Proc. Natl. Acad. Sci. USA
Vol. 84, pp. 920–924, February 1987
Biochemistry

Sequence analysis of carcinoembryonic antigen: Identification of glycosylation sites and homology with the immunoglobulin supergene family

(tumor antigens/microsequence analysis/fast atom bombardment mass spectrometry/deglycosylation/reverse-phase HPLC)

RAYMOND J. PAXTON, GREGORY MOOSER*, HEMA PANDE, TERRY D. LEE, AND JOHN E. SHIVELY†

Division of Immunology, Beckman Research Institute of the City of Hope, Duarte, CA 91010

Communicated by M. Frederick Hawthorne, October 8, 1986

ABSTRACT **A direct method for the determination of N-linked glycosylation sites in highly glycosylated proteins is described. Carcinoembryonic antigen (CEA) and a nonspecific crossreacting antigen (NCA) were chemically degycosylated, and peptide maps were prepared by reverse-phase HPLC. The peptides were sequenced on a gas-phase microsequencer, and glycosylation sites were identified as the phenylthiohydantoin derivative of *N*-acetylglucosaminylasparagine. The sequences were confirmed by fast atom bombardment mass spectrometry. Highly homologous, extended amino-terminal sequences were determined for CEA and two NCAs, NCA-95 and NCA-55. Cysteine-containing sequences for CEA and NCA-95 show up to 95% sequence homology, and the CEA sequences also show internal sequence homologies. A comparison of the CEA sequences with known protein sequences suggests that CEA may be a member of the immunoglobulin supergene family. The protein sequence data have been used to identify a genomic DNA clone for one of the NCA antigens [Thompson, J., Pande, H., Paxton, R. J., Shively, L., Padma, A., Simmer, R. L., Todd, C. W., Riggs, A. D. & Shively, J. E. (1987) *Proc. Natl. Acad. Sci. USA*, in press] and a cDNA clone for CEA [Zimmermann, W., Ortlieb, B., Friedrich, R. & von Kleist, S. (1987) *Proc. Natl. Acad. Sci. USA*, in press].**

Carcinoembryonic antigen (CEA), first described by Gold and Freedman (1, 2), is one of the most widely investigated human tumor-associated antigens. CEA (M_r, 180,000) was originally detected in colonic adenocarcinoma and fetal gut, but has since been detected in other malignant and nonmalignant tissues. The chemistry, biochemistry, immunology, tissue distribution, and clinical aspects of CEA have been reviewed (3). CEA-related antigens have been detected by immunological methods in various tissues. Nonspecific crossreacting antigen (NCA) was isolated from normal spleen and lung (4, 5). Kessler *et al.* (6) isolated an NCA-like antigen from liver metastases of colonic adenocarcinoma that was sufficiently different from spleen NCA (7) to be given the provisional designation tumor-extracted antigen. Buchegger *et al.* (8) have identified two NCA antigens; NCA-55 (M_r, 55,000) corresponds to the originally described molecule and is present in granulocytes and epithelial cells, whereas NCA-95 (M_r, 95,000) is present only in granulocytes. Grunert *et al.* (9) have detected these two antigens in colon tumor and normal lung and also a 75-kDa antigen in colon tumor. Other CEA-related antigens are NCA2 (M_r, 160,000) from meconium (10), biliary glycoprotein I (M_r, 85,000) from bile (11), and a group of antigens from feces (12). Neumaier *et al.* (13, 14) have also detected a 128-kDa antigen in colonic adenocarcinoma, a 100-kDa antigen in meconium, and several antigens in normal plasma.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

CEA contains 50–60% carbohydrate by weight, and the NCAs contain 20–50% carbohydrate by weight. The amino acid compositions of CEA and the NCA antigens are comparable. The amino-terminal sequences of tumor-extracted antigen (6) and NCA (7) are identical but differ from the CEA sequence (15) in having alanine rather than valine at position 21. The further structural analysis of CEA has been hampered by the high degree of glycosylation, which renders it extremely resistant to proteolytic cleavage. Shively *et al.* (16) digested CEA with trypsin in the presence of Triton X-100; however, extended sequences of the resulting peptides were not obtained, probably due to the presence of carbohydrate in these peptides. CEA has also been deglycosylated with anhydrous HF (17). Preliminary studies showed that deglycosylated CEA was more amenable than native CEA to proteolytic cleavage and subsequent sequence analysis (18).

In the present study, we describe a general method for the structural analysis of highly glycosylated proteins and apply this method to CEA, NCA-95, and NCA-55. This method, which includes chemical deglycosylation, peptide mapping by reverse-phase HPLC, gas-phase microsequence analysis, and fast atom bombardment (FAB) mass spectrometry (MS), allowed the direct determination of numerous N-linked glycosylation sites in CEA and NCA-95 peptides. We also report highly homologous, extended amino-terminal sequences for CEA, NCA-95, and NCA-55 and homologous cysteine-containing sequences for CEA and NCA-95. The CEA cysteine-containing sequences show internal homology, suggesting that CEA evolved by a series of gene duplication events. In addition, CEA cysteine-containing sequences are homologous to sequences found in several members of the immunoglobulin supergene family (19).

