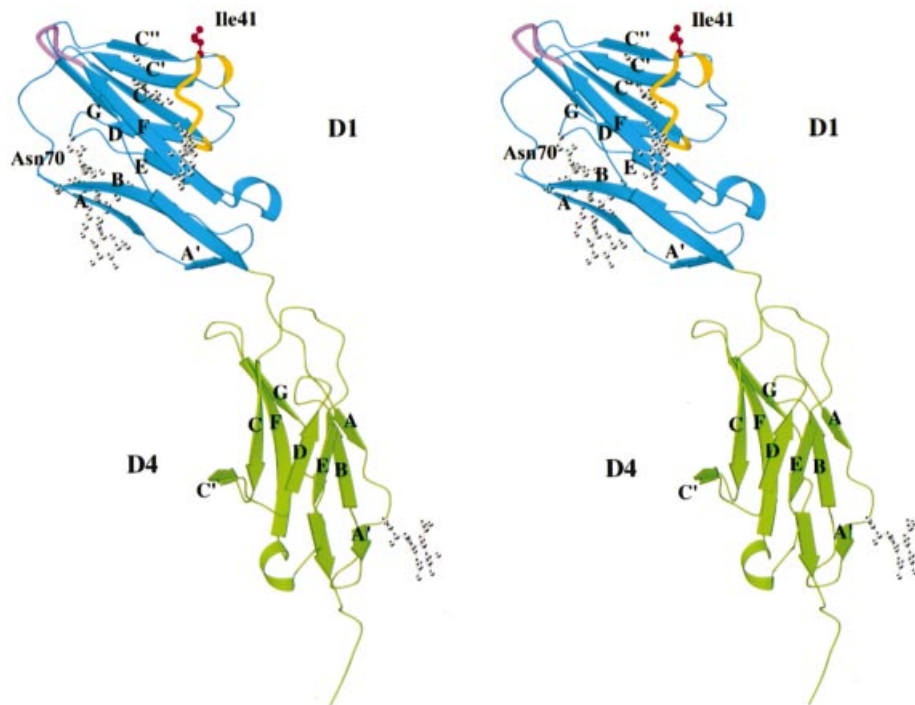
Carcinoembryonic antigen (CEA; CD66e) was initially discovered as a tumor antigen (Gold and Freedman, 1965). A large group of related glycoproteins within the Ig superfamily (IgSF) is now called the CEA family. These anchored or secreted glycoproteins are expressed by epithelial cells, leukocytes, endothelial cells and placenta (Hammarstrom, 1999). In humans, the CEA family contains 29 genes and pseudogenes. The revised nomenclature of this family of glycoproteins has recently been summarized (Beauchemin *et al.*, 1999). The CEA family consists of the CEA-related cell adhesion molecule (CEACAM) and pregnancy-specific glycoprotein (PSG) subfamilies, whose proteins share many common structural features (Hammarstrom, 1999).

CEACAM1 (CD66a) is the most highly conserved member of the CEA family. Most species have only one CEACAM1 gene, but mice have two closely related genes called CEACAM1 and CEACAM2 (Beauchemin *et al.*, 1999). CEACAM1 has many important biological functions. It is a potent vascular endothelial growth factor (Ergun *et al.*, 2000) and a growth inhibitor in tumor cells (Izzi *et al.*, 1999), plays a key role in differentiation of mammary glands (Huang *et al.*, 1999), is an early marker of T-cell activation and modulates the functions of murine T lymphocytes (Morales *et al.*, 1999; Nakajima *et al.*, 2002). Human CEACAM1 is one of several human CEACAM proteins that serve as receptors for virulent strains of *Neisseria gonorrhoeae*, *Neisseria meningitidis* and *Haemophilus influenzae* (Bos *et al.*, 1999; Virji *et al.*, 1999, 2000).

In mice, four isoforms of CEACAM1 generated by alternative mRNA splicing have either two [D1,D4] or four [D1–D4] Ig-like domains on the cell surface, a transmembrane segment and either a short or a long cytoplasmic tail (Beauchemin *et al.*, 1999). The long tail contains a modified immunoreceptor tyrosine-based inhibition motif (ITIM)-like motif. Tyrosine phosphorylation of this motif is associated with signaling (Huber *et al.*, 1999), but the natural ligands for the ecto-domain and the modulation of gene expression by CEACAM1 signaling are not well understood.

All four isoforms of murine CEACAM1a, as well as murine CEACAM2, can serve as receptors for mouse hepatitis virus (MHV) strain A59 (MHV-A59) when the recombinant murine proteins are expressed at high levels in a hamster cell line (BHK) (Dveksler *et al.*, 1991, 1993a; Nedellec *et al.*, 1994). MHVs are large, enveloped, positive-stranded RNA viruses in the Coronaviridae family in the order Nidovirales. Various MHV strains cause diarrhea, hepatitis, and respiratory, neurological and immunological disorders in mice. Infection is initiated by binding of the 180 kDa spike glycoprotein (S) on the viral envelope to a CEACAM glycoprotein on a murine cell membrane. Most inbred mouse strains are highly susceptible to MHV infection, but SJL/J mice are highly resistant. Susceptible strains are homozygous for the CEACAM1a allele that encodes the principal MHV receptor, while SJL/J mice are homozygous for the CEACAM1b allele. CEACAM1b proteins have weaker MHV binding and receptor activities than CEACAM1a proteins (Ohtsuka *et al.*, 1996; Rao *et al.*, 1997; Wessner *et al.*, 1998).

Until now, extensive N-linked glycosylation has hampered crystallization of any CEA protein for structural analysis. This article reports the crystal structure of the soluble ecto-domain of an isoform of murine CEACAM1a that consists of domains 1 and 4 (designated msCEACAM1a[1,4]) and has MHV-neutralizing activity. We



**Fig. 1.** Stereo view of the ribbon drawing of msCEACAM1a[1,4], which contains two Ig-like domains. The CC' loop in the N-terminal domain (D1), which is involved in binding of MHV and other ligands, is highlighted in yellow. The predicted key virus-binding residue Ile41 on the CC' loop is shown in red in ball-and-stick representation. The FG loop of D1, another biologically important element, is shown in violet. The carbohydrate moieties are drawn in gray in ball-and-stick representation. The glycan at Asn70 that is conserved in the whole CEA family is labeled. The figure was prepared using MOLSCRIPT (Kraulis, 1991).

have analyzed the relationship of the structure of the msCEACAM1a[1,4] glycoprotein to its MHV-binding and -neutralizing activities. Based on the structure of msCEACAM1a[1,4], we predict the structures of other CEA family members, and discuss their biological significance.

## Results and discussion

### ***Molecular structure of msCEACAM1a[1,4]***

The msCEACAM1a[1,4] protein analyzed in this paper contains the 202 extracellular amino acids of the naturally expressed CEACAM1a[1,4] protein plus a His<sub>6</sub> tag connected to the C-terminus by a thrombin cleavage peptide. This soluble murine CEACAM1a[1,4] protein has strong virus-neutralizing activity at 37°C, pH 7.2, and readily induces an irreversible conformational change in the MHV-A59 spike glycoprotein under these conditions (Zelus *et al.*, 1998; B.D. Zelus and K.V.Holmes, in preparation). The His-tagged protein was expressed by an adenovirus vector in the Chinese hamster ovary Lec3.2.8.1 (CHO lec<sup>-</sup>) cell line, which stably expresses recombinant CAR, the receptor for Coxsackie B and adenoviruses (Stanley, 1989; Bergelson *et al.*, 1997; Zelus *et al.*, 1998). These cells were readily transduced by the adenovirus vector, and they produce proteins with more homogeneous glycans than normal CHO cells. Analysis of the protein secreted by the lec<sup>-</sup>, CAR<sup>+</sup> CHO cells led to the final refined model for the structure of msCEACAM1a[1,4]. The structure was determined using multi-wavelength anomalous diffraction (MAD) phases in combination with molecular replacement (MR). The structure was refined to

3.32 Å with  $R_{\text{work}}/R_{\text{free}}$  of 29.5/32.9%. The relatively high  $R$ -factors are probably caused by disordered C-terminal residues and carbohydrate moieties.

