J Mol Evol (1989) 29:126–134

Journal of
Molecular Evolution

Intra- and Interspecies Analyses of the Carcinoembryonic Antigen (CEA) Gene Family Reveal Independent Evolution in Primates and Rodents

Fritz Rudert, Wolfgang Zimmermann, and John A. Thompson

Institut für Immunbiologie, Universität Freiburg, Stefan-Meier-Straße 8, D-7800 Freiburg, FRG

Summary. Various rodent and primate DNAs exhibit a stronger intra- than interspecies cross-hybridization with probes derived from the N-terminal domain exons of human and rat carcinoembryonic antigen (CEA)-like genes. Southern analyses also reveal that the human and rat CEA gene families are of similar complexity. We counted at least 10 different genes per human haploid genome. In the rat, approximately seven to nine different N-terminal domain exons that presumably represent different genes appear to be present. We were able to assign the corresponding genomic restriction endonuclease fragments to already isolated CEA gene family members of both human and rat. Highly similar subgroups, as found within the human CEA gene family, seem to be absent from the rat genome. Hybridization with an intron probe from the human nonspecific cross-reacting antigen (NCA) gene and analysis of DNA sequence data indicate the conservation of noncoding regions among CEA-like genes within primates, implicating that whole gene units may have been duplicated. With the help of a computer program and by calculating the rate of synonymous substitutions, evolutionary trees have been derived. From this, we propose that an independent parallel evolution, leading to different CEA gene families, must have taken place in, at least, the primate and rodent orders.

Key words: Carcinoembryonic antigen — Evolution — Gene family — Human — Rat — Synonymous substitutions — Silent molecular clock — Evolutionary trees

Offprint requests to: J.A. Thompson

Introduction

Since its discovery as an oncodevelopmentally regulated glycoprotein (Gold and Freedman 1965a,b), carcinoembryonic antigen (CEA) has become one of the most widely used human tumor markers. Immunological characterization and isolation of various cross-reacting antigens already indicated the existence of a CEA gene family (reviewed in Thompson and Zimmermann 1988). These preliminary data could be confirmed at the protein (Engvall et al. 1978; Kessler et al. 1978; Paxton et al. 1987) and DNA levels (Oikawa et al. 1987c; Thompson et al. 1987; Neumaier et al. 1988). Southern analyses indicated the existence of 9–11 genes in humans (Thompson et al. 1987). Sequencing of genomic and cDNA clones has revealed the complete primary protein structure of CEA (Beauchemin et al. 1987; Oikawa et al. 1987c; Zimmermann et al. 1987), NCA (Oikawa et al. 1987b; Thompson et al. 1987; Neumaier et al. 1988; Tawaragi et al. 1988), and a pregnancy-specific β_1 glycoprotein (PSβG) (Watanabe and Chou 1988a). The highly conserved domains shared by each are a 34-amino-acid leader peptide, a 108–110-amino-acid N-terminal domain, a 178–180-amino-acid repeating unit of which three copies are present in CEA, whereas only one and a half can be found in PSβG and one in NCA, and a 26-amino-acid carboxyl region (CEA) that is 2 amino acids shorter in NCA and degenerate in PSβG. Closer analysis of the deduced primary and secondary structure showed that these proteins can be grouped within the immunoglobulin superfamily (Oikawa et al. 1987a; Williams 1987; Neumaier et al. 1988). Despite the high conservation of the so far identified CEA gene family members, immunological studies were not able to identify unequivocally their coun-

terparts in nonprimate mammals or birds (Wahren et al. 1983). Up until recently, CEA-related molecules could only be immunologically detected in higher primates (Haagensen et al. 1982; Jantscheff et al. 1986). However, we have succeeded in the isolation of CEA-related genes from a rat genomic library (Kodelja et al. 1989). The available DNA sequence data do not allow the direct assignment of analogous counterparts between human and rat CEA-like genes. Therefore, it appeared necessary to analyze total genomic DNA from these species for the existence of unidentified genes that might have a higher degree of interspecies similarity than those already isolated. The rat genomic clones and the parallel isolation of other members of the human CEA gene family (Thompson et al. 1989) provided us with suitable restriction fragments for hybridization with total genomic DNA from a number of animal species in order to help clarify this as well as the general evolutionary pathways of the CEA gene family.

Here, we present the data of extensive inter- and intraspecies DNA/DNA hybridization analyses together with comparative analyses of DNA sequence data from the human and rat CEA gene families that have led us to a hypothetical model for the evolution of these genes.

Materials and Methods

Species Examined and Origins of Genomic DNA. High molecular weight DNA was prepared from *Rattus norvegicus* (rat) liver, *Mus musculus* (mouse) liver, *Tupaia belangeri* (tree shrew) peripheral leukocytes (whole blood was a kind gift from D. von Holst, Bayreuth), *Pipistrellus pipistrellus* (bat) muscle, liver, and heart, *Cercopithecus aethiops* (marmoset) using the transformed kidney cell line Tc-7 derivative of CV-1, *Pan troglodytes* (chimpanzee) peripheral leukocytes (whole blood purchased from TNO Primate Center, Netherlands), and *Homo sapiens* (human) peripheral leukocytes.

Isolation and Southern Analysis of Genomic DNA. DNA was isolated essentially as described before (Zimmermann et al. 1988). Tissues were first pulverized in liquid nitrogen and then subjected to proteinase K digestion, followed by removal of excess protein and undissolved material by centrifugation in a CsCl gradient (ρ = 1.695 g/cm^3) at 10,000 *g* and 20°C for 30 min.

Hybridization was carried out as previously described (Zimmermann et al. 1988) with minor modifications. The hybridization temperature was lowered to 37°C in the presence of 40% formamide and the low stringency wash was generally performed at 60°C, in 2 × SSPE (1 × SSPE = 0.18 M NaCl, 0.01 M NaH_2PO_4, 0.001 M EDTA, pH 7.4).

Calculation of the Number of Nucleotide Substitutions. The rates of synonymous (nonsynonymous) substitutions were calculated separately in coding regions. The calculation of these K_S (K_A) values was done using the computer program LWL85 for nuclear genes (Li et al. 1985).

