Volume 301, number 2, 207–214 FEBS 10937 April 1992

A predicted three-dimensional structure for the carcinoembryonic antigen (CEA)

Paul A. Bates, Jingchu Luo* and Michael J.E. Sternberg

Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, 44 Lincoln's Inn Fields, PO Box 123, London CW2A 3PX, UK

Received 7 February 1992; revised version received 3 March 1992

A three-dimensional model for the carcinoembryonic antigen (CEA) has been constructed by knowledge-based computer modelling. Each of the seven extracellular domains of CEA are expected to have immunoglobulin folds. The N-terminal domain of CEA was modelled using the first domain of the recently solved NMR structure of rat CD2, as well as the first domain of the X-ray crystal structure of human CD4 and an immunoglobulin variable domain REI as templates. The remaining domains were modelled from the first and second domains of CD4 and REI. Link conformations between the domains were taken from the elbow region of antibodies. A possible packing model between each of the seven domains is proposed. Each residue of the model is labelled as to its suitability for site-directed mutagenesis.

Cell surface antigen; Tumour marker; Computer model; Site-directed mutagenesis

1. INTRODUCTION

Carcinoembryonic antigen (CEA), a highly glycosylated cell surface protein of molecular mass about 180,000 [1], is a widely distributed tumour marker (for recent reviews see [2,3]). Although it is expressed in high proportions on a variety of human tumours, notably of the colon but also of the breast and lung [4], studies also reveal the presence of CEA in normal tissues [5]. Thus the distribution and possibly the functional role of CEA is complex. Nevertheless monoclonal antibodies (Mab's) against CEA have been used successfully for radioimmuno-localization of tumors [6–8]. There is now a substantial panel of Mab's to CEA and an attempt has been made to place them into non-interacting groups [9]. However, little progress has been made towards mapping the exact locations of these Mab's to the surface of CEA. A carefully constructed epitope map should help to probe the function(s) and help in the search for Mab's with greater binding affinity to this important tumour marker.

The primary structure of CEA has been obtained from the gene sequence [10]. The mature CEA (starting at Lys) consists of 668 amino acids, with the first 643 being extracellular. The extracellular region consists of seven domains, each having sequence similarity with immunoglobulin (Ig) domains, thus CEA is a member of the Ig superfamily [11].

At present, there is no crystal structure for CEA that could guide structure/function studies, such as site-directed mutagenesis, however, the sequence relationship with the immunoglobulins suggests that the complete extracellular region can be modelled with some confidence from Ig folds whose structures are known from X-ray and NMR analysis. There are now a number of fully determined Ig folds from the Fab fragments of antibodies, and recently for two cell surface antigens, the first two domains of human CD4 (X-ray) [13], and the first domain of rat CD2 (NMR) [14]. In this paper we report the prediction for each of the seven extracellular domains based on multiple sequence alignments [15], and a knowledge-based approach for fragment fitting [16–19]. As the links between each of the seven domains show some similarity to the links between the variable and constant domains of Ig folds from Fab fragments we also suggest how these seven domains are packed together.

2. EXPERIMENTAL

2.1. *Alignment of CEA with immunoglobulins*

The basic structure of an Ig domain is a tertiary fold consisting of a stacked pair of β-sheets. In the known structure of the Ig constant domains (C), there are seven component β-strands (lettered A–G), with strands A, B, E and D lying antiparallel in one sheet and strands G, F and C being antiparallel in the other. The variable domains (V) have an analogous fold but between strands C and D there are extra strands, C' and C''. The C'' strand is very short. In all antibody domains of these types there is a conserved disulphide bridge between strands B and F, however, for some members of the Ig superfamily, e.g. CD2, this is not always the case [11]. A third type of Ig domain,

*On leave from: Department of Biology, Peking University, Beijing 100871, People's Republic of China.

Correspondence address: P.A. Bates, Biomolecular Modelling Laboratory, Imperial Cancer Research Fund, 44 Lincoln's Inn Fields, PO Box 123, London WC2A 3PX, UK. Fax: (44) (71) 430 2666.

