# Structural models for carcinoembryonic antigen and its complex with the single-chain Fv antibody molecule MFE23

Mark K. Boehm[a,b], Stephen J. Perkins[a,*]

[a]*Department of Biochemistry and Molecular Biology, Royal Free Campus, Royal Free and University College Medical School, University College London, Rowland Hill St., London NW3 2PF, UK*
[b]*Department of Oncology, Royal Free Campus, Royal Free and University College Medical School, University College London, Rowland Hill St., London NW3 2PF, UK*

**Abstract** MFE23 is a single chain Fv antibody that has a high affinity for carcinoembryonic antigen (CEA). A full homology model for CEA based on V-type, I-type and C2-type immunoglobulin folds, 28 oligosaccharides and the interdomain angle of CD2 was validated using solution scattering data. The superimposition of the intermolecular contacts observed in our recent crystal structure of MFE23 with the N-terminal domain of CEA permitted the MFE23–CEA complex to be modelled. Good surface and electrostatic complementarity and carbohydrate-unhindered access of MFE23 with the indentation between the first two CEA domains was observed. The model is supported by biochemical data and provides insight on the high affinity of MFE23 for CEA. © 2000 Federation of European Biochemical Societies. Published by Elsevier Science B.V. All rights reserved.

*Key words:* Antibody; Carcinoembryonic antigen; Homology model; Structure prediction; Scattering; Single-chain Fv

## 1. Introduction

Antibodies and cell surface proteins represent two major groups of the immunoglobulin (Ig) superfamily [1]. While antibodies are constructed from V- and C1-type Ig folds, cell-surface proteins such as the carcinoembryonic antigen (CEA) are constructed from combinations of V-, C1-, C2- and/or I-type Ig folds [2]. Structures by both crystallography and constrained scattering modelling show that Ig folds are arranged as extended single chains or form associations with other chains via domain pairs [3–5]. This superfamily is frequently involved in protein–protein interactions. The single chain Fv antibody molecule MFE23 has a nanomolar affinity for CEA, an important tumour marker in colorectal cancer [6,7]. We have determined the crystal structure of unliganded MFE23 [8]. In this structure, there is a large region of contact between the three antigen-binding loops *H1*, *H2* and *H3* of MFE23 and the β-sheet of an adjacent MFE23 molecule.

These intermolecular contacts displayed remarkably high surface and electrostatic complementarity. Several individual MFE23 loop residues that participate in these contacts are involved in CEA binding as supported by site-specific mutagenesis [8].

A structure for CEA is required for modelling the structure of its complex with MFE23. As its 50% carbohydrate content has precluded attempts to crystallise it, an alternative strategy based on small-angle X-ray and neutron scattering data and modelling based on V- and C2-type Ig folds showed that its structure is well represented by an extended zig-zag domain arrangement [4]. This curve modelling was based on a recently developed automated strategy directly constrained using known atomic structures [9], which is advantageous compared to the recent use of spherical harmonics or genetic algorithms for modelling proteins which do not make use of these known structures in curve fits. Since then, an I-type Ig domain was shown to exist that is structurally distinct from the C2-type domain, of which three copies occur in CEA [2,10]. I-type crystal structures for two cell-surface proteins have been determined [11–14]. These developments enable significantly improved homology models for each of the seven glycosylated domains of CEA to be constructed. Accordingly we present a full model for CEA and validate it with small-angle X-ray and neutron scattering data.

The combination of experimentally deduced structures for MFE23 and CEA with the observed intermolecular lattice contacts seen in the crystal structure of MFE23 enabled a model for their complex to be constructed by superimposition based on structural homology between the MFE23 and CEA domains. The model was well-supported by good surface and electrostatic complementarity and carbohydrate-unhindered access. It is also supported by biochemical data and clarifies reasons for the high affinity of the complex. Such a modelling strategy may offer a new approach for structural predictions of protein–protein complexes prior to crystal structure determinations, which is otherwise computationally complex to achieve [15].

## 2. Materials and methods

### 2.1. Seven homology models for the CEA domains

Homology models for the CEA-1 to CEA-7 domains (also known as N, IA, IB, IIA, IIB, IIIA and IIIB in that order) were constructed using INSIGHT II, BIOPOLYMER, HOMOLOGY and DISCOVERY (MSI, San Diego, CA, USA) on Silicon Graphics INDY workstations. Loops were built using a database derived from 349 crystal structures at 0.2 nm resolution or better [16]. Energy refinements based on the consistent valence force field were performed at the

*Corresponding author. Fax: (44)-20-7794 9645.
E-mail: steve@rfhsm.ac.uk

Abbreviations:* CEA, carcinoembryonic antigen; CEA-1 to CEA-7, domain numbering in CEA; H1-H120 and L1-L106, sequence numbering of MFE23 heavy and light chains respectively; *H1-H3* and *L1-L3*, antigen-binding loops of MFE23; ICAM-2, intercellular cell adhesion molecule-2; Ig, immunoglobulin; PDB, protein data bank; $V_H$, variable heavy chain domain; $V_L$, variable light chain domain; VCAM-1, vascular cell adhesion molecule

loop splice junctions, then on the sidechain atoms of all residues in the structurally conserved regions, and all the atoms of both types of loop residues. Iterations were made using the steepest descent algorithm to improve connectivities and minimise bad contacts or stereochemistries. The secondary structure backbone was retained by fixing the mainchain atoms in the conserved regions, and tethering these in the loop regions. Models were verified using PROCHECK [17]. Solvent accessibilities were calculated using a water molecule probe of 0.14 nm diameter in COMPARER [18,19]. Secondary structures were identified using DSSP [20].