The theoretical basis for the calculation of the rates of synonymous (nonsynonymous) substitutions is described in Li et al. (1985). The rates of synonymous substitutions for noncoding regions (K_N) were calculated simply as the number of substitutions divided by the number of sites compared and were corrected for potential multiple substitutions (Miyata et al. 1980; Sakoyama et al. 1987) using:

$$K_N^C = -\tfrac{3}{4} \ln(1 - \tfrac{4}{3}K_N) \quad (1)$$

Gaps were excluded from the comparisons.

Evolutionary Trees. The creation of the cladograms was achieved by independently using the evolutionary inference package PHYLIP (Felsenstein 1985) for the direct comparison of whole nucleotide sequences and by computing distance matrices with the UPG (unweighted pair group) method [for the program and more detailed information see Li (1981)] using the K_S values. For the computation with PHYLIP, gaps were excluded from the comparisons in a way that all sequences were aligned, and by removing all insertions and deletions a minimum consensus was achieved. For optimal sequence alignment, the computer program ALIGN by R. Friedrich, Gießen, and M. Trippel, Freiburg (unpublished), was used, which is based on an algorithm by Needleman and Wunsch (1970).

Results

Southern Analyses of Primate and Nonprimate Genomic DNAs

In order to examine the degree of interspecies sequence conservation and the extent of gene amplification, we used restriction endonuclease fragments from the N-terminal domain exons of two human (Thompson et al. 1989) and four rat genomic clones (Kodelja et al. 1989) as well as one human cDNA clone (Neumaier et al. 1988) to probe total genomic DNAs from various species. The exact location of these probes is depicted in Fig. 1.

From Fig. 1a–c it can be seen that under nonstringent conditions, the human-derived probes hybridize strongly with multiple DNA fragments within the primate order. The rat probes hybridize preferentially with rodent DNAs (Fig. 1d–f). However, only weak or marginal hybridization signals were obtained with rat probes versus primate DNAs or human probes with rodent DNAs, respectively. Assuming that the N-terminal domain is present only once per gene, as appears to be the case in humans, and that unknown exons do not contain a recognition sequence for the restriction endonucleases used for the digestion of the genomic DNAs, we estimate that approximately seven to nine CEA-like genes exist in the rat by counting the observed cross-hybridizing DNA fragments identified with the different rat probes. For this calculation, it has been taken into account that the N-terminal domain exon for rnCGM2 has a recognition sequence for SstI (Kodelja et al. 1989).

By combining the results obtained from hybridization experiments under stringent conditions using probes from various genes, and taking into ac-

