Structure of the carbonic anhydrase VI (CA6) gene: evidence for two distinct groups within the α -CA gene family

Weiping JIANG¹ and Dwijendra GUPTA²

Department of Biochemistry and Molecular Biology, The Pennsylvania State University College of Medicine, Hershey, PA 17033, U.S.A.

The secreted carbonic anhydrase (CA VI) is believed to be one of the oldest mammalian CAs in evolutionary terms. To elucidate its gene structure and compare it with other members of the α -CA family, we cloned genomic fragments encoding the bovine CA6 gene and determined its exon/intron organization. The gene spans approx. 25 kb and consists of eight exons and seven introns. Exon 1 encodes the 5' untranslated region, the signal peptide and the N-terminus of the mature enzyme. Exon 8 encodes the 3' untranslated region and the C-terminal extension that is unique to CA VI. Exons 2–7 encode the CA domain, which shows significant sequence similarity to other CAs. Two distinct groups exist in the α -CA family on the basis of a

comparison of the known gene structures. One group consists of the cytoplasmic (CA I, II, III and VII) and mitochondrial (CA V) members. The other group consists of the membrane-bound (CA IV and IX) and secreted (CA VI) members. In particular, the seven exon/intron boundaries in the CA domain of the CA6 gene are conserved in the CA9 gene, which encodes the multidomain protein CA IX that is overexpressed in tumours and has transforming potential.

Key words: evolution, exon/intron organization, membrane-bound carbonic anhydrase, secreted carbonic anhydrase.

INTRODUCTION

Carbonic anhydrase (CA, EC 4.2.1.1) catalyses the reversible hydration of CO₂ [1]. There are three independent CA gene families: α -CA, β -CA and γ -CA [2]. Within the α -CA gene family, nine CA isoforms and three CA-related proteins have been discovered from mammals and designated CA I to CA XII [1,2]. They differ in many aspects such as activity, subcellular localization, tissue distribution and domain structure. For example, these proteins are cytoplasmic (CA I, II, III and VII), mitochondrial (CA V), membrane-bound (CA IV, IX and XII) or secreted from cells (CA VI). However, all twelve proteins share significant sequence similarity in the common CA domain of approximately 260 amino acid residues. Only the membrane-bound (CA IX and XII) and secreted proteins have additional domains.

CA VI is specifically expressed in the salivary gland of a number of mammalian species. The amino acid sequences are highly conserved across the species [3,4]. For example, the mature bovine CA VI is 87% and 68% identical with the sheep and human proteins respectively [5]. In addition to the three histidine residues conserved in all active members of the CA family, the mature CA VI consists of the following unique structural features (see Figure 1A). The CA domain has an internal disulphide bond and two N-linked oligosaccharides that can also be sulfated. The C-terminal extension is rich in hydrophilic amino acids. The proposed functions of CA VI include the maintenance of pH and bicarbonate levels in saliva or even in the upper alimentary canal [6]. It has recently been shown that CA VI is the same as gustin, a Zn-metalloprotein constituting approx. 3% of human parotid saliva protein, which is decreased in patients with loss of taste and is associated with pathological morphology of taste buds [7].

The growing number of members of the α -CA gene family offer a great opportunity to investigate their evolution with

regard to their structure, function and regulation. Especially with recent discoveries of overexpression of CA IX and XII in certain cancers [8–10], it is necessary to elucidate the gene structures of individual members as one of the first steps towards understanding the regulation of these CA genes in both physiological and pathological processes. Here we describe the first gene structure for CA VI and compare it with other known gene structures of the α -CA family.

EXPERIMENTAL

Library screening

The *SphI*–*XhoI* fragment of the cloned bovine CA VI cDNA [5] was labelled with ³²P by using a Prime-A-Gene kit (Promega). The resulting probe was used to screen a λ DASH II bovine genomic library (Stratagene) under the conditions recommended by the supplier. Putative plaques were verified by secondary screening. Positive clones were subjected to Southern blot analysis by using the cloned cDNA as a probe to identify the fragments that contained the exon sequences of the CA6 gene. These fragments were subcloned into plasmid pBluescript (Stratagene) by either PCR or restriction enzyme digestion.

PCR cloning

Most of the templates used in the PCR cloning were DNA from five GenomeWalker libraries. The construction of the GenomeWalker libraries and general conditions for PCR amplifications were performed as described previously [11]. A CA6-specific primer (Primer 1; see Table 1) and an adaptor primer (AP1) were used in the initial round of PCR. The resulting PCR products were diluted and used as templates in subsequent PCR with a second CA6 primer (Primer 2; see Table 1) and AP2. The final PCR products were separated by agarose-gel electrophoresis,

Abbreviation used: CA, carbonic anhydrase.

¹ To whom correspondence should be addressed at present address: R & D Systems Inc., 614 McKinley Place N. E., Minneapolis, MN 55413, U.S.A. (e-mail weipingj@rndsystems.com).