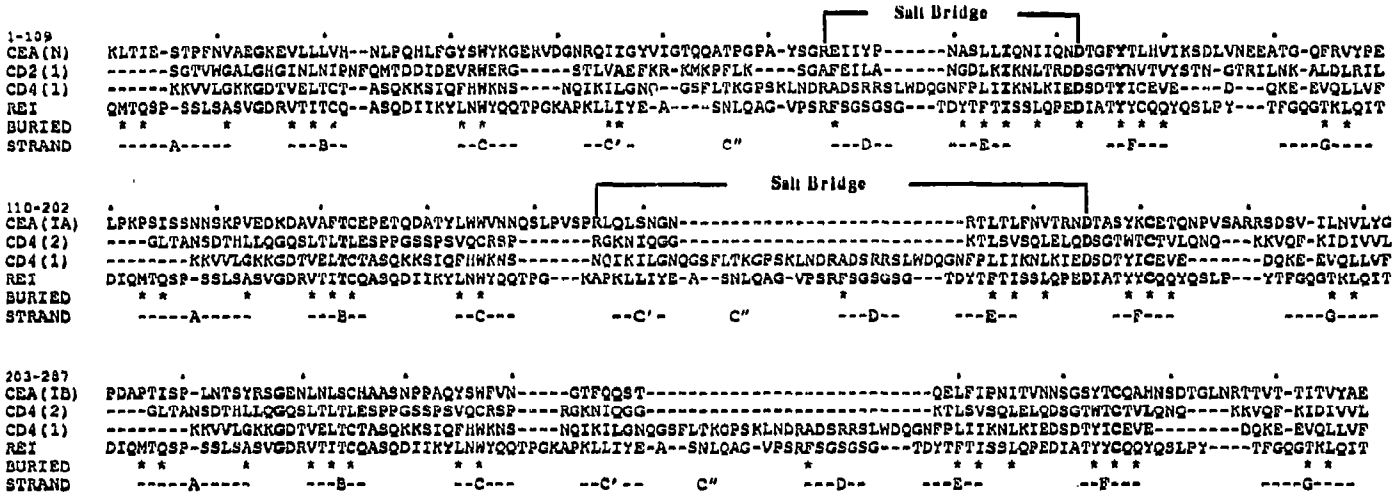

Fig. 1. Sequence alignment of the first three domains of CEA with the templates used for modelling. Every tenth residue of CEA is marked by a dot. Identical residues are highlighted, gaps are denoted by dashes. In the row denoted BURIED asterisks indicate those residues that are greater than or equal to 90% buried in all templates. These percentages were calculated by a local program, TEMPLATE, which measures the solvent accessibility of each residue relative to an extended Gly-X-Gly chain. STRAND denotes the averaged secondary structure elements for the templates; strands are labelled A to G for each fold. Conserved hydrophobic residues and key residues, such as the conserved salt bridge in the N-terminal domains (not CD2), help to anchor the alignment in place. For the C2 domains, A and B repeats, the C″ and D strands are predicted to be missing relative to a full V domain.

termed C2, is not found in antibodies but occurs in other members of the Ig superfamily [11]. These domains have a greater sequence similarity and structural topology to V domains but are of similar length, and even shorter, in sequence to that of C domains. The recent X-ray crystal structure of a C2 domain, the second domain of human CD4 [13], shows that in this C2 domain the C″ and D strands are missing compared to a V domain.

The first step in prediction by homology is to establish the most suitable parent proteins of known structure on which to base the modelling [16]. The N-terminal Ig domain has been predicted to be V type [11]. This is followed by three gene duplication repeats, each of which consists of two C2-like domains, labelled 'type A' and '-B' [12]. Table I gives the results of sequence alignments of each of these domains with known (X-ray and NMR) Ig folds using the ALIGN program [20]. The analysis agrees with the predictions made above and also identifies the best templates from which to build each domain. The templates chosen to build the N-domain were from the crystal coordinates of REI (Bence-Jones dimer VL kappa) [21], CD4(domain 1) and the NMR coordinates from CD2(domain 1). The only example of a fully characterized C2 domain is from the crystal coordinates of the second domain of CD4, thus this was the main template used to model the C2 repeats. However, the C2 domain from CD4 does not have the conserved disulphide link so the variable domains of CD4(1) and REI were also included as templates.

Automatic sequence alignments such as those reported in Table I do help to establish the particular class to which each domain belongs. However, refined alignments still require manual intervention. The best possible alignment for the N-terminal domain of CEA against each of the V templates and for the first repeat against its templates is shown in Fig. 1. Points considered whilst aligning the sequences were: (i) keep the key markers of the fold aligned, such as the conserved Trp in the C strand, conserved salt bridges and disulphide links; (ii) keep the conserved hydrophobic core of each fold; (iii) maintain the β-strand structure, i.e. all insertions and deletions should be in the loop regions; and (iv) where possible the above points should be consistent with an alignment between the sequence to be modelled and other close members of that family (see Fig. 2).

2.2. *Modelling the main chain*

The alignment of Fig. 1 defined the conserved β-strand framework for the templates used. To obtain a structural equivalencing for the residues of the framework the relevant templates were first superimposed [17]. The average strand lengths and their positions were then assigned by visual inspection. Main-chain atoms for different strands from different templates, depending on the sequence correspondence with CEA, were used as indicated in Table II.

The next step is to obtain a model for the main chain conformation

Table I

Sequence comparisons of CEA domains N, IA and IB with other Ig-folds

	CD2 (R1)	CD4 (H1)	CD4 (H2)	1REI (V)	3FAB (VH)	3FAB (VL)
CEA(N)	**3.11**	**2.98**	0.41	**4.09**	3.32	0.77
CEA(IA)	2.45	**2.24**	**2.28**	**4.53**	2.81	2.97
CEA(IB)	3.85	**2.40**	**4.19**	**2.49**	2.74	3.75

The table gives the number of standard deviations between the alignment score and the mean for 500 randomized sequences. The program ALIGN [20] was used with a gap penalty of 6, a weight of 6 and the MD250 matrix. Highlighted scores indicate the templates used in the modelling process. CD2(R1), rat CD2 first domain, V; CD4(H1), human CD4 first domain, V; CD4(H2), human CD4 second domain, C2; 1REI(V), Bence-Jones immunoglobulin, V; 3FAB(VH), human immunoglobulin (New) heavy chain, V; and 3FAB(LH), human immunoglobulin (New) light chain, V [37]. The table helps to identify domains belonging to a particular class, for example CEA(N) scores badly against CD4(H2) (a C2 domain) but well against most V domains. This table also helped to select the templates from which to model each domain. The highest scores were not necessarily taken to represent the best templates, for example CEA(IB) scores reasonably well with both CD2(R1) and CD4(H1) with CD2(R1) being slightly higher. However, CD4(H1) was used as a template for CEA(IB) because it has a conserved disulphide link between the B and F strands of the Ig fold and CD2(R1) does not. As CEA(IB) also has the conserved disulphide it makes the modelling process more accurate to choose the CD4(H1) template.