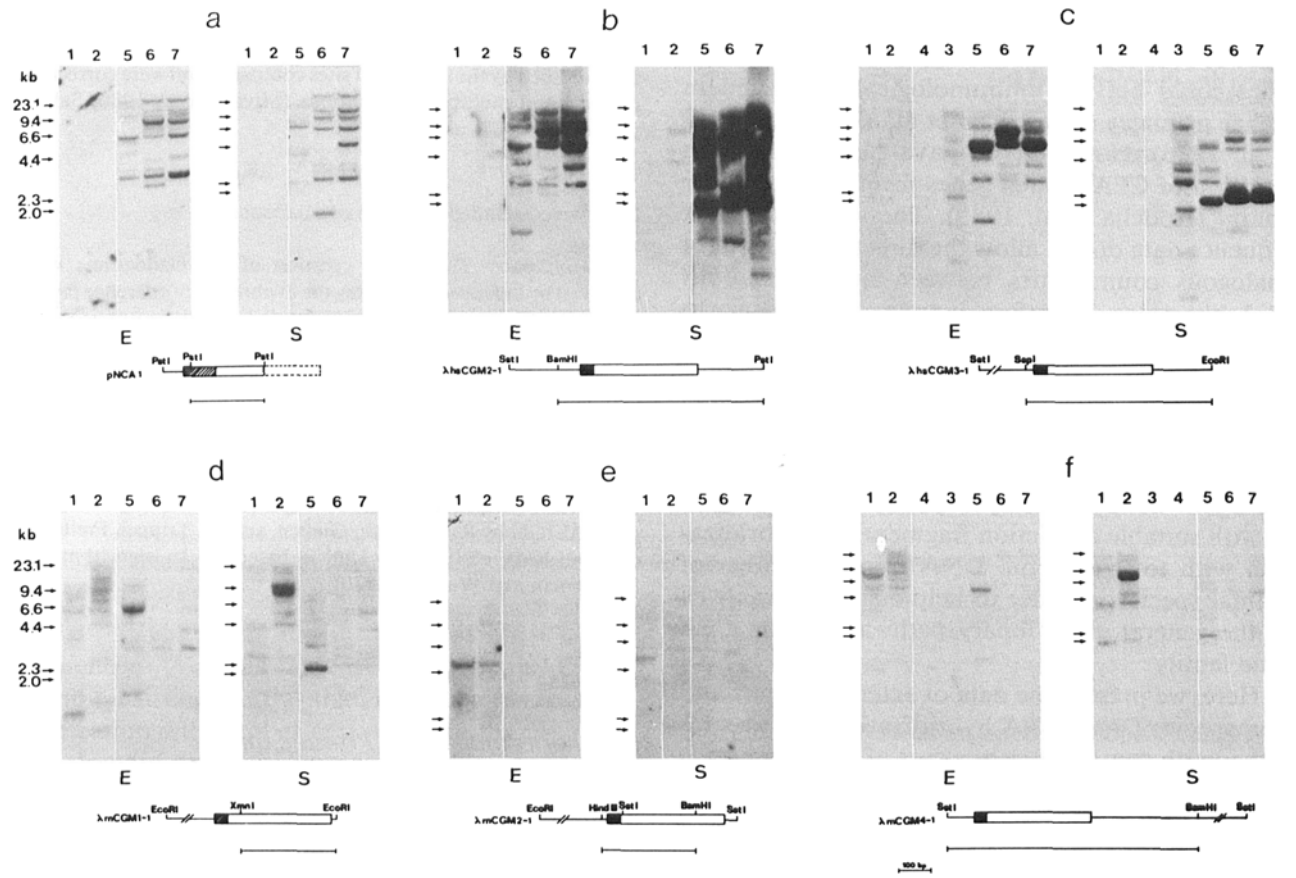

Fig. 1. Hybridization of total genomic DNA from several mammalian species using probes containing different N-terminal domain exon regions of human (a, b, c) and rat (d, e, f) CEA gene family members. The filters were washed at 60°C, 2× SSPE (b–f) or at 65°C, 2× SSPE (a). Lane 1, *Rattus norvegicus*; lane 2, *Mus musculus*; lane 3, *Pipistrellus pipistrellus*; lane 4, *Tupaia belangeri*; lane 5, *Cercopithecus aethiops*; lane 6, *Pan troglodytes*; lane 7, *Homo sapiens* (2.5 μg DNA per lane). The locations of the restriction endonuclease fragments used as probes for hybridization are depicted below the autoradiographs. References for probe sequences: a (Neumaier et al. 1988); b, c (Thompson et al. 1989); d–f (Kodelja et al. 1989). The probes in a–c represent one member of each known subgroup of the human CEA gene family. The blot hybridized with the probe from the N-terminal domain exon of hsCGM2 (b) is overexposed to visualize the weak hybridization signals in rodent DNA. E, EcoRI; S, SstI digests of the genomic DNAs. Arrows indicate the positions of λ/HindIII restriction endonuclease fragments.

count restriction endonuclease fragment lengths as determined by sequencing human and rat CEA-like genes (Thompson et al. 1987; Kodelja et al. 1989; Thompson et al. 1989), we have been able to assign different DNA fragments to specific N-terminal domain exons (Fig. 2). It can be seen from the unidentified restriction endonuclease fragments that more N-terminal domains exist than have been isolated so far in both human and rat.

The human probes derived from NCA, hsCGM2, and hsCGM3 gave multiple distinct hybridization signals with additional mammalian species tested (bat: Fig. 1c, lane 3; mole: data not shown). Among the rat genes, only the probe derived from rnCGM3 showed an identifiable hybridization signal with DNA from mole and bat, whereas the other probes detected weakly hybridizing fragments only in bat genomic DNA (Fig. 1f and data not shown). The tree shrew did not reveal a clear hybridization signal with N-terminal domain exon probes derived from either human or rat CEA-like genes (Fig. 1c and f). A control hybridization using a probe from the 3′ end of a β-tubulin cDNA clone from *Drosophila melanogaster* (Bialojan et al. 1984) that is known to be highly conserved throughout mammals revealed no significant differences between species in the amounts of DNA on the filters (Fig. 3).

Under high stringency conditions, the rat and human probes showed differences in their degrees of intraspecies sequence conservation. For example, using the human N-terminal domain exon probe from NCA (Fig. 1a), apart from its corresponding genomic fragment the size of which is known from a genomic NCA clone (Thompson et al. 1987), one additional EcoRI fragment and two additional SstI fragments still hybridize in humans even under stringent conditions (data not shown). This implies the existence of two additional genes that are closely related to the NCA gene. In the rat, only single DNA fragments could be detected under high stringency

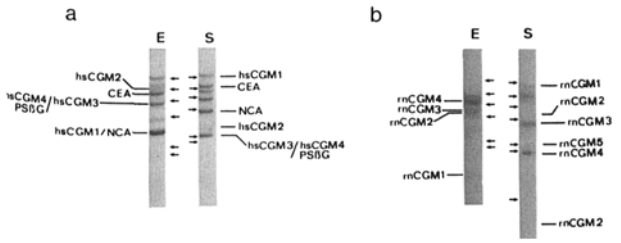

Fig. 2. Assignment of genomic restriction endonuclease fragments comprising the N-terminal exons to already isolated members of the human (a) and rat (b) CEA gene families. This was achieved either by analysis of genomic clones, or by hybridization with corresponding probes under high stringency conditions. In (a) the NCA N-terminal probe (see Fig. 1a) and in (b) the probe from the N-terminal exon of rnCGM4 (see Fig. 1f) was used for hybridization of total genomic DNA at low stringency. In rat genomic DNA (b) not all assigned restriction endonuclease fragments are clearly visible, because of weak cross-hybridization due to the lower degrees of nucleotide sequence similarities among the rat genes. E, EcoRI; S, SstI digests of the genomic DNAs. The arrows indicate the positions of λ/HindIII restriction endonuclease fragments.

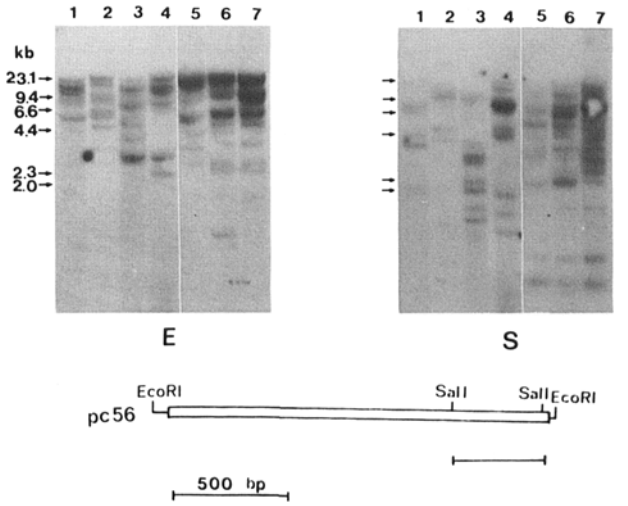

Fig. 3. Control hybridization of total genomic DNAs from all species tested with a highly conserved probe from a β-tubulin cDNA from *Drosophila melanogaster* (Bialojan et al. 1984). The filters were washed at 65°C, 2× SSPE. For the numbering above the lanes see legend to Fig. 1. The location of the probe is shown in the lower part of the figure. The filters are the same as used for the experiment shown in Fig. 1b. E, EcoRI; S, SstI digests of the genomic DNAs. Arrows indicate the positions of λ/HindIII restriction endonuclease fragments.

