EXON-INTRON ORGANIZATION OF A GENE FOR PREGNANCY-SPECIFIC ß$_1$-GLYCOPROTEIN, A SUBFAMILY MEMBER OF CEA FAMILY: IMPLICATIONS FOR ITS CHARACTERISTIC REPETITIVE DOMAINS AND C-TERMINAL SEQUENCES

Shinzo Oikawa[1], Chikako Inuzuka[1], Goro Kosaki[2] & Hiroshi Nakazato[1]

[1]Suntory Institute for Biomedical Research, Shimamoto-cho,
Mishima-gun, Osaka 618, Japan
[2]Tokyo Metropolitan Komagome Hospital, Bunkyo-ku, Tokyo 113, Japan

Received August 18, 1988

A fragment of human gene for pregnacy-specific ß$_1$-glycoprotein(s), recently identified CEA family member(s), has been cloned. Analyses of nucleotide and deduced amino acid sequences revealed that it carried, from 5' to 3' direction, exons IA, IB, IIA, IIB, C3, C1 and C2, the first four encoding peptides distinct from but highly similar to domains of PSßGs. The lack of consensus 3' splice site sequence ahead of IB indicated that it was an abortive exon, which would explain the peculiar domain construction of PSßGs, *i.e.* N-IA-IIA-IIB-C1, 2 or 3. Apparently, the multiple C-terminal sequences for a PSßG were generated by alternative splicing among C1, C2 and C3 exons. Furthermore, sequences which overlapped partly with Cexons, were found to be similar to parts of 3'-UTR of CEA and NCA, indicating further the close relationship of CEA/NCA and PSßG subfamily genes. © 1988 Academic Press, Inc.

CEA (1) is one of the most widely used human tumor markers although it lacks absolute tumor specificity because of the presence of a number of immunologically closely related glycoprotein antigens, which comprise CEA family.

Recent success in cloning cDNAs and parts of genomic sequences revealed the existence of multiple genes of highly similar sequences for CEA family (2,3,4,5,6,7,& 8). The characteristic domain structures (2,7) are evident for CEA and NCA of which the former is composed of 108-residue N-domain, three repetitive 178-residue domains I, II and III, and 26-residue hydrophobic M-domain, the latter is composed of 108-residue N-domain, 178 residue domain I and 24-residue M-domain. Domains I, II and III are further subdivided into 92-residue A and 86-residue B subdomains (9). Domain N and subdomains A and B respectively, have been shown to belong to Ig superfamily (9), *i.e.* CEA family belongs to Ig superfamily (3,9,10).

Abbreviations: CEA, carcinoembryonic antigen; NCA, nonspecific crossreacting antigen; PSßG, pregnacy-specific ß$_1$-glycoprotein; UTR, untranslated region; -b, -bases.

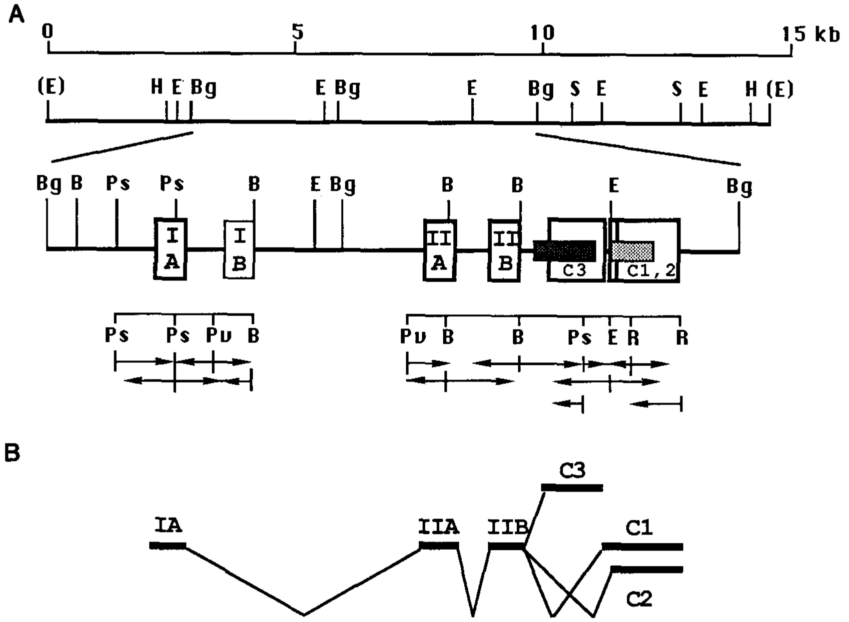

Fig. 1 (A) Sequencing strategy and exon-intron structure of CGM35. The scale is shown at the top in kilobases. Only restriction endonucleases relevant to the present work are shown: B, *Bam*HI; Bg, *Bgl* II; E, *Eco*RI; H, *Hind* III; Ps, *Pst* I; Pv, *Pvu* II; R, *Rsa* I; S, *Sac* I; (E), *Eco*RI linker. Extent and direction of sequencing are shown by arrows. □ , exon, but the sequence shown in thin lines may not be an exon; ■ , ▨ ; sequences similar to parts of CEA and NCA cDNA, respectively. (B) Mode of splicing. Thick horizontal lines denote exons, which correspond to those shown in (A).