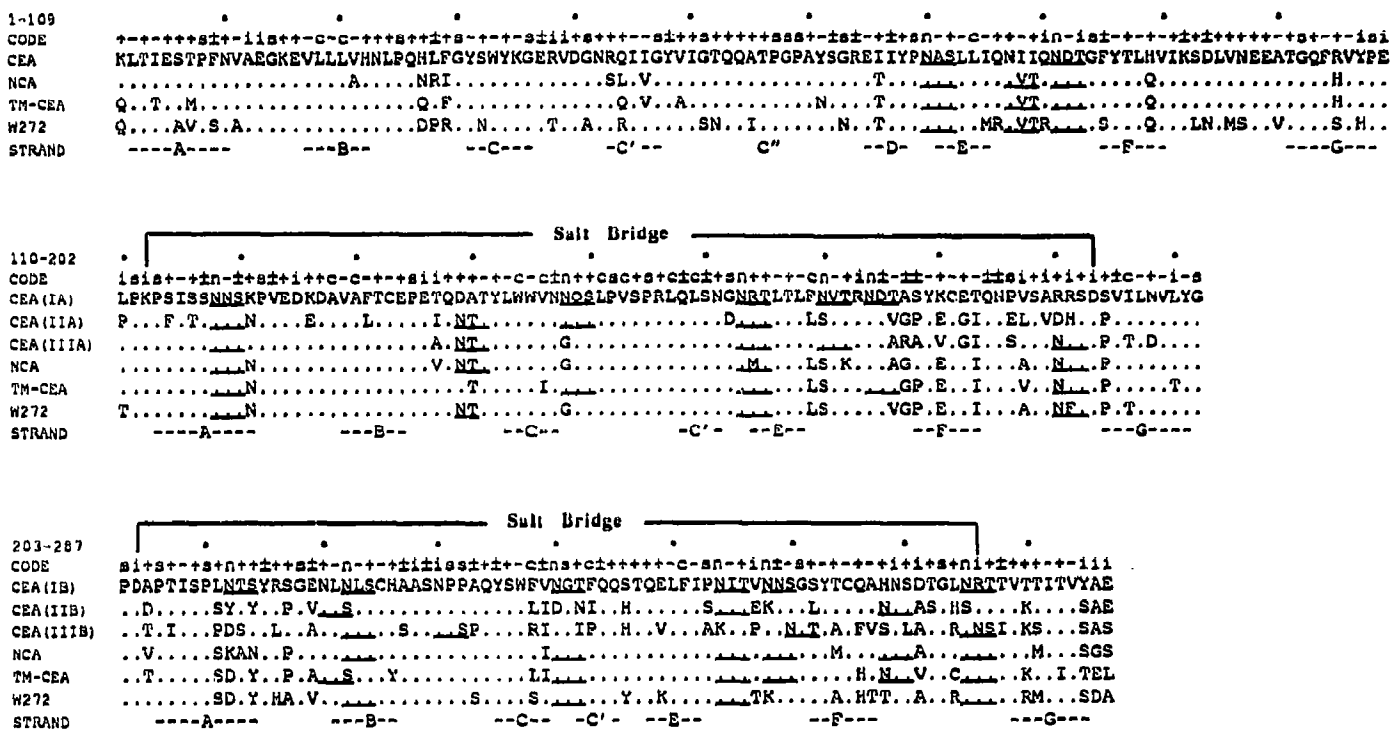

Fig. 2. Sequence alignments for the N-terminal domain of CEA, the repeating A, B units of CEA and other members of the CEA family. NCA, human non-specific crossreacting antigen [43]; TM-CEA, human transmembrane CEA [44]; W272, a new species of human NCA [45]. In general there are more sequence differences in the loop regions (between strands). There are, however, no insertions or deletions even between various family members suggesting strong evolutionary pressure to maintain, certainly a similar molecular topology for each molecule, but also similarity in protein function. Many of the potential *N*-glycosylation sites, underlined, are also conserved. CODE refers to the individual residues of CEA as to their suitability for site-directed mutagenesis experiments (see text): +, exposed; −, 90–100% buried; ±, semi-buried 60–90%; s, a proline or glycine residue, residues essential for the folds stability; c, predicted to be covered by carbohydrate; i, interface residue between domains; n, Asn-linked glycosylation site. Most carbohydrate-covered residues are hydrophobic and these are highly conserved both within repeats and within family members. Predicted interface residues are also well conserved, for example the salt bridges residues K(112) to D(193) and D(204) to R(276) (although not last B repeat). The first salt bridge is also conserved at the interface of VL-to-CL domains in 3FAB (immunoglobulin (New)) [42].

of the regions, mainly loops between β-strands, that connect the conserved sections of the multiple template. One standard approach [18,19] to model a connection of n residues that spans fixed regions, in this case usually β-strands, is to search through a database of the known proteins for a fragment of the appropriate number of residues that roughly span the connection. A modified version of this approach was used and candidate fragments were obtained from a data base consisting of the atomic coordinates for all the Ig folds characterised by X-ray crystallography. The database was compiled from twelve Fab fragments, Brookhaven codes: 3FAB, 4FAB, 2MCP, 2FB4, 3HFM, 3MCG, 1MCW, 2HFL, 1REI, 2FBJ, 2RHE, and 2IG2, and two cell surface molecules, 2HLA (the Ig folds of class II HLA) and CD4, plus the NMR coordinates of CD2. Thus by using different β-strands from the templates and different fragments from an Ig database (see Table II) the best framework for each fold was built.

2.3. *Completing each domain*

The CEA side chains were placed on the model and their orientations selected by the following criteria:

(i) the conformation of the equivalent residue in the framework was used if there was an identical residue in CEA;

(ii) for conserved substitutions the CEA side chain was positioned in an analogous location to the side chain of the framework [23];

(iii) for non-conserved substitutions, the CEA side chain was set to the most commonly observed conformation [24];

(iv) finally, visual inspection removed any serious steric clashes by selecting an alternate, allowed side chain conformation.

Each domain was energy-minimized using the CHARMM program [25] with default values as incorporated into QUANTA in the POLYGEN software. Initially 25 steps of steepest descent were performed to remove the major steric clashes. This was followed by 500 steps of adopted basis Newton–Raphson minimization. The minimization serves to obtain good peptide geometry and to remove all steric clashes. A stable negative energy was obtained for each domain. However, a low energy is not necessarily an indication that a model is correct [26]. Each model was therefore tested by a program called POL_DIAGNOSTICS_88 [27] that compares key characteristic quantities of a model with those observed for a wide range of proteins in the Brookhaven database, such as the ratio of hydrophobic and hydrophilic residues exposed on the surface of the protein compared with those of the core. All domains passed each test set by the program.