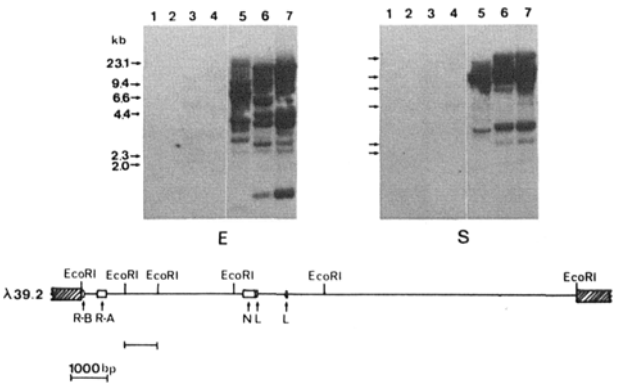

Fig. 4. Genomic DNA from different mammalian species hybridized with a probe from an intron region of the NCA 50 gene (Thompson et al. 1987). The filters were washed at 65°C, 2× SSPE. For the numbering above the lanes see legend to Fig. 1. The exact location of the probe within the NCA gene is depicted in the lower part of the figure. E, EcoRI; S, SstI digests of the genomic DNAs. The boxes represent exons coding for leader (L), N-terminal domain (N), half repeat A (R-A), and half repeat B (R-B).

conditions with all of the probes used, apart from the one from rnCGM2 that hybridized with two SstI fragments for the above-mentioned reason (data not shown). The extent of cross-hybridization among the rat genes is also weaker (compare Fig. 1d–f) as compared to the human genes at lower stringency (compare Fig. 1a–c).

Hybridization of the various genomic DNAs with a probe from the immunoglobulin-like repeating unit of CEA revealed similar results to the N-terminal domain, i.e., multiple hybridization signals in primates and very weak signals in rodents (data not shown).

Apart from the coding regions, we also tested a fragment from an NCA gene intron. Multiple strongly hybridizing fragments can be seen only among the primates under low stringency conditions (Fig. 4). At high stringency, one strong and two or three weakly hybridizing DNA fragments remain visible in human and chimpanzee DNA (data not shown).

Determination of the Rates of Nucleotide Substitutions in Exons and Introns

We determined the rate of synonymous (K_S) and nonsynonymous substitutions (K_A) for the N-terminal domain exons of CEA-like genes in both human and rat. Furthermore, we determined the rate of substitutions in the intron regions (K_N) flanking the N-terminal domain exons of several human CEA gene family members. The value of K_S can be used to estimate the absolute or relative branching points of phylogenetically related sequences (Busslinger et al. 1982; Sakoyama et al. 1987). Usually, the values for K_S and K_N found in the literature are approximately equal (Miyata et al. 1980; Sakoyama et al. 1987). This is consistent with the assumption that there should be negligible selective pressure on both kinds of sites (Miyata et al. 1980; Perler et al. 1980; Busslinger et al. 1982).

In comparisons of human CEA gene family members, we found K_S values double to three times higher than the corresponding values of K_N for the adjacent intron regions. This tendency is most obvious for the more distantly related genes hsCGM2 and hsCGM3 (see Fig. 5). When looking at K_A, i.e.,

	K_A^C	K_S^C	K_A^C/K_S^C	K_N^C		
				5'	Mean	3'
NCA vs. CEA	0.047±0.014	0.146±0.047	0.322	0.107	0.122	0.136
NCA vs. BGP1	0.046±0.013	0.132±0.043	0.348	nd	nd	nd
NCA vs. hsCGM1	0.034±0.011	0.161±0.049	0.211	0.167	0.118	0.068
NCA vs. hsCGM2	0.206±0.031	0.269±0.069	0.766	0.190	0.201	0.212
NCA vs. hsCGM3	0.284±0.039	0.505±0.109	0.562	0.189	0.193	0.197
NCA vs. hsCGM4	0.294±0.040	0.615±0.135	0.478	nd	nd	nd
NCA vs. PSβG	0.297±0.040	0.476±0.104	0.624	nd	nd	nd
CEA vs. BGP1	0.048±0.014	0.102±0.038	0.471	nd	nd	nd
CEA vs. hsCGM1	0.051±0.014	0.075±0.032	0.680	nd	nd	nd
CEA vs. hsCGM2	0.207±0.031	0.248±0.064	0.835	nd	nd	nd
CEA vs. hsCGM3	0.248±0.039	0.442±0.097	0.643	nd	nd	nd
CEA vs. hsCGM4	0.284±0.039	0.508±0.110	0.559	nd	nd	nd
CEA vs. PSβG	0.297±0.040	0.392±0.088	0.758	nd	nd	nd
BGP1 vs. hsCGM1	0.049±0.014	0.077±0.031	0.636	nd	nd	nd
BGP1 vs. hsCGM2	0.200±0.030	0.255±0.067	0.784	nd	nd	nd
BGP1 vs. hsCGM3	0.298±0.040	0.370±0.092	0.805	nd	nd	nd
BGP1 vs. hsCGM4	0.302±0.040	0.473±0.118	0.638	nd	nd	nd
BGP1 vs. PSβG	0.311±0.041	0.345±0.089	0.901	nd	nd	nd
hsCGM1 vs. hsCGM2	0.220±0.032	0.244±0.067	0.902	nd	nd	nd
hsCGM1 vs. hsCGM3	0.293±0.039	0.431±0.098	0.680	nd	nd	nd
hsCGM1 vs. hsCGM4	0.297±0.040	0.533±0.122	0.557	nd	nd	nd
hsCGM1 vs. PSβG	0.306±0.041	0.404±0.094	0.757	nd	nd	nd
hsCGM2 vs. hsCGM3	0.342±0.044	0.609±0.130	0.562	0.218	0.231	0.245
hsCGM2 vs. hsCGM4	0.357±0.045	0.664±0.145	0.538	nd	nd	nd
hsCGM2 vs. PSβG	0.336±0.043	0.529±0.113	0.635	nd	nd	nd
hsCGM3 vs. hsCGM4	0.042±0.013	0.142±0.045	0.300	0.084	0.067	0.050
hsCGM3 vs. PSβG	0.057±0.015	0.113±0.039	0.504	0.049	0.052	0.055
hsCGM4 vs. PSβG	0.060±0.015	0.145±0.046	0.414	0.111	0.074	0.037
rnCGM1 vs. rnCGM2	0.455±0.054	0.997±0.221	0.456	nd	nd	nd
rnCGM1 vs. rnCGM3	0.134±0.024	0.338±0.080	0.396	nd	nd	nd
rnCGM1 vs. rnCGM4	0.165±0.028	0.486±0.104	0.340	nd	nd	nd
rnCGM1 vs. rnCGM5	0.357±0.046	1.009±0.245	0.354	nd	nd	nd
rnCGM2 vs. rnCGM3	0.526±0.061	1.465±0.392	0.359	nd	nd	nd
rnCGM2 vs. rnCGM4	0.458±0.055	1.130±0.257	0.405	nd	nd	nd
rnCGM2 vs. rnCGM5	0.428±0.051	1.185±0.259	0.361	nd	nd	nd
rnCGM3 vs. rnCGM4	0.259±0.037	0.759±0.172	0.341	nd	nd	nd
rnCGM3 vs. rnCGM5	0.429±0.052	1.084±0.252	0.396	nd	nd	nd
rnCGM4 vs. rnCGM5	0.376±0.048	1.048±0.301	0.359	nd	nd	nd