Figure 1 shows the ribbon diagram of the molecular structure of soluble murine msCEACAM1a[1,4]. The two Ig-like domains of msCEACAM1a[1,4] are arranged in tandem. When the membrane-proximal domain (D4) was oriented vertically as if it were perpendicular to the cell membrane, the virus-binding domain (D1) had a bending angle of ~60° from the vertical direction, with its A'GFCC'C'' β-sheet (called the CFG face hereafter) facing upwards, away from the cell membrane (Figure 1). The rotation angle between D1 and D4 is ~170°, which places the CFG face of D4 on the opposite side of the molecule from the CFG face of D1, like many other IgSF proteins on the cell surface (Wang and Springer, 1998). Although there are five potential N-linked glycosylation sites on this protein, the crystal structure showed that only four of these sites are utilized: three in D1 and one in D4. One or more sugar moieties were seen clearly at each of these sites (Figure 1), but no electron density was visible to indicate the presence of a possible glycan at Asn161 in the Asn-Asn-Ser motif in the DE loop of D4. The only observed glycan in D4 is at Asn119 (Figure 1) near the bottom of the molecule, pointing downward towards the cell membrane. This glycan may play a role in holding the rod-like molecule erect on the membrane, as has been shown for CD2 (Jones *et al.*, 1992), ICAM-2 (Casasnovas *et al.*, 1997) and CD4 (Wu *et al.*, 1997; Wang *et al.*, 2001).

The N-terminal domain (D1) of msCEACAM1a[1,4] belongs to the V set Ig-like fold. Within the IgSF, the CEA family and the CD2 family are unique in that their

N-terminal domains lack the usually conserved inter-sheet disulfide bond between  $\beta$ -strands B and F. In a DALI search for structures homologous to D1 of msCEACAM1a[1,4] (using the website <http://www2.ebi.ac.uk/dali/>), D1 of CD2 was one of the top hits. There are, however, three important structural elements that distinguish D1 of msCEACAM1a[1,4] from CD2 D1. One striking feature of D1 of msCEACAM1a[1,4] is its uniquely structured, prominently protruding CC' loop (highlighted in yellow in Figure 1) that points upwards. The unique and intricate structure of the CC' loop will be described in detail below. D1 of msCEACAM1a[1,4], like other V set Ig-like folds, should retain a salt bridge between Arg64 at the beginning of the D strand and Asp82 at the beginning of the F strand. This salt bridge may help to strengthen the interactions between the two anti-parallel  $\beta$ -sheets of D1. In contrast, CD2 D1 does not have this salt bridge between the  $\beta$ -sheets (Jones *et al.*, 1992). Another difference between the D1s of msCEACAM1a[1,4] and CD2 is found at the A–A' kink. As a structural hallmark in both V and I set Ig folds, the A strand in one sheet runs midway through the domain, and then crosses over to join the opposite sheet, becoming the A' strand. This may stabilize the membrane-distal domain that is the usual site for ligand binding (Wang and Springer, 1998). A *cis*-proline is usually located at the kink position. In D1 of msCEACAM1a[1,4], the A' strand is significantly shorter than that of most other Ig-like molecules, whereas D1 of CD2 and some other CD2 family members have a relatively long A' strand, with no A strand at all. These features might reflect differences in the biological functions of CD2 and CEACAM1a.

Structural analysis shows that the C-terminal domain (D4) of msCEACAM1a[1,4] falls into the I1 set category (Harpaz and Chothia, 1994; Wang and Springer, 1998), rather than the C2 set as widely believed. Compared with the I set Ig-like domains of most other IgSF members, D4 of msCEACAM1a[1,4] has an unusually long CD loop of 10 residues (amino acids 146–155). The long CD loop in D4 of msCEACAM1a[1,4] is probably quite stable because it has a  $\beta$ -turn at each end and Leu150 and Leu152 in the middle of the loop point inward, joining the molecule's hydrophobic core.

msCEACAM1a[1,4] has a linker between D1 and D4. The last residue of D1 is His107, and the A strand of the following domain D4 starts at Phe114. The peptide segment in between does not appear to have main chain–main chain hydrogen bonds to the D4 domain. No significant interactions were observed between D1 and D4. The surface buried area between these two domains is 530 Å<sup>2</sup>, with a 1.7 Å probe. These observations indicate that the D1–D4 junction of msCEACAM1a[1,4] might be quite flexible.

#### **The unique CC' loop of the N-terminal domain is an MHV-binding site**

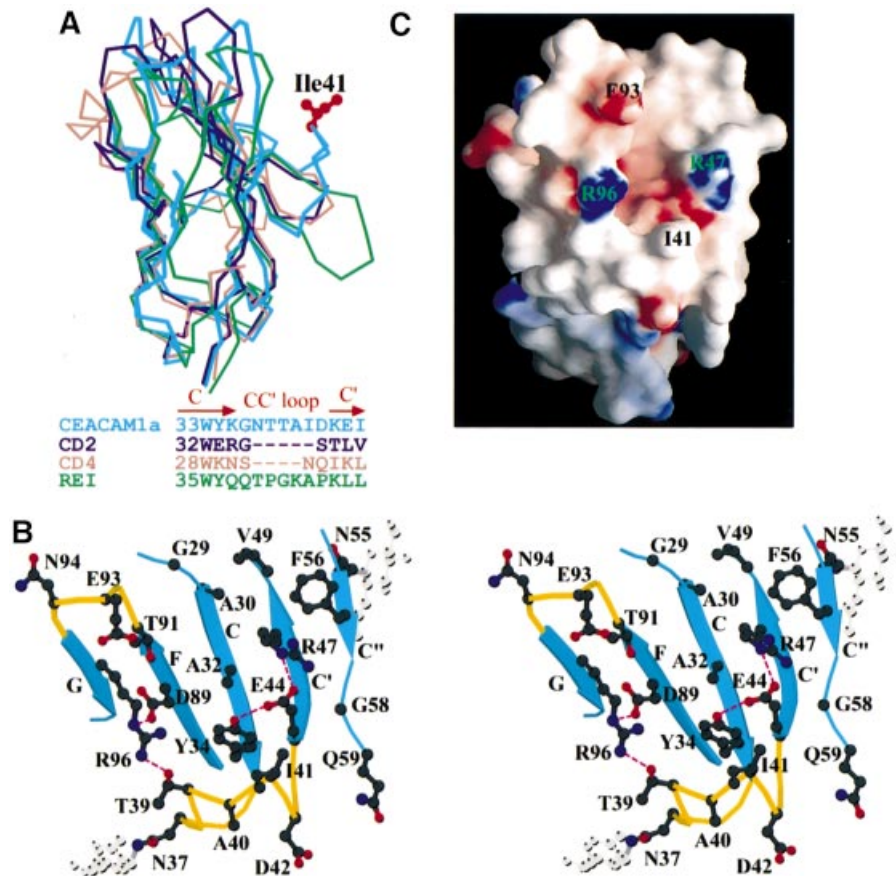
Both the spike glycoprotein of MHV virions and mAb-CC1, a monoclonal antibody to murine CEACAM1a that blocks the binding of the virus to the receptor, were shown to bind to D1 of murine CEACAM1a (Dveksler *et al.*, 1993b). Mutational analyses of murine CEACAM1a show that the peptide segments between amino acids 38 and 43 (Rao *et al.*, 1997) or between amino acids 34 and 52

(Wessner *et al.*, 1998) are involved in binding to the MHV spike glycoprotein, virus receptor activity and binding of mAb-CC1. Our structure for msCEACAM1a[1,4] shows that this segment is in the CC' loop and the C' strand.

Compared with the N-terminal domains of other IgSF members, D1 of msCEACAM1a[1,4] has an unusual CC' loop, highlighted in yellow in Figure 1. Figure 2A shows an overlay onto D1 of msCEACAM1a[1,4] of the N-terminal domains of three other representative IgSF proteins: CD2 (Jones *et al.*, 1992), CD4 (Wang *et al.*, 1990) and Bence-Jones protein REI (Epp *et al.*, 1975), a typical variable domain of an antibody. The N-terminal domains of both CD2 and CD4 have shorter CC' loops than that of msCEACAM1a[1,4] and REI. Although the CC' loops of D1 of REI and msCEACAM1a[1,4] are the same length, that of REI is only slightly curved, while, remarkably, the CC' loop of msCEACAM1a[1,4] folds back onto the CFG face.

The convoluted conformation of the CC' loop in D1 of msCEACAM1a[1,4] is unique among IgSF molecules. The loop, from Lys35 to Glu44, is well structured (Figure 2B) and probably maintained in a rigid conformation. Within the C-terminal portion of the loop residues 40–44 may form one and a half turns of a  $3_{10}$  helix. A particularly interesting structural element is the packing of the mid-portion backbone of the CC' loop (from Thr39 to Ile41) against the aromatic ring of Tyr34 on the C strand (Figure 2B). Several potential hydrogen bonds may help to maintain the unique conformation of this region, as shown in Figure 2B. Although a tyrosine equivalent to Tyr34 is conserved in the variable domains of most antibody light chains, nevertheless the CC' loop in antibodies assumes a  $\beta$ -hairpin structure (see REI in Figure 2A) probably because the conserved Pro-Gly sequence motif of antibodies (Figure 2A) favors a sharp turn at the tip of the loop. This might prevent the CC' loop of REI from assuming a convoluted conformation like that seen in D1 of msCEACAM1a[1,4].