² Present address: Institute of Biotechnology, Guru Ghasidas University, Bilaspur (M.P.), India.

Figure 1 Overview of the bovine CA6 gene

(A) The cDNA and protein structure. The 5' and 3' untranslated regions (UTR) are shown as horizontal lines. The coding sequence is divided into three regions (boxes), indicated as follows: S, signal peptide absent from the mature, secreted protein; carbonic anhydrase, the region showing sequence similarity to other CAs (see Figure 3); C, the C-terminal extension that is unique to CA VI. The numbers along the coding sequence indicate the positions of the amino acid residues at the beginning and end of each region as well as those involved in the following: residues 23 and 205, cysteine residues that form an internal disulfide bond; residues 48 and 237, asparagine residues responsible for N-linked glycosylation (Y-shaped structure); residues 92, 94 and 119, three histidine residues that bind zinc. The numbering system is the same as that used previously [5]. (B) The exons of mRNA; (A) and (B) are drawn to approximately the same scale. (C) The exon/intron organization. The exons and introns are numbered and represented by boxes and lines respectively. The sizes of the introns are indicated below individual introns. (D) The genomic clones. Clones a to l were obtained by PCR and arrows represent the directions of the cloning. Clone m was derived from the library screening. Clones m-1 and m-2 are subclones of m that were generated by PCR (m-1) and restriction enzyme digestion (m-2). #, number.

Table 1 PCR primers used to obtain genomic clones

Clone	Size (kb)	Primer 1	Position	Primer 2	Position
a	1.7	5'-CAGCAGGAATAGCAGGGTGATCAT-3'	Exon 1	5'-GGGCCGATGCTGCGGAGGCTCCTCTT-3'	Exon 1
b	2.2	5'-GCGGGTGTGGGGGCGTGCTATT-3'	Exon 1	5'-GGATCAGCCAGCCAGACCCGGTAAA-3'	Exon 1
c	1.5	5'-GAATAGTGAGGAGAGGAGGAGAA-3'	Intron 1	5'-GTGTTTTTCATCCAGCACCCCT-3'	Exon 2
d	0.2	5'-AGAGGGGGTGCTGGATGAA-3'	Exon 2	5'-TACCGTGTGGCCGTGTTGG-3'	Exon 2
e	5.0	5'-CCCTTCCCTGAGGGCTCTGAA-3'	Exon 2	5'-TACTGCGAGCCATCGGAGGTTG-3'	Exon 3
f	1.8	5'-CTCAGAGCCGCTGATCTCCGAGGAAT-3'	Exon 3	5'-GTGGAAGTGCATCTGCTTGGCTAAGT-3'	Exon 3
g	0.6	5'-CAGATCAGCCTGCCATCCAGC-3'	Exon 3	5'-CCACCGCAGTGAGAGCCCAACCAC-3'	Intron 3
h	0.9	5'-GGGGTTGCAAGAGTCAGACATGA-3'	Intron 3	5'-GCGACTTGGCAAGCACCCTCCTA-3'	Intron 3
i	1.8	5'-CCAGGGCCGCGCAGCACAGCCAAACCA-3'	Exon 4	5'-TCCGGCTCATTCTGGGCTTCCTCATA-3'	Exon 4
j	3.8	5'-GGCAGCTATGAGGAAGCCAGAATGAG-3'	Exon 4	5'-CGGATGGTTGGCTGTGC-3'	Exon 4
k	0.9	5'-GCCCAAAACAAGGTCACTGGGG-3'	Intron 4	5'-CCGAGGCAGAGAGCAGGGAGG-3'	Intron 5
l	4.0	5'-TGGATGGTCTCGTTCTGGTGATTCA-3'	Exon 7	5'-ATGAGAGGAGAGGCACTGGGTAAC-3'	Intron 6
m-1	4.2	5'-AATTAACCTCACTAAAGGG-3'	Vector	5'-GTGATTCGAGAGGAGGTAG-3'	Exon 8

isolated by the GeneClean procedure (Bio 101), and ligated into a TA-cloning vector, pCR2.1 (Invitrogen) or pGEM-T Easy (Promega). Bovine genomic DNA was also used as a template in PCR to amplify CA6 fragments directly with primers 1 and 2 (see Table 1).

Sequence analysis

Plasmid DNA was sequenced at the Molecular Genetics Core Facility of our institution with an ABI Prism automated sequencer. The determined genomic sequences were compared with the cDNA sequence to identify exons, introns and their boundaries. The exons and exon/intron boundaries were sequenced on both strands. The sizes of introns were estimated by agarose-gel electrophoresis of restriction fragments. The

amino acid sequences of CA with known gene structures were aligned with Clustal W [12] and adjusted manually. The Transfac database containing known transcription-factor-binding sites was searched with the 5' flanking region by using MatInspector [13].