The CEA-1 domain (residues 1–108) was modelled using the N-terminal V-type structures from the human cell surface proteins CD2, CD4 and CD8 as templates in order to allow for insertions or deletions in the loop regions (residues 4–104 in human CD2; residues 1–99 in human CD4; residues 2–114 in human CD8; protein data bank (PDB) codes: 1hnf, 3cd4 and 1cd8 respectively) [21–23]. The conserved β-strands (Fig. 1a) were superimposed using the conserved Trp residue in β-strand C as a reference. Each of the 10 β-strands and designated loops in the CEA-1 homology model was based on the template with the highest sequence similarity to CEA-1 (underlined in Fig. 1a). Eight structurally-conserved β-strands (37 residues) and four designated conserved loops (39 residues) were identified, while five loops that displayed insertions and deletions were reconstructed from database searches (32 residues: not underlined in Fig. 1a).

Homology models for the CEA-2 to CEA-7 domains were constructed as follows. Residues 3–94 of the CEA-2, CEA-4 and CEA-6 domains were modelled using residues 1–90 in human vascular cell adhesion molecule-1 (VCAM-1) which is an I-type domain (PDB code: 1vca-A) [11]. This utilised seven structurally conserved β-strand regions (38 residues), seven designated loops (50 residues) and one searched loop (four residues) that corresponded to an insertion between β-strands F and G at residues 78–81 (Fig. 1b). Residues 2–84 of the CEA-3, CEA-5 and CEA-7 domains were modelled using residues 104–181 of human CD2 (PDB code: 1cd8) [23]. This utilised seven structurally conserved β-strand regions (29 residues), six designated loops (38 residues) and two searched loops that corresponded to short insertions or deletions (16 residues) between β-strands A and B and β-strands F and G (Fig. 1c).

Two full models of CEA were created by superimposition of the seven homology models onto each of the best-fit and CD2-derived CEA models determined by small-angle X-ray and neutron scattering [4]. The inter-domain linker residues (nine residues) were represented

by searched loops. Triantennary carbohydrate structures $Man_3$-$GlcNAc_6Gal_3Fuc_3NeuNAc_1$ was added to each of the 28 putative N-linked sites on each CEA model as before. Electrostatic surfaces were calculated using DELPHI (MSI, San Diego, CA, USA). Red represents a potential of less than −4 kT (acidic), blue a potential of more than +4 kT (basic) and white as 0 kT (neutral).

### 2.2. Scattering curve fits for CEA

Debye sphere models were created by placing the atomic coordinates of the two CEA models in a three-dimensional grid of cubes of side 0.571 nm. A sphere of volume equal to a single cube was placed at the centre of each cube if a specified number of atoms was present in the cube. The sphere cutoff was based on the requirement that the total volume of spheres was that of the dry volume [24], and this model contained 970 spheres. Hydration is not detectable by neutron scattering and was not considered. The hydration shell is observed in X-ray scattering experiments, so a hydrated sphere model was created by uniformly expanding the dry model in order to include the additional volume of hydration of 0.3 g $H_2O$/g of glycoprotein, based on an electrostricted volume of 0.0245 $nm^3$ per bound water molecule [24]. Scattering curves for comparison with the experiment were calculated by the Debye equation assuming a uniform sphere scattering density in the program SCT [25]. This procedure has been tested with crystal structures in a molecular weight range of 23 000–127 000 [26–28]. A full-width-half-height wavelength spread of 10% for $\lambda$ of 1.0 nm and a beam divergence of 0.024 radian were used to correct the calculated curve [27]. The agreement between the modelled and experimental curves was determined using the $R_G$ value derived from the calculated curve in the same $Q$ range used for experimental Guinier fits and the $R$-factor for the $Q$ range extending to 2 $nm^{-1}$ [27].