In D1 of msCEACAM1a[1,4], the consequence of the folding back of the well structured CC' loop against the CFG face is that the side chain of Ile41 at the center of the loop is prominently exposed, pointing away from the membrane (Figures 1 and 2A). Mutational evidence suggests that the Thr38–Thr39–Ala40–Ile41 sequence motif in murine CEACAM1a[1,4] is important for binding to the MHV spike glycoprotein (Wessner *et al.*, 1998). Two glycans, one at Asn37 and the other at Asn55, flank this important virus-binding motif (Figures 1 and 2B), which might help delineate the region for viral spike glycoprotein docking. Based on our structural data, we speculate that Ile41 might be the energetic 'hot spot' for binding to the MHV spike. A widely accepted model for the interaction of cell surface receptors with their ligands is that a central hydrophobic contact provides the major binding energy, while surrounding hydrophilic interactions contribute the specificity of binding (Clackson and Wells, 1995; Kim *et al.*, 2001). This also appears to be the case for receptor–virus interactions, as shown for the binding of gp120 glycoprotein of HIV-1 to CD4 (Kwong *et al.*, 1998). Figure 2B and C shows a view looking down on the CFG face of D1 of msCEACAM1a[1,4], which is likely to be the surface accessible to the MHV virus spike protein. The protruding hydrophobic Ile41 is surrounded



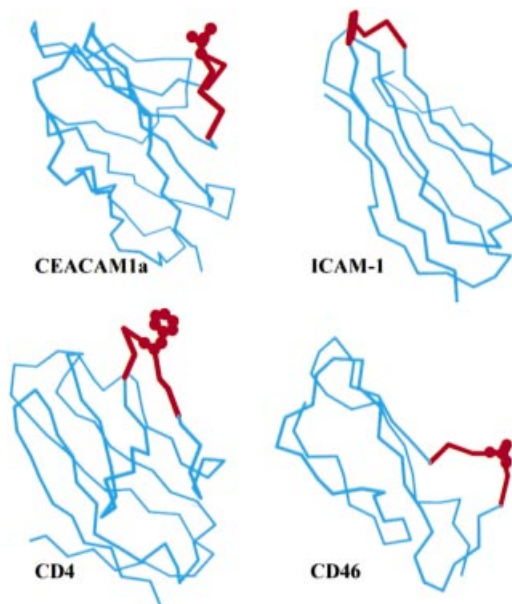
**Fig. 2.** (A) Superposition of D1 of msCEACAM1a[1,4], CD2, CD4 and Bence-Jones protein REI using  $C_{\alpha}$  atoms of  $\beta$ -sheets. Each molecule is shown in  $C_{\alpha}$  trace, with msCEACAM1a in cyan, CD2 in purple, CD4 in brown and REI in green, respectively. The unique convoluted conformation of the CC' loop in msCEACAM1a[1,4] is striking. The sequence alignment of the CC' loop regions of these four molecules are shown using the same color code. (B) Stereo view of the exposed residues on the CFG face of D1 of msCEACAM1a[1,4]. The  $C_{\alpha}$  trace of the CC' loop is highlighted in yellow. Displayed side chains and carbohydrates are drawn in ball-and-stick representation. (C) Electrostatic potential surface representation of the same view as (B). The electrostatic potential is colored blue for positive and red for negative, and was calculated in the absence of carbohydrates and solvent molecules. (A) and (B) were prepared with MOLSCRIPT (Kraulis, 1991), and (C) with GRASP (Nicholls *et al.*, 1991).

by a number of surface-exposed, charged residues, including Asp42, Glu44, Arg47, Asp89, Glu93 and Arg97. Ile41 might insert into a hypothetical hydrophobic pocket in the viral spike glycoprotein, and charged residues that surround the pocket could stabilize the MHV-binding interaction and contribute to virus binding specificity. No structures are as yet available for any coronavirus spike glycoproteins. Strains of MHV that differ in virulence and tissue tropism show considerable variation in the amino acid sequences of their S glycoproteins, yet all MHV strains tested can use murine CEACAM1a as a receptor. The observation that there is no single anti-S mAb that blocks infection by all strains of MHV (Talbot and Buchmeier, 1985) supports the hypothesis that murine CEACAM1a may bind to a conserved pocket in S that is not accessible to antibodies. The protruding Ile41 and the charged residues that surround it on the surface of the virus receptor are targets for further mutational analyses.

Cell adhesion molecules might be particularly suitable candidates for virus binding because their physiological ligand-receptor binding affinities are very low, and adhesion is an avidity-driven process. Viruses evolve to

have a stronger binding affinity for the receptor (usually ~100–1000 times stronger) to compete with the weakly bound physiological ligand (Wang, 2002). Uniquely exposed surface features of the cell adhesion molecules are selected for virus binding. Figure 3 compares the virus-binding domain of msCEACAM1a[1,4] with those of several other virus receptors, with the key virus-binding elements highlighted. We propose that the projecting Ile41 on the unique CC' loop of D1 of msCEACAM1a[1,4] is the key topological feature for MHV binding. In CD4, the key HIV gp120-binding Phe43 is located at the protruding ridge-like C'C'' corner of D1 (Wang *et al.*, 1990). This structural element inserts into a recess in the surface of HIV gp120 (Kwong *et al.*, 1998). Compared with most IgSF members, ICAM-1, the receptor for the major group of rhinoviruses, has a unique, tapering tip that inserts into the narrow 'canyon' on the rhinovirus surface, where the conserved receptor-binding residues lie (Kolatkar *et al.*, 1999). The measles virus receptor CD46 belongs to the complement control protein (CCP) superfamily. The center of the virus-binding epitope of CD46 is a well-structured, protruding DD' loop consisting of a small group of hydrophobic residues with the key Pro39





**Fig. 3.** A comparative view of structures of several virus receptors, including: msCEACAM1a, the receptor for murine coronavirus MHV; ICAM1, the receptor for the major group of rhinoviruses; CD4, the primary receptor for HIV; and CD46, the receptor for measles virus. Only their N-terminal domains are shown here. The key virus-binding motifs with unique topological features are highlighted in red.

extending furthest out (Figure 3) (Casasnovas *et al.*, 1999). Thus, unique protruding hydrophobic residues on cell adhesion molecules might be prime targets for virus binding.

### **MHV receptor activities of murine CEACAM isoforms, chimeras and mutants**

The various natural isoforms of the murine CEACAM1a, CEACAM1b and CEACAM2 glycoproteins differ markedly in their virus binding, neutralization and virus receptor activities (Dveksler *et al.*, 1993a; Ohtsuka *et al.*, 1996; Gallagher, 1997; Zelus *et al.*, 1998). A series of soluble or anchored mutant murine CEACAM proteins with various point mutations, deletions or domain exchanges with other CEA-related glycoproteins has been tested for virus binding and receptor activities (Rao *et al.*, 1997; Wessner *et al.*, 1998). Several particularly intriguing observations were made. MHV-A59 and soluble spike protein bound better to D1 of murine CEACAM1a from MHV-susceptible mice than to CEACAM1b from MHV-resistant mice. Soluble murine CEACAM1b[1–4] had 4- to 10-fold less virus neutralization activity for MHV-A59 than msCEACAM1a[1–4]. Surprisingly, msCEACAM1b[1–4] failed to neutralize the neurotropic JHM strain of MHV, and msCEACAM1b[1,4] failed to neutralize either MHV-A59 or MHV-JHM (Zelus *et al.*, 1998). While the naturally occurring two-domain CEACAM1a[1,4] isoform neutralized MHV-A59 nearly as well as the four-domain isoform CEACAM1a[1–4], the CEACAM1a[1,2], consisting of D1 and D2 domains, had only minimal MHV-A59-neutralizing activity. Thus, there is virus strain specificity in the interactions of MHV with various CEACAM1 proteins, and regions of CEACAM1

outside of the virus-binding domain (D1) can affect virus–receptor activity.