Fig. 5. Comparison of the mutational rates for coding and noncoding regions within human and rat CEA gene families. The rates of synonymous (K_S) and nonsynonymous (K_A) substitutions (±standard error) in the N-terminal domain exons of CEA-like genes have been calculated in pairwise comparisons (see Materials and Methods). The first two and the last nucleotide were excluded from the comparisons because the corresponding codons are split by the exon/intron boundaries. For several human gene pairs the rate of substitutions in intron regions (K_N) adjacent to the N-terminal domain exon has also been determined. The introns compared correspond to 444 nucleotides upstream and 173 nucleotides downstream of the NCA N-terminal domain exon. References for the sequences used: NCA (Thompson et al. 1987), CEA (Oikawa et al. 1987c, and unpublished data), BGP1 (Hinoda et al. 1988), hsCGM1-4 (Thompson et al. 1989), PSβG (Thompson et al. 1989; Watanabe and Chou 1988a), rnCGM1-5 (Kodelja et al. 1989).

the rate of substitutions that lead to amino acid exchanges, another interesting feature can be seen. This rate is quite high, being most obvious again for the more distantly related human genes, resulting in values for K_A/K_S between 0.2 and 0.9 (Fig. 5). Among the rat genes this ratio ranges around 0.4 (Fig. 5).

Discussion

Is There Only Weak Selective Pressure on the Primary Structure of the N-terminal Domain?

The relatively high values for the ratio between amino acid replacement sites and silent sites found for most gene pairs (Fig. 5) reflect weak selective pressure on the primary protein sequence of the N-terminal domain. Comparably high values are also found for immunoglobulins (Miyata et al. 1980). In contrast, coding regions of genes with rigid functional constraints on the amino acid sequence such as histone genes or cytochrome c genes have K_A/K_S values that are 10–50 times lower (Miyata et al. 1980). Therefore, the primary structure seems to be of minor importance for the function of the CEA-related proteins or, alternatively, at least some of the CEA-like genes might represent functionless pseudogenes. However, most of the observed amino acid substitutions in the N-terminal domains are nonrandom and comprise conservative exchanges. Despite a low degree of similarity at the amino acid level, a strong structural resemblance exists between the deduced secondary structures of the N-terminal domains of all CEA-like proteins and the immunoglobulin V domains (Thompson et al. 1989). In fact, only a few critical and a number of conservatively exchanged amino acids seem to be sufficient to guarantee the formation of the characteristic immunoglobin-like fold (Kodelja et al. 1989; Thompson et al. 1989). Therefore, it appears to be the secondary and/or tertiary structure, where a common selective pressure for all of these molecules comes into play.

Conservation of Intron Regions among Human CEA-like Genes

Conservation of intron regions directly flanking the N-terminal domain exons has been found from se-

quence data for several CEA-like genes (Thompson et al. 1989). Southern data obtained with a probe from an intron region that is relatively distant from the next coding region (Fig. 4) also indicate conservation of intron sequences. Several partially characterized genomic clones for human CEA-like genes that cross-hybridize with probes from the N-terminal and repeat domain regions but that represent different gene loci, also hybridize with this intron probe (data not shown). This shows that the observed cross-hybridization is connected with CEA gene family members rather than unrelated genes or noncoding regions. Using the same method, Zimmermann et al. (1988) recently presented strong evidence for conservation of 3′ nontranslated regions, too.

We propose that recent gene amplification, by which whole gene units are duplicated, may explain the observed similarity of intron and 3′ nontranslated regions among members of the human CEA gene family. The clustering of human CEA-related genes on chromosome 19 (Zimmermann et al. 1988) and the occurrence of simple sequences that could serve as potential recombination sites between nonhomologous genes (Slightom et al. 1980; Cohen et al. 1982) in the introns of several human and rat CEA-related genes (Kodelja et al. 1989; Thompson et al. 1989) support this assumption. Such putative recombination sites could also facilitate exon shuffling (Gilbert 1978), whereby genes varying in the number of repeated domains may have arisen (cf. CEA and NCA). When the rate of silent substitutions of the N-terminal domain exon is compared with the exchange rate in adjacent introns (see Fig. 5), especially in more distantly related human CEA-like genes, it can be seen that the exons have mutated approximately twice as much as the corresponding introns. Although it cannot be excluded fully that the observed K_S/K_N ratio is due to an elevated mutation rate in the exon, we believe that it reflects the selective conservation of certain intron regions. In this context, there could be unknown functional constraints on these regions, contributing to the observed degree of sequence similarity. Conservation of functional elements in intron (Keller and Noon 1984) as well as 3′ nontranslated regions (Miyata et al. 1980; Harlow et al. 1988) has been reported.

Estimation of the Size of the Human and Rat CEA Gene Families

Through hybridization of genomic DNA with probes from various genes under high stringency conditions and analysis of genomic clones, individual members of the CEA gene families could be assigned to restriction endonuclease fragments of human and rat DNA. Using the NCA N-terminal domain exon probe in Southern analyses of genomic DNA, additional fragments to that shown to correspond to the NCA gene still hybridize even under high stringency conditions. They could be assigned to the highly similar genes of CEA and hsCGM1 that are members of a subgroup within the human CEA gene family, as already discussed. Evidence for another subgroup in the human genome has also been reported (Thompson et al. 1989; Watanabe and Chou 1988b). So far, three members of this subgroup have been found: hsCGM3, hsCGM4 (Thompson et al. 1989), and PSβG (Thompson et al. 1989; Watanabe and Chou 1988a,b). Their N-terminal domain exons exhibit over 90% nucleotide sequence similarity (Thompson et al. 1989).

