Origins and Molecular Evolution of the Carbonic Anhydrase Isozymes[a]

DAVID HEWETT-EMMETT, PENELOPE J. HOPKINS, and RICHARD E. TASHIAN

Department of Human Genetics
University of Michigan Medical School
Ann Arbor, Michigan 48109

JOHN CZELUSNIAK

Department of Anatomy
Wayne State University School of Medicine
Detroit, Michigan 48201

INTRODUCTION AND HISTORICAL BACKGROUND

In the 50 years or so since carbonic anhydrase (CA) was identified as an erythrocyte enzyme activity distinct from hemoglobin,[1] our understanding has undertaken several dramatic turns as described so well in papers by Davenport,[2] Edsall,[3] and others in this volume. For those of us interested in the origins and evolution of what we now realize is a multigene family, several discoveries serve to warn us about drawing our latest conclusions about the extent of this family in too dogmatic a fashion. Thus in the early 1970s, it seemed clear from the apparent absence of more than one CA isozyme in the erythrocytes of a marsupial, the red kangaroo, chicken, and other avian species that the gene duplication that gave rise to both the high-activity, sulfonamide-sensitive CA isozyme, CA II, and the low-activity, sulfonamide-sensitive isozyme, CA I, occurred about 100 million years ago, shortly before the radiation of the placental mammals.[4] Most debate at this time centered around the similar rates of evolution of CA I and CA II. Despite its evolutionary conservatism, the role of CA I was by no means clear; the observation that erythrocyte CA I is about 90% inhibited by prevailing concentrations of chloride ions,[5] and that a homozygous deficiency of erythrocyte CA I in a Greek family is apparently asymptomatic[6] only added to this paradox.

This view of CA evolution changed radically in the late 1970s following two major discoveries. Firstly, Holmes[7,8] suggested that skeletal muscle contained a third form of carbonic anhydrase that could be described as low activity and sulfonamide resistant. This finding was confirmed by others,[9,10] and it transpired that this muscle CA, termed CA III, had in fact been purified much earlier by Scopes[11] and Noltmann's group.[12] A further twist was provided by Carter *et al.*[13] who showed that a form of CA with properties indistinguishable from skeletal muscle CA III was expressed in the livers of male rats, thus explaining earlier reports by Garg[14] and King *et al.*[15] of a sulfonamide-resistant CA in male rat livers. The second finding was that turtle red blood cells contained two forms of

[a] This work was supported by NIH grant GM 24681 (R.E.T.) and an NSF predoctoral fellowship (J.C.).

CA with properties generally similar to the CA I and CA II isozymes of mammals.[16] Sequencing studies of the low-activity, sulfonamide-sensitive CA showed that it was clearly a CA I-like isozyme[17,18] (see also below). About this time, a low-activity, sulfonamide-resistant CA was purified from chicken skeletal muscle and partial sequence studies demonstrated that it was a CA III-like isozyme.[17,19] These studies caused successive revisions to the earlier evolutionary trees.[4,17–20]

As a result of these developments during the last 7 years, together with our deepening understanding of the way genes evolve provided by studies at the DNA level, we are now prepared to recognize the existence of pseudogenes, intervening sequences and gene conversion,[21] and in addition to *expect* multigene families to be commonplace rather than exceptions. We should not, therefore, in light of the missed clues regarding CA III, be quick to dismiss the possibility that the recently described carbonic anhydrases in ovine parotid gland,[22] bovine lung membranes,[23] mammalian kidney membrane,[24] and mitochondria[25] are indeed separate isozymes. Results presented in this volume on mouse CA II genomic sequences by Venta *et al.*[26] suggest that there might be several CA II-like genes, but whether they are functional remains to be seen.

CLASSIFYING THE CARBONIC ANHYDRASES

As partial amino acid and gene sequence data on these new carbonic anhydrases become available, it will be increasingly important to classify them as CA I-, CA II-, or CA III-like isozymes or to recognize them as representatives of new isozyme classes. In the case of carbonic anhydrases from primitive vertebrates or invertebrates, we may be examining *species* that diverged before the gene duplications that resulted in the different CA gene lineages. This classification can be achieved by constructing evolutionary trees; however, this requires the use of computer algorithms.