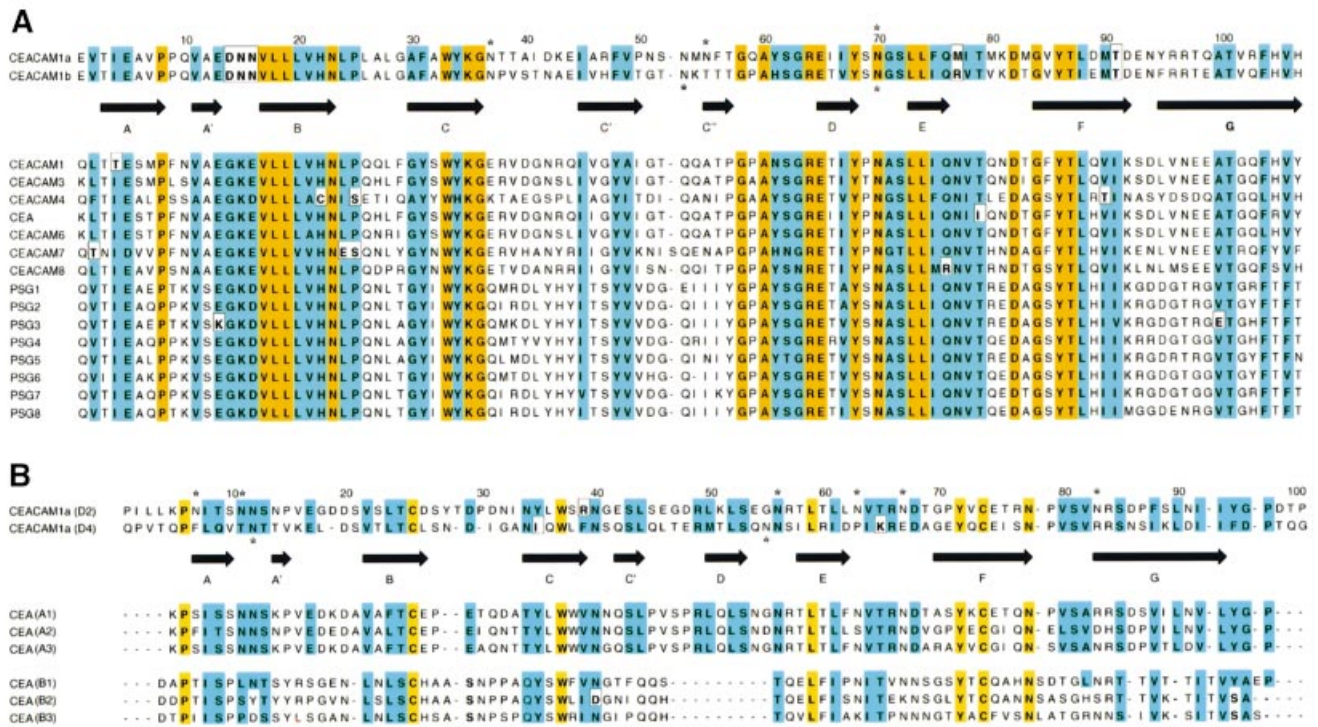
The amino acid sequences of murine CEACAM1a and CEACAM1b differ, principally in the N-terminal virus-binding domain (Dveksler *et al.*, 1993a). The lengths of the 1a and 1b proteins are the same, and all of the structurally important residues are the same or similar. The overall folding of murine CEACAM1b isoforms is therefore believed to be the same as or similar to that of the corresponding CEACAM1a isoforms. Figure 4A (top) shows the sequence alignment of D1 from murine CEACAM1a and CEACAM1b. The most extensive differences between CEACAM 1a and 1b are in the peptide segment from the CC' loop to the end of the C' strand, which plays a role in virus binding. In D1 of CEACAM1b, residue Ile41 is replaced by a threonine, which may account for its low virus-binding activity relative to CEACAM1a. It is possible that a projecting Val39 on the CC' loop of CEACAM1b might provide an alternative but weaker virus-binding hot spot as Ile41 does for CEACAM1a.

An intriguing question is why the C-terminal deletion mutant msCEACAM1a[1,2] has very little virus-neutralizing activity, while the soluble form of the naturally occurring murine CEACAM1a[1,4] isoform neutralizes virus as well as the msCEACAM1a[1–4] isoform (Zelus *et al.*, 1998). Analysis of the sequence alignment of domains 2 (D2) and 4 (D4) of CEACAM1a reveals two major differences (Figure 4B, top). The BC loop of D2 is two residues longer than that of D4, and D2 has four more potential *N*-glycosylation sites than D4 (marked with asterisks in Figure 4B). The longer BC loop of D2 and the possible glycan attached to Asn192 at the beginning of the G strand of D2 may both restrict inter-domain flexibility between D1 and D2 in msCEACAM1a[1,2] in comparison with the junction between D1 and D4 in msCEACAM1a[1,4]. Moreover, model building (data not shown) suggests that there might be a potential hydrogen bond between His107 of D1 and Asn141 of D2, while no such hydrogen bond is possible at this site in the junction of D1 and D4. All of these structural differences could cause the D1–D2 junction to be less flexible than the highly flexible junction between D1 and D4 that was revealed by X-ray crystallography. The four domain isoform CEACAM1a[1–4] has two more interdomain junctions than the truncated CEACAM1a[1,2] protein, and may therefore be more flexible.

### **Predicted structures of other CEA family members and conservation of a glycan-shielded surface hydrophobic patch in the N-terminal domain**

CEA family members are all composed of several Ig-like domains in tandem. Following the N-terminal domain, two similar types of domains, called A and B, alternate along the chain. For example, CEA (CD66e), encoded by the CEACAM5 gene, has the N-A1-B1-A2-B2-A3-B3 domain structure (Hammarstrom, 1999).

BLAST search (<http://www.ncbi.nlm.nih.gov/BLAST/>) of D1 of murine CEACAM1a found sequences of N-terminal domains of all mammalian CEA members. Five residues appear to be absolutely conserved: Trp33, Arg64, Leu73, Asp82 and Tyr86. The sequence alignment of N-terminal domains of human CEA family members is



**Fig. 4.** Sequence alignment of D1 and D4 of murine CEACAM1 with corresponding domains of human CEA family members. Residues invariant throughout all sequences shown are colored yellow, while physico-chemically conserved residues (with no more than two exceptions) are colored blue. The  $\beta$ -strands are indicated with arrows below the appropriate sequence. (A) D1 of murine CEACAM1a is aligned with D1 of murine CEACAM1b (top), as well as the human CEA members found in the SWISSPROT database (bottom). (B) The I1 set Ig-fold of D4 of murine CEACAM1a is aligned with D2 of the same molecule (top). These sequences are compared with the presumed I1 set of domains A1, A2, A3 and presumed I2 set of domains B1, B2, B3 of CEA (CD66e) (bottom). The asterisks indicate potential *N*-glycosylation sites.

shown in Figure 4A (bottom). No significant deletions or insertions were found in D1 of human CEA-related proteins, except for a few cases in which the length of the C'C' loop varied slightly. Like D1 of murine CEACAM1a, the N-terminal domains of the human CEA family members in Figure 4A can be classified as V set Ig-like fold, as predicted previously (Bates *et al.*, 1992). This is determined by these key conserved structural features (Chothia *et al.*, 1998): Pro8 at the A-A' kink point; Trp33 on the C strand that acts as the center of a hydrophobic core; a salt bridge between Arg64 and Asp82; and the tyrosine-corner motif (Hemmingsen *et al.*, 1994) D\*G\*Y86 at the beginning of the F strand.

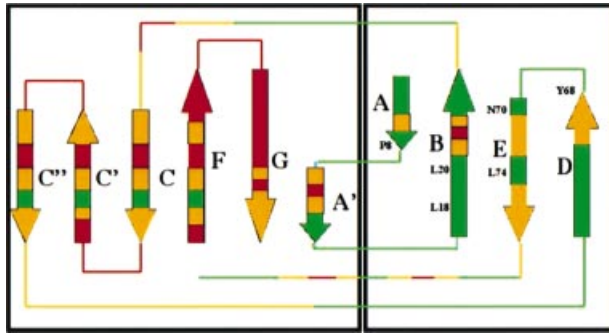
One newly recognized, highly conserved structural feature of msCEACAM1a[1,4] that appears to be unique to CEA family members (listed in Figure 4A) is the glycosylation site at Asn70, on the opposite side of D1 from the proposed virus-binding surface (Figure 1). In the crystal structure of msCEACAM1a[1,4], the glycan at Asn70 is better ordered than other glycans. Beneath the presumably large glycan at Asn70 lies a group of hydrophobic residues, including Val7 and Pro8 of the A strand, Leu18 and Leu20 of the B strand, Leu74 of the E strand, and probably also Tyr68 and Ile66 of the D strand. The area covers  $\sim 650 \text{ \AA}^2$ . The glycan at Asn70 appears to stabilize the protein by preventing the exposure of this large surface hydrophobic patch. Most of these protected amino acid residues are either invariant (Pro8

and Leu18) or very conserved (Leu20, Tyr68 and Leu74) among CEA proteins (Figure 4A). It is well known that glycans stabilize protein folding. Nevertheless, to our knowledge, msCEACAM1a[1,4] provides the first structure example for a large, glycan-shielded surface hydrophobic patch that is conserved in a protein family. The biological significance of this remarkable structural feature of the CEA family is not yet clear.