## 3. Results and discussion

### 3.1. Homology modelling of the CEA domains

The seven domains of CEA had been previously classified and modelled on the basis of Ig folds in the order V-C2-C2-C2-C2-C2-C2 [1,4,29]. Since then, the second and fourth CEA domains (CEA-2 and CEA-4) had been reassigned as I-type domains [10]. I-type structures are intermediate between V-type and C1-type structures. Using the standard Ig nomen-

```
(a) V-type domain                  5   10   15   20   25   30      35   40      45   50   55      60   65   70   75   80   85   90   95  100  105
                                   ....|....|....|....|....|....|   ....|....|   ....|....|.. ..|.. ..|....|....  |....|....|....|....|. ..|....|....|..
Human CEA-1  (N)     : KLTIESTPFNVAEGKEVLLLVHNLPQHLFG  YSWYKGERVD   GNRQIIGYVIGT QQATP  GPAYSGREIIYP  NASLLIQNIIQNDTGFYTLHVI KSDLVNEEATGQFRVYP
Human CD2    (domain 1): TNALETWGALGQDINLDIPSFQMSDDIDDIKWEKTS   DKKKIAQFRKEKETFKEK  DTYKLF  KNGTLKIKHLKTDDQDIKVSIYDTKGKNVLEKIFDLKIQE
Human CD4    (domain 1): KKVVLGKKGDTVELTCTASQKKSIQ  FHWKNSN   QIKILGNQGSF  LTKGPSKLNDRADSRRSLWDQGNFPLILIKNLKIEDSDTYICEVEDQK   EEVQLLVFG
Human CD8    (domain 1): SQFRVSPLDRTWNLGETVELKCQVLLSNPTSG  CSWLFQPRGAAASPTFLLYLSGNR PKAAEGLDTQRFSGKRLG   DTFVLTLSDFRRENEGYYFCSAL SNSIMYFSHFVPVFLPA
Beta strands         : <-A-> <-A'-> <--B-->    <----C--->      <--C'-> <C"->     <--D->    <-E->     <----F----> <------G----->
Human CD2    (1hnf 1): ...EEEEETTS.EEE..TT....SSEEEEEEEGG   G..EEEEE.GGG.EEES.   TTEEE.    TTS.EEE.S..GGG.EEEEEEEETTS.EEEEEEEEEEE.
Human CD4    (3cd4 1): .EEEEEETTS.EEE..B.SSSSS..  EEEEETT        ..EEEEEETTE EE..SSTTTTTEE..GGGGGGTB..EEE.S..GGG.EEEEEEETTEE      EEEEEEEE
Human CD8    (1cd8 1): ..EEEE..S....TT..EEEEEEE..SS.SS. EEEEEESSSSS...EEEEEE.SS. .EE.TT..TTTEEEEEET   TEEEEEEE.S..GGG.EEEEEEEE ETTEEEE...EEE....
------------------------------------------------------------------------------------------------------------------------------------
(b) I-type domain                  5   10   15   20   25   30   35   40   45   50   55   60   65   70   75   80   85   90
                                   ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....
V-frame profile      : .................*..*.*.*........*.*........*.*.....**.......*.*.*..*.*......*.
Human CEA-2  (IA)    : ELPKPKPSISSNNSKPVEDKDAVAFTCEPETQDATYLWWVNNQSLPVSPRLQLSNGNRTLTLFNVTRNDTASYKCETQNPVSARRSDSVILNVLYG
Human CEA-4  (IIA)   : EPPKPKFITSNNSNPVEDEDAVALTCEPEIQNTTYLWWVNNQSLPVSPRLQLSNDNRTLTLLSVTRNDVGPYECGIQNELSVDHSDPVILNVLYG
Human CEA-6  (IIIA)  : ELPKPKPSISSNNSKPVEDKDAVAFTCEPEAQNTTYLWWVNNQSLPVSPRLQLSNGNRTLTLFNVTRNDARAYVCGIQNSVSANRSDPVTLDVLYG
Human VCAM   (domain 1): FKIETTPESRYLAQIGDSVSLTCSTTGCESPFFSWRTQIDSPLNGKVTNEGTTSTLTMNPVSFGNEHSYLCTATCES  RKLEKGIQVEIYS
Beta strands         : <-A-> <-A'>  <---B--->    <--C->      <-D-> <--E->     <----F-->  <-----G---->
Human VCAM   (1vca 1): .EEEEESSSEEEEETTS.EEEEEEEES.SS.EEEEEETT....SSEEEEETTEEEEEESS..GGG.SEEEEEEEETT EEEEEEEEEEEEB
------------------------------------------------------------------------------------------------------------------------------------
(c) C2-type domain                 5   10   15   20   25   30   35   40   45   50   55   60   65   70   75   80
                                   ....|....|....|....|....|....|....|....|....|....|....|....|....|....|....|....
Human CEA-3  (IB)    : PDAPTISPLNTSYRSGENLNLSCHAASNPPAQYSWFVNGTFQQSTQELFIPNITVNNSGSYTCQAHNSDTGLNRTVTTITVYA
Human CEA-5  (IIB)   : PDDPTISPSYTYYRPGVNLSLSCHAASNPPAQYSWLIDGNIQQHTQELFISNITEKNSGLYTCQANNSASGHSRTTVKTITVSA
Human CEA-7  (IIIB)  : PDTPIISPPDSSYLSGANLNLSCHSASNPPSPQYSWRINGIPQQHTQVLFIAKITPNNNGTYACFVSNLATGRNNSIVKSITVSA
Human CD2    (domain 2): ERVSKPKISWT  CINT TLTCEVMNGTDPELNLYQDGKHLKLSQRVITHKWTTSLSAKFKCTAGNKVS KESSVEPVSCPEK
Beta strands         : <-A->     <-B->    <--C->  <-C'>  <-E>    <----F---->    <----G---->
Human CD2    (1hnf 2): ......EEEEE  TTTT EEEEE.SS.SS.EEEEEESS.EEEEESS.EEEE..S.EEEEEEEEEEE.SS. EEEEEEEEEE...
```

Fig. 1. Sequence alignment used for the modelling of CEA based on homologous crystal structures. The 28 oligosaccharide sites in CEA are denoted by bold underlining. Sequences are labelled with their PDB code, and underlined regions indicate the structures used to construct the CEA homology models. The β-strands (E) identified by the DSSP program are labelled A to G. a: The V-type domain of CEA is compared with the first V-type domain in CD2, CD4 and CD8. b: The three I-type domains of CEA are compared with that found in the first domain of VCAM-1. The V-frame profile that resulted in the assignment to an I-type domain is shown asterisked above the alignment. c: The three C2-type domains of CEA are compared with that found in the second domain of the cell-surface protein CD2.

clature for the sequential lettering of β-strands from A to F, the I-type structure contains DEBA and A′GFCC′ β-sheets, in distinction to the V-type structure with DEBA and A′GFCC′C″ β-sheets, the C1-type fold with DEBA and GFCC′ β-sheets, and the C2-type fold with EBA and GFCC′ β-sheets [2,30]. The determination of I-type crystal structures for VCAM-1 and intercellular cell adhesion molecule-2 (ICAM-2) [11–14] verified the existence of this structure. The previous alignment of the CEA-2, CEA-4 and CEA-6 sequences with those of C2-type domains had resulted in large deletions and insertions [4,29]. Fig. 1b showed that the CEA-2, CEA-4 and CEA-6 sequences are well aligned with the V-frame profile used to identify I-type domains [10]. The V-frame matches encompasses residues from β-strands B to G, but not from β-strand A which has a distinct structure in different domains (see below) [10]. Only position 18 in the CEA-2, CEA-4 and CEA-6 domains and position 69 in the CEA-6 domain are minor exceptions to the 18 residues that define this profile.