exons, four of them being translatable into peptides (Fig. 2) highly similar in size and sequence to subdomain As and Bs of CEA family (Fig. 3). However, it should be noted that 3'-splice site preceding "1B" was not conforming to the consensus sequence, suggesting that "1B" might not be processed into mRNA. Other three exons, two of them almost entirely overlapping except for 86-b, apparently encoded C-terminal sequences of this CEA family member. There was no N-domain coding region within entire length of the DNA insert, for CEA N-domain probe did not hybridize to any of the restriction fragments in Southern blot hybridization analysis.

When peptide sequences encoded by each exon were aligned with those of subdomain As and Bs of the members of CEA family deduced from the cloned cDNAs (Fig. 3), and sequence similarities were calculated (Table 1), high similarity, especially to those of PSßGs was evident. Each peptide was 43 to 63% and more than 82% similar to the corresponding subdomains of CEA and NCA, and PSßG, respectively. Apparently, the present gene was of a member of PSßGsubfamily rather than of CEA/NCA subfamily. Furthermore, it is significantly more related to PSßG16/93 and C/Dthan to E. Interestingly, sequence similarity between As or Bs belonging to different repetitive domains, *i.e.* I and II, of PSßGs was only 44 to 49% while that between As or Bs belonging to the same repetitive domains was 80 to 95%. In contrast, subdomains

Recently, primary structure of precursors to two kinds of PSßGwas deduced from cloned cDNAs (11), and it was found that the PSßGs belonged to CEA family (12), its domain construction and amino acid sequences being highly similar to CEA and NCA. The precursors comprised, consecutively, 143-residue N-terminal domain including a signal peptide, 93-residue domains IA, IIA and 86 or 88-residue domain IIB and lacked the hydrophobic M-domain. Apparently, domains IA and IIA, and IIB, respectively, are homologous to subdomains A and B of CEA or NCA. The characteristics peculiar to the PSßG is the lack of domain B between IA and IIA, in addition to the lack of M-domain and the presence of two kinds of C-terminal sequences. An apparently identical PSßG with different C-terminus and another PSßG, which was composed of domains N, IA and IIB were recently reported (13).

In this report we will describe cloning and nucleotide sequence of a genomic DNA segment containing introns and exons which encoded sequences highly similar to subdomain As and Bs of PSßGs. The possible mechanisms generating the peculiar domain construction and different C-termini of PSßGs will be discussed.

MATERIALS AND METHODS

Human Genomic Library —— A human genomic library which had been prepared from placental DNA using bacteriophage charon 4A vector as described by Lawn *et al.* (14) was kindly provided by Dr. Masabumi Shibuya of Tokyo University.

Screening, Subcloning and DNA Sequence Determination —— Approximately one million clones were plated and screened using two ^{32}P-labeled *Pvu* II fragments of CEA cDNA which corresponded to the repetitive domains of CEA (2). Two positive clones were obtained and the one termed λCGM35, was characterized by restriction endonuclease analysis (Fig. 1A). Before subcloning, Southern blot hybridization analysis (15) was performed to locate sequences related to those of cDNA for CEA and NCA using ^{32}P-labeled probes described below. Only the fragments containing sequences related to those were recloned into M13mp18 or M13mp19 (16) and sequenced by the chain termination method (17).

Probes —— *Pvu* II-*Acc* I and *Pvu* II-*Pvu* II fragments of pCEA55-2 clone (2) were for N-domain and repetitive domains, respectively. *Rsa* I-*Eco*R I fragment of λKr40 (2) and *Eco*R I-*Hin*d III fragment of NCA15 (7) were for the sequences related to 3'-UTR of CEA and NCA, respectively. Probes were ^{32}P-labelled by the nick translation method (18).