To assess the pattern of sequence conservation for all members of the mammalian CEA family in the SWISSPROT database, we calculated the variability in sequence using Shannon's entropy (Stewart *et al.*, 1997). Figure 5 shows a topology diagram of D1 of msCEACAM1a[1,4], colored to indicate the relative degree of conservation of residues calculated for 42 CEA family members. The green, yellow and red colors represent the most to the least conserved residues, respectively. This figure shows a striking difference in the extent of amino acid conservation between the two faces of D1 among CEA family members. The ABED face containing the glycan-shielded hydrophobic patch is much more conserved than the CFG face. The CFG faces of the N-terminal domains of IgSF proteins are frequently used for cell surface recognition (Stuart and Jones, 1995; Wang and Springer, 1998). The variability in this face among CEA members probably confers their unique binding specificities.

At the bottom of Figure 4B, the sequences of the six A and B type domains of the human CEA protein are aligned





**Fig. 5.** Topology diagram for D1 of msCEACAM1a with  $\beta$ -strands shown as arrows. The diagram is colored according to the degree of variability in sequence of N-terminal domain for all available mammalian CEA molecules. The variability was measured using Shannon's entropy value ( $H$ ) (Stewart *et al.*, 1997). The least variable, or most conserved, residues ( $H < 1$ ) are colored green, while the most variable ones ( $H > 2$ ) are colored red. Residues in between ( $1 < H < 2$ ) are colored yellow. The difference in the degree of sequence conservation between the ABED and CFG faces is striking. On the ABED face, the glycan at Asn70 and the shielded hydrophobic residues are marked.

with D2 and D4 of murine CEACAM1a. The three A type domains of human CEA, and probably also the A domains of other CEA members, are structurally very homologous to D4 of murine CEACAM1a, an I1 set of Ig fold. The B type domains of human CEA appear to have no D strand, but probably a C' strand that connects directly to the E strand, as observed for I2 set of Ig fold (Wang and Springer, 1998). Both I1 and I2 sets differ from the C set by having the A–A' kink, and they are distinct from the V set in not having the C'' strand (Wang and Springer, 1998). In summary, our data suggest that the general architecture of all CEA family members consists of a V set N-terminal domain followed by alternating I1 and I2 set Ig-like domains.

#### **The CC' and FG loops of the N-terminal domains of various CEA family members may mediate biologically important molecular interactions**

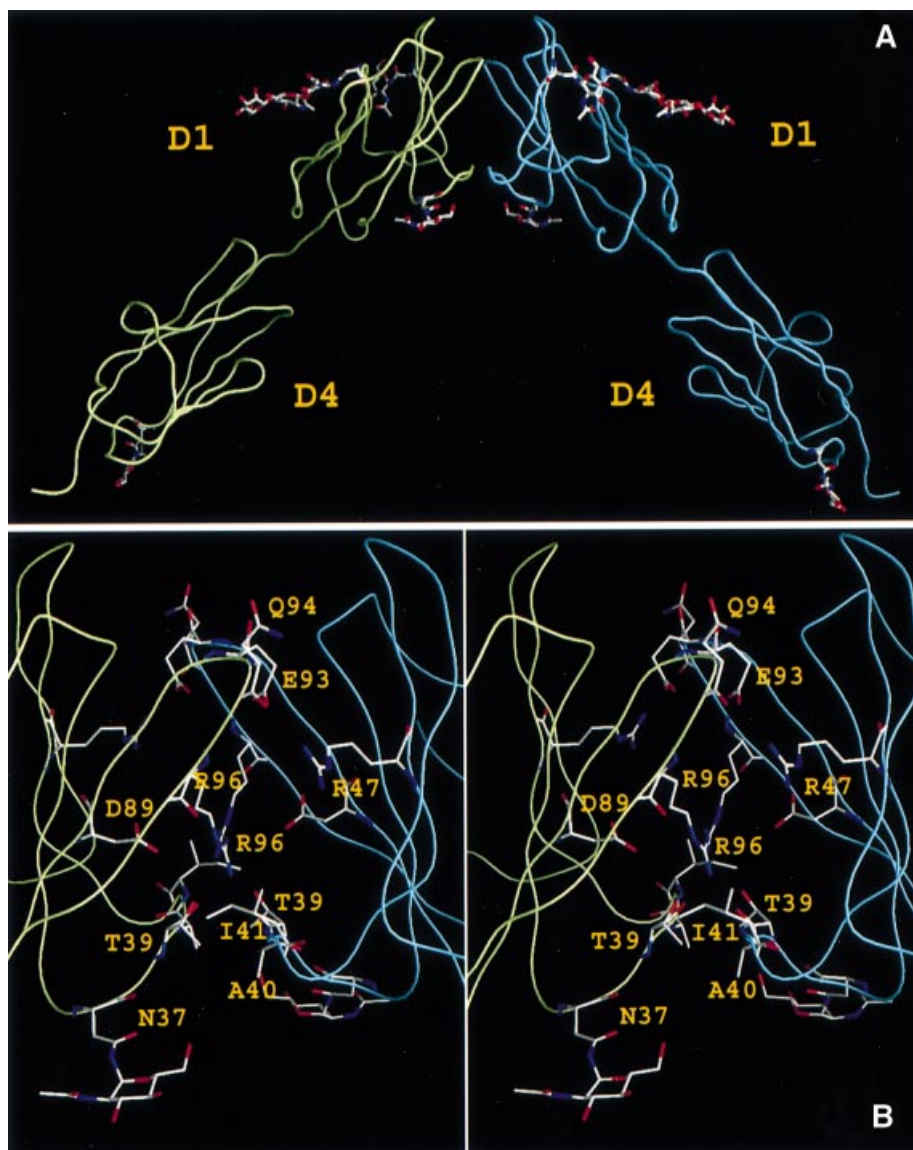
Given the high structural homology, the structure of murine CEACAM1a can be used to elucidate other molecular interactions of CEA family members including bacterial binding, immunomodulation, and homophilic and heterophilic adhesion.

Certain human CEA family members are subverted as receptors for bacterial pathogens, including *H.influenzae*, *N.meningitidis* and *N.gonorrhoeae*. The N-terminal domains of many human CEA members are recognized by multiple opacity-associated (Opa) proteins on the surface of pathogenic strains of *Neisseria* (Bos *et al.*, 1999; Virji *et al.*, 1999). Homolog scanning mutagenesis revealed that Phe29, Ser32 and Gly41 (and to a lesser extent Gln44) of CEA (CD66e) are required for maximal Opa protein binding activity (Bos *et al.*, 1999). Tyr34 and Ile91 (and to a lesser extent Val39 and Gln89) of human CEACAM1 (CD66a) are critical residues for most Opa protein interactions (Virji *et al.*, 1999). Since the N-terminal domains of CEA and human CEACAM1 are the same length as that of murine CEACAM1a (Figure 4A), we show in Figure 2B that the *Neisseria*-binding residues on CEA and human CEACAM1 are on

the C strand through the CC' loop and on the F strand. Two points are worth noting. Val39 and Gly41 of human CEACAM1 and CEA, respectively (corresponding to Thr39 and Ile41 in msCEACAM1a[1,4]; Figure 2B), are on the tip of the CC' loop. If the CC' loops of CEA and CEACAM1 were as flat as that of the Bence-Jones protein REI (Figure 2A), then Val39 and Gly41 would not be close enough to other important Opa-binding residues to form an integrated binding site. This probably also explains why the Y34A mutation of human CEACAM1 abrogated binding of the majority of Opa proteins (Virji *et al.*, 1999), since the aromatic ring of this conserved Tyr34 is the key to maintaining the convoluted structure of the CC' loop, as shown for msCEACAM1a[1,4]. Thus, the CC' loops of CEA and human CEACAM1 probably assume a convoluted conformation like that of msCEACAM1a[1,4]. The second point is that the area around Phe29 of CEA and Ile91 of human CEACAM1 (corresponding to Gly29 and Thr91 in msCEACAM1a[1,4]; Figure 2B) is highly hydrophobic and might be an important determinant of binding energy. Knowing the structure of msCEACAM1a[1,4] makes it possible to rationally design mutations to elucidate the molecular basis of the specific interactions between bacterial Opa proteins and CEA members on human cell membranes.