Homology models were created for the seven domains of CEA (CEA-1 to CEA-7). In general, as expected for Ig folds, the sequence identities with the Ig crystal structures used for modelling were low at 12–20%, and this is increased to 28–37% if residue similarities were considered [4]. There is a high conservation of buried hydrophobic residues within β-strands. The CEA-1 model was constructed from the superimposition of the V-type domains from human CD2, CD4 and CD8 (Section 2). The five loops that had to be rebuilt were located either at the N-terminal tip of the CEA-1 domain (i.e. in the same location as the antigen-binding site in antibody V-type domains) or at the exposed edge C″ β-strand in the A′GFCC′C″ β-sheet (Fig. 1a). The CEA-2, CEA-4 and CEA-6 models were constructed from the I-type domain of VCAM-1, whose sequence is very similar in length. As the alignment of Fig. 1b indicated that only one small insertion in a loop was needed, this provided the most straightforward modelling template. A proline sometimes occurs between the A and A′ β-strands of I-type folds in VCAM-1 and ICAM-2 and defines the switch between these [11,14]. Prolines are con-

served in CEA-2, CEA-4 and CEA-6 in this region at positions 3, 5 and 14, together with a conserved putative carbohydrate site at position 10. As a choice of structure here was not possible, the modelling of this segment is approximate. There is also a one-residue displacement of the V-frame hydrophobic residues in β-strands C and G of VCAM-1. In summary, compared to the C2-type derived modelling, the changeover to an I-type structure repositioned residues presently assigned to β-strands A′ and C′ from the opposite face of the β-sheet sandwich. This has the effect of topologically relocating the carbohydrate chains at Asn-118, Asn-297 and Asn-474. The CEA-3, CEA-5 and CEA-7 models were based on the C2-type domain from human CD2 and required only two small insertions in loop regions (Fig. 1c).

### 3.2. Determination of full structural models for CEA

In the structure determination of CEA by small-angle X-ray and neutron scattering, an automated analysis of 4096 arrangements of one V-type and six C2-type Ig domains taken unmodified from the crystal structure of CD2 were rotated relative to each other in 15° steps. This had shown that only elongated zig-zag domain arrangements with 28 extended oligosaccharide chains fitted the data [4]. The mean interdomain rotation between two domains in the best-fit CEA model was determined to be within error of that seen in the CD2 crystal structure [4,21,31]. Here, two full structures for CEA were created by superimpositions of the new homology models for CEA-1 to CEA-7 onto the Ig folds of both these previous CD2-derived seven-domain models (Section 2). Both models were assessed using small-angle neutron and X-ray scattering fits using the data of [4]. Previously, to achieve a good fit, different scattering densities for each of the protein and carbohydrate components had to be used in addition to the inclusion of a hydration shell in the X-ray modelling. Here, both CEA homology models resulted in improved X-ray and neutron curve fits to the extent that only a single scattering density sphere model was needed to obtain a good fit. The best-fit CEA model had a modelled X-ray $R_G$ value of 7.9 nm, in good agreement with the experimental X-ray $R_G$
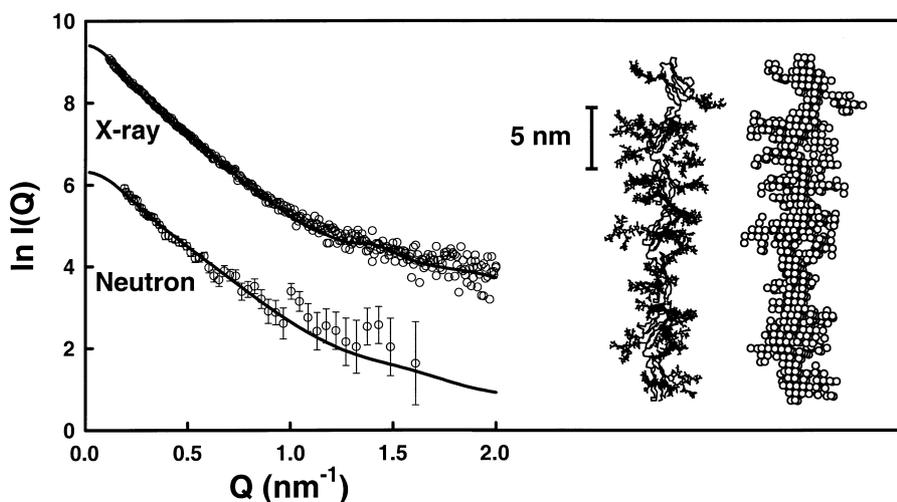