MATERIALS AND METHODS

Purification of Antigens. CEA, NCA-95, and NCA-55 were isolated from liver metastases of colon tumors and purified as described for CEA (20). Final purifications of NCA-95 and NCA-55 were achieved by reverse-phase HPLC (R.J.P. and J.E.S., unpublished data).

Deglycosylation and Peptide Mapping. CEA and NCA-95 were chemically deglycosylated with trifluoromethanesulfonic acid (TFMSA):anisole, 2:1 (vol/vol), as described (21) and desalted by dialyzing against 10% (vol/vol) aqueous

Abbreviations: CEA, carcinoembryonic antigen; NCA, nonspecific crossreacting antigen; FAB, fast atom bombardment; >PhNCS, phenylthiohydantoin; GlcNAc, *N*-acetylglucosamine; Asn(GlcNAc), *N*-acetylglucosamine attached to the side chain of asparagine; Asn(GlcNAc)>PhNCS, phenylthiohydantoin derivative of Asn(GlcNAc); N-CAM, neural cell-adhesion molecule; MS, mass spectrometry; TFMSA, trifluoromethanesulfonic acid.

*Permanent address: School of Dentistry, University of Southern California, Los Angeles, CA 90089.
†To whom reprint requests should be addressed.

pyridine. Deglycosylated CEA and deglycosylated NCA-95 were reduced with dithiothreitol and carboxymethylated with iodoacetic acid as described (22). The carboxymethylated antigens (1–5 nmol) were dialyzed against 0.2 M NH_4HCO_3, pH 7.8, and digested with L-1-tosylamido-2-phenylethyl chloromethyl ketone-treated trypsin or chymotrypsin (Worthington) at a protein:enzyme ratio of 50:1 (wt/wt) at 37°C for 18 hr. The peptides were separated by reverse-phase HPLC on a Vydac C_{18} column using a linear gradient of acetonitrile in 0.1% (vol/vol) trifluoracetic acid.

Microsequence Analysis. Proteins and peptides (0.1–2 nmol) were subjected to automated Edman degradation on a gas-phase microsequencer built at the City of Hope (23), and phenylthiohydantoin (>PhNCS) amino acid derivatives were separated and identified by reverse-phase HPLC (23). Amino acid compositions were determined using a Beckman 121 MB amino acid analyzer.

Synthesis of the >PhNCS Derivative of *N*-Acetylglucosaminylasparagine [Asn(GlcNAc)>PhNCS]. The starting amino acid Asn(GlcNAc) [10 μmol; listed as 2-acetamido-1-β(L-aspartamido)-1,2-dideoxy-D-glucose (Sigma)] was dissolved in 0.05 ml of triethylamine/pyridine/H_2O, 10:40:50 (vol/vol). Phenylisothiocyanate (20 μmol; Pierce sequanal grade) was added, and the reaction was stirred at 43°C for 20 min. The reaction mixture was extracted four times with 0.05 ml of benzene and dried in a vacuum centrifuge. The yellow oil was dissolved in 0.1 ml of 25% (vol/vol) aqueous trifluoroacetic acid and stirred at 50°C for 20 min at which time the product precipitated from solution. Asn(GlcNAc)>PhNCS was solubilized in 50% (vol/vol) aqueous acetonitrile and analyzed as follows. A single peak was obtained by reverse-phase HPLC, aspartic acid/asparagine and glucosamine were detected by amino acid analysis, and FAB-MS showed the expected mass-to-charge ratio (m/z) of 453, $(M + H)^+$. Carboxymethylcysteine>PhNCS, Glu>PhNCS, and Asn>PhNCS were synthesized using the same procedure.

FAB-MS. Samples (0.1–0.5 nmol) were concentrated to dryness in polypropylene microcentrifuge tubes using a vacuum centrifuge, redissolved in 1–2 μl of 5% (vol/vol) aqueous acetic acid, and added to 3 μl of glycerol or 3 μl of a dithiothreitol/dithioerythritol, 5:1 (wt/wt) mixture on a 1.5 × 5 mm stainless steel sample stage. FAB spectra were taken with a JEOL HX-100HF mass spectrometer utilizing a 6-kV xenon atom primary beam, and data were collected with a JEOL DA500 data system.