RESULTS AND DISCUSSION

Fig. 1 is a schematic representation of the exon-intron organization of the human genomic clone λCGM35, also depicted is the subcloning and sequencing strategy of the DNA fragment. Amino acid sequences translatable from the three frames of the DNA fragments were deduced and compared with those of CEA and NCA to identify exons. The consensus sequences at 5'- and 3'-splice sites (19, 20) were also referred to in order to locate the exon-intron boundaries. As is shown in Fig. 1, there were seven putative

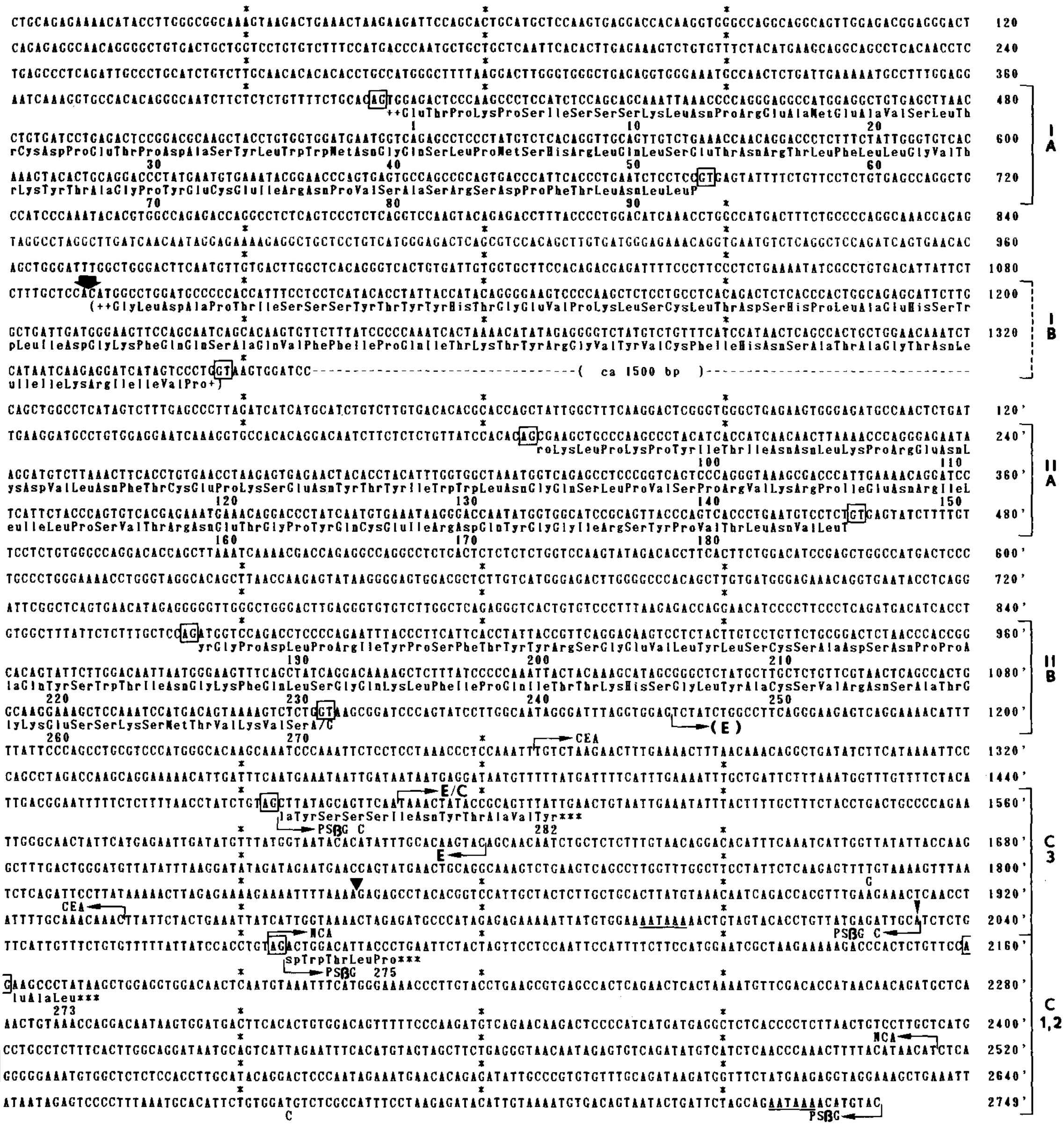

Fig. 2 Nucleotide sequence and deduced amino acid sequence of CGM35. Nucleotides are numbered beginning from *Pst* I site (1-1358) and *Pvu* II site (1'-2749') of the sequenced fragments (Fig. 1A). Sequences containing exons are shown on the right by]. Consensus splice site sequences are boxed, with ⬇ being non-conforming sequence. IB which may be an abortive exon is shown by broken], and amino acids deduced are parenthesized. Amino acids are numbered beginning from the first residue of exon IA, throgh the C-terminal residues of exon C1, C2 or C3. ┌→ , ←┘; start and end of the sequence resembling to that of a part of the cDNA indicated; PSßG, PSßG16/93 (9); E, C, PSßGE and C(13), respectively; (E), not highly but moderately similar to E, ▼ , *Alu* family insert seen in the case of CEA; ▼ , poly A addition site. Putative poly A signals are underlined. Nucleotides 1788' and 2676' are ambiguous.

```
                        *         *         *         *         *         *
CEA         I   ELPKPSISSNNSKPVEDKDAVAFTCEPETQDATYLWWVNNQSLPVSPRLQLSNGNRTLTL
CEA        II   -P---F-T----N----E----L-----I-NT---------------------D------
CEA       III   ----------------------------A-NT-------G--------------------
NCA         I   ------------N---------------V-NT-------G---------------M----
CGM35       I   -T-------SKLN-A-AME--SL--D---P--S----M-G----M-H-----ET----F-
CGM35      II  PK----Y-TI--L--A-N--VLN-----KSENY--I--L-G--------VKRPIE--I-I-
PSβG        I  L-T-------S-LN-A-TME--SL--D---P--S----M-G----MTHS-K--ET----F-
PSβG       II  PK----Y-TI--LN-A-N--VLN-----KSENY--I--L-G--------VKRPIE--I-I-
PSβG E      I  L-T-------S-LN-A-AMET-IL--D---P-TS-Q--M-G----MTH-F---ET----F-
CONSENSUS         PKP I      P E       TC P      Y WW N QSLP          N  L L

                        *         *         * ↓       *         *         *
CEA         I   FNVTRNDTASYKCETQNPVSARRSDSVILNVLYGPDAPTISPLNTSYRSGENLNLSCHAA
CEA        II   LS-----VGP-E-GI--EL-VDH--P----------D-----SY-Y--P-V--S------
CEA       III   -------ARA-V-GI--S---N---P-T-D------T-I---PDS--L--A-------S-