Fig. 2. Comparison of the calculated small-angle X-ray and neutron scattering curves for the homology model of CEA with experimental small-angle X-ray and neutron data. The seven CEA domains in each model are shown as α-carbon traces, whereas the carbohydrate chains are represented in full. The corresponding sphere model (sphere diameter, 0.571 nm) is also shown. The calculated neutron curve was corrected for wavelength resolution and beam divergence. For the experimental X-ray curve (○), the $R_G$ value is 7.9 nm and the $R_{2.0}$ value for the X-ray data is 4.4%. For the experimental neutron curve (○), the $R_{1.7}$ value is 8.1%.
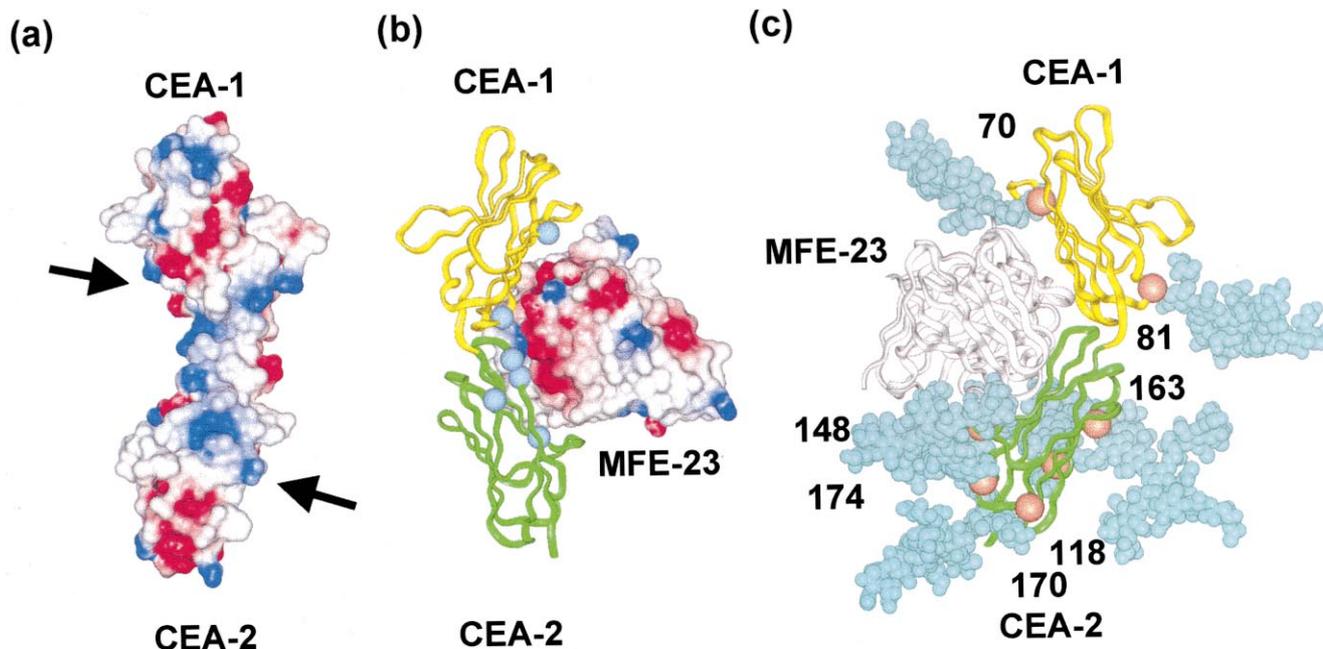
Fig. 3. The predicted model of the MFE23 complex with CEA-1 and CEA-2. a: An electrostatic view of the CEA-1 and CEA-2 domains is shown, where a stripe of basic residues (Lys-15, Arg-64, Lys-112, Lys-180, Arg-190 and Arg-191) at the interface between the CEA-1 and CEA-2 domains is arrowed. b: The α-carbons of these six basic residues are shown as blue spheres in the ribbon view of the CEA homology model (V-type, yellow; I-type, green) which is rotated by 180° about the vertical axis relative to the electrostatic view of the CEA-1 and CEA-2 domains. These basic residues are complementary to a stripe of acidic residues seen on the MFE23 electrostatic surface. These acidic residues are located on the *H1* and *H2* loops of MFE23 and at the N-terminus of the $V_L$ domain, and involve Asp-H31, Asp-H52, Glu-H53, Asp-H56, Glu-H58 and Glu-L1. c: Comparison of the oligosaccharide arrangement in CEA relative to the modelled MFE23 binding site on CEA. A ribbon diagram of MFE23 (white) is shown attached to the CEA-1 and CEA-2 domains (yellow and green respectively), in which the CEA-1 and CEA-2 domains are shown in the same orientation as the two domains in a. Seven oligosaccharide chains at Asn-70 and Asn-81 in CEA-1 and at Asn-118, Asn-148, Asn-163, Asn-170 and Asn-174 in CEA-2 are shown in blue solid representations. These do not prevent MFE23 binding at the interface between the CEA-1 and CEA-2 domains.

value of 8.0 nm. The modelled X-ray $R_{XS}$ value was 1.9 nm, in good agreement with the experimental X-ray $R_{XS}$ value of 2.1 nm. The small-angle X-ray and neutron *R*-factors were improved to 4.3 and 8.2% respectively in place of the previous values of 4.7 and 8.7%. The CD2-derived CEA model (Fig. 2) had modelled X-ray $R_G$ and $R_{XS}$ values of 7.9 and 2.0 nm respectively, and the small-angle X-ray and neutron *R*-factors were also improved at 4.4 and 8.1% respectively. These improvements justify the incorporation of the seven domain homology models with extended carbohydrate chains to represent the CEA structure in solution.

### 3.3. A model for the complex between MFE23 and CEA

The prediction of a structural model for the complex between MFE23 and CEA was based on the intermolecular contacts seen in the MFE23 crystal structure [8]. There, the antigen-binding loops *H1*, *H2* and *H3* in the variable heavy chain domain ($V_H$) formed extensive interactions with the side of the variable light chain domain ($V_L$) of a neighbouring MFE23 molecule. When this neighbouring $V_L$ MFE23 domain was superimposed with the N-terminal domain in the two-domain crystal structures of CD2, CD4, ICAM-2 and VCAM-1 [11,14,21,22] based on residues within the invariant β-strands B, C, E and F, the six antigen binding loops of MFE23 were seen in all four cases to be positioned within the angled surface formed between the N-terminal and C-terminal domains. All four superimpositions showed that the six loops were positioned between the lower part of the DEBA face of the N-terminal domain and the upper part of the GFC

face of the C-terminal domain. This is a different topology from that seen within the MFE23 crystal structure, where only the *H1*, *H2* and *H3* loops form good contacts. This difference is explained by noting that the two independent $V_L$ and $V_H$ domains in adjacent MFE23 molecules in the lattice are orientated antiparallel to each other, so only half the expected contacts would be expected to be formed, in contrast to the parallel arrangement of the Ig fold domains in CD2 that is enforced by their covalent connection and which permits a full set of contacts with MFE23 to be formed.