2.4. *Domain packing*

Comparison of the sequences between domains show two types of similar linkers: type I between the N-terminal V domain and the first IA repeat and between B-to-A repeats; type II between all A-to-B repeats. Fig. 3 shows an alignment of all these CEA linker regions with the elbow regions between V and C domains of an Fab fragment. There are some sequence similarities such as conserved prolines and hydrophilic residues. As a tentative suggestion we modelled the linker conformations based on the elbow joint conformation from this antibody. The resulting packing arrangement between each domain is supported by the fact that many of the predicted interface residues are conserved within repeats including a conserved salt bridge at all but

```
CEA(V-IA)          FRVYP-ELPK----PSI
CEA(IB-IIA)        ITVYA-EPPK----PFI      Type I linker
CEA(IIB-IIIA)      ITVSA-EPPK----PSI
CEA(IA-IB)         LNVLY-GPDA----PTI
CEA(IIA-IIB)       LNVLY-GPDD----PTI      Type II linker
CEA(IIIA-IIIB)     LDVLY-GPDT----PII
3FAB(VL-CL)        TKLTVLRQPKAAPSVTL
BURIED              * *        * *
STRAND            --G->      <-A--
```

Fig. 3. Sequence alignment of all the linker regions of CEA with that of an antibody elbow region from 3FAB (immunoglobulin (New)) [42]. Buried residues in the β-strands across the interface of the VL and CL domains of 3FAB are indicated by asterisks.

one interface. These salt bridges lie between the linking region and the F-G loop of the following domain (see Fig. 2). A salt bridge is also conserved in the elbow region of many antibodies. However it should be noted that other packing arrangements are possible and that confirmation can only come from an X-ray crystal structure of two or more domains of CEA or possibly by extensive Mab epitope mapping as recently described for a related Ig superfamily protein ICAM-1 (intercellular adhesion molecule-1) [28].

The complete model for the extracellular portion of CEA was then subjected to the same energy minimization and testing protocol as described above for the individual domains. The model passed all empirical tests set by program POL_DIAGNOSTICS_88.

2.5. *Placement of N-linked oligosaccharides*

The high molecular weight of CEA, compared with that for the molecular weight calculated from just the un-glycosylated protein, indicates that most if not all potential N-linked glycosylation sites are occupied [2,3].

There are now a few X-ray crystal structures for which coordinates have been deposited in the Brookhaven database and that have N-linked carbohydrates, for example hemagglutinin [29], and an Fc antibody fragment [30]. Both the above examples show that the common core carbohydrate covers hydrophobic residues on the surface of the protein. The single carbohydrate site on the Fc fragment is more relevant to the model for CEA as it is located on the D-E loop of an Ig (C) domain and covers hydrophobic residues on the B and E strands. Clipping this sugar out from the Fc fragment and placing it onto the CEA model at the first potential N-linked glycosylation site, also on a D-E loop on the N-terminal V domain, without adjusting any of the torsional parameters of the system shows that this sugar can also cover nearby hydrophobic residues: in fact, the equivalent hydrophobic residues as found in Fc, those on the B and E strands. These hydrophobic residues would normally point out into solvent. Fig. 2 shows that these hydrophobic residues are conserved for other members of the CEA family. The remaining potential N-linked glycosylation sites were linked to this same sugar fragment and the free rotating bonds of the Asn and Asn–sugar complex were adjusted to cover nearby, exposed, hydrophobic residues. Because of the large number of potential glycosylation sites, the ability to judge which hydrophobic residues are covered by which sugars soon becomes difficult. However, Fig. 2 indicates which of the normally exposed residues, mostly hydrophobic, are predicted to be covered by sugars. The prediction does assume that most carbohydrates cover the surface of the protein like a net as found for the Fc fragment, however, carbohydrates can also point directly away from the surface of the protein as found in the X-ray crystal structure of bovine pancreatic DNAse I [31]. Nevertheless hydrophobics close to carbohydrate sites should receive some cover. This exercise, although not particularly accurate, does indicate that there is only one region that carbohydrates do not cover completely and that is the CFG face (β-strands C', C", C, F, and G), see Fig. 5, of the N-terminal V domain.

2.6. *The transmembrane region*

There has been some speculation as to whether the string of hydrophobic residues at the C-terminus of CEA is a transmembrane region [2,3]. There is a pattern of Gly in this region, every fifth residue, that is similar to that for the transmembrane region of the alpha and beta chains of HLA class II proteins. Fig. 4 shows a helical wheel for the transmembrane regions of the alpha [32] and beta [33] chains of a class II molecule, DR1, and the putative transmembrane region of CEA. All Gly residues are located on the same side of the helix. Class II is strongly predicted to be a heterodimer on the cell surface [34] and the pattern of Gly residues is expected to aid a close association of the transmembrane regions [35]. This implies that CEA could also be a dimer on the cell surface. There is also experimental evidence for CEA existing as a homodimer from SDS-polyacrylamide gel electrophoresis and cross-linking experiments [36].