NCA         I   LS-K---AG--E--I---A--N---P-T--------V-----SKAN--P-----------
CGM35       I   LG--KYTAGP-E--IR-----S---PFT--L-H-L------SSY-Y-HT--VPK---LTD
CGM35      II   PS----E-GP-Q--IRDQYGGI--YP-T--------L-A-Y-SF-Y-----V-Y---S-D
PSβG        I   LG--KYTAGP-E--IR-----S---P-T--L-
PSβG       II   PS----E-GP-Q--IRDRYGGV---P-T--------L-A-Y-SF-Y-----V-Y---S-D
PSβG E      I   -G--KYTAGP-E--IR-SG--S---P-T--L-
PSβG E     II                                   H---L-A-H-SY-N----D--Y---F-N
CONSENSUS         V       Y C          S    L  L G D P I      Y  G    LSC

                        *         *         *         *         *         *
CEA         I   SNPPAQYSWFVNGTFQQSTQELFIPNITVNNSGSYTCQAHNSDTGLNRTTVTTITVYA
CEA        II   ---------LID-NI--H------S---EK---L-----N--AS-HS----K----S-
CEA       III   ---SP----RI--IP--H--V---AK--P--N-T-A-FVS-LA--A-NSI-KS---S-
NCA         I   ----------I------------------------M------A---------M---SG
CGM35       I   -H-L-EH--LID-K----A-VF---Q--KTYR-V-V-FI---A-AGTNLIIKR-I-P-
CGM35      II   ---------TI--K--L-G-K----Q--TKH--L-A-SVR--A--KESSKSM-VK-SD/E
                                                                          /A
PSβG       II   ---------TI-EK--LPG-K---RH--TKH--L-V-SVR--A--KESSKSM-VE-SD/E
                                                                          /A
PSβG E     II   ---------TI--K----G-N----Q--TKH--L-V-SVR--A--EESSTSL-VK-S-
CONSENSUS       S P    SW      Q   Q  FI  IT    G Y C   N              V
```

Fig. 3 Comparison of the domains of CGM35 with those of CEA family. Only amino acids different from those of CEA I are shown in single notation. Dashes mean identity. Arrow indicates boundary between A and B subdomains. Underlined sequence is probably not expressed in the protein. The last residues of CGM35 II and PSβG are D, E or A due to the presence of three kinds of C-terminal coding sequences. Possible N-glycosylation sites are boxed. PSβGC/D are different from PSβG only at amino acid 82 in II. PSβG stands for PSβG16/93.

belonging to different repetitive domains of CEA/NCA subfamily, were more than 72% similar with the exception of CEA IIIB which were about 60% similar to CEA IB, IIB and NCA IB (Fig. 3). These results clearly indicate that divergence among repetitive domains are greater in PSβG subfamily than in CEA/NCA subfamily. In view of this, CGM35 IB had no counterpart among PSβGs whose primary structure had been deduced from the cloned cDNAs. It was noted that there was no putative N-glycosylation site in subdomain Bs of PSβG, albeit the presence of several of ones in other subdomains (Fig. 3).