The corresponding genomic clones that we recently isolated contain EcoRI (5.1 kb) and SstI (2.2 kb) restriction endonuclease fragments of the same length that hybridized with the hsCGM3 N-terminal domain exon probe (data not shown). On hybridization with the same probe, no additional hybridizing DNA fragments with a length different from that representing hsCGM3 can be observed in a genomic blot when high stringency washing conditions are employed. The hybridization signal with the homologous probe (hsCGM3 N-terminal domain exon) as well as with the heterologous probe (NCA) is unexpectedly strong in human genomic DNA (see Figs. 1 and 2). This can be explained by the presence of multiple genomic fragments of the same length corresponding to the highly similar gene duplication products hsCGM3/4 and PSβG that constitute the above-mentioned subgroup. Therefore, all three members of this subgroup can be assigned to genomic restriction endonuclease fragments of identical size for EcoRI and SstI digests (see Fig. 2).

Due to the presence of such highly similar subgroups that can result in multiple genomic DNA fragments of the same length, it is difficult to estimate the absolute size of the human CEA gene family. There must be at least 10 different genes, but this number may be even higher. This has to be clarified by genomic and cDNA cloning. In this context, a possible explanation for the nonreciprocal cross-hybridization of the human N-terminal probes with rat genomic DNA is suggested. Due to the presence of closely related subgroups in humans, and presumably in other higher primates, some of the DNA fragments visualized could represent more than one member of such a subgroup and thus give a more intensive hybridization signal with the rat probes. However, due to the higher degree of sequence divergence within the rat CEA family (Kodelja et al. 1989) and the absence of highly similar subgroups, as indicated by Southern analyses under high stringency conditions, the individual restric-

tion endonuclease fragments represent only single N-terminal domain exons, and thus give weaker hybridization signals. From Southern analyses in the rat we estimated approximately seven to nine different N-terminal domain exons that are presumed to represent different CEA-like genes. From these data, it looks as if the CEA gene family from the rat is of similar complexity to that in humans.

Evolutionary Trees of the CEA-like Gene Families in Primates and Rodents

Orthologous genes (originating through speciation) are usually more closely related than paralogous genes (originating through gene duplication) (Wu and Li 1985; Harlow et al. 1988). Therefore, it would be expected that the individual rat CEA-like genes should be more similar to their putative human counterparts than to other CEA-like rat genes, assuming that a common gene family originated before primate/rodent divergence. This situation has been found for all proto-oncogene subfamilies studied so far in hexapods and vertebrates, where the major gene duplication events must have occurred some 800 million years (Myr) ago, i.e., before the two animal classes branched (Shilo 1987). The striking similarities within the human subgroups but divergence to members of the rat CEA gene family indicate the opposite case. The degrees of nucleotide sequence similarities for the N-terminal domain exons of the human genes lie between 67% and 95%. In the rat these values are between 56% and 84%, whereas in an interspecies comparison they drop down to the narrow range of 52–66% (Kodelja et al. 1989). Therefore, a significant loss of similarity can be observed between species. Such high intraspecies similarities as found for several human CEA-like genes would be characteristic for genes that have duplicated separately after the speciation of rodents and primates, i.e., after mammalian radiation took place.

By analyzing the nucleotide sequences of the N-terminal domain exons using computer programs (see Materials and Methods), relative evolutionary trees were derived for the CEA gene families in primates and rodents. We then applied a silent molecular clock (Sakoyama et al. 1987), using the rate of silent mutations in the N-terminal domain exon to introduce absolute branching points into this model. However, the application of this method to estimate branching points over larger time intervals (>80 Myr) can lead to statistical errors for several inherent reasons, such as saturation effects due to undetectable multiple substitutions (Perler et al. 1980). This problem can be circumvented partly by statistical correction (Miyata et al. 1980). Moreover, it could be shown that the evolutionary rate has significantly slowed down during primate evolution (Chang and Slightom 1984; Goodman et al. 1984; Li et al. 1985; Koop et al. 1986; Sakoyama et al. 1987), whereas rodents appear to be one of the most rapidly evolving mammalian orders (Li et al. 1985; Wu and Li 1985; Harlow et al. 1988). Because of these intra- as well as interorder differences, no universal silent molecular clock for mammals is applicable [for a detailed evaluation of the molecular clock in mammals see Li et al. (1987)], and we therefore decided to use regional values of V_S (mean neutral evolutionary rate) for different time spans of primate evolution to take the different rate problem, as much as possible, into account.

We calculated the upper limit of a K_S^C value that could be achieved between two diverging sequences in the time interval since the assumed separation of the new world monkeys [i.e., ≈45 Myr ago, where prosimians and new world monkeys are thought to have split off (Britten 1986)], with an average evolutionary rate of 1.3×10^{-9} substitutions/site/year, as extrapolated for higher primates (Britten 1986), to $K_S^C = 0.117$. This value actually reflects an absolute point in time (45 Myr). The evolutionary rate for lower primate species, i.e., prosimians, appears to be approximately equal to that of rodents, $V_S = 6.6 \times 10^{-9}$ substitutions/site/year (Britten 1986). Thus, we subtracted the calculated K_S^C of 0.117 as a constant from our K_S^C values (Fig. 5) found for human gene pairs. From the resulting values we calculated the additional time span that would have been necessary to generate the mutations occurring with the higher V_S of 6.6×10^{-9} substitutions/site/year using the equation $t = (K_S^C - 0.117)/2V_S$. These time intervals were added to the branching point of higher primates (45 Myr). The resulting value is believed to represent a roughly adequate time point at which the compared sequences actually diverged during primate evolution. For rodents, where a constantly high rate of $V_S = 6.6 \times 10^{-9}$ substitutions/site/year is assumed (Britten 1986), these corrections were not necessary. We calculated the intraspecies branching points by using K_S^C values from Fig. 5 directly in the equation $t = K_S^C/2V_S$.

This led us to a hypothetical model for the evolution of the human and rat CEA gene families depicted in Fig. 6. From the available data it looks as if possibly two or more common primordial genes were present before mammalian radiation took place (see Fig. 6). Due to the controversy regarding the time of mammalian radiation (conservative estimations place it between 75 and 110 Myr), and the relative uncertainty of the applied V_S values (even more statistical analyses are necessary to gain reasonable absolute values), it is not quite clear as to the number of primordial genes that were distributed among the newly arising mammalian species.