RESULTS AND DISCUSSION

Amino Acid Composition and Amino-Terminal Sequence Analysis. CEA, NCA-95,‡ and NCA-55 were isolated from liver metastases of colonic adenocarcinoma. Table 1 shows amino acid compositions for these antigens. The most notable difference is the apparent absence of methionine in CEA. In some CEA samples a trace of methionine is found; however, the mol% is always less than that determined for NCA-95 and NCA-55. It has been reported that several NCA proteins were not cleaved by treatment with cyanogen bromide and, hence, did not contain methionine (24). We have not tried this experiment, but we have detected ≈1 mol% of methionine in reverse-phase HPLC purified NCA-95 and NCA-55 and have identified a methionine residue in an NCA-95 tryptic peptide by microsequence analysis.

‡This laboratory has reported the isolation of a NCA-like antigen from colonic adenocarcinoma (6). It is now clear that two forms of NCA (M_rs, 95,000 and 55,000) can be isolated from tumor. We have previously designated the larger molecule TEX, but we will now adopt the convention of Buchegger *et al.* (8) and refer to it as NCA-95.

Table 1. Amino acid compositions of CEA, NCA-95, and NCA-55

Amino acid	Amino acid mol%		
	CEA	NCA-95	NCA-55
Cys/2	1.7	1.6	1.4
Asx	13.4	12.7	12.2
Thr	8.7	8.4	8.3
Ser	10.8	9.4	9.3
Glx	10.2	11.3	11.6
Pro	8.8	7.7	7.4
Gly	5.6	6.9	8.1
Ala	5.5	6.2	6.2
Val	8.1	7.0	7.0
Met	0	1.0	1.3
Ile	4.2	4.4	4.3
Leu	8.3	8.8	8.6
Tyr	4.5	4.8	4.8
Phe	2.6	2.7	2.5
His	1.7	1.9	1.3
Lys	2.5	2.8	3.2
Arg	3.5	2.6	2.8

Samples were hydrolyzed in 6 M HCl containing 0.2% 2-mercaptoethanol at 110°C for 48 hr. Cysteine was determined as cysteic acid in a separate analysis after performic acid oxidation. Percent carbohydrate by weight for CEA is 52%, for NCA-95 is 46%, and for NCA-55 is 30%. These values are estimates based on earlier studies (3) and on the amount of glucosamine determined during amino acid analysis.

Studies in this laboratory determined amino-terminal sequences, up to residue 24, for CEA, NCA-95, and NCA-55 (6, 7, 15). Fig. 1 shows extended sequences for the three proteins. Also shown is the translated protein sequence of the amino-terminal coding exon for an NCA genomic DNA clone (25). All of the peptides necessary to extend the CEA sequence were found in a single chymotryptic map of deglycosylated CEA (see below). There are 10 amino acid differences between the CEA sequence and the translated sequence of the NCA clone. There are no differences between the NCA-95 and NCA-55 protein sequences and the translated sequence; hence, the exact identity of this clone cannot yet be established.

```
                    10        20        30        40
CEA      KLTIESTPFNVAEGKEVLLLVHNLPQHLFGYSWYKGERVD
NCA-95   KLTIESTPFNVAEGKEVLLLAHNLPQNRIGYSWYK  VD
NCA-55   KLTIESTPFNVAEGKEVLLLAHNLPQNRIGYSWYK
NCA EXON KLTIESTPFNVAEGKEVLLLAHNLPQNRIGYSWYKGERVD

                    50        60        70        80
CEA      GNRQITGYVIGTQQATPGPAYSGRETIYPNASLLIQNVTQ
NCA-95   GNSLIVGYVIGTQQATPGPAYSGR
NCA EXON GNSLIVGYVIGTQQATPGPAYSGRETIYPNASLLIQNVTQ

                    90        100       110
CEA      NDTGFYTLHVIKSDLVNEEATGQFRVYPELPKPSISS...
NCA-95             VIKSDLVNEEATGQFHVYPELPKPSISS...
NCA EXON NDTGFYTLQVIKSDLVNEEATGQFHVY
```

FIG. 1. Sequence homologies at the amino termini of CEA, NCA-95, and NCA-55. Amino-terminal sequencer runs identified the first 25, 35, and 35 residues for CEA, NCA-95, and NCA-55, respectively. The CEA and NCA-95 sequences were confirmed and extended by sequencing tryptic and chymotryptic peptides prepared from the deglycosylated antigens. These peptides were aligned using sequence overlaps and the translated protein sequence of an NCA genomic clone isolated in this laboratory (25). The first 12 amino acids coded for by this exon are part of the signal peptide and are not shown. Differences between CEA and the NCA sequences are indicated with boxes. The bold asparagine residues (N) were identified as Asn(GlcNAc) residues during sequence analysis and, therefore, are sites of glycosylation.

The first three sites of glycosylation in CEA, asparagine residues 70, 77, and 81, have also been determined (see below). The clustering of glycosylation sites is found in other parts of the CEA molecule and generally occurs near cysteine residues. Hence, this region at one time may have contained a cysteine residue that mutated to a different amino acid during evolution.