RESULTS

Isolation of CA6 genomic clones

Two positive clones were isolated from the screening of the λ DASH II bovine genomic library with the *SphI*-*XhoI* fragment of the cloned bovine CA VI cDNA [5]. Southern analysis of the two clones showed that clone m (Figure 1) contained more exon sequences than the other clone. However, clone m did not contain the entire CA6 gene.

TAACCTTAAG	TCTTATTAAA	CTGGAGAAAT	GGAGGCAAA	AGCAGTTGAA	TAACCTTGCTG	AAGGTCACAC	AGCTAGTAAG	TGGTAGAGCA	90	
GGATCTCGAA	CTCTGTGTGT	CTGGCCACAG	AATCTCTTCT	CTCCACCAGA	GTPAAATCAT	GCTACTGTAC	TGCCATTGCA	GCATCCCTTA	180	
GACTTCCAGG	AAAGAGTAGC	CATGAGATTA	AGGAAGCTGC	AGAGGGCCCT	CGCTCTGTTC	TACAGACACG	GACGGTCCCT	TGTTTCAGGT	270	
CAGGAACATT	TCTGCCCTTT	GCTTGTGTCC	TCTGTGAAGG	CGCTGGGAGG	TCCGAGCCCT	TACAGCAATG	TAACTTTGAA	AATCTCTCTC	360	
AGCTTAACCT	AAGGACTGTT	TACCAAGTGT	GACACAGGAG	GATTAGSCAA	TTCAITGATC	TGGCATGTAG	GATGGGTTTG	GGTGTGACCA	450	
AGTGCATGTT	GCTGTCTAGA	GTAATCCCAT	TCCATAGAGC	AACTCAGACA	CCCACAGCAG	CTGGGATAGC	CCAGGATACA	TGTTGGGATG	540	
TGGCTGCTGA	GGGTGGACTC	AGCATCCACC	AGTGAATCAA	GAGTTAGGAT	GAGGAACCTC	CCTGTGGTGC	CAGTGGTTAA	AATCTTGGCT	630	
TCCAAACAGC	GGGTGGACGC	CCTAGGCGAG	CCCTAAGTCT	GCATTATGAG	CAGGTGTCTA	CTGTCTGAGC	AGGCAGGAAA	TACCACGTAG	720	
CATCTGGCTG	GGAGTTGATG	ATGACACAGG	AGGCTTAGCC	TGCTTCCGAT	CATGGGGTGG	CAAGAGATCG	GACATGACTG	AGGCAGCTAT	810	
CTGATCTGGG	AGCTGAAGATC	CTCTAGTCTT	TGTGGCCGGA	AAACCAAAAC	ATCAAAAAC	GGAGGCAATG	TACCAAAATC	AATAAGACTG	900	
TTAAAAATGG	TCCCCCCCCA	CACACACACA	AAATAGSCTT	AAGAAAAAAC	GAGGATGAGA	ACAGAACAGC	CTACAGTTGG	ATTCTCCATT	990	
TCATGCTCAA	ACATGGGGCT	TCCCCAGTGA	CTCAGCAGAT	AAAGATATCA	CCTGTAATGC	AGGAGACAGA	GAGAGCTGGG	TTTGTACCTC	1080	
NF-E2 p45										
GGGTCCAGGA	GATCCCCCTG	AGGAGGAAGT	GGCAACCCAC	TCCAGTATTC	TTACCCGGAG	AATCCCATGG	ACAGAGGAGC	CTGGGGGGCT	1170	
ACAGTCCATG	GAGTCTCGAA	AGAGTCTGACA	CCACTGACCA	CGCATGGGGC	CCAGGCGCTC	TCTCCAACTG	TCTGGCTGAG	GGCAGCCCTA	1260	
AGTCTGCATT	AGTGGCAGGT	GTCTACTGCT	GAGGCAGGCA	GGAAATACCA	CGCAGCATCC	GGCTGGGGTT	TAGGCACTG	GGCATGCTCC	1350	
TCACCGGTTT	CCTTCTCTGG	CCCTGCTTGC	CAGGAACGAG	GTACACAGAA	ATGCCATTGT	TTTGGAAAAA	CTTCCCACTT	CTGCAGCCCT	1440	
MZF1										
CTGCTTGGGG	CCATCTCTCT	GGCCCGAGGT	AGTTGCTTGC	CCACAAAGGC	TGCATCTGTT	GACCCCTCTG	GTGAGTTTTC	<u>ATAAAT</u> TCAGA	1530	
TATA box										
CAGCAACAT	CCGAACCTTC	AGGCGAAGTG	GACTCAGTGC	TTTGTGTCAGT	GGTGATGGAC	AGGTGCGCGT	CACCTGGCCA	GGGAAGCGGG	1620	
TGTGGGGGGG	GTGCTATTTT	TCAGTGGGCA	CAGCGAAGGG	TGATCAGCCG	AGCCAGACCC	GGTAAATTTG	TATTTACAGT	TTCTGAGCCT	1710	
TTAAAGAGAG	GGCTTCCGCA	GCATCGGCC	TGACTCCAGA	CCTGCTCGAG	CATGATCACC	CTGCTAATCC	TGCTCGTGGT	GGGAGCCGAC	1800	
M I T L L F L L V V G A Q										
GGCCAGCACG	AGTGGACCTA	CTCAGtgtaag	gcccccccca	tgggcagtgctg	ccactctctg	agggtgaagc	-----3.5 kb-----	1890		
A Q H E W T Y S E										
cccccttgge	tctctccctg	gtcagAGGGG	GTGCTGGATG	AAAAACCTG	GAGGCTGCG	TACCCGAGCT	GTGGGGGGAC	AAGACAGTCG	1980	
G V L D E K H W R L Q Y P D C G G T R Q S										
CCCATTGACC	TGAAGATGAA	GAAAGTTCGG	TACAACCTTT	CCCTGAGGGC	TCTGAACCTC	ACAGGCTACG	GACTCCGGCA	AGGGAGATTC	2070	
P I D L K M K K V R Y N P S L R A L N L T G Y G L R Q G E F										
CCCATGACCA	ACAACGGCC	ACACGGgttaag	gatggcgagc	agaccggtct	ctgtccctcg	gggaagcggg	-----5.0 kb-----	2160		