Table II

β-strands and loop fragments used in CEA model building

Domain	Res. range	Template	Fragment
CEA(N)	1–4	REI	
	5–8		2HLA(A)
	9–20	CD2(1)	
	21–23		3MCG(1)
	24–35	CD2(1)	
	36–42		4FAB(H)
	43–48	CD4(1)	
	49–52		2MCP(L)
	53–58	CD4(1)	
	59–63	REI	
	64–68	CD4(1)	
	69–70	CD2(1)	
	71–75	CD4(1)	
	76–92	CD2(1)	
	93–98		2HFL(H)
	99–106	CD2(1)	
LINK 1	107–112		3FAB(L)
CEA(IA)	113–132	CD4(2)	
	133–144	CD4(1)	
	145–148	REI	
	149–155		2HFL(L)
	156–166	CD4(2)	
	167–181	CD4(1)	
	182–185	CD4(2)	
	186–190		2HFL(L)
	191–199	CD4(2)	
LINK 2	200–205		3FAB(L)
CEA(IB)	206–209	CD4(2)	
	210–211		3HFM(H)
	212–224	CD4(2)	
	225–237	CD4(1)	
	238–243		1FBJ(L)
	244–245	REI	
	246–251	CD4(2)	
	252–265	CD4(1)	
	266–270	CD4(2)	
	271–275		1MCW(M)
	276–284	CD4(2)	
LINK 3	285–290		3FAB(L)

The table shows the templates and fragments used for the first three domains only. The column headed Template indicates from which template protein the β-strands were taken and the column headed Fragment the proteins from which loops that join the β-strands were taken. Selection criteria depended not just on sequence compatibility with CEA but also on good peptide geometry between joining regions and the minimization of steric clashes between atoms. A local program was used to select the correct sections of Templates and Fragments and join them together [22]. The remaining A and B repeats plus linkers were generated by simple rotation and translation of these building blocks to produce a rod-shaped molecule.

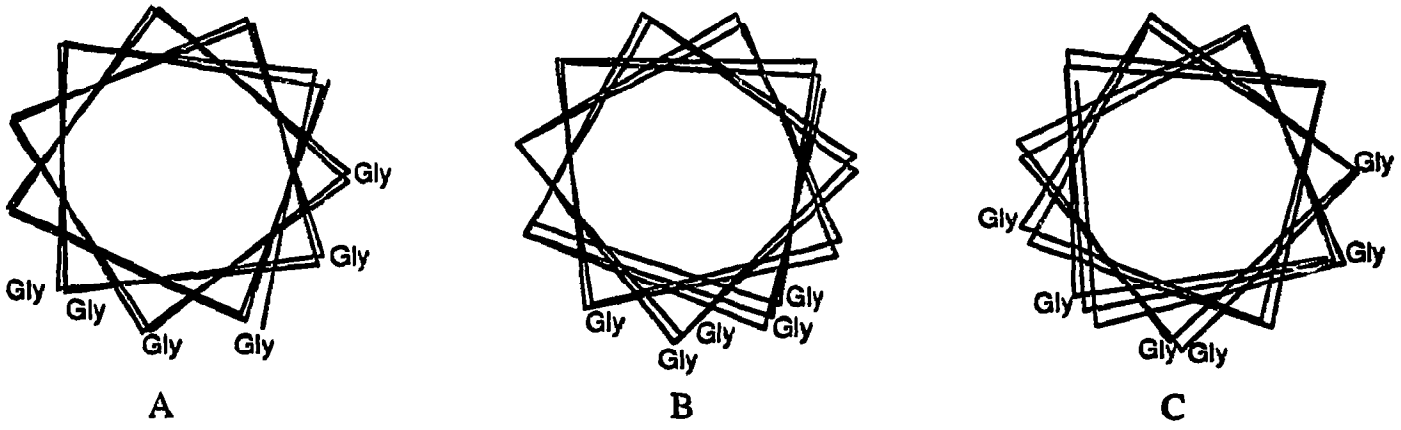

Fig. 4. Helical wheel diagrams for the transmembrane regions of (A) CEA, (B) HLA class II, DR1 α and (C) HLA class II, DR1 β. All Gly residues are on one side of the helix, suggesting close packing between α and β chains of class II, a known heterodimer. The implication is that CEA may have the potential to form homodimers or interact with a second molecule on the same cell.

3. DISCUSSION AND CONCLUSION

Although the Ig domain-like structure of CEA has been established by others [11,12], an accurate alignment of CEA to structural templates and the subsequent construction of a full three-dimensional model (see Fig. 5) has hitherto not been attempted. This approach has several advantages over simpler alignments. Residues along the sequence can be predicted with more confidence as to whether they are exposed, part of the conserved core of an Ig fold, or buried at the interface between domains. This information is of importance in guiding the construction of site-directed mutants to CEA in order to map Mab epitopes and ultimately the function(s) of CEA.

Fig. 2 shows the relative degree of exposure for each of the residues. Several classes are used:

(i) residues that are 90–100% buried forming the conserved cores of the folds or residues buried at the predicted domain-domain interfaces. Such residues would not be changed in probing structure/function activities for CEA as they are essential for the correct folding and packing of the domains;

(ii) residues that are partially buried, 60–90%. Such residues may be changed but care should be taken to not change too many at once or to make non-conservative changes, e.g. the conserved salt bridge that is partially buried in the N-terminal domain;

(iii) residues that are covered by an adjacent sugar site. These residues are exposed on the un-glycosylated protein and tend to be hydrophobic. Mutations to these residues may indirectly alter functional activity by altering sugar conformation;

(iv) residues that are essentially exposed on the surface and can be changed quite freely.

Site-directed mutational studies have not been carried out on CEA. However, one mutation study has been carried out for a closely related cell surface molecule NCA [37]. This molecule, although only consisting of one internal repeat, is closely homologous to CEA. The alignment of the first three domains of CEA and NCA are shown in Fig. 2. The mutants for NCA lie in a region corresponding to the exposed E–F loop at the base of the N-terminal domain in which a number of sequence substitutions are found relative to CEA.

A second region that shows differences in sequence between CEA and other members of its family, as well as between repeats, is in the F–G loop of each fold. These loops are also exposed but on the opposite side of the molecule to the E–F loops. Therefore this suggests that monoclonal antibodies raised to the F–G loops should fall into different groups from the antibodies raised to the E–F loops.