Deglycosylation and Peptide Mapping. CEA and NCA-95 were deglycosylated using TFMSA. Fig. 2 shows the expected decrease in molecular weight for the two proteins. These results are similar to those described for deglycosylation with anhydrous HF (17, 18). Because the gel was overloaded, several minor bands are visible in the deglycosylated CEA and deglycosylated NCA-95 lanes. These bands may represent variant forms of CEA and NCA-95 or trace amounts of unrelated glycoproteins that were also deglycosylated during the procedure. Another explanation is that peptide bond cleavage occurred during deglycosylation. Following reduction and carboxymethylation, the deglycosylated antigens were digested with trypsin or chymotrypsin, and the peptides were separated by reverse-phase HPLC. Fig. 3 shows a chymotryptic map for deglycosylated CEA. Equivalent results were obtained for deglycosylated NCA-95. The tryptic maps for deglycosylated CEA and NCA-95 showed well-separated peptides eluting early in the gradient; however, the later-eluting peptides were not as well resolved as those in Fig. 3. This was potentially due to their larger size and hydrophobic character.

Deglycosylation of CEA and NCA-95 has greatly facilitated their structural analysis. In general, highly glycosylated proteins are resistant to proteolytic enzymes, and those peptides that are generated are difficult to separate (27, 28). Previous studies by Shively *et al.* (16) documented this for CEA. Conversely, the deglycosylated antigens are amenable to enzymatic proteolysis and peptide mapping by reverse-phase HPLC, and chemical deglycosylation leaves a single residue of N-linked *N*-acetylglucosamine (GlcNAc) at each site of glycosylation. The modified asparagine residues have been used to directly identify several sites of glycosylation in CEA and NCA-95 (see below).

Internal Sequence Homology Within CEA Cysteine-Containing Sequences. Fig. 4 shows 13 CEA cysteine-containing sequences segregated into four groups based on sequence homologies. The homologies within groups 1–4 are 84%, 68%, 86%, and 73%, respectively. Homologies between groups are also evident. Groups 2 and 4 have tyrosine at cysteine-2 (i.e., two residues before cysteine), aspartic acid or asparagine at cysteine-6, asparagine at cysteine-7, threonine at cysteine-9, leucine at cysteine-15, asparagine at cysteine +4 (i.e., four residues after cysteine), and serine or threonine at cysteine +7. The homologies between other pairs of groups are not as striking, although with conservative amino acid changes homologies can be found. The intergroup and intragroup homologies suggest that CEA evolved from a primordial gene, which coded for one or two cysteine-containing sequences, by a series of tandem gene duplications and divergencies (further discussed below).

Four NCA-95 cysteine-containing sequences are included in Fig. 4 to illustrate the interprotein sequence homologies. The NCA-95 sequences are most homologous to the CEA sequences immediately preceding them in the figure. The sequence homologies between CEA and NCA-95 suggest that

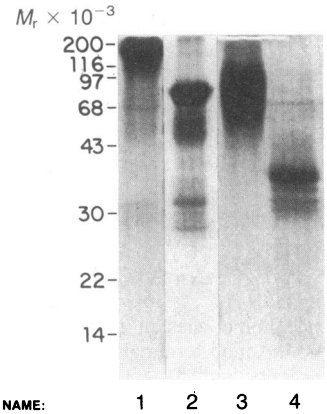

FIG. 2. $NaDodSO_4$/PAGE analysis of CEA (lanes 1 and 2) and NCA-95 (lanes 3 and 4) before (lanes 1 and 3) and after (lanes 2 and 4) deglycosylation. The antigens were deglycosylated, and aliquots were concentrated to dryness for $NaDodSO_4$/PAGE. CEA and NCA-95 (100 μg each), deglycosylated CEA (34 μg), and deglycosylated NCA-95 (13 μg) were reduced with 5% (vol/vol) 2-mercaptoethanol, electrophoresed on a 12% polyacrylamide gel, and stained with Coomassie blue R-250 (26).

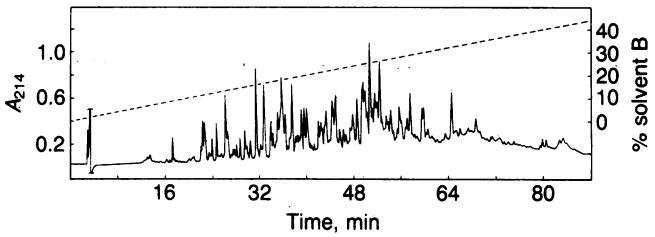

FIG. 3. Chymotryptic map of TFMSA-treated CEA. A 5-nmol sample of carboxymethylated deglycosylated CEA was digested with 2% (wt/wt) chymotrypsin at 37°C for 18 hr. The digest was applied to a Vydac C_{18} column, and the peptides were eluted with a linear gradient of 100% (vol/vol) buffer A to 40% (vol/vol) buffer A/60% (vol/vol) buffer B in 120 min. Buffer A is 0.1% trifluoroacetic acid in H_2O. Buffer B is 0.1% trifluoroacetic acid/9.9% (vol/vol) H_2O/90% (vol/vol) acetonitrile. A_{214}, solid line. % buffer B, dashed line.