P M T N N G H T V										
ccctgcccc	tgttctctct	tctagTACAG	ATCAGCCCTG	CATCCAGCAT	GGCATGACA	ACTTCCGATG	GCTCCAGTGA	CTTAGCCAAG	2250	
Q I S L P S S M R M T T S D G S Q Y L A K										
CAGATGACAT	TCCACTGGGG	CGGAGCATCC	TGCGAGTACA	GGGCTCTGTA	GCACACTGTG	GATGGGATGC	GATACATAAT	TGAGgtacct	2340	
Q M H F H W G G D S S E I S G S E H T V D G M R Y I I E										
tcaggtctct	ctttcttcca	aaccccatag	-----3.0 kb-----	tgagtgactt	tcaactccctt	tcttgettcc	acagATTTCAC	2430		
I H										
GTGCTCCACT	ACAATCTCAA	ATATGGCAGC	TATGAGGAAG	CCCAGAATGA	CCCGGATGGT	TTGGCTGTGC	TGGCGGCCCT	GOTTGAGgta	2520	
V V H Y H S K Y G S Y E E A Q N E P D G L A V L A A L V E										
ggctgaacct	acacactcga	aattataaac	-----4.0 kb-----	ctctcccttag	gttaactget	cegtgtttcc	ccctcagGTA	2610		
V										
AGGATTATG	CTGAAAATAC	TTACTACAGC	AACCTCATTT	CTCACCTGGA	GGCATCAGG	TATGAGgta	agagaaacca	acagaaagcca	2700	
K D Y A E N T Y Y S N F I S H L E D I R Y A G										
caccaggcca	gagcaatgag	-----1.5 kb-----	gactttgctt	ctcatccctt	ctattagGAC	AAAGCAGAGT	TCTGAGGGAC	2790		
Q S T V L R D										
CTTGACATTC	AGGACATGCT	ACCTGGGGAC	CTCCGACTAT	ACTACAGCTA	CTTGGGGTCA	CTCACCCTTC	CTTCTGTGAC	AGAAACCTTC	2880	
L D I Q D M L P G D L R Y Y Y S Y L G S L T T P S C T E N V										
CAGTCTGTTG	TGGTAGCAGA	TACTGTCAAG	CTCTCTAAGA	CACAGgttat	gtacaagaat	ggctttatat	-----3.0 kb-----	2970		
H W F V A D T V K L S K T Q										
aaccagtgcc	tctctctcca	ttaattcctt	tcagtgtgct	ggcagATTGA	GAAGCTAGAG	AATCTCTTAC	TGAATCACCA	GAACGAGACC	3060	
I E K L E N S L L N H Q N E T										
ATCCAAAATA	ATTACCGCAG	CACCTCAGCCC	CTGAACCCACA	GAGTGGTAGA	AGCCAACTTC	GTGCTCTACC	CTCATCAGG	tgagtgagat	3150	
I Q N N Y R S T Q P L N H R V V E A N F V S H P H Q E										
caactctcgt	ctgttcttgg	-----3.0 kb-----	aaagttctga	cttttctctg	ttaactcttc	tttttataga	ATACACTTTA	3240		
Y T L										
GGCTCCAGC	TCCACTTTTA	TCTAAACAAC	ATTGACCAAA	ACCTTGAGTA	CCTGAGAAGA	TTTATTGAA	AGAAGATAAC	AAAGAGGAAA	3330	
G S K L H F Y L N N I D Q N L E Y L R R F I E Q K I T K R K										
AAAGAAAGT	ATTGGCTTGG	AAGAGTGGAG	CCTGCTCTCT	TTTCAAGAAG	CCCTGGGTGT	CCAGATCAT	CCATTTCTAC	CGAGGACGCT	3420	
K E K Y W P										
GCTTAATTTA	CGATTAAAC	AGGAGAAACC	ATATCATCTG	AACGGAAGTG	AGATGGCTTA	AATGCATGAG	AAGGGTTTGA	GCGAGATGAT	3510	
ACCAATGGGA	ATTGACTGGA	ATAGAAATTT	AACCTACCTC	CTCTCGAATC	ACATACATTT	ATTGTGCTTA	ATTAGAAATC	AACCTGAAAT	3600	
CCTGATAGCA	TTTAAATGAA	ATAAAATAAC	TTTGGGAAAA	TTTGTATTTT	AAACCTTCTC	TTTCTATPCT	CCTTGGTGCC	ATTTCAAAAA	3690	
TTCCTATATC										
TCGGTTTTTT	TGCTCTATCC	GGAGGACCCA	GTTGGGAGGG	ATTTCACGCT	TTTTCAAAAG	TTCCACTGAC	GTTATTGAAA	GGTTAAAAAT	3780	
GAAAGTTTTCC	AATCAACAAC	TGATATTFTT	ATAACTTTTA	TGGAACCTGT	AAGTACTATT	ATTGTGTTGT	ATTATTATTG	TTAATTATTG	3870	
CTTCCAACTT	TTCACTATTA	CAACAATGCA	GAGTAGGAAT	TTGGCTGTGC	TCAGTCACTT	CATGGTGTCC	GACTCCCTCG	GACCTTATGG	3960	
ACCACTGCCG	CCAGGCTCCT	CTGTCCATGT	GATPCTCCAG	GCAAGAATAC	TAGAGTGGGT	TGCCATTTCC	TTCTCCAGTG	GTAAGGTGCG	4050	
AACTGAGTGA	AGTGAAGTCG	CTCAGTCTGT	TCTGACTCTT	TGCGACCCAT	GGACTGTAGC	CTACAGGCTT	CCTCTGTCCA	TGGGATTTC	4140	
TAGGCAAGAG	TACTGGAGTG	GTTTGCCATT	TCCTTCTCCA	GGATTGGGCT	GAGTTAAGCT	TAAAAAACCT	GACTTCCCTG	GTGTATCAGA	4230	
CCAGGAAGAA	TTC									4243