Watanabe and Chou isolated two cDNA clones, PSβG16 and 93, encoding human PSβGs of 417 and 419 amino acids, respectively (11). The sequenced portions of these cDNAs were identical with the exception that PSβG93 contained an additional 86-b at the end of the common 3'-coding region. This resulted in the generation of two C-terminal sequences after common 414 amino acids, which were EAL and DWTVP, for PSβG16 and 93, respectively. More recently, three PSβGs deduced from cloned cDNAs were reported (13). PSβGD was virtually identical to PSβG93 with only three amino acids

Table I. Amino acid similarities between subdomain As and Bs of CGM 35, PSßG 16/93(12), PSßG E(13), CEA(2) and NCA(7). PSßG stands for PSßG 16/93 which is virtually identical to PSßG C/D (13). The number of matches is expressed as per cent of the similarity length.

A Subdomain

	CEA		NCA	CGM35		PSßG		PSßG E
	II	III	I	I	II	I	II	I
CEA I	73.9	83.7	81.5	59.8	53.3	58.7	53.3	55.4
CEA II		76.1	79.3	57.6	53.3	56.5	55.4	55.4
CEA III			85.9	58.7	55.4	57.6	55.4	58.7
NCA I				63.0	56.5	62.0	58.7	59.8
CGM35 I					45.7	93.5	47.8	88.0
CGM35II						46.7	95.7	46.7
PSßG I							48.9	88.0
PSßG II								48.9

B Subdomain

	CEA		NCA	CGM35		PSßG	PSßG E
	II	III	I	I	II	II	II
CEA I	72.1	60.5	86.0	46.5	57.0	53.5	59.3
CEA II		58.1	73.3	47.7	57.0	55.8	60.5
CEA III			60.5	43.0	50.0	47.7	51.2
NCA I				46.5	58.1	54.7	60.5
CGM35 I					47.7	44.2	48.8
CGM35II						93.0	82.6
PSßG II							77.9

differences. PSßGCshared 414 N-terminal amino acids with PSßGD but followed by an entirely different C-terminal sequence of 14 amino acids, AYSSSINYTSGNRN. PSßGEwas composed of N-domain and subdomains IA and IIB which were distinct from but highly similar to corresponding domains of other PSßGs.

Interestingly, three exons of CGM35 contained sequences which could generate three kinds of mRNA by differential splicing (Fig. 1B). As is shown in Fig. 2, C3 exon at nucleotides 1475'-2034' encoded 12-residue C-terminal sequence, whose first 9 residues were identical to those of PSßGC. C1 and C2 exons starting at nucleotides 2076' and 2162', respectively, were identical except for the extra 86-nucleotides at 5'-terminus of C1 exon. C1 and C2 exons, respectively, would encode C-terminal sequences, DWTLP and EAL, which were virtually identical to those of PSßG93/D(11, 13) and 16 (11), respectively. In addition to the C-terminal amino acid sequence similarities, C1, C2 and C3 exons are highly similar, *i.e.* >93%, in nucleotide sequence to 3'-end of PSßG93/D, 16 and C, respectively (Fig. 4). The sequence related to 3'-end of PSßGEwas also found but