The MFE23–CEA complex was modelled by superimposing MFE23 onto the N-terminal domain of the CD2-derived CEA model. The CD2-derived structure was used in preference to CD4, ICAM-2 and VCAM-1 structures for three reasons: (i) the two CD2 domains formed the most favourable contacts with MFE23 in the superimpositions, followed by those of CD4; (ii) the length of the linker peptide between the two CD2 domains is similar to those in CEA, while that for CD4 is four residues shorter; (iii) the domain pair in ICAM-2 and VCAM-1 corresponded to an I-C2 arrangement instead of the V-C2 arrangement of CD2 and CD4. The model for the complex assumed that no conformational changes occur in the CEA interdomain angles or the MFE23 antigen-binding loops on complex formation, which is considered likely in view of the nanomolar affinity of the complex.

### 3.4. Assessment of the model for the complex between MFE23 and CEA

The formation of antibody–antigen complexes involves a

combination of electrostatic and hydrophobic interactions at the interface between the two proteins as well as good surface complementarity [15,32–35]. Accordingly the electrostatic interaction between MFE23 and CEA in the model was examined. In its antigen-binding loops, MFE23 presents negatively charged residues at Asp-H31 (*H1*), and at Asp-H52, Glu-H53, Asp-H56 and Glu-H58 (*H2*) (Kabat numbering is used throughout), four of which are highly exposed. Together with Glu-L1 in the MFE23 framework, these form a stripe of six acidic residues across the surface of MFE23 (Fig. 3b). The CEA homology model demonstrates a complementary stripe of six basic residues (arrowed in Fig. 3a) across the CEA-1 and CEA-2 domains. These include Lys-15 and Arg-64 on the DEBA face of CEA-1 and Lys-112, Lys-180, Arg-190 and Arg-191 on the GFC face of CEA-2, the positions of which were well-defined in the CEA homology modelling described above. The α-carbon distances between the two surfaces showed that eight pairs of MFE23 and CEA residues (Asp-H31/Lys-15, Asp-H31/Arg-64, Asp-H52/Lys-15, Glu-H53/Lys-15, Asp-H56/Lys-15, Glu-H58/Lys-112, Glu-H58/Arg-190 and Glu-H58/Arg-191) had separations between 1.0 and 1.4 nm. These are similar to the α-carbon separations of 1.1–1.4 nm (16 values) in Glu–Arg salt bridges between the $V_H$ residues Glu-H35 and Glu-H50 and the lysozyme residues Arg-45 and Arg-68 seen in crystal structures of Fab–lysozyme complexes (PDB codes: 3hfl, 1mlc and 2iff) [32,33]. Of particular interest is the proximity of Lys-15 and Arg-64 of CEA-1 to the *H2* loop of MFE23 in the model of the complex. These CEA residues are structurally homologous to Lys-L18 and Arg-L77 in the neighbouring molecule of MFE23 that form intermolecular contacts with the *H2* residues Asp-H52 and Glu-H53 of MFE23 [8]. This offers an explanation of how the MFE23 intermolecular contacts can be used to model the MFE23–CEA complex.

The shape complementarity of the MFE23 and CEA surfaces at their interface in the complex was examined using a set of 3Å-thick slices of the Connolly surfaces for MFE23 and CEA. The surfaces were satisfactorily accommodated at the antibody–antigen interface. No unusual contacts were found using the bump check facility in the WHATIF program available from the Internet at EMBL Heidelberg, when two atoms are said to 'bump' if they are closer than the sum of their van der Waals radii minus 0.4 Å. The total surface area lost in complex formation is 738 Å$^2$, which is within the range of the corresponding values of 632–916 Å$^2$ reported in six antibody–antigen crystal structures [32]. In the MFE23–CEA complex, the percentage surface areas lost for each of the six loops *H1*, *H2*, *H3*, *L1*, *L2* and *L3* were 13, 5, 42, 25, 4 and 12% in that order, and these values are as expected from known antibody–antigen crystal structures [34,35]. The CEA sidechains of Lys-15, Arg-64, Lys-180, Arg-190 and Arg-191 showed decreased accessible surfaces of 11–38 Å$^2$ in the complex, and likewise the MFE23 sidechains of Asp-H31, Asp-H52 and Glu-H53 showed decreased accessible surfaces, as expected from their α-carbon separations.

The 28 carbohydrate chains of CEA may obstruct the MFE23 binding site. Scattering indicates that the carbohydrate is on average extended freely into solution [4]. Fig. 3c showed that the glycosylation site at Asn-70 was located at the top of the CEA-1 domain, while Asn-81 was located at the GFC face on the opposite side of the CEA-1 domain to where MFE23 is located. The five glycosylation sites on the CEA-2
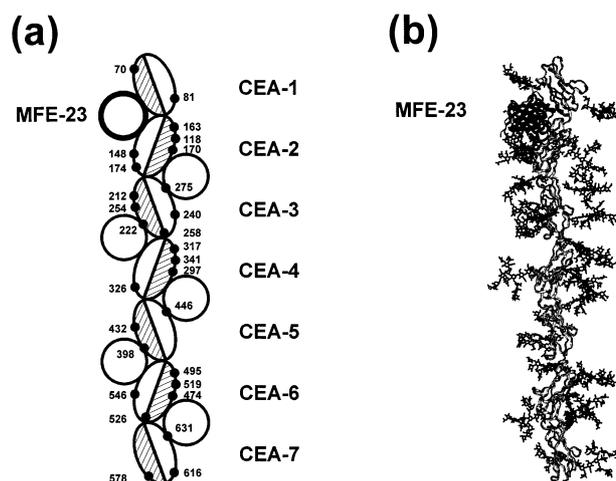