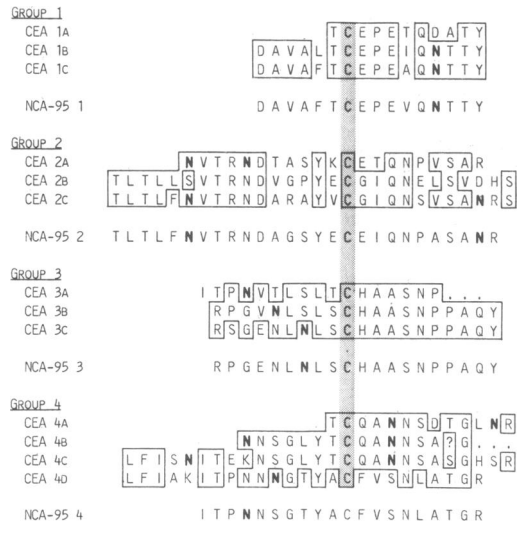

FIG. 4. Internal sequence homologies within the cysteine-containing regions of CEA. Tryptic and chymotryptic peptide sequences from CEA were ordered using sequence overlaps, and the cysteine-containing regions were segregated into four groups based on sequence homologies. In a few instances, the peptides were ordered using CEA cDNA sequence data (29). The homologies within each group are indicated by boxes, and the cysteine residues of each group have been aligned to illustrate intergroup homologies. The NCA-95 sequences were obtained using the same methods. The bold asparagine residues (N) were identified during sequence analysis as Asn(GlcNAc) residues and, therefore, are sites of glycosylation. Sequences 1B, 1C, 2B, 2C, 3B, 4C, and 4D have been confirmed by CEA cDNA sequence. Sequences 3A, 4A, and 4B are incomplete, and the five carboxyl-terminal amino acids of sequence 2B have been omitted.

during evolution entire genes, or portions thereof, may have duplicated and then diverged to generate the CEA gene family, which contains at least 7, and potentially as many as 10, members (3, 25).

Identification of Glycosylation Sites. Fig. 5*A* shows the reverse-phase HPLC chromatograms for sequencer cycles 1–4 of a deglycosylated CEA tryptic peptide. The >PhNCS observed at cycle 1 eluted between Glu>PhNCS and Asn>PhNCS and was initially unidentified; however, cycles 2–7 showed the sequence Asp-Thr-Ala-Ser-Tyr-Lys. A comparison of this sequence with the amino acid composition suggested the possibility of aspartic acid/asparagine and GlcNAc in cycle 1. This was confirmed by back hydrolysis (6 M HCl, 0.2% 2-mercaptoethanol, 110°C, 24 hr) of the cycle 1 >PhNCS, and it was proposed that this derivative was Asn(GlcNAc)>PhNCS. This assignment and the sequence above were confirmed by FAB-MS of the intact peptide (Fig. 5*B*), which showed the expected mass-to-charge ratio (m/z) of 1001, $(M + H)^+$. The first three cycles also satisfy the obligate acceptor sequence, Asn-Xaa-Ser/Thr, for N-glycosylation (30). A final confirmation of the Asn(GlcNAc) >PhNCS derivative was obtained by synthesis. The elution position for synthetic Asn(GlcNAc)>PhNCS, between Glu>PhNCS and Asn>PhNCS (Fig. 5*C*), was the same as that observed for the sequencer-generated derivative.

Three sites of glycosylation are present in the amino terminus of CEA (Fig. 1). Chymotryptic peptides corresponding to residues 62–74 and residues 75–86 were sequenced and analyzed by FAB-MS. Asn(GlcNAc)>PhNCS