(A)

```
                                       *         *         *         *         *
CGM35                            AGTCTATCTGGCCTTCAGGGAAGAGTCAGGAAAACATTTTTATTCCC
PSβG E                           CT...TC..CT....A.TCC..C.TAGCA.CTGTG..G.CAT...TG
                   *         *         *         *         *         *         *
CGM35     AGCCTGCGTCCCATGGGCACAAG-CAAATCCCAAATTCTCCTCCTA-AACCCTCCAA---ATTTGTCTAA
PSβG E    TATT.CA.GAAG.CT....GG..ATTT..GGA..GG....T.A.A.GG..T..TG..TAC.AGCTC..G.
CEA                                                                .....G.
                   *         *         *         *         *         *         *
CGM35     GAACTTTGAAAACTTTAACAAACAGGCTGATATC-TTCATAA----AATTCCCAGCCTAGACCAAGCAG-
PSβG E    T...-..C..G.-.CAT..C.CTG.A..A.G.A.T....A..TTTT...GAA...G..GATA.CTT..T-
CEA       ...T..CC..........TG...TAA....C.G.-.....G.----..C.GT.CA..A...T.......A
                   *         *         *         *         *         *         *
CGM35     GAAAAAC-ATTGATTTCAATGAAATAATTGATAATAAT-GAGGATAATGTTTTTATGA-TTTTCATTTGA
PSβG E    ....TT.A.GAC.AAGA.GAA.....C.CA..GT..T.G..CT.A.TAA.CAAA.G..-.AA.G....C.
CEA       .....TA-...A......TG.G.C...A...-.C....-.........A....C..A.T....T......
                   *         *         *         *         *         *         *
CGM35     AAATTTGCTGATTCTTTAAATGGTTTGTTTTCTACA-TTGACGGAA-TTTTTCTCTTTTAACCTATCTGT
PSβG E    T.....T..-...-.GA.....TGC..A..CT.GG.-A..-TTTC.-..C.C.AGA....TGAACAT.T.
CEA       ..T...................TC......C.C.G.T..C.G.A..C.....T........G.....CAC
                   *         *         *         *         *         *         *
CGM35     AGCTTATAGCAGTTCAATAAACTATACCGCAGTTTATTGAACTGTAATTGAAATATTTACTTTTGCTTTC
PSβG E    TT...-G....-A.TGG....G...-----.C...TG.A...AAA........C....C.........C.
CEA       ......C....A...TG....A...-----.C...TG.....AAA......G.C......A...T..CC.
PSβG C      .......................-----.C..CTGG....C..........C................
                   *         *         *         *         *         *         *
CGM35     TACCTGACTGCCCCAGAATTGGGCAACTATTCATGAGAATTGATATGTTTATGGTAATACACATATTTGC
PSβG E    ..T....G...............A.T...........T...C.................A.GC.......
CEA       ..TG..GTC..T.....C.....A............AT...T....-.G..........T.GT..-....
PSβG C     ......................................................................
                   *         *         *         *         *         *         *
CGM35     ACAAGTACAGCAACAATCTGCTCTCTTTGTAACAGGACACATTTCAAATCATTGGTTATATTACCAAGGC
PSβG E    ......T.[Poly A]                   ↓↓
CEA       ......T..AT..A..........T.G.A.G....A.T......G...A...................A.
PSβG C    ..........T...........T...........T.........G......................T..
                   *         *         *         *         *         *         *
CGM35     TTTGACTGGGATGTTATATTTAAGGATATAGA----TA--GAATGAACCAGTATGAACTGCAGGCAAAGT
CEA       .......A.A....CG.....G........A.CCCA..GGT...A....CAC.G.T...A..AA......
PSβG C    .....T.C..........-....AA.C.....----..--..........A...................
                   *         *         *         * (G)     *         *         *
CGM35     CTGAAGTCAGCCTTGGTTTGGCTTCCTATTCTCAA-GAGTTTTGTAAAAGTTTAATCTCAGATTCCTTAT
CEA       ............................G.G....TT.AAC..C..............G...........
PSβG C     ...................................-...G.....G..GA....................
                   *         *         *         *         *         *         *
CGM35     AAAAACTTAGAGAAAAGAAAATTTTAAAA[---Gap---]GAGAGCCTACACGGTCCATTGCTACTCTTG
CEA       ........CC..C....-C..C.......[Alu 303-b]A.A..T...TGT....AG.CA.........
PSβG C    ..........................G..[---Gap---]..C........T..................
                   *         *         *         *         *         *         *
CGM35     CTGCACTTATGTAAACAATCAGACCACGTTTGAAGAAACTCAACCTATTTTGCAAACAAACTTATTCTAC
CEA       .....G.....A...G...G..G...A..C...T....A.A...T.......A....[Poly A]
PSβG C    ...........................A.............C............................
                   *         *         *         *         *         *         *
CGM35     TGAAATTATCATTGGTAAAACTAGAGATGCCCATAGAGAGAAAAATTATGTGGAAAATAAAAACTGTAGT
PSβG C    ....................G.................G...............................
                   *         *         *         *         *         *         *
CGM35     ACACCTGTTATGAGATTGC
PSβG C    .T.............C...[Poly A]
```