Due to the large number of potential glycosylation sites, the amount of exposed protein is expected to be relatively small. If indeed CEA is expressed as a homodimer on the cell surface (section 2.6) or associates with a second cell surface molecule on the same cell, as for example a collagen receptor [38], then even less protein would be exposed. This indicates that many of the monoclonals raised against CEA probably interact, at least in part, with the sugars. Other than the loops described above few other bare protein sites can be identified from the model as being sites for antibody attachment. The most obvious choice would be from the very top of the first domain, in the loops equivalent to the CDR loops of an antibody, and the CFG face of the N-terminal domain.

Some of the sugars could mediate protein–protein interactions. For example the first highly conserved, putative glycosylation site (see Fig. 2) is analogous in position to the carbohydrate in Fc fragments (section 2.5). This sugar mediates CH2-to-CH2 interactions across the dimer interface [30].

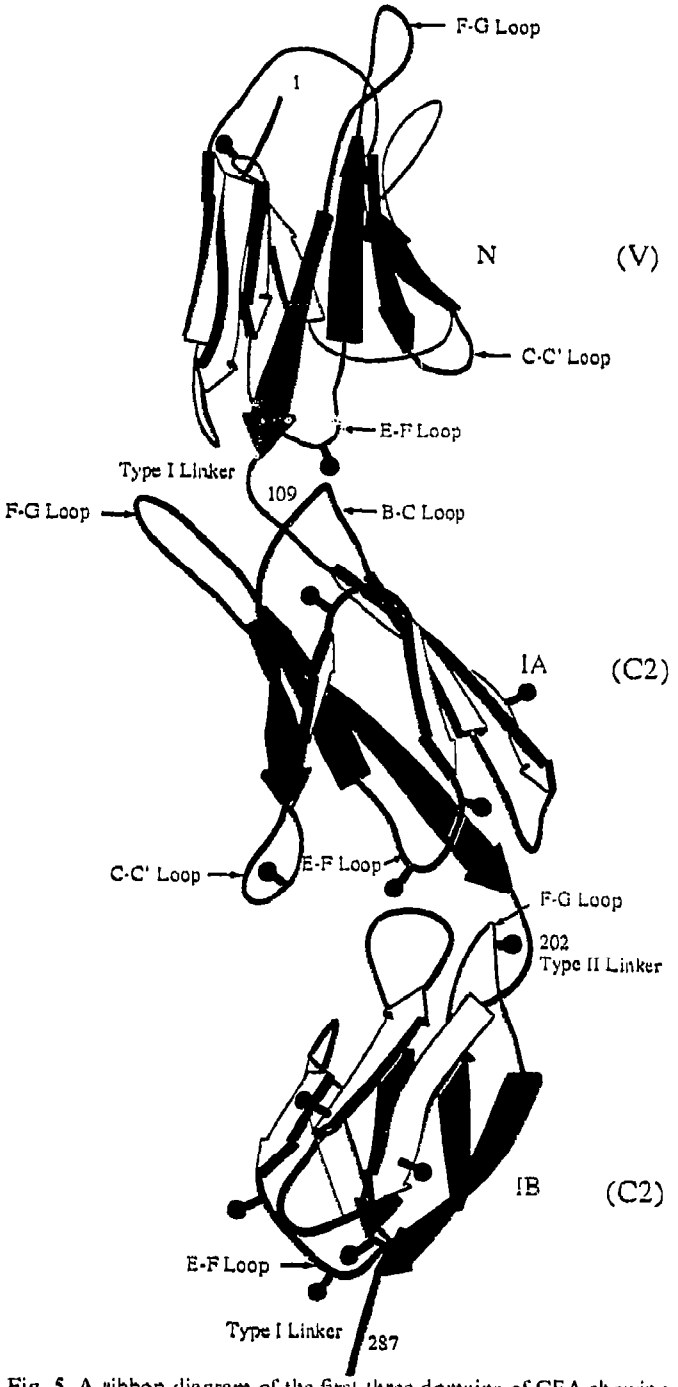

Fig. 5. A ribbon diagram of the first three domains of CEA showing the postulated packing arrangement. The complete molecule, due to rotation and translation of the A and B repeats, would be rod-like in appearance. The CFG face of each domain is shaded. (This includes the C' strand for V.) These CFG faces alternate in orientation, approx. 180°, about the central axis of the molecule. This same alternation of CFG faces is found for VL-to-VH packing geometries in antibodies but also in other cell surface antigens such as CD4 [13] and ICAM-1 [28]. Asn-linked carbohydrate positions are shown as solid knobs. In general the CFG faces are less well covered by carbohydrates than the AB(D)E faces; this is particularly noticeable for the N-terminal domain. A few of the loops are labelled. Particularly exposed and variable loops (see Fig. 2) are the E–F and the F–G loops, respectively. These loops are predicted to be on alternate sides of the molecule. The figure was produced by the aid of a plotting program called RIBBON [46].