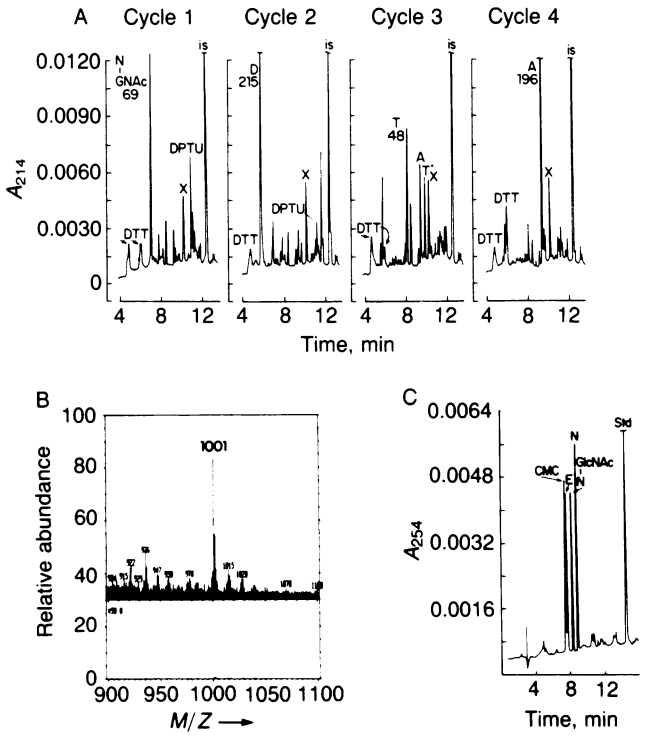

FIG. 5. Identification of a CEA glycosylation site. (*A*) Reverse-phase HPLC chromatograms for sequencer cycles 1–4 of a glycosylated tryptic peptide from TFMSA-treated CEA (40% injected). Cycle 1 was determined to be Asn(GlcNAc)>PhNCS and, hence, defines a site of glycosylation. Picomoles of each residue are indicated under the assigned residues. is, Internal standards; DTT, dithiothreitol; DPTU, diphenylthiourea; X, an uncharacterized product observed during sequence analysis. (*B*) FAB mass spectrum of the intact peptide. The expected mass-to-charge ratio (m/z) of 1001 for the molecular ion $(M + H)^+$ was observed. (*C*) Chromatogram showing the elution position of synthetic Asn(GlcNAc)>PhNCS relative to three closely eluting >PhNCS. Approximately 25 pmol of each was analyzed. Std, internal standard.

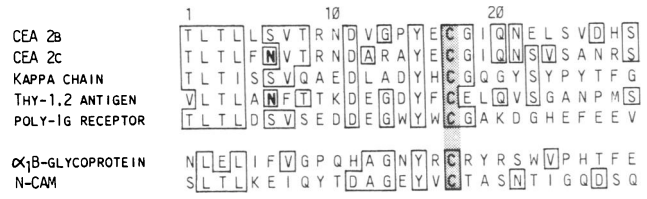

FIG. 6. Sequence homologies between CEA and immunoglobulin supergene family members. CEA sequences from Fig. 4 are aligned with: a mouse κ-chain variable region (MOPC-21) (37), residues 72–97; mouse Thy-1.2 antigen, residues 70–97; rabbit polyimmunoglobulin receptor, residues 522–549; α_1B-glycoprotein, residues 433–460; and chicken N-CAM, residues 242–269. Homologies between CEA and the other proteins are boxed. The bold asparagine residues (**N**) are sites of glycosylation. Isoleucine in the κ chain (position 3) is considered equivalent to leucine.

was easily identifiable in cycle 9 for the first peptide and in cycles 3 and 7 for the second peptide (data not shown). FAB-MS showed the expected molecular ions for the two peptides, m/z = 1636 and m/z = 1806, respectively. Hence, the ability to identify Asn(GlcNAc) during microsequence analysis is not dependent on its relative location in the peptide. The yield of Asn(GlcNAc)>PhNCS is ≈30% of that expected. This is perhaps due to a lower solubility of the aminothiazolinone derivative of Asn(GlcNAc) in the extracting solvent, butyl chloride, relative to other amino acid aminothiazolinone derivatives. A few peptides containing Asn(GlcNAc-GlcNAc) residues (due to incomplete deglycosylation) have also been isolated. When these peptides were sequenced, no >PhNCS for the modified asparagine residue was detected; however, the presence of Asn(GlcNAc-GlcNAc) was confirmed by FAB-MS of the intact peptide. In these cases the additional GlcNAc probably renders the aminothiazolinone derivative insoluble in butyl chloride.

To our knowledge, this is the first time N-linked glycosylation sites have been directly identified by automated microsequence analysis, although an O-linked glycosylation site in human interleukin-2 has been directly identified (31). Takahashi *et al.* (32) identified Asn(GlcNAc)>PhNCS in a four-residue glycopeptide; however, several milligrams of peptide and manual Edman degradation were used. Although the sequences of numerous glycoproteins have been determined, the identification of glycosylation sites is generally accomplished by indirect methods (27, 28, 33, 34).

Sequence Homology to the Immunoglobulin Supergene Family. The sequences in Fig. 4 were compared to the National Biomedical Research Foundation Protein Sequence Database[§] using the BIONET data base searching program IFIND. For CEA-2B, 10 of the top-scoring 20 proteins (SD from the mean, ≥6.2) were members of the immunoglobulin supergene family. For CEA-2C, 16 of the top-scoring 31 proteins (SD, ≥5.7) were members of this family. The 26 immunoglobulin supergene family members included 16 κ-chain variable regions, the mouse and rat Thy-1 antigens (34), the rabbit polyimmunoglobulin receptor (35), and the human secretory component (36), which is the proteolytically cleaved form of the polyimmunoglobin receptor. Six of the top-scoring sequences were duplicated. Fig. 6 shows the sequence homologies between the CEA peptides and immunoglobulin supergene family members. The majority of sequence matches occur in the region preceding the cysteine residues. For the mouse κ chain, this sequence is located within the highly conserved immunoglobulin light chain FR3 (framework region 3, residues 57–88) (37). The sequence

[§]Protein Identification Resource (1986) Protein Sequence Database (Natl. Biomed. Res. Found., Washington, DC), Release 8.0.