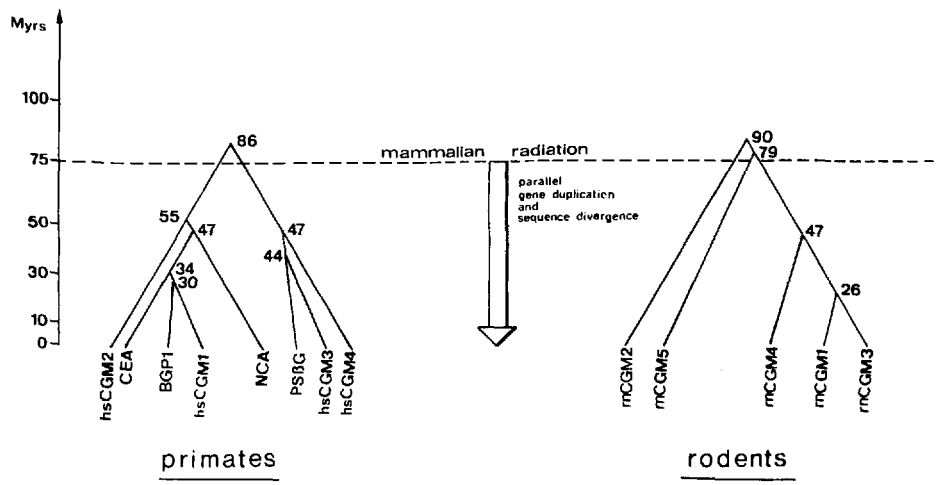

Fig. 6. Hypothetical model for the independent evolution of different CEA gene families in primates and rodents. The numbers denote the estimated divergence times (Myr) of the sequences and are calculated as described (see Discussion) from K_S^* values (Fig. 5). A single value usually represents the mean of the comparisons of all genes below this node, except the first node for the human genes. It is calculated as the mean of the comparisons from hsCGM4 with all genes on the left side of the tree and hsCGM2 with all genes on the right side of the tree.

However, as indicated by Southern analyses and sequence comparisons, it becomes obvious that two nonidentical human and rat CEA gene families must have evolved independently by parallel gene duplication followed by sequence divergence.

Such a parallel and animal order-independent evolution, leading to separate systems, has also been shown for the immunoglobulin C_H and C_L chain domains for human and mouse (Barker et al. 1980), where there is a different level of amplification in the two species. In the case of the immunoglobulins this is to be seen more as an amplification step of functionally determined genes in already existing systems, because the C domains of the contemporary light and heavy chains were present long before mammalian radiation took place (Barker et al. 1980). In contrast to this, the CEA family(-ies) apparently represents a very young system, having originated from a small number of primordial genes. From the proposed model, an interesting question arises as to whether these independently evolving genes have convergently acquired the same or related functions in the different animal orders. To answer this, the structure, expression, and function of the rat and human genes is currently under more detailed investigation.

Acknowledgments. We thank Sabine Barnert and Martina Weiss for excellent technical assistance and Indra Sarah Weber for helpful discussions. The computer programs LWL85 and MUPG were kindly provided by W.-H. Li from the Center for Demographic and Population Genetics, University of Texas. The computer program package PHYLIP was kindly provided by George D.F. Wilson from the Scripps Institution of Oceanography, La Jolla. This work was supported by grants from the Dr. Mildred Scheel-Stiftung für Krebsforschung and the Deutsche Forschungsgemeinschaft.

References

Barker WC, Ketcham LK, Dayhoff MO (1980) Origins of immunoglobulin heavy chain domains. J Mol Evol 15:113–127

Beauchemin N, Benchimol S, Cournoyer D, Fuks A, Stanners CP (1987) Isolation and characterization of full-length functional cDNA clones for human carcinoembryonic antigen. Mol Cell Biol 7:3221–3230

Bialojan S, Falkenburg D, Renkawitz-Pohl R (1984) Characterization and developmental expression of βtubulin genes in *Drosophila melanogaster*. EMBO J 3:2543–2548

Britten RY (1986) Rates of DNA sequence evolution differ between taxonomic groups. Science 231:1393–1398

Busslinger M, Rusconi S, Birnstiel ML (1982) An unusual evolutionary behaviour of a sea urchin histone gene cluster. EMBO J 1:27–33

Chang L-YE, Slightom JL (1984) Isolation and nucleotide sequence analysis of the β-type globin pseudogene from human, gorilla and chimpanzee. J Mol Biol 180:767–784

Cohen JB, Effron K, Rechavi E, Ben-Neriah Y, Zakut R, Givol D (1982) Simple DNA sequences in homologous flanking regions near immunoglobulin V_H genes: a role in gene interaction? Nucleic Acids Res 10:3353–3370

Engvall E, Shively JE, Wrann M (1978) Isolation and characterization of the normal crossreacting antigen: homology of its NH_2-terminal domain with that of carcinoembryonic antigen. Proc Natl Acad Sci USA 75:1670–1674

Felsenstein J (1985) Confidence limits on phylogenies: an approach using the bootstrap. Evolution 39:783–791

Gilbert W (1978) Why genes in pieces? Nature 271:501

Gold P, Freedman SO (1965a) Demonstration of tumor-specific antigens in human colonic carcinomata by immunological tolerance and absorption techniques. J Exp Med 121:439–462

Gold P, Freedman SO (1965b) Specific carcinoembryonic antigens of the human digestive system. J Exp Med 122:467–480

Goodman M, Koop BF, Czelusniak J, Weiss ML, Slightom JL (1984) The η-globin gene: its long evolutionary history in the β-globin family of mammals. J Mol Biol 180:803–823

Haagensen DE Jr, Metzgar RS, Swenson B, Dilley WG, Cox CE, Davis S, Murdoch J, Zamcheck N, Wells SA Jr (1982) Carcinoembryonic antigen in nonhuman primates. J Natl Cancer Inst 69:1073–1076

Harlow P, Litwin S, Nemer M (1988) Synonymous nucleotide substitution rates of β-tubulin and histone genes conform to high overall genomic rates in rodents but not in sea urchins. J Mol Evol 27:56–64

Hinoda, J, Neumaier M, Hefta SA, Drzeniek Z, Wagener C, Shively L, Hefta LJF, Shively JE, Paxton RJ (1988) Molecular cloning of a cDNA coding biliary glycoprotein I: primary structure of a glycoprotein immunologically crossreac-

tive with carcinoembryonic antigen. Proc Natl Acad Sci USA 85:6959–6963

Jantscheff P, Indzhiia V, Micheel B (1986) Search for CEA-like molecules in polymorphonuclear leukocytes of nonhuman primates using monoclonal antibodies. Arch Geschwulstforsch 56:113–116