Fig. 4. Predicted association of MFE23 with the domains of CEA. a: The seven CEA domains are shown as ellipsoids with MFE (circles) bound at all six possible interfaces between adjacent domain pairs. The MFE23 binding site between the CEA-1 and CEA-2 domains is boldface. The DEBA face of each CEA domain is shown hatched, and the GFC face is shown open. The 28 carbohydrate sites are represented by ●, where each one is identified by its Asn residue number. The number is positioned outside the MFE23 circle if the Asn residue offers no steric hindrance to MFE23 binding, and inside if there is steric overlap. b: MFE23 is shown as a black ribbon at its predicted binding site in a ribbon view of the full CD2-derived model of CEA. The 28 oligosaccharide chains are shown as black outlines.

domain at Asn-118, Asn-148, Asn-163, Asn-170 and Asn-174 were located around the base of the CEA-2 domain. All seven Asn residues had unchanged solvent accessibilities in the model of the complex. Visual inspection with stereo glasses showed no steric overlap of the seven carbohydrate chains with MFE23 (Fig. 3c). This was confirmed using the bump check facility in WHATIF. The inspection of the carbohydrate sites at the five remaining interdomain junctions in CEA did not favour a MFE23 binding site at any of these (Fig. 4a). Sixteen sites were located to the DEBA faces, and 10 to the GFC faces. MFE23 was successively superimposed onto the N-terminal domain at each junction in the CD2-derived full CEA model. At each position, the relationship of MFE23 to nearby carbohydrate chains was checked using stereo glasses. This showed that each of these five potential sites would be blocked by each of Asn-275, Asn-222, Asn-446, Asn-398, or Asn-631 in turn (Fig. 4a). It is possible that the role of carbohydrate at these sites may maintain CEA as an extended zig-zag structure that protrudes perpendicular from the cell surface, and that Asn-578 and Asn-616 may perform a similar function at the cell surface.

The interaction between MFE23 and CEA has been examined using four $V_H$ mutated forms of MFE23 [36]. A Tyr-H100b to Pro mutation abolished MFE23 binding to CEA; the sidechain surface accessibility of Tyr-H100b is 62% in MFE23 and 0% in the model of the MFE23–CEA complex. A Glu-H53 to Lys mutation decreased MFE23 binding to CEA; the corresponding accessibility of Glu-H53 is 60% in MFE23 and 31% in the MFE23–CEA complex. A Thr-H98 to Ala mutation led to a minor improvement in MFE23 binding to CEA; the accessibility of Thr-H98 was 99% in MFE23 and 3% in the MFE23–CEA complex. A Tyr-H100a to Ala mutation had no effect on MFE23 binding to CEA; the accessi-

bility of Tyr-H100a is unchanged at 6% in both MFE23 and the MFE23–CEA complex. The outcome of these experimental mutagenesis studies are fully consistent with the model of the MFE23–CEA complex.

## 4. Conclusions

In this study, recent advances in structural analyses of the Ig superfamily has resulted in an improved full homology model for CEA that was validated by means of solution scattering curve fits (Fig. 2). This model has been deposited in the PDB (code 1e07). MFE23 was developed by phage technology as an antibody molecule with a high nanomolar affinity for its target molecule CEA [6]. By combining the CEA model with the observed intermolecular contacts in our crystal structure of MFE23 [8], a plausible molecular model for the complex between MFE23 and CEA could be constructed (Fig. 4b). The model for the complex between MFE23 and CEA is supported by mutagenesis using MFE23 [36] and by experimental evidence that showed that MFE23 binds to a double-domain CEA-1 and CEA-2 fragment expressed in *E. coli* (Thornton, J.D., Keep, P.A., Chester, K.A. and Begent, R.H.J., unpublished results). In particular, the model showed that all six antigen-binding loops interacted with CEA, thus offering an explanation of the high affinity of MFE23 for CEA. It also indicated six potential electrostatic ion-pairs that may stabilise the affinity of MFE23 for CEA, two of which were homologous to those seen in the intermolecular contacts in the MFE23 crystal structure. Furthermore, the extensive buried surface area of this model is consistent with what is seen in antigen–antibody crystal structures; and no steric conflict was observed with any of the 28 carbohydrate chains in CEA. As far as is known, a docking procedure for the modelling of an antibody–antigen complex based on antibody crystallographic intermolecular contacts has not been previously described [15]. A related approach was largely successful in predicting the structure of the cell-surface complex formed between CD2 and CD58, which is stabilised by five salt bridges [37,38]. In an independent approach using mutagenesis, electrostatic forces have been shown to be important for stabilising the complex between the monoclonal antibody 2E8 and apolipoprotein E [39]. It is not possible to verify the model for the MFE23–CEA complex by scattering, as MFE23 is too small compared to CEA to permit its localisation relative to CEA (Fig. 4). We are presently attempting crystal structure strategies for this complex, which if successful will make possible the critical evaluation of the success of this modelling procedure in prediction work. At the present time, the model makes possible the rational interpretation of current biochemical and molecular biology experiments designed to improve the understanding of the binding affinity of MFE23 to CEA and the tumour targeting properties of MFE23.