Figure 2 Exon/intron structure of the CA6 gene

The 5' and 3' flanking regions as well as the exon sequences are in capitals; the intron sequences are in lower-case. The amino acid sequence is below the coding sequence. Potential transcription-factor-binding sites are underlined and the names of the factors are indicated. The potential TATA box is italicized and underlined. The nucleotide corresponding to the 5' end of the largest cDNA clone is doubly underlined, as is the nucleotide corresponding to the end of the cDNA before the poly(A) tail. The polyadenylation signal is underlined.

To obtain genomic fragments covering the remainder of the CA6 gene, a genome walking approach was taken. Templates, either the GenomeWalker libraries or the bovine genomic DNA, were amplified by PCR with primers synthesized on the basis of the determined sequences (Table 1). For the GenomeWalker libraries, two rounds of PCR were necessary to obtain pure products with significant yields. Twelve PCR fragments were obtained and the resulting plasmid clones (clones a to l in Figure 1) were sequenced. The twelve PCR clones, together with the λ DASH II clone m, covered the entire CA6 gene (Figure 1).

Exon/intron structure of the CA6 gene

The bovine CA6 gene spans approx. 25 kb and consists of eight exons and seven introns (Figures 1 and 2). Exon 1 encodes the 5' untranslated region, the signal peptide and the N-terminus of the

mature enzyme. Exon 8 encodes the 3' untranslated region and the C-terminal extension that is unique to CA VI. Exons 2–7 encode the CA domain, which shows significant sequence similarity to other CAs. The splice donor and acceptor sites of all seven introns conform to the consensus sequences GT-AG [14]. The exons are small, ranging from 70 bp (exon 5) to approx. 420 bp (exon 8) in size. In contrast, the introns are large, ranging from 1.5 kb (intron 5) to approx. 5 kb (intron 2) in size. Overall, approx. 6% (1.5/25 kb) of the CA6 gene corresponds to the exon sequences.