The PSG subfamily of the CEA family appears to be essential for a successful pregnancy, although the functions of PSGs are not yet fully understood. One hypothesis is that PSGs may attenuate the mother's immune response to her semi-allogeneic fetus (Hammarstrom, 1999). The N-terminal domains of most human PSGs, but not baboon or rodent PSGs, contain an Arg-Gly-Asp (RGD) motif (Zhou and Hammarstrom, 2001). The RGD motif is known to be associated with integrin binding, and mediates a wide variety of cell adhesion events. For example, in human fibronectin (FN), an integrin-binding RGD motif is located on a type II' turn at the tip of a protruded FG loop of the tenth FN domain (Leahy *et al.*, 1996). Figure 4A shows that in D1 of the human PSGs the RGD motifs are aligned at the very tip of the FG loop (highlighted in violet in Figure 1). The corresponding sequence in msCEACAM1a[1,4] is Glu92-Asn93-Tyr94 (Figure 4A), which assumes a type II  $\beta$ -turn. It is conceivable that those PSG proteins with an RGD motif can slightly change the conformation at the tip of the FG loop to adopt a type II' turn more suitable for integrin binding. The heterophilic binding of soluble PSGs to integrins might cause local immunosuppression in the uterus by shielding the integrins on cell membranes (Hammarstrom, 1999). In other species, PSGs lacking the RGD motif may still use one acidic residue (Glu or Asp) in the protruding FG loop (Zhou and Hammarstrom, 2001) to bind integrin, as demonstrated for leukocyte integrin ligands (Wang and Springer, 1998) and E-cadherin (Taraszka *et al.*, 2000).

CEA family members can mediate intercellular adhesion *in vitro* and *in vivo* through binding interactions that involve the N-terminal domain (Hammarstrom, 1999). Mutational analyses of the N-terminal domain (D1) of human CEACAM1 (Watt *et al.*, 2001) and CEA (Taheri *et al.*, 2000) showed that residues on the CFG face, and especially residues on the CC' loop of D1, are directly engaged in homophilic cell adhesion. Mutations V39A and



**Fig. 6.** Backbone worm representation of the 'parallel' interaction between the dyad-related msCEACAM1a[1,4] molecules seen in the crystal structure, prepared with GRASP (Nicholls *et al.*, 1991). (A) Two monomers related by a crystallographic 2-fold axis are shown in blue and green, respectively. Carbohydrates are drawn in ball-and-stick representation. (B) Stereo view of the close-up view across the dimer interface. The residues potentially involved in interactions are shown in ball-and-stick representation.

D40A in the CC' loop abolished homophilic adhesion of human CEACAM1 (Watt *et al.*, 2001).

To study possible mechanisms for homophilic binding of msCEACAM1a[1,4], we examined all molecular interactions observed in the crystal lattice of msCEACAM1a[1,4]. We found only two major contact areas between symmetry-related molecules: one through D1 by a 2-fold axis, and the other through D4 by a 3<sub>1</sub>-fold axis. The D1–D1 contact is more likely to have physiological significance than the screw-axis related D4–D4 contact. Figure 6 shows how the CC' and FG loops in D1s of two dyad-related molecules made contact in the crystal structure of msCEACAM1a[1,4]. Hydrophilic interactions appear to dominate the adhesive interface, like that between CD2 and CD58 (Wang *et al.*, 1999). As discussed above, the uniquely convoluted conformation of the CC' loop of msCEACAM1a[1,4] is likely to be similar for

human CEA members. The fact that Y34A, but not Y34F, mutation abrogated homophilic adhesion of CEA (Taheri *et al.*, 2000) shows the importance of the hydrophobic aromatic ring for maintaining the structure of the convoluted CC' loop, and the role of the CC' loop in homophilic adhesion. A convoluted, protruding CC' loop would likely prevent CEA molecules from adopting the 'hand-shaking' type of adhesion seen between CD2 and CD58. Figure 6B shows that Val39 of one human CEACAM1 molecule (corresponding to Thr39 in msCEACAM1a[1,4]) might have hydrophobic contact with Val39 from its symmetry mate, while Asp40 of CEA (corresponding to Ala40 of msCEACAM1a[1,4]; Figure 6B) might have electrostatic interaction with Arg38 (not shown in Figure 6) of the symmetry mate. This may explain why mutations V39A and D40A in CEACAM1 disrupt homophilic cell adhesion.



**Table I.** Data collection, structure determination and refinement

Data collection				
Data set	Pt peak <sup>a</sup>	Pt-inflection <sup>a</sup>	Pt-remote <sup>a</sup>	Native
Space group	<i>P</i> 3 <sub>1</sub> 21			<i>P</i> 3 <sub>1</sub> 21
Unit cell (Å)	<i>a</i> , <i>b</i> = 111.85, <i>c</i> = 66.34			<i>a</i> , <i>b</i> = 111.26, <i>c</i> = 65.64
X-ray source	APS			APS
Wavelength (Å)	1.0715	1.0718	1.0534	1.100
Resolution (Å)	20–3.85	20–3.85	20–3.85	30–3.32
Observations (unique)	49 179 (8681) <sup>a</sup>	50 389 (8645) <sup>a</sup>	45 774 (8566) <sup>a</sup>	123 640 (7127)
<i>I</i> / $\sigma$ overall	16.0 (3.1) <sup>b</sup>	15.2 (3.3) <sup>b</sup>	13.2 (2.3) <sup>b</sup>	17.3 (3.7) <sup>b</sup>
Completeness (%)	99.2 (91.8) <sup>b</sup>	99.6 (96.3) <sup>b</sup>	97.6 (82.9) <sup>b</sup>	99.7 (100.0) <sup>b</sup>
<i>R</i> <sub>merge</sub> (%)	7.5 (45.4) <sup>b</sup>	6.9 (42.3) <sup>b</sup>	8.0 (55.4) <sup>b</sup>	7.3 (37.1) <sup>b</sup>
Structure determination				
Figure of merit	0.49			
Phasing power	1.92	1.86	1.79	
<i>R</i> <sub>cullis</sub> (anomalous)	0.82	0.84	0.88	
<i>R</i> <sub>cullis</sub> (isomorphous)	0.60	0.61	0.61	
Structure refinement				
Resolution (Å)	15–3.32			
No. of work/test reflections	6144/754			
Non-hydrogen protein/carbohydrate/solvent atoms	1692/81/26			
<i>R</i> <sub>work</sub> / <i>R</i> <sub>free</sub> (%)	29.5/32.9			
Bond length (Å)/angle (°) r.m.s.d. from ideal geometry	0.011/2.325			
Ramachandran statistics (%) favourable/additional/generous/forbidden	68.5/23.4/8.2/0			
Protein atoms average <i>B</i> -value (Å <sup>2</sup> ), main chain/side chain	55.12/64.15			

<sup>a</sup>Bijvoet pairs are both counted.<sup>b</sup>Last resolution bin.

The ‘parallel’ mode of adhesion could occur between molecules on the same cell or opposing cells. The numerous inter-domain junctions of long CEA members may render them flexible enough to permit a *trans* interaction between opposing cells using this parallel mode. CHO cells transfected with human CEACAM1-1s, which has only the D1 domain as its extracellular portion, showed negligible adhesion despite a high level of protein (Watt *et al.*, 2001). Perhaps there was not enough flexibility in this short molecule to allow the parallel mode of binding. Further crystallographic studies and mutational analysis are needed to characterize *cis* or *trans* adhesion mechanisms between CEA family members.

## Materials and methods

### Protein expression and purification

Nucleotide sequences encoding the first 236 amino acids of murine CEACAM1a[1,4], including the natural 34 amino acid signal sequence, were amplified by PCR using an oligonucleotide that added an *Xba*I site in-frame at the 3' end. This DNA was ligated in-frame into a previously described construct encoding a thrombin cleavage peptide followed by six histidine residues and a stop codon (Zelus *et al.*, 1998), and inserted into the pShuttle CMV vector (He *et al.*, 1998). This construct was inserted into the pAd-Easy adenovirus vector, and adenoviruses that contained the cDNA were plaque purified and amplified in 293 cells as described previously (He *et al.*, 1998). Lec- CHO cells stably transfected with CAR, the Coxsackie/adenovirus receptor, were transduced with the CEACAM1a[1,4]-containing adenovirus. The soluble, His-tagged murine CEACAM1a[1,4] protein from the supernatant medium was purified by nickel affinity chromatography on a Pharmacia HiTrap chelating column, and eluted with imidazole. Fractions containing the protein were identified by immunoblotting with polyclonal rabbit antibody directed

against murine CEACAM1a, and the pooled fractions were dialyzed against 25 mM Tris buffer pH 9.0 with 5% glycerol. The protein was further purified by ion-exchange chromatography on a HQ20 (Poros) column and eluted in a NaCl gradient. Fractions containing the protein were pooled, dialyzed against 25 mM Tris pH 7.6, 150 mM NaCl, 5% glycerol, and stored at -80°C. The purity of the proteins was determined by silver staining of SDS-PAGE gels and by western blotting with anti-CEACAM1a antibody. The medium of 40 T150 flasks of adenovirus transduced lec-, CAR+ CHO cells yielded ~0.5–1 mg of purified msCEACAM1a[1,4] protein.