A simpler method was described in an earlier review[19] whereby the new sequence is aligned with all of the mammalian CA isozyme sequences and compared only to those residues that are unique to a particular isozyme but invariant. Thus all mammalian CA Is have Asp at residue 8, all CA IIs have Gly and all CA IIIs have Ala. At present, these unique, invariant residues comprise 20 for CA I (based on human, chimpanzee, orangutan, rhesus macaque, rabbit, ox, and horse), 23 for CA II (based on human, rhesus macaque, cebus monkey, rabbit, ox, sheep, and horse) and 32 for CA III (based on human, gorilla, and ox). These residues are shown in TABLE 1. In TABLE 2, we show how some recently determined partial and complete amino acid sequences match these unique invariant residues. Clearly the skeletal muscle CAs from horse (H. F. Deutsch, personal communication) and chicken are CA III-like, and the mouse CA sequence inferred from the cDNA sequence derived from anemic spleen mRNA[27] is CA II-like. The data on the wallaby[28] and chicken erythrocyte high-activity CA isozymes[29] are too limited to be classified by this method, but when evolutionary trees are constructed (see below), they are indeed CA II-like.

We can anticipate that the number of unique invariant residues for CA I and CA II will drop as more mammalian CA isozymes are sequenced. The mouse CA II sequence has, for instance, decreased the number of such unique invariant CA II residues from 23 to 16. However, in the case of CA III (so few of which have been characterized in more than one species), inclusion of the horse CA III sequence actually increases the number of such sites from 32 to 40. The amended

TABLE 1. Unique and Conserved Residues in Mammalian CA Isozymes and Demonstration that Anemic Mouse Spleen cDNA Encodes a CA II-Like Isozyme[a]

	1	3	4	8	14	19	20	26	27	32	33	34	36	38	54	56	58	62	*64	*65	*67
CA I	Ala		Asp	Asp			Tyr	Asn				Lys	Ser				Glu	Val			
CA II	-	*His*	*His*	Gly		*Asp*			*Arg*							*Ser*	Arg				*Asn*
CA III	-		Glu	Ala	Asp					Glu	Leu			Ile	Gly		Thr		Lys	Thr	Arg
Mouse "II"	-	*His*	*His*	Ser	Glu	*Asp*	Phe	Asp	*Arg*	Asp	Ile	Asp	Ala	Ala	Ala	*Ser*	Ser	Asn	His	Ser	*Asn*
Match		II	II			II			II							II					II

	*69	71	73	74	75	76	78	80	87	*91	95	112	125	[b]126	129	130	132	136	138	147	153
CA I		Glu		Asp	Asn				*Ser*					Ala				Ser		Leu	
CA II	*Glu*					Lys			Thr			*Lys*	*Thr*	-	*Gly*	*Asp*	*Gly*		*Pro*		
CA III	Val		Thr	Tyr			Met	Arg	Pro	Arg	Leu			-	Asn						Glu
Mouse "II"	*Glu*	Asp	Ser	Gln	Asp	Asn	Val	Lys	*Ser*	Ile	Phe	*Lys*	*Thr*	-	*Gly*	*Asp*	*Gly*	Gln	*Pro*	Phe	Ala
Match	II								I			II	II		II	II	II		II		

	155	156	157	166	173	175	187	198	*200	221	225	226	230	231	232	235	241	251	255
CA I							Ser		His		Gln								Thr
CA II				*Ser*	Ser	Asp							*Asn*	*Phe*					Gln
CA III	Gly	Glu	Phe		Glu			Phe		Asp		Leu			Ser	Asn	Leu	Ile	
Mouse "II"	Gln	Gly	Leu	*Ser*	Arg	Ala	Gly	Leu	Thr	Glu	His	Phe	*Asn*	*Phe*	Asn	Gly	Met	Leu	Lys
Match				II									II	II					

[a] Unique, conserved residues are those found only in one isozyme and that are invariant in that isozyme. Based on seven sequences for CA I (human, chimp, orangutan, rhesus macaque, rabbit, ox, and horse), seven for CA II (human, rhesus macaque, cebus monkey, rabbit, ox, and horse) and three for CA III (human, gorilla, and ox). Sequence numbering based on CA I. Sequences used are referenced in Tashian *et al.*[19] except mouse cDNA,[27] horse CA II,[42] and ox CA III.[20] - indicates residue not present in CA II or CA III. * indicates active-site residues.