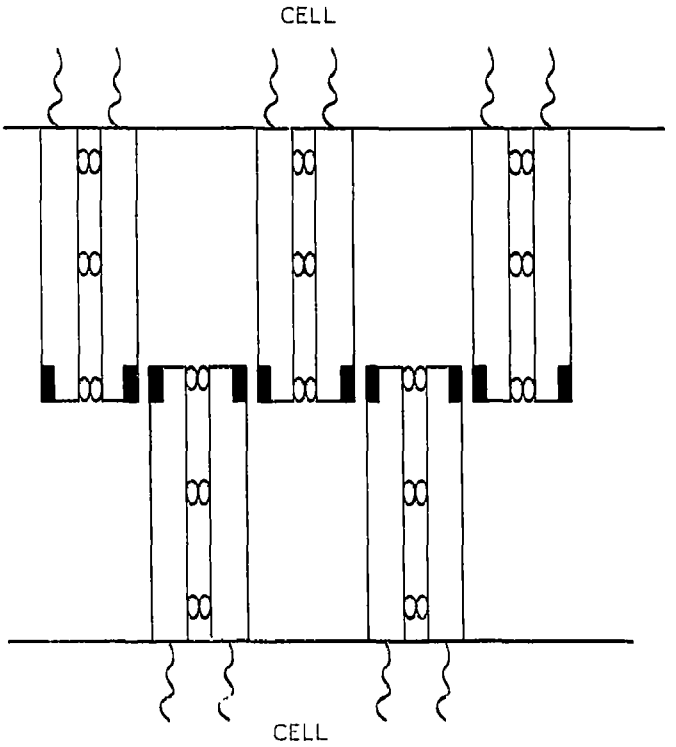

Fig. 6. A highly schematic diagram for how CEA chains, shown as oblong boxes, may interact to form both homodimers on the same cell and homotypic interactions between cells. Squiggly lines indicate the transmembrane region of each chain. Ovals represent some of the carbohydrates that would be trapped at the interface between chains of each dimer pair. The shaded boxes represent the CFG face on each N-terminal domain that could mediate cell–cell homotypic adhesion. Such arrays of chains would form a stronger bond between cells than single-chain CEA homotypic adhesion.

One of the functions of CEA is to mediate cell-to-cell association via homotypic adhesion [39]. This interaction may involve the first one or two N-terminal domains of CEA, as expected for a closely related cell surface adhesion molecule NCAM (neural cell adhesion molecule) [40]. These interactions are similar to cell–cell association mediated by two Ig superfamily proteins, CD2 and LFA-3. There is some evidence to suggest that the first N-terminal from CD2 and the first N-terminal domain from LFA-3 mediate most of the homotypic interaction and that this occurs on their CFG faces [14]. This face is also the packing face for each VL and VH domain in Fab fragments. The implication is that a sticky patch has evolved on this face of an Ig fold. For the N-terminal domain of CEA this is also the face that seems to be void of carbohydrate. Indeed if the CFG face of the first N-terminal domain of CEA contributes all or most of the binding interface for homotypic interaction, then this face would be the prime candidate for mutational studies.

In addition to the homotypic interactions discussed above, there also exists the possibility of homodimerization of CEA chains on the same cell (section 2.6). For

a homodimerization model an analogy with antibody can once again be taken. The two chains of the dimer could run parallel to each other with some carbohydrates trapped between the dimer interface as found in Fc fragments [30]. If the top N-terminal domain was covered on its ABDE face by the homodimer interface, the CFG face would still be free for homotypic interactions between cells. Therefore, as shown in Fig. 6, if both homodimerization and homotypic interactions occur at the same time an extensive array of CEA interactions can build up between cells producing a stronger adhesion than just single-chain CEA homotypic interactions on their own.

In our model the linkers between domains are a few residues shorter than in Fab fragments (see Fig. 3). Therefore we tentatively suggest that the angles between domains are less acute and the packing tighter than in Fab fragments. Tighter packing between Ig folds with similar twist angles between domains as those predicted for CEA has been found between the first two N-terminal domains of CD4 and is postulated for the first two N-terminal domains of ICAM-1 [28]. Thus the complete molecule is expected to be rod shaped, Fig. 5, with no obvious kink between any of the seven domains. An electron micrograph study of CEA [41] also indicates a rod-shaped molecule. This contrasts with the cell surface proteins NCAM [40] and ICAM-1 [28] which have been shown by electron micrograph studies to contain hinge regions.

Residues suitable for site-directed mutagenesis studies have been indicated on the sequence alignments (Fig. 2). It is intended that this model for CEA, particularly the question of precisely how the domains are packed together, can be refined as more site-directed mutagenesis experiments are performed.

Acknowledgements: We thank Drs. Paul Driscoll, Jason Cyster, Iain Campbell and Alan Williams for a pre-release of the coordinates of CD2, and Sir Walter Bodmer, Helga Durbin and Dr. Fritz Wrba for helpful discussions.