Fig. 4 Nucleotide alignments of similarities in the 3'-end between CGM 35 and members of CEA family. (A) CGM 35 (nucleotides 1162'-2033'), PSßGE(982-1440) (13), CEA (2499-3468) ((2), nucleotides 2930-3468 are our unpublished data) and PSßGC(1244-1796) (13) are aligned. (B) CGM 35 (2074'-2749'), PSßG16/93 (1310-1906, with additional 86-b seen in PSßG93, which are underlined) (11) and NCA (1074-1520) (7) are aligned. PSßGD(1242-1928) is almost identical to PSßG16/93 except for the portion shown in last line. Identical residues and deletions are shown by dots and dashes, respectively. ↓ , poly A addition site. Poly A signals are underlined.

```
(B)
                   *         *         *         *         *         *         *
CGM35     AGACTGGACATTACCCTGAATTCTACTAGTTCCTCCAATTCCATTTTCTTCCATGGAATCGCTAAGAAAA
PSβG16    T.........G.T.................A..................C..........A......GC.
NCA       ...T.....CAG............T....C.........C.......A.C........C.A....A..C.
                   *         *         *         *         *         *         *
CGM35     AGACCCACTCTGTTCCAGAAGCCCTATAAGCTGGAGGTGGACAACTCAATGTAAATTTCATGGGAAAACC
PSβG16    ......................................................................
NCA       ..GT.TG.....C...T...........T.......A..............A......A.A.........
                   *         *         *         *         *         *         *
CGM35     CTTGTACCTGAAGCGTGAGCCACTCAGAACTCACTAAAATGTTCGACACCATAACAACAGATGCTCAAAC
PSβG16    ..............A...................C.....A.............................
NCA       ..CAGG.....G.T...T..........---G...----.------....-....T.G...CAG......
                   *         *         *         *         *         *         *
CGM35     TGTAAACCAGGACAATAAGTGGATGACTTCACACTGTGGACAGTTTTTCCCAAGATGTCAGAACAAGACT
PSβG16    ...............C......................................................
NCA       ..C......T.GTG.G..A.T..C...........A.......C................A.........
                   *         *         *         *         *         *         *
CGM35     CCCCATCATGATGAGGCTCTCACCCC-TCTTAAC-TGTCCTTGCTCATGCCTGCCTCTTTCACTTGGCAG
PSβG16    .....................C....-.......-...................................
NCA       ..T.........A.......T.....C.T....TT..........T...............G........
                   *         *         *         *         *         *         *
CGM35     GATAATGCAGTCATTAGAATTTCACATGTAGTAGCTTCTGAGGGTAAC--AATAGAGTGTCAGATATGTC
PSβG16    ..........................................A.....--..C.................
NCA       ...G....T........T........A.A.........A.........TT..C............C.A..
                   *         *         *         *         *         *         *
CGM35     ATCTCAA-CCCAAACTTTTACATAACATCTCAGGGGGAAATGTGGCTCTCTCCACCTTGCATACAGGACT
PSβG16    .......C.T.........................................................G..
NCA       T.G....T.....CG..........A..AAGAGATCCTTTA
                   *         *         *         *         *         *         *
CGM35     CCCAATAGAAATGAACACAGAGATATTGCCCGTGTGTTTGCAGATAAGATGGTTTCTATGAAGAGGTAGG
PSβG16    ..............................T.............G..........G........C.....
                   *         *         *         *         *         *         *
CGM35     AAAGCTGAAATTATAATAGAGTCCCCTTTAAATGCACATTCTGTGGATGTCTC--GCCATTTCCTAAGAG
PSβG16    .................................C......G........G...TT...G...........
                   *         *         *         *         *         *
CGM35     ATACATTGTAAAATGTGACAGTAATACTGATTCTAGCAGAATAAAACATGTAC
PSβG16    .............C..........G..--........................CACCTCCC[EcoR I]
PSβG D    .............C..........G..--.... .......................T...TTGCT[Poly A]
```

Figure 4 - Continued.

it was separable into two portions, the 5'- and 3'-end sequences comprising nucleotides 1164'-1489' and 1490'-1620' were 47.6% and 80.2% similar to corresponding PSβGE sequences, respectively. The 3'-end sequence was where similarity between 3'-end sequences of PSβGC and PSβGE was found (13). Although 5'-end sequence could encode 35-residue peptide if it was processed into mRNA, the significance of these findings is not clear at this time.

In addition to the similarities found between 3'-end sequences of members of PSβG subfamily, CGM35 contained sequences highly similar, *i.e.* >76%, to 3'-UTR of CEA and NCA. The CEA and NCA like sequences overlapped partly with exons C3 and C1,2, respectively (Fig. 2). The corresponding CEA sequence started about 40-residue downstream of the first *Alu* sequence, extended beyond the poly A addition site of the shorter cDNA (2) and ended at the poly A addition site of the longer cDNA (Fig. 4A). The second *Alu* sequence of 303-b found in the longer cDNA (details will be published elsewhere) was missing in the CGM35 sequence (Fig. 2 & 4A). The NCA sequence started from 40-b downstream of the stop codon, *i.e.* where similarity between NCA and CEA cDNA ceased (7). The sequence similarity of parts of 3'-UTR of cDNAs of PSβGD and NCA was already noted (13). In spite of these similarities, sequences similar to those corresponding to M-domains of CEA or NCA were not found. These results, along with the finding that none of the known PSßGs including CGM 35 has C-terminal hydrophobic M-domain, might indicate that M-domains of CEA/NCA subfamily are encoded by separate exons.