FIG. 7. Sequence homologies between CEA, α_1B-glycoprotein, and N-CAM. The CEA sequence from Fig. 4 is aligned with residues 397–411 of α_1B-glycoprotein and residues 196–214 of chicken N-CAM. Identical residues are boxed, and a glycosylated asparagine in CEA is indicated as in Fig. 6. The three aromatic amino acids are considered equivalent.

after the cysteine residue is located within the CDR3 (complementarity-determining region 3, residues 89–97). This region makes contact with various antigenic determinants and is not expected to be as conserved as the framework regions. Five residues are conserved in the CEA sequences and the immunoglobulin supergene family sequences. Three of these, the cysteine at position 17, the tyrosine at position 15, and the leucine at position 2, are invariant in all variable region immunoglobulin light chains (37). The aspartic acid at position 11 is probably invariant,¶8 and the threonine at position 3 is invariant in some light-chain subclasses but not in others. The CEA sequences are perhaps most related to the mouse Thy-1 antigen. In addition to the invariant residues, the Thy-1 antigen and the CEA-2C sequence match at the glycosylated asparagine and at three residues following the cysteine. It has been proposed that the Thy-1 antigens may be closely related to the primordial immunoglobulin domain (34). If this is true, the CEA family may also be related to this same domain.

The sequences of α_1B-glycoprotein (38) and chicken N-CAM (neural cell-adhesion molecule) (39) have been shown to be related to the immunoglobulin supergene family. Fig. 6 shows that sequences from these proteins are also related to the CEA sequences. Fig. 7 shows cysteine-containing sequences for α_1B-glycoprotein and N-CAM that occur before, but not adjacent to, the sequences shown in Fig. 6. The homology between these sequences and the CEA-1B sequence is evident. It has been proposed that α_1B-glycoprotein, which includes five homologous domains and 10 cysteine residues (each domain comprises one disulfide loop), evolved by gene duplication from a primordial gene that encoded a single domain and two cysteine residues. Similarly, it has been suggested that the carbohydrate-rich central portion of N-CAM, which consists of four homologous domains and eight cysteine residues, also evolved by gene duplication.

Based on the sequence data in Fig. 4, a similar proposal of gene duplication could be made for CEA and the CEA family members. For CEA there appear to be four cysteine residues in each domain; however, the homology between groups 2 and 4 (Fig. 4) suggests that the original duplicating unit may have contained two cysteine residues. The gene duplication theory is particularly attractive for the CEA family. Not only does it provide an explanation for the homologous sequences in CEA, but it can be used to account for the number of and sizes of the CEA gene family members. For example, NCA-55 may contain one domain (four cysteine residues), NCA-95 two domains, and CEA three or four domains. In addition, each protein contains an amino-terminal domain (Fig. 1) devoid of cysteine residues. The origin of this domain is unclear, but it is possible that it originated from the cysteine-containing domain, followed by mutation of the cysteine residues to other amino acids. The amino-terminal domain would have a different three-dimensional structure and function from the repeating disulfide loop-containing domains.

¶In some of the reported sequences a distinction between aspartic acid and asparagine was not made.

We gratefully thank Kristen Haaga for amino acid analysis and Kassu Legesse for mass spectral analysis. This work was supported by Grant CA37808 from the National Institutes of Health.