5' flanking region

The 5' end of the published bovine CA VI cDNA [5] corresponded to nt 1696 (Figure 2). Sequencing of the several additional cDNA clones, however, revealed that the 5' end of the largest

Figure 3 Alignment of amino acid sequences of the bovine CA VI and the human carbonic anhydrases with known gene structures

Numbers above the aligned sequences correspond to numbering of amino acid residues in CA I [8]. The CA VI sequence is also numbered at the right in accordance with Jiang et al. [5]. The positions of introns are represented by vertical bars and the introns in the CA6 gene are numbered below the bars as in Figure 1. The phases of introns are indicated by superscript numbers before the introns. The invariant amino acids in all the sequences are indicated by stars and the conserved substitutions by dots. The three histidine residues involved in zinc binding are indicated by arrows. The two cysteine residues conserved in the CA IV, VI and IX sequences and involved in disulphide bond formation in CA VI are indicated by #. The two N-glycosylation motifs in the CA VI sequence are underlined. The transmembrane domain of the CA IX sequence is doubly underlined.

cDNA corresponded to nt 1576 (doubly underlined in Figure 2). The new 5' end might correspond to the start site for transcription because the size of the largest cDNA (1.45 kb) matched that of the mRNA (1.4 kb). A sequence of CATAAA showing similarity to the TATA box resides 50 bp upstream of the new 5' end. A total of 131 matches to the known transcription factor-binding sites were found from the Transfac databases in the 5' flanking region. Two of these, GTGACTCAGCA and TCCCCACT, are underlined in Figure 2, which were 100% identical with both the core and matrix sequences of DNA motifs recognized by transcription factors NF-E2 p45 [15] and MZF1 [16].

Comparison of the CA6 gene structure with other members of the α -CA family

The gene structures of the following human CAs have been reported: I [17], II [18], III [19], IV [20], V [21], VII [22] and IX [10]. The comparison of these gene structures with CA6 clearly showed that the members of the α -CA gene family evolved from the same ancestor (Figure 3). All the α -CA genes with the exception of CA4 contain six introns in the CA domain, whereas CA4 has an additional intron. Three introns in the known gene structures corresponding to introns 3, 4 and 6 of the CA6 gene are present in identical positions and are all phase 0 introns, lying between two codons [23]. Furthermore, the comparison also revealed that the members of the α -CA gene family can be divided into two distinct groups. One group consists of the cytoplasmic (CA I, II, III and VII) and mitochondrial (CA V) members, which have three more introns in identical positions and phases. The second group consists of the membrane-bound

(CA IV and IX) and secreted (CA VI) members, which have two more introns in identical positions and phases. The two introns correspond to introns 2 and 5 of the CA6 gene. Interestingly, CA6 is more closely related to CA9 than to CA4 because the positions of the first introns are also conserved between the CA6 and CA9 genes, whereas the CA4 gene has two introns in the adjacent area. Between the first and second groups, the positions of the introns corresponding to intron 5 of the CA6 gene are similar, but not identical because the introns in the first group are phase 0 introns, whereas the introns in the second group are phase 1 introns, lying between the first two nucleotides of a codon [23].

The C-terminal extensions of CA VI and CA IX are encoded by additional exons that do not show significant sequence similarity between the two genes. Although there is one exon (exon 8) in the CA6 gene, there are two exons in the CA9 gene. The first CA9 exon encodes the transmembrane domain; the second encodes the cytoplasmic tail.

DISCUSSION

Two distinct groups within the α -CA gene family

We have defined the gene structure of the bovine CA6 gene encoding CA VI, the secreted isoenzyme of the α -CA family (Figures 1 and 2). The structure of the human CA6 gene has not been reported. Because of the high similarity between the cDNA sequences of the bovine and human CA VI [5] and the conservation between the exon/intron structures of the bovine CA6 gene and other human CA genes, it would be a surprise if the structure of the human CA6 gene were not similar to that of the

bovine gene. The comparison of the known gene structures clearly divided the members of the CA family with known gene structures into two distinct groups: CA1–3, 5 and 7 being the first group and CA4, 6 and 9 being the second (Figure 3). In addition, the comparison revealed for the first time a closer evolutionary relationship of CA6 to CA9 than to CA4 in the second group. The second group might also include the CA12 gene encoding a second tumour-associated isoenzyme CA XII, which is closely related to the CA6 and CA9 genes in the CA domains [8,9]. In conjunction with the phylogenetic analysis [2], the members of the second group represent the oldest mammalian CAs evolutionarily.

The closer relationship between CA1–3, 5 and 7 in the first group is also reflected at the chromosomal locations of these genes. The CA1–3 genes are clustered on chromosome 8q22 [24]. The CA5 and 7 genes are in different regions of another chromosome, 16q24.3 and 16q21–23 respectively [21,22]. In contrast, the three genes of the second group are not on chromosomes 8 and 16. The CA4 and 9 genes map to chromosome 17q23 and 17q21.2 respectively [9,20], whereas the human CA6 gene maps to chromosome 1p36.22–33 [25].