### Crystallization and X-ray data collection

Single crystals of msCEACAM1a[1,4] were grown from a crystallization buffer containing 10% PEG 8000, 0.2 M magnesium acetate and 0.1 M cacodylate at pH 6.4 using the vapor-diffusion hanging drop method. For data collection at cryogenic temperature, the crystals were treated with a cryoprotectant solution (25% glycerol, 10% PEG 8000 and 0.1 M cacodylate), and then frozen and stored in liquid nitrogen. Platinum derivatives were prepared by soaking the crystals overnight in the same cryo-protectant solution containing 0.5 mM K<sub>2</sub>PtBr<sub>4</sub>.

X-ray diffraction data were collected from pre-frozen crystals at APS SBC 19ID at a temperature of 100 K. A native crystal diffracted to a resolution of 3.32 Å, with one molecule in one asymmetric unit. A MAD data set of the platinum derivative was obtained to a resolution of 3.85 Å. All the raw data were indexed and reduced with HKL2000 (Otwinowski and Minor, 1997) (Table I).

### Structure determination and refinement

The msCEACAM1a[1,4] structure was solved using the MAD phases in combination with MR. Using programs in the CCP4 suite (CCP4, 1994), we located one platinum binding site in one asymmetric unit in both difference and anomalous difference Patterson maps. Heavy atom parameters were refined at 4 Å resolution with the program MLPHARE in the CCP4 suite, and an additional platinum site was identified. Phase extension was performed using the native data set to 3.32 Å by solvent flattening and histogram matching with DM. The resulting phases were used to carry out a phased molecular replacement with ROTPTF on the

Bronx X-ray server for the two separate domains. The N-terminal domains of human CD2 (PDB code 1HNF) and Fc- $\gamma$  receptor III (PDB code 1E4J) were used as search models for the D1 and D4 domains of msCEACAM1a[1,4], respectively. The model was traced with XtalView (<http://www.scripps.edu/pub/dem-web/>) on the basis of the MAD phases, using the MR solutions as a guideline.

After cycles of model building using program O (Jones *et al.*, 1991) and refinement, the final model was refined at 3.32 Å resolution to an  $R_{\text{free}}$  factor of 32.9% and  $R_{\text{work}}$  of 29.5% (Table 1) using X-PLOR (Brünger, 1992). At 1.5 $\sigma$  contour level ( $\sigma = 0.125 e/\text{\AA}^3$ ) in the  $2F_o - F_c$  map, there was continuous density for the main chain backbone. The final model contains 202 residues (from Glu1 to Pro202) of msCEACAM1a plus one amino acid (Ser) from the cloning construct and a total of six sugar residues associated with four of the five potential glycosylation sites. There was no interpretable electron density beyond residue Ser203, where 13 residues, including an inserted Arg204, a thrombin cleavage site and a His<sub>6</sub> tag are present in the expression construct. These C-terminal residues are apparently disordered. The current model also includes a total of 26 water molecules. The coordinates have been deposited in the PDB data bank under the accession code 1L67.

## Acknowledgements

The authors gratefully acknowledge Ellis Reinherz, Jerome Schaack, David Wentworth, Jamie Breslin, Larissa Thackray and Brian Turner for helpful discussions and critical reading of the manuscript. We thank Yuting Yang for help with data collection. This work was supported in part by NIH grants GM56008 and HL48675 to J.-h.W., AI 26075 and AI 25231 to K.V.H., and HL 54734 to J.M.B.

## References

- Bates,P.A., Luo,J. and Sternberg,M.J. (1992) A predicted three-dimensional structure for the carcinoembryonic antigen (CEA). *FEBS Lett.*, **301**, 207–214.
- Beauchemin,N. *et al.* (1999) Redefined nomenclature for members of the carcinoembryonic antigen family. *Exp. Cell Res.*, **252**, 243–249.
- Bergelson,J.M., Cunningham,J.A., Droguett,G., Kurt-Jones,E.A., Krithivas,A., Hong,J.S., Horwitz,M.S., Crowell,R.L. and Finberg,R.W. (1997) Isolation of a common receptor for Coxsackie B viruses and adenoviruses 2 and 5. *Science*, **275**, 1320–1323.
- Bos,M.P., Hogan,D. and Belland,R.J. (1999) Homologue scanning mutagenesis reveals CD66 receptor residues required for neisserial Opa protein binding. *J. Exp. Med.*, **190**, 331–340.
- Brünger,A.T. (1992) *X-PLOR. Version 3.1: A System for Crystallography and NMR*. Yale University Press, New Haven, CT.
- Casasnovas,J.M., Springer,T.A., Liu,J.H., Harrison,S.C. and Wang,J.-H. (1997) Crystal structure of ICAM-2 reveals a distinctive integrin recognition surface. *Nature*, **387**, 312–315.
- Casasnovas,J.M., Larvie,M. and Stehle,T. (1999) Crystal structure of two CD46 domains reveals an extended measles virus-binding surface. *EMBO J.*, **18**, 2911–2922.
- Chothia,C., Gelfand,I. and Kister,A. (1998) Structural determinants in the sequences of immunoglobulin variable domain. *J. Mol. Biol.*, **278**, 457–479.
- Clackson,T. and Wells,J.A. (1995) A hot spot of binding energy in a hormone–receptor interface. *Science*, **267**, 383–386.
- CCP4 (1994) The CCP4 suite: programs for protein crystallography. *Acta Crystallogr. D*, **50**, 760–763.
- Dveksler,G.S., Pensiero,M.N., Cardellichio,C.B., Williams,R.K., Jiang,G.S., Holmes,K.V. and Dieffenbach,C.W. (1991) Cloning of the mouse hepatitis virus (MHV) receptor: expression in human and hamster cell lines confers susceptibility to MHV. *J. Virol.*, **65**, 6881–6891.
- Dveksler,G.S., Dieffenbach,C.W., Cardellichio,C.B., McCuaig,K., Pensiero,M.N., Jiang,G.S., Beauchemin,N. and Holmes,K.V. (1993a) Several members of the mouse carcinoembryonic antigen-related glycoprotein family are functional receptors for the coronavirus mouse hepatitis virus-A59. *J. Virol.*, **67**, 1–8.
- Dveksler,G.S., Pensiero,M.N., Dieffenbach,C.W., Cardellichio,C.B., Basile,A.A., Elia,P.E. and Holmes,K.V. (1993b) Mouse hepatitis virus strain A59 and blocking antireceptor monoclonal antibody bind to the N-terminal domain of cellular receptor. *Proc. Natl Acad. Sci. USA*, **90**, 1716–1720.
- Epp,O., Lattman,E.E., Schiffer,M., Huber,R. and Palm,W. (1975) The molecular structure of a dimer composed of the variable portions of the Bence-Jones protein REI refined at 2.0-Å resolution. *Biochemistry*, **14**, 4943–4952.
- Ergun,S. *et al.* (2000) CEA-related cell adhesion molecule 1: a potent angiogenic factor and a major effector of vascular endothelial growth factor. *Mol. Cell*, **5**, 311–320.
- Gallagher,T.M. (1997) A role for naturally occurring variation of the murine coronavirus spike protein in stabilizing association with the cellular receptor. *J. Virol.*, **71**, 3129–3137.
- Gold,P. and Freedman,S.O. (1965) Specific carcinoembryonic antigens of the human digestive system. *J. Exp. Med.*, **122**, 467–481.
- Hammarstrom,S. (1999) The carcinoembryonic antigen (CEA) family: structure, suggested functions and expression in normal and malignant tissues. *Semin. Cancer Biol.*, **9**, 67–81.
- Harpaz,Y. and Chothia,C. (1994) Many of the immunoglobulin superfamily domains in cell adhesion molecules and surface receptors belong to a new structural set which is close to that containing variable domains. *J. Mol. Biol.*, **238**, 528–539.
- He,T.C., Zhou,S., da Costa,L.T., Yu,J., Kinzler,K.W. and Vogelstein,B. (1998) A simplified system for generating recombinant adenoviruses. *Proc. Natl Acad. Sci. USA*, **95**, 2509–2514.
- Hemmingsen,J.M., Gernert,K.M., Richardson,J.S. and Richardson,D.C. (1994) The tyrosine corner: a feature of most Greek key  $\beta$ -barrel proteins. *Protein Sci.*, **3**, 1927–1937.
- Huang,J., Hardy,J.D., Sun,Y. and Shively,J.E. (1999) Essential role of biliary glycoprotein (CD66a) in morphogenesis of the human mammary epithelial cell line MCF10F. *J. Cell Sci.*, **112**, 4193–4205.
- Huber,M., Izzi,L., Grondin,P., Houde,C., Kunath,T., Veillette,A. and Beauchemin,N. (1999) The carboxyl-terminal region of biliary glycoprotein controls its tyrosine phosphorylation and association with protein-tyrosine phosphatases SHP-1 and SHP-2 in epithelial cells. *J. Biol. Chem.*, **274**, 335–344.
- Izzi,L., Turbide,C., Houde,C., Kunath,T. and Beauchemin,N. (1999) *cis*-determinants in the cytoplasmic domain of CEACAM1 responsible for its tumor inhibitory function. *Oncogene*, **18**, 5563–5572.
- Jones,E.Y., Davis,S.J., Williams,A.F., Harlos,K. and Stuart,D.I. (1992) Crystal structure at 2.8 Å resolution of a soluble form of the cell adhesion molecule CD2. *Nature*, **360**, 232–239.
- Jones,T.A., Zou,J.-Y., Cowan,S.W. and Kjeldgaard,M. (1991) Improved methods for building protein models in electron density maps and location of errors in these models. *Acta Crystallogr. A*, **47**, 110–119.
- Kim,M., Sun,Z.Y., Byron,O., Campbell,G., Wagner,G., Wang,J. and Reinherz,E.L. (2001) Molecular dissection of the CD2–CD58 counter–receptor interface identifies CD2 Tyr86 and CD58 Lys34 residues as the functional ‘hot spot’. *J. Mol. Biol.*, **312**, 711–720.
- Kolattkar,P.R., Bella,J., Olson,N.H., Bator,C.M., Baker,T.S. and Rossmann,M.G. (1999) Structural studies of two rhinovirus serotypes complexed with fragments of their cellular receptor. *EMBO J.*, **18**, 6249–6259.
- Kraulis,P. (1991) MOLSCRIPT: a program to produce both detailed and schematic plots. *J. Appl. Crystallogr.*, **24**, 924–950.
- Kwong,P.D., Wyatt,R., Robinson,J., Sweet,R.W., Sodroski,J. and Hendrickson,W.A. (1998) Structure of an HIV gp120 envelope glycoprotein in complex with the CD4 receptor and a neutralizing human antibody. *Nature*, **393**, 648–659.
- Leahy,D.J., Aukhil,I. and Erickson,H.P. (1996) 2.0Å crystal structure of a four-domain segment of human fibronectin encompassing the RGD loop and synergy region. *Cell*, **84**, 155–164.
- Morales,V.M. *et al.* (1999) Regulation of human intestinal intraepithelial lymphocyte cytolytic function by biliary glycoprotein (CD66a). *J. Immunol.*, **163**, 1363–1370.
- Nakajima,A. *et al.* (2002) Carcinoembryonic antigen cell adhesion molecule 1 functions as an inhibitory activation molecule on mouse T lymphocytes. *J. Immunol.*, **168**, 1028–1035.
- Nedellec,P., Dveksler,G.S., Daniels,E., Turbide,C., Chow,B., Basile,A.A., Holmes,K.V. and Beauchemin,N. (1994) Bgp2, a new member of the carcinoembryonic antigen-related gene family, encodes an alternative receptor for mouse hepatitis viruses. *J. Virol.*, **68**, 4525–4537.
- Nicholls,A., Sharp,K.A. and Honig,B. (1991) Protein folding and association: insights from the interfacial and thermodynamic properties of hydrocarbons. *Proteins*, **11**, 281–296.
- Ohtsuka,N., Yamada,Y.K. and Taguchi,F. (1996) Difference in virus-binding activity of two distinct receptor proteins for mouse hepatitis virus. *J. Gen. Virol.*, **77**, 1683–1692.
- Otwinowski,Z. and Minor,W. (1997) Processing of X-ray diffraction data collected in oscillation mode. In Carter,C.W., Jr and Sweet,R.M.