1. Gold, P. & Freedman, S. O. (1965) *J. Exp. Med.* **121,** 439–462.
2. Gold, P. & Freedman, S. O. (1965) *J. Exp. Med.* **122,** 467–481.
3. Shively, J. E. & Beatty, J. D. (1985) *CRC Crit. Rev. Oncol./Hematol.* **2,** 355–399.
4. von Kleist, S., Chavanel, G. & Burtin, P. (1972) *Proc. Natl. Acad. Sci. USA* **69,** 2492–2494.
5. Mach, J.-P. & Pusztaszeri, G. (1972) *Immunochemistry* **9,** 1031–1034.
6. Kessler, M. J., Shively, J. E., Pritchard, D. G. & Todd, C. W. (1978) *Cancer Res.* **38,** 1041–1048.
7. Engvall, E., Shively, J. E. & Wrann, M. (1978) *Proc. Natl. Acad. Sci. USA* **75,** 1670–1674.
8. Buchegger, F., Schreyer, M., Carrel, S. & Mach, J.-P. (1984) *Int. J. Cancer* **33,** 643–649.
9. Grunert, F., AbuHarfeil, N., Schwarz, K. & von Kleist, S. (1985) *Int. J. Cancer* **36,** 357–362.
10. Burtin, P., Chavanel, G. & Hirsch-Marie, H. (1973) *J. Immunol.* **111,** 1926–1928.
11. Svenberg, T., Hammarstrom, S. & Hedin, A. (1979) *Mol. Immunol.* **16,** 245–252.
12. Kuroki, M., Kuroki, M., Koga, Y. & Matsuoka, Y. (1984) *J. Immunol.* **133,** 2090–2097.
13. Neumaier, M., Fenger, U. & Wagener, C. (1985) *J. Immunol.* **135,** 3604–3609.
14. Neumaier, M., Fenger, U. & Wagener, C. (1985) *Mol. Immunol.* **22,** 1273–1277.
15. Terry, W. D., Henkart, P. A., Coligan, J. E. & Todd, C. W. (1972) *J. Exp. Med.* **136,** 200–204.
16. Shively, J. E., Kessler, M. J. & Todd, C. W. (1978) *Cancer Res.* **38,** 2199–2208.
17. Glassman, J. N. S., Todd, C. W. & Shively, J. E. (1978) *Biochem. Biophys. Res. Commun.* **85,** 209–216.
18. Shively, J. E., Simmer, R. L., Pande, H., Yang, Y. H. J., Wagener, C., Riggs, A. D. & Todd, C. W. (1984) in *Progress in Cancer Research and Therapy*, eds. Wolman, S. R. & Mastromarino, A. J. (Raven, New York), Vol. 29, pp. 47–57.
19. Hood, L., Kronenberg, M. & Hunkapiller, T. (1985) *Cell* **40,** 225–229.
20. Pritchard, D. G. & Todd, C. W. (1976) *Cancer Res.* **36,** 4699–4701.
21. Edge, A. S. B., Faltynek, C. R., Hof, L., Reichert, L. E., Jr., & Weber, P. (1981) *Anal. Biochem.* **118,** 131–137.
22. Waxdal, M. J., Konigsberg, W. H., Henley, W. L. & Edelman, G. M. (1968) *Biochemistry* **9,** 1959–1966.
23. Hawke, D. H., Harris, D. C. & Shively, J. E. (1985) *Anal. Biochem.* **147,** 315–330.
24. AbuHarfeil, N., Grunert, F. & von Kleist, S. (1984) *Tumour Biol.* **5,** 339–350.
25. Thompson, J., Pande, H., Paxton, R. J., Shively, L., Padma, A., Simmer, R. L., Todd, C. W., Riggs, A. D. & Shively, J. E., (1987) *Proc. Natl. Acad. Sci. USA*, in press.
26. Laemmli, U. K. (1970) *Nature (London)* **227,** 680–685.
27. Mort, A. J. & Lamport, D. T. A. (1977) *Anal. Biochem.* **82,** 289–309.
28. Tetaert, D., Takahashi, N. & Putnam, F. W. (1982) *Anal. Biochem.* **123,** 430–437.
29. Zimmermann, W., Ortlieb, B., Friedrich, R. & von Kleist, S. (1987) *Proc. Natl. Acad. Sci. USA*, in press.
30. Marshall, R. D. (1972) *Annu. Rev. Biochem.* **41,** 673–702.
31. Robb, R. J., Kutny, R. M., Panico, M., Morris, H. R. & Chowdhry, V. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 6489–6490.
32. Takahashi, N., Yasuda, Y., Kuzuya, M. & Murachi, T. (1969) *J. Biochem.* **66,** 659–667.
33. Lozier, J., Takahashi, N. & Putnam, F. W. (1983) *J. Chromatogr.* **266,** 545–554.
34. Williams, A. F. & Gagnon, J. (1982) *Science* **216,** 696–703.
35. Mostov, K. E., Friedlander, M. & Blobel, G. (1984) *Nature (London)* **308,** 37–43.
36. Eiffert, H., Quentin, E., Decker, J., Hillemeier, S., Hufschmidt, M., Klingmuller, D., Weber, M. H. & Hilschmann, N. (1984) *Hoppe-Seylers Z. Physiol. Chem.* **365,** 1489–1495.
37. Kabat, E. A., Wu, T. T., Bilofsky, H., Reid-Miller, M. & Perry, H. (1983) *Sequences of Proteins of Immunological Interest* (National Institutes of Health, Bethesda, MD).
38. Ishioka, N., Takahashi, N. & Putnam, F. W. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 2363–2367.
39. Hemperly, J. J., Murray, B. A., Edelman, G. M. & Cunningham, B. A. (1986) *Proc. Natl. Acad. Sci. USA* **83,** 3037–3041.