In conclusion, CGM 35 clone carried a sequence which contained most of a gene for a new member of PSßG subfamily within CEA family. As is summarized in Fig. 1A, the N-domain truncated sequence consisted of exons IA, IB, IIA, IIB, C3, C1 and C2, from 5' to 3' direction. As discussed above, exon IB was apparently an abortive exon which would not be processed into mRNA. Alternative splicing will generate at least three kinds of mRNA which encode PSßGs having three different C-terminal sequences. Thus, at least three PSßGs, (N)-IA-IIA-IIB-C1, (N)-IA-IIA-IIB-C2 and (N)-IA-IIA-IIB-C3, which are distinct from but highly similar to PSßG93/D, 16 and C, respectively, will be produced. In addition, it is possible that the fourth PSßG having C-terminal sequence derived from the E-like sequence would be found. Another implication of the present findings is that, PSßGE(13) having N-IA-IIB-CE construction might be encoded by a gene having two consecutive abortive exons, namely "IB" and "IIA", although other mechanisms such as alternative splicing can not be excluded.

Finally, considering the highly conserved domain structures among CEA family members, it is conceivable that genes for the members are similary constructed, *i.e.* each domain and subdomain are encoded by separate exons like in immunoglobulin and T cell receptors, introns between A and B are rather short and those between N and A, and B and A are rather long.

REFERENCES

1. Gold, P. & Freedman, S. O. (1965) J. Exp. Med. 121, 439-462
2. Oikawa, S. Nakazato, H, & Kosaki, G. (1987) Biochem. Biophys. Res. Commun. 142, 511-518
3. Zimmermann, W., Ortlieb, B., Friedrich, R. & von Kleist, S. (1987) Proc. Natl. Acad. Sci. USA. 84, 2960-2964
4. Beauchemin, N., Benchimol, S., Cournoyer, D., Fuchs, A. & Stanners, C. P. (1987) Mol. Cell. Biol. 7, 3221-3230
5. Thomson, J. A., 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. 84, 2965-2969
6. Oikawa, S., Kosaki, G. & Nakazato, H. (1987) Biochem. Biophys. Res. Commun. 146, 464-469
7. Tawaragi, Y., Oikawa, S., Matsuoka, Y., Kosaki, G. & Nakazato, H. (1988) Biochem. Biophys. Res. Commun. 150, 89-96
8. Neumaier, M., Zimmermann, W., Shively, L., Hinoda, Y., Riggs, A. D. & Shively, J. E. (1988) J. Biol. Chem. 263, 3202-3207
9. Oikawa, S., Imajo, S., Noguchi, T., Kosaki, G. & Nakazato, H (1987) Biochem. Biophys. Res. Commun. 144, 634-642

10. Paxton, R. J., Mooser, G., Pande, H., Lee, T. D., & Shively, J. E. (1987) Proc. Natl. Acad. Sci. USA. 84, 920-924

11. Watanabe, S. & Chou, J. Y. (1988) J. Biol. Chem. 263, 2049-2054

12. Watanabe, S. & Chou, J. Y. (1988) Biochem. Biophys. Res. Commun. 152, 762-768

13. Streydio, C., Lacka, K., Swillens, S. & Vassart, G. (1988) Biochem. Biophys. Res. Commun. 154, 130-137

14. Lawn, R. M., Fritsch, E. F., Parker, R. C., Blake, G. & Maniatis, T. (1978) Cell 15, 1157-1174.

15. Southern, E. M. (1975) J. Mol. Biol. 98, 503-517

16. Yanisch-Perron, C., Vieira, J. and Messing, J. (1985) Gene, 33, 103-119

17. Sanger, F., Nicklen, S. & Coulsen, A. R. (1975) Proc. Natl. Acad. Sci. USA, 72, 3961-3965

18. Rigby, P. W. J., Dieckmann, M., Rhodes, C and Berg, P. (1977) J. Mol. Biol. 113, 237-251

19. Chambon, P. & Breathnach, R. (1981) Ann. Rev. Biochem. 50, 349-383

20. Shapiro, M. B. & Senapathy, P. (1987) Nucl. Acids Res. 15, 7155-7174