- (eds), *Macromolecular Crystallography*. Academic Press, San Diego, CA, pp. 307–326.
- Rao,P.V., Kumari,S. and Gallagher,T.M. (1997) Identification of a contiguous 6-residue determinant in the MHV receptor that controls the level of virion binding to cells. *Virology*, **229**, 336–348.
- Stanley,P. (1989) Chinese hamster ovary cell mutants with multiple glycosylation defects for production of glycoproteins with minimal carbohydrate heterogeneity. *Mol. Cell. Biol.*, **9**, 377–383.
- Stewart,J.J., Lee,C.Y., Ibrahim,S., Watts,P., Shlomchik,M., Weigert,M. and Litwin,S. (1997) A Shannon entropy analysis of immunoglobulin and T cell receptor. *Mol. Immunol.*, **34**, 1067–1082.
- Stuart,D.I. and Jones,E.Y. (1995) Recognition at the cell surface: recent structural insights. *Curr. Opin. Struct. Biol.*, **5**, 735–743.
- Taheri,M., Saragovi,U., Fuks,A., Makkerh,J., Mort,J. and Stanners,C.P. (2000) Self recognition in the Ig superfamily. Identification of precise subdomains in carcinoembryonic antigen required for intercellular adhesion. *J. Biol. Chem.*, **275**, 26935–26943.
- Talbot,P.J. and Buchmeier,M.J. (1985) Antigenic variation among murine coronaviruses: evidence for polymorphism on the peplomer glycoprotein, E2. *Virus Res.*, **2**, 317–328.
- Taraszka,K.S., Higgins,J.M., Tan,K., Mandelbrot,D.A., Wang,J.H. and Brenner,M.B. (2000) Molecular basis for leukocyte integrin  $\alpha_E\beta_7$  adhesion to epithelial (E)-cadherin. *J. Exp. Med.*, **191**, 1555–1567.
- Virji,M., Evans,D., Hadfield,A., Grunert,F., Teixeira,A.M. and Watt,S.M. (1999) Critical determinants of host receptor targeting by *Neisseria meningitidis* and *Neisseria gonorrhoeae*: identification of Opa adhesiotopes on the N-domain of CD66 molecules. *Mol. Microbiol.*, **34**, 538–551.
- Virji,M., Evans,D., Griffith,J., Hill,D., Serino,L., Hadfield,A. and Watt,S.M. (2000) Carcinoembryonic antigens are targeted by diverse strains of typable and non-typable *Haemophilus influenzae*. *Mol. Microbiol.*, **36**, 784–795.
- Wang,J.-H. (2002) Protein recognition by cell surface receptors: physiologic receptors versus virus interactions. *Trends Biochem. Sci.*, **27**, 122–126.
- Wang,J.-H. and Springer,T.A. (1998) Structural specializations of immunoglobulin superfamily members for adhesion to integrins and viruses. *Immunol. Rev.*, **163**, 197–215.
- Wang,J.-H. *et al.* (1990) Atomic structure of a fragment of human CD4 containing two immunoglobulin-like domains. *Nature*, **348**, 411–418.
- Wang,J.-H., Smolyar,A., Tan,K., Liu,J.-H., Kim,M., Sun,Z.-Y.J., Wagner,G. and Reinherz,E.L. (1999) Structure of a heterophilic adhesion complex between human CD2 and CD58 (LFA-3) counter-receptors. *Cell*, **97**, 791–803.
- Wang,J.-H., Meijers,R., Xiong,Y., Liu,J.-H., Sakihama,T., Zhang,R., Joachimiak,A. and Reinherz,E. (2001) Crystal structure of the human CD4 N-terminal two domain fragment complexed to a class II MHC molecule. *Proc. Natl Acad. Sci. USA*, **98**, 10799–10804.
- Watt,S.M. *et al.* (2001) Homophilic adhesion of human CEACAM1 involves N-terminal domain interactions: structural analysis of the binding site. *Blood*, **98**, 1469–1479.
- Wessner,D.R., Shick,P.C., Lu,J.H., Cardellicchio,C.B., Gagneten,S.E., Beauchemin,N., Holmes,K.V. and Dveksler,G.S. (1998) Mutational analysis of the virus and monoclonal antibody binding sites in MHVR, the cellular receptor of the murine coronavirus mouse hepatitis virus strain A59. *J. Virol.*, **72**, 1941–1948.
- Wu,H., Kwong,P.D. and Hendrickson,W.A. (1997) Dimeric association and segmental variability in the structure of human CD4. *Nature*, **387**, 527–530.
- Zelus,B.D., Wessner,D.R., Williams,R.K., Pensiero,M.N., Phibbs,F.T., deSouza,M., Dveksler,G.S. and Holmes,K.V. (1998) Purified, soluble recombinant mouse hepatitis virus receptor, Bgp1(b) and Bgp2 murine coronavirus receptors differ in mouse hepatitis virus binding and neutralizing activities. *J. Virol.*, **72**, 7237–7244.
- Zhou,G.Q. and Hammarstrom,S. (2001) Pregnancy-specific glycoprotein (PSG) in baboon (*Papio hamadryas*): family size, domain structure and prediction of a functional region in primate PSGs. *Biol. Reprod.*, **64**, 90–99.

Received January 17, 2002; revised February 21, 2002;  
accepted March 13, 2002