REFERENCES

1 Terry, W., Henkart, P., Coligan, J. and Todd, C. (1974) Transplant. Rev. 20, 100–129.
2 Thomas, P., Toth, C.A., Saini, K.S., Jessup, J.M. and Steele, G. (1990) Biochim. Biophys. Acta 177–189.
3 Thompson, J.A., Grunert, F. and Zimmermann, W. (1991) J. Clin. Lab. Anal. 5, 344–366.
4 Shively, J.E. and Beatty, J.D. (1985) CRC Crit. Rev. Oncol. Hematol. 2, 355–399.
5 Loewenstein, M.S. and Zamcheck, N. (1977) Gastroenterology 72, 161–164.
6 Mach, J.P., Chatal, J.F., Lumbroso, J.D., Buchegger, F., Forni, M., Ritschard, J., Berche, C., Douillard, J.Y., Carrel, S., Herlyn, M., Steplewski, Z. and Koprowski, H. (1983) Cancer Res. 43, 5593–5600.
7 Wahren, B., Hammarstrom, S., Harmenberg, V., von Krusenstierna, S., Larsson, S., Lundell, G., Ruden, U. and Schnell, P.O. (1986) Tumor Biol. 7, 361–369.
8 Beatty, J.D., Duda, R.B., Williams, L.E., Sheibani, K., Paxton, R.J., Beatty, B.G., Philben, V., Werner, J., Shively, J.E., Vlahos, W., Hohal, W., Riihimaki, D., Terz, J.J. and Wagman, L.D. (1986) Cancer Res. 46, 6494–6502.
9 Hammerstrom, S., Shively, J.E., Paxton, R.J., Beatty, B.G., Larsson, A., Ghosh, R., Bormer, O., Buchegger, F., Mach, J.P., Burtin, P., Seguin, P., Darbouret, B., Degorce, F., Sertour, J., Jolu, J.P., Fuks, A., Kalthoff, H., Schmiegel, W., Arndt, R., Kloppel, G., von Kleist, S., Grunert, F., Schwarz, K., Matsuoka, Y., Kuroki, M., Wagener, C., Weber, T., Yachi, A., Imai, K., Hishikawa, N. and Tsujisaki, M. (1989) Cancer Res. 49, 4852–4858.
10 Beauchemin, N., Benchimol, S., Cournoyer, D., Fuks, A. and Stanners, C.P. (1987) Mol. Cell Biol. 7, 3221–3231.
11 Williams, A.F. (1987) Immunol. Today 8, 298–303.
12 Oikawa, S., Imajo, S., Noguchi, T., Kosaki, G. and Nakazato, H. (1987) Biochem. Biophys. Res. Commun. 144, 634–642.
13 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 348, 411–418.
14 Driscoll, P.C., Cyster, J.G., Campbell, I.D. and Williams A.F. (1991) Nature 353, 762–765.
15 Barton, G.J. and Sternberg, M.J.E. (1987) J. Mol. Biol. 198, 327–337.
16 Blundell, T.L., Sibana, M.J., Sternberg, M.J.E. and Thornton, J.M. (1987) Nature 326, 347–352.
17 Sutcliffe, M.J., Hayes, F.R.F. and Blundell, T.L. (1987) Prot. Eng. 1, 377–384.
18 Jones, T.A. and Thirup, S. (1986) EMBO J. 5, 819–822.
19 Claessens, M., Van Cutsem, E., Lasters, I. and Wodak, S. (1989) Prot. Eng. 2, 335–345.
20 Dayhoff, M.O., Barker, W.C. and Hunt, L.T. (1983) Methods Enzymol. 91, 524–545.
21 Epp, O., Lattman, E.E., Shiffer, M., Huber, R. and Palm, W. (1975) Biochemistry 14, 4943–4952.
22 Bates, P.A., Sternberg, M.J.E. (1992) in: Protein Engineering, A Practical Approach (Rees, A.R., Wetzel, D. and Sternberg, M.J.E., eds.) IRL Press, Oxford, in press.
23 Summers, L.J., Blundell, T.L., Gause, G.G. and Tomarev, S.I. (1986) FEBS Lett. 208, 11–16.
24 McGregor, M.J., Islam, S.A. and Sternberg, M.J.E. (1987) J. Mol. Biol. 198, 295–310.
25 Brooks, B.R., Bruccoleri, R.E., Olafson, B.D., States, D.J., Swaminathan, S. and Karplus, M. (1983) J. Comput. Chem. 4, 187–217.
26 Novotny, J., Bruccoleri, R.E. and Karplus, M. (1984) J. Mol. Biol. 177, 787–818.
27 Baumann, G., Frommel, C. and Sander, C. (1989) Prot. Eng. 2, 329–334.
28 Berendt, A.R., McDowall, A., Craig, A.G., Bates, P.A., Sternberg, M.J.E., Marsh, K., Newbold, C.I. and Hogg, N. (1992) Cell 68, 71–81.
29 Weis, W., Brown, J.H., Cusack, S., Paulson, J.C., Skehel, J.J. and Wiley, D.C. (1988) Nature 333, 426–431.
30 Deisenhofer, J. (1981) Biochemistry 20, 2361–2370.
31 Suck, D., Oefner, C. and Kabsch, W. (1984) EMBO J. 3, 2423–2430.
32 Das, H.K., Lawrance, S.K. and Weissman, S.M. (1983) Proc. Natl. Acad. Sci. USA 80, 3543–3547.
33 Tonnelle, C., Demars, R. and Long, E.O. (1985) EMBO J. 4, 2839–2847.
34 Brown, J.H., Jardetzky, T., Saper, M.A., Samraoui, B., Bjorkman, P.J. and Wiley, D.C. (1988) Nature 332, 845–850.
35 Travers, P., Blundell, T.L., Sternberg, M.J.E. and Bodmer, W.F. (1984) Nature 310, 235–238.
36 Lisowska, E., Krop-Watorek, A. and Sedlaczek, P. (1983) Biochem. Biophys. Res. Commun. 115, 206–211.
37 Schwarz, K., Bruckel, N., Schwaibold, H., von Kleist, S. and Grunert, F. (1989) Mol. Immun. 26, 467–475.
38 Pignatelli, M., Durbin, H. and Bodmer, W.F. (1990) Proc. Natl.

Acad. Sci. USA 87, 1541–1545.
39 Benchimol, S., Fuks, A., Jothy, S., Beauchemin, N., Shirota, K. and Stanners, C.P. (1989) Cell 57, 327–334.
40 Becker, J.W., Erickson, H.P., Hoffman, S., Cunningham, B.A. and Edelman, G.M. (1989) Proc. Natl. Acad. Sci. USA 86, 1088–1092.
41 Slayter, H.S. and Coligan, J.E. (1975) Biochemistry 14, 2323–2330.
42 Poljak, R.J., Amzel, L.M., Avey, H.P., Chen, B.L., Phizackerley, R.P. and Saul, F. (1973) Proc. Natl. Acad. Sci. USA 70, 3305–3310.
43 Tawaragi, Y., Oikawa, S., Matsuoka, Y., Kosaki, G. and Nakazato, H. (1988) Biochem. Biophys. Res. Commun. 150, 89–96.
44 Barnett, T.R., Kretschmer, A., Austen, D.A., Goebel, S.J., Hart, J.T., Elting, J.J. and Kamarck, M.E. (1989) J. Cell Biol. 108, 267–276.
45 Arakawa, F., Kuroki, M., Misumi, Y., Oikawa, S., Nakazato, H. and Matsuoka, Y. (1990) Biochem. Biophys. Res. Commun. 166, 1063–1071.
46 Priestle, J.P. (1988), J. Appl. Crystallogr. 21, 572–576.