## References

[1] Williams, A.F. and Barclay, A.N. (1988) Annu. Rev. Immunol. 6, 381–405.
[2] Chothia, C. and Jones, E.Y. (1997) Annu. Rev. Biochem. 66, 823–862.
[3] Harris, L.J., Larson, S.B., Skaletsky, E. and McPherson, A. (1998) Immunol. Rev. 163, 35–43.
[4] Boehm, M.K., Mayans, M.O., Thornton, J.D., Begent, R.H.J., Keep, P.A. and Perkins, S.J. (1996) J. Mol. Biol. 259, 718–736.
[5] Boehm, M.K., Woof, J.M., Kerr, M.A. and Perkins, S.J. (1999) J. Mol. Biol. 286, 1421–1447.
[6] Chester, K.A., Begent, R.H.J., Robson, L., Keep, P.A., Pedley, R.B., Boden, J.A., Boxer, G., Green, A., Winter, G., Cochet, O. and Hawkins, R.E. (1994) Lancet 343, 455–456.
[7] Begent, R.H.J., Verhaar, M.J., Chester, K.A., Casey, J.L., Green, A.J., Napier, M.P., Hope-Stone, L.D., Cushen, N., Keep, P.A., Johnson, C.J., Hawkins, R.E., Hilson, A.J.W. and Robson, L. (1996) Nat. Med. 2, 979–984.
[8] Boehm, M.K., Corper, A.L., Wan, T., Sohi, M., Sutton, B.J., Thornton, J.D., Keep, P.A., Chester, K.A., Begent, R.H.J. and Perkins, S.J. (2000) Biochem. J. 346, 519–528.
[9] Perkins, S.J., Ashton, A.W., Boehm, M.K. and Chamberlain, D. (1998) Int. J. Biol. Macromol. 22, 1–16.
[10] Harpaz, Y. and Chothia, C. (1994) J. Mol. Biol. 238, 528–539.
[11] Jones, E.Y., Harlos, K., Bottomly, M.J., Robinson, R.C., Driscoll, P.C., Edwards, R.M., Clements, J.M., Dudgeon, T.J. and Stuart, D.I. (1995) Nature (Lond.) 373, 539–544.
[12] Wang, J.H., Pepinsky, R.B., Stehle, T., Liu, J.H., Karpusas, M., Browning, B. and Osborn, L. (1995) Proc. Natl. Acad. Sci. USA 92, 5714–5718.
[13] Wang, J.H., Stehle, T., Pepinsky, B., Liu, J.H., Karpusas, M. and Osborn, L. (1996) Acta Cryst. D52, 369–379.
[14] Casasnovas, J.M., Springer, T.A., Liu, J.H., Harrison, S.C. and Wang, J.H. (1997) Nature (Lond.) 387, 312–315.
[15] Sternberg, M.J.E. (Ed.) (1996) Protein structure Prediction, IRL, Oxford.
[16] Hobohm, U. and Sander, C. (1994) Protein Sci. 3, 522–524.
[17] Laskowski, R.A., MacArthur, M.W., Moss, D.S. and Thornton, J.M. (1993) J. Appl. Crystallogr. 26, 283–291.
[18] Lee, B. and Richards, F.M. (1971) J. Mol. Biol. 55, 379–400.
[19] Šali, A. and Blundell, T.L. (1990) J. Mol. Biol. 212, 403–428.
[20] Kabsch, W. and Sander, C. (1983) Biopolymers 22, 2577–2637.
[21] Bodian, D.L., Jones, E.Y., Harlos, K., Stuart, D.I. and Davis, S.J. (1994) Structure 2, 755–766.
[22] Wang, J., Yan, Y., Garrett, T.P.J., Liu, J., Rodgers, D.W., Garlick, R.L., Tarr, G.E., Husain, Y., Reinherz, E.L. and Harrison, S.C. (1990) Nature (Lond.) 348, 411–418.
[23] Leahy, D.J., Axel, R. and Hendrickson, W.A. (1992) Cell 68, 1145–1162.
[24] . Perkins, S.J. (1986) Eur. J. Biochem. 157, 169–180.
[25] Perkins, S.J. and Weiss, H. (1983) J. Mol. Biol. 168, 847–866.
[26] Ashton, A.W., Boehm, M.K., Gallimore, J.R., Pepys, M.B. and Perkins, S.J. (1997) J. Mol. Biol. 272, 408–422.
[27] Smith, K.F., Harrison, R.A. and Perkins, S.J. (1990) Biochem. J. 267, 203–212.
[28] Perkins, S.J., Smith, K.F., Kilpatrick, J.M., Volanakis, J.E. and Sim, R.B. (1993) Biochem. J. 295, 87–99.
[29] Bates, P.A., Luo, J. and Sternberg, M.J.E. (1992) FEBS Lett. 301, 207–214.
[30] Bork, P., Holm, L. and Sander, C. (1994) J. Mol. Biol. 242, 309–320.
[31] Jones, E.Y., Davis, S.J., Williams, A.F., Harlos, K. and Stuart, D.I. (1992) Nature (Lond.) 360, 232–239.
[32] Braden, B.C. and Poljak, R.J. (1995) FASEB J. 9, 9–16.
[33] Davies, D.R. and Cohen, G.H. (1996) Proc. Natl. Acad. Sci. USA 93, 7–12.
[34] Wilson, I.A. and Stanfield, R.L. (1994) Curr. Opin. Struct. Biol. 4, 857–867.
[35] Stanfield, R.L., Fieser, T.M., Lerner, R.A. and Wilson, I.A. (1990) Science 248, 712–719.
[36] Read, D.A., Chester, K.A., Keep, P.A., Begent, R.H.J., Pedersen, J.T. and Rees, A.R. (1995) Br. J. Cancer 71 (Suppl. XIV: abstr. P132), 57–57.
[37] Ikemizu, S., Sparks, L.M., van der Merwe, P.A., Harlos, K., Stuart, D.I., Jones, E.Y. and Davis, S.J. (1999) Proc. Natl. Acad. Sci. USA 96, 4289–4294.
[38] Wang, J.H., Smolyar, A., Tan, K., Liu, J.H., Kim, M., Sun, Z.Y.J., Wagner, G. and Reinherz, E.L. (1999) Cell 97, 791–803.
[39] Raffai, R., Weisgraber, K.H., MacKenzie, R., Rupp, B., Rassart, E., Hirama, T., Innerarity, T.L. and Milne, R. (2000) J. Biol. Chem. 275, 7109–7116.