The closer relationship between CA IV, VI and IX in the second group is also reflected at the levels of amino acid sequences. For example, the positions of two cysteine residues in CA IV and IX are identical with those in CA VI (indicated by # in Figure 3) that form an internal disulphide bond, a unique feature that was previously known only in CA VI. It is possible that in CA IV and IX these two cysteine residues also form a similar internal disulphide bond. Considering that the CA domains of the three enzymes are extracellular, the folding or the structure of these domains could also be similar.

There are two competitive hypotheses concerning the origin of introns, 'intron early' and 'intron late' [26–28]. Our observation of the identical positions and phases for the three introns corresponding to introns 3, 4 and 6 of the CA6 gene in the known α -CA gene structures favours the 'intron early' hypothesis. The presence/absence and phase change of other introns could be the results of selective intron loss and base slippage through evolution. Our conclusion seems to be different from that of Hewett-Emmett and Tashian [2], who found support for the 'intron late' hypothesis. However, in their analysis of intron positions, they included the α -CA genes not only from human (CA1–5 and 7) but also from other species (*Drosophila*, *Caenorhabditis* and *Chlamydomonas*).

Regulation of the CA genes

The 5' flanking region of the CA6 gene contains the sequence CATAAA, which is similar to the consensus sequence of the TATA box (Figure 2). The CATAAA sequence is also present in the CA3 gene [19]; a similar sequence, CATAAG, is in the CA1 gene [17]. Although a typical TATA box (TATAAA) is present in both CA2 [29] and CA4 [20] genes, it is not present in other CA genes (CA5, 7 and 9) [10,21,22]. Similarly, binding sites for several transcription factors such as SP1 are present only in a subset of the CA genes. The presence or absence of the TATA box and certain transcription-factor-binding sites in the different CA genes is not surprising because these genes all have distinct tissue expression patterns. Further investigation is required for the identification of control elements (promoter, enhancer and repressor) and cognate transcription factors that are responsible for regulation of the differential expression of the CA genes.

Many potential transcription-factor-binding sites were identified from the Transfac database in the 5' flanking region of the CA6 gene; two of those that show 100% identity with the

consensus sequences for binding NF-E2 p45 and MZF1 are shown (Figure 2). NF-E2 (nuclear factor-erythroid 2) is an obligate heterodimer between haemopoietic-specific 45 kDa (p45) and widely expressed 18 kDa (p18) subunits, both members of the basic region-leucine zipper (bZip) family of transcription factors [15]. MZF1 (myeloid zinc finger 1) recognizes two distinct consensus sequences and has a critical role in haemopoiesis by modulating the expression of genes involved in this process [16]. The functional significance of the NF-E2 p45 and MZF1 binding sites in the CA6 gene remains to be determined. Nevertheless, the 100% identities of the sites in the CA6 gene with the consensus sequences suggest the following possibilities. First, these two transcription factors might also regulate the expression of genes in the salivary gland in addition to those in the haemopoietic cells. Secondly, the salivary gland might contain transcription factors that are similar to NF-E2 p45 and MZF1. Thirdly, the CA6 gene might also be expressed in the haemopoietic cells, joining other members of the CA family known to have a role in haemopoiesis. For example, CA I is the second most abundant protein in human erythrocytes after globin and is a very early marker for erythroid differentiation [17].

Functions of the CA6-related genes

The secreted CA VI has been proposed to have a role in pH regulation in saliva and to have a protective effect in the oesophagus and the stomach [1,6]. The comparison of the gene structures in this report has shown that CA VI is closely related to the two membrane-bound isoenzymes (CA IV and IX), especially that (CA IX) overexpressed in certain cancers. This relationship raises two interesting possibilities. One is that more CA genes, especially those associated with cancer, will be discovered and will show close relationships to the CA6 gene. The other is that the expression of the CA6 gene itself might be altered in certain diseases and cancer, especially those associated with the salivary gland. For example, Thatcher et al. [7] suggested that CA VI might be involved in taste bud development and normal taste function because decreased levels of gustin/CA VI in human salivary samples were correlated with both loss of taste (hypogeusia) and pathological changes in taste buds. Therefore further investigation into the regulation of CA6 and related genes is necessary for an understanding of the diverse functions of the members of the α -CA family.

We thank Professor Veer P. Bhavanandan for his continuous support, Brian Le for his help during the course of this work, and Michelle Benson for her critical reading of the manuscript. D.G. is a recipient of a J. William Fulbright Award under the Indo-American Cultural Exchange Program.