Keller EB, Noon WA (1984) Intron splicing: a conserved internal signal in introns of animal pre-mRNAs. Proc Natl Acad Sci USA 81:7417–7420

Kessler MJ, Shively JE, Pritchard DG, Todd CW (1978) Isolation, immunological characterization, and structural studies of a tumor antigen related to carcinoembryonic antigen. Cancer Res 38:1041–1048

Kodelja V, Lucas K, Barnert S, von Kleist S, Thompson JA, Zimmermann W (1989) Identification of a carcinoembryonic antigen (CEA) gene family in the rat: analysis of the N-terminal domains reveals immunoglobulin-like, hypervariable regions. J Biol Chem (in press)

Koop BF, Goodman M, Xu P, Chan K, Slightom JL (1986) Primate η-globin sequences and man's place among the great apes. Nature 319:234–238

Li W-H (1981) Simple method for constructing phylogenetic trees from distance matrices. Proc Natl Acad Sci USA 78: 1085–1089

Li W-H, Wu Ch-I, Luo C-C (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol Biol Evol 2:150–174

Li W-H, Tanimura M, Sharp PM (1987) An evaluation of the molecular clock hypothesis using mammalian DNA sequences. J Mol Evol 26:330–342

Miyata T, Yasunaga T, Nishida T (1980) Nucleotide sequence divergence and functional constraint in mRNA evolution. Proc Natl Acad Sci USA 77:7328–7332

Needleman SB, Wunsch CD (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48:443–453

Neumaier M, Zimmermann W, Shively L, Hinoda Y, Riggs AD (1988) Characterization of a cDNA clone for the nonspecific crossreacting antigen (NCA) and a comparison of NCA and carcinoembryonic antigen (CEA). J Biol Chem 263:3202–3207

Oikawa S, Imajo S, Nogouchi T, Kosaki G, Nakazato H (1987a) The carcinoembryonic antigen (CEA) contains multiple immunoglobulin-like domains. Biochem Biophys Res Commun 144:634–642

Oikawa S, Kosaki G, Nakazato H (1987b) Molecular cloning of a gene for a member of carcinoembryonic antigen (CEA) gene family; signal peptide and N-terminal domain sequences of nonspecific crossreacting antigen (NCA). Biochem Biophys Res Commun 146:464–469

Oikawa S, Nakazato H, Kosaki G (1987c) Primary structure of human carcinoembryonic antigen (CEA) deduced from cDNA sequence. Biochem Biophys Res Commun 142:511–518

Paxton RJ, Mooser G, Pande H, Lee TD, Shively JE (1987) Sequence analysis of carcinoembryonic antigen: identification of glycosylation sites and homology with the immunoglobulin supergene family. Proc Natl Acad Sci USA 84:920–924

Perler F, Efstratiadis A, Lomedico P, Gilbert W, Kolodner R, Dodgson J (1980) The evolution of genes: the chicken preproinsulin gene. Cell 20:555–566

Sakoyama Y, Hong K-J, Byun SM, Hisajima H, Ueda S, Yaoita Y, Hayashida H, Miyata T, Honjo T (1987) Nucleotide sequences of immunoglobulin ϵ genes of chimpanzee and orangutan: DNA molecular clock and hominoid evolution. Proc Natl Acad Sci USA 84:1080–1084

Shilo B-Z (1987) Proto-oncogenes in *Drosophila melanogaster*. Trends Genet 3:69–72

Slightom JL, Blechl AE, Smithies O (1980) Human G_γ- and A_γ-globin genes: complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. Cell 21:627–638

Tawaragi Y, Oikawa S, Matsuoka Y (1988) Primary structure of nonspecific crossreacting antigen (NCA), a member of carcinoembryonic antigen (CEA) gene family, deduced from cDNA sequence. Biochem Biophys Res Commun 150:89–96

Thompson J, Zimmermann W (1988) The carcinoembryonic antigen (CEA) gene family: structure, expression and evolution. Tumor Biol 9:63–83

Thompson JA, Pande H, Paxton RJ, Shively L, Padma A, Simmer RL, Todd CW, Riggs AD, Shively JE (1987) Molecular cloning of a gene belonging to the carcinoembryonic antigen gene family and discussion of a domain model. Proc Natl Acad Sci USA 84:2965–2969

Thompson JA, Mauch E-M, Chen F-S, Hinoda Y, Schrewe H, Ortlieb B, Barnert S, von Kleist S, Shively JE, Zimmermann W (1989) Analyses of the size of the carcinoembryonic antigen (CEA) gene family: isolation and sequencing of N-terminal domain exons. Biochem Biophys Res Commun 158: 996–1004

Wahren B, Gadler F, Gahrten G, Hammerström S, Hareland Y, Hyden N, Ljungdahl E, Mahlen A, Ruden U, Wiklund M (1983) NCA: a differentiation antigen of myelopoietic cells in humans and hominoid monkeys. Ann NY Acad Sci USA 417:344–358

Watanabe S, Chou JY (1988a) Isolation and characterization of complementary DNAs encoding human pregnancy-specific β_1-glyprotein. J Biol Chem 263:2049–2054

Watanabe S, Chou JY (1988b) Human pregnancy-specific β_1-glycoprotein: a new member of the carcinoembryonic antigen gene family. Biochem Biophys Res Commun 152:762–768

Williams AF (1987) A year in the life of the immunoglobulin superfamily. Immunol Today 8:298–303

Wu C-I, Li W-H (1985) Evidence for higher rates of nucleotide substitution in rodents than in man. Proc Natl Acad Sci USA 82:1741–1745

Zimmermann W, Ortlieb B, Friedrich R, von Kleist S (1987) Isolation of cDNA clones encoding the human carcinoembryonic antigen reveal a highly conserved repeating structure. Proc Natl Acad Sci USA 84:2960–2964

Zimmermann W, Weber B, Ortlieb B, Rudert F, Schempp W, Fiebig H-H, Shively JE, von Kleist S, Thompson JA (1988) Chromosomal localization of the carcinoembryonic antigen gene family and differential expression in various tumors. Cancer Res 48:2250–2254

Received December 15, 1988