REFERENCES

- 1 Sly, W. S. and Hu, P. Y. (1995) *Annu. Rev. Biochem.* **64**, 375–401
- 2 Hewett-Emmett, D. and Tashian, R. E. (1996) *Mol. Phylogenet. Evol.* **5**, 50–77
- 3 Fernley, R. T., Wright, R. D. and Coghlan, J. P. (1988) *Biochemistry* **27**, 2815–2820
- 4 Aldred, P., Fu, P., Barrett, G., Penschow, J. D., Wright, R. D., Coghlan, J. P. and Fernley, R. T. (1991) *Biochemistry* **30**, 569–575
- 5 Jiang, W., Woitach, J. T. and Gupta, D. (1996) *Biochem. J.* **318**, 291–296
- 6 Parkkila, S. and Parkkila, A.-K. (1996) *Scand. J. Gastroenterol.* **31**, 305–317
- 7 Thatcher, B. J., Doherty, A. E., Orvisky, E., Martin, B. M. and Henkin, R. I. (1998) *Biochem. Biophys. Res. Commun.* **250**, 635–641
- 8 Türeci, Ö., Sahin, U., Vollmar, E., Siemer, S., Göttert, E., Seitz, G., Parkkila, A.-K., Shan, G. N., Grubb, J. H., Pfeundschoh, M. and Sly, W. S. (1998) *Proc. Natl. Acad. Sci. U.S.A.* **95**, 7608–7613
- 9 Ivanov, S. V., Kuzmin, I., Wei, M.-H., Pack, S., Geil, L., Johnson, B. E., Stanbridge, E. J. and Lerman, M. I. (1998) *Proc. Natl. Acad. Sci. U.S.A.* **95**, 12596–12601
- 10 Opavsky, R., Pastorek, S., Zelnik, V., Gibadulinova, A., Stanbridge, E. J., Zavada, J., Kettmann, R. and Pastorek, J. (1996) *Genomics* **33**, 480–487
- 11 Jiang, W., Woitach, J. T., Gupta, D. and Bhavanandan, V. P. (1998) *Biochem. Biophys. Res. Commun.* **251**, 550–556

- 12 Thompson, J. D., Higgins, D. G. and Gibson, T. J. (1994) *Nucleic Acids Res.* **22**, 4673–4680
- 13 Quandt, K., Frech, K., Karas, H., Wingender, E. and Werner, T. (1995) *Nucleic Acids Res.* **23**, 4878–4884
- 14 Mount, S. M. (1982) *Nucleic Acids Res.* **10**, 459–472
- 15 Shivdasani, R. and Orkin, S. H. (1995) *Proc. Natl. Acad. Sci. U.S.A.* **92**, 8690–8694
- 16 Perrotti, D., Melotti, P., Skorski, T., Casella, I., Peschle, C. and Calabretta, B. (1995) *Mol. Cell. Biol.* **15**, 6075–6087
- 17 Lowe, N., Brady, H. J. M., Barlow, J. H., Sowden, J. C., Edwards, M. and Butterworth, P. H. W. (1990) *Gene* **93**, 277–283
- 18 Venta, P. J., Welty, R. J., Johnson, T. M., Sly, W. S. and Tashian, R. E. (1991) *Am. J. Hum. Genet.* **49**, 1082–1090
- 19 Lloyd, J., Brownson, C., Tweedie, S., Charlton, J. and Edwards, Y. H. (1987) *Genes Dev.* **1**, 594–602
- 20 Okuyama, T., Batanian, J. R. and Sly, W. S. (1993) *Genomics* **16**, 678–684
- 21 Nagao, Y., Batanian, J. R., Clemente, M. F. and Sly, W. S. (1995) *Genomics* **28**, 477–484
- 22 Montgomery, J. C., Venta, P. J., Eddy, R. L., Fukushima, Y.-S., Shows, T. B. and Tashian, R. E. (1991) *Genomics* **11**, 835–848
- 23 Patthy, L. (1987) *FEBS Lett.* **214**, 1–7
- 24 Lowe, N., Edwards, Y. H., Edwards, M. and Butterworth, P. H. W. (1991) *Genomics* **10**, 882–888
- 25 Sutherland, G. R., Baker, E., Fernandez, K. E. W., Callen, D. F., Aldred, P., Coghlan, J. P., Wright, R. D. and Fernley, R. T. (1989) *Cytogenet. Cell Genet.* **50**, 149–150
- 26 Jiang, W., Kumar, J. M., Flannery, A. V. and Bond, J. S. (1997) in *Proteolysis in Cell Functions* (Hopsu-Havu, V. K., Järvinen, M. and Kirschke, H., eds.), pp. 187–195, IOS Press, Washington, DC
- 27 Gibert, W. and Glynias, M. (1993) *Gene* **135**, 137–144
- 28 Palmer, J. D. and Logsdon, J. M. (1991) *Curr. Opin. Genet. Dev.* **1**, 470–477
- 29 Shapiro, L. H., Venta, P. J. and Tashian, R. E. (1987) *Mol. Cell. Biol.* **7**, 4589–4593

Received 4 June 1999/2 August 1999; accepted 8 September 1999