[b] Position 126 is included because of its presence only in CA I. It is not, however, unique and invariant as horse CA I is Ser.

TABLE 2. Judging Homologous Relationships of New Carbonic Anhydrases by Matches with Residues Unique and Invariant in CA Isozymes I, II, and III[a]

Source and Putative Isozyme	Residues	CA I	CA II	CA III
Chicken "III"	112	0/8	1/7	11/16
Horse "III"	227	1/16	1/20	26/27
Mouse "II"	259	1/20	16/23	0/32
Chicken "II"	112	1/11	6/12	2/22
Wallaby "II"	53	0/1	1/4	0/7
Turtle "I"	233	10/15	5/18	1/27

[a] Sequences are from the following sources: Chicken "III" (D. Hewett-Emmett, unpublished); Horse "III" (H. F. Deutsch, personal communication); Mouse "II;"[27] Chicken "II" (C. M. Yoshihara and J. B. Dodgson, personal communication and this volume;[29] and D. Hewett-Emmett, cited in ref. 19); Wallaby "II;"[28] Turtle "I."[18]

table (incorporating the mouse CA II and horse CA III data) that we recommend for future use is shown in TABLE 3. There are now 18 unique conserved CA I residues in mammals, 15 CA II residues, and 40 CA III residues.

If the lung membrane CA turns out to represent CA IV,[23] we might anticipate that a comparison of its sequence with these unique invariant residues will show relatively few matches, and those that do match may be scattered fairly evenly between CA I, CA II, and CA III. If the mitochondrial CA[25] turns out to be different from CA II, but still CA II-like, we might anticipate it sharing some unique invariant residues with CA II and rather fewer with CA I or CA III.

Clearly this method is useful for giving a rapid preliminary glimpse of the evolutionary relationship of newly characterized CAs, but the construction of evolutionary trees is preferred where sequence data are limited.

THE ACTIVE SITE OF THE CARBONIC ANHYDRASES

The determination of the 3-D structures of CA I and CA II,[30] and the subsequent refinement of the x-ray diffraction data described at this meeting by Kannan *et al.*[31] have shown that the two structures are very similar. Since CA III is almost equally as divergent from CA I and CA II as they are from each other, it seems very likely that the 3-D structure and active site of CA III will be generally very similar to those of CA I and CA II. From the x-ray work and from the active-site studies in a number of laboratories,[17,19,30] it has been possible to identify 30 active-site residues. The amino acid present at these sites in the better-characterized CA I, CA II, and CA III isozymes is shown in TABLE 4. Sixteen of the 30 residues are invariant in all three isozymes, while certain of the other sites fall into the category of unique invariant residues described earlier. Particularly interesting is residue 200, which is His in CA I but Thr in all CA II and CA III sequences. At the entrance to the active site, a cluster of five residues are found to be unique and invariant in CA III (Lys-64, Thr-65, Arg-67, Val-69, and Arg-91). These include the three basic residues that may relate to the weak acid phosphatase activity of CA III,[32] and its low esterase activity.[7,19,32,33] In addition, Val-69 represents the

TABLE 3. Unique and Invariant Residues for the Different Mammalian CA Isozymes Revised from TABLE 1 after Inclusion of Mouse CA II and Horse CA III Sequences[a]

	1	3	4	8	14	18	19	26	27	32	33	34	36	38	54	56	58	62	*64	*65	*67	*69
CA I	Ala		Asp	Asp				Asn				Lys	Ser				Glu	Val				
CA II	-	His	His				Asp		Arg							Ser					Asn	Glu
CA III	-		Glu	Ala	Asp	Glu				Glu	Leu			Ile	Gly		Thr		Lys	Thr	Arg	Val
	71	73	74	75	78	80	87	*91	95	112	125	[b]126	129	130	135	136	138	147	153	155	156	157
CA I	Glu		Asp	Asn								Ala				Ser		Leu				
CA II										Lys	Thr	-	Gly	Asp			Pro					
CA III		Thr	Tyr		Met	Arg	Pro	Arg	Leu			-	Asn		Leu				Glu	Gly	Glu	Phe
	166	173	183	187	189	198	*200	*207	208	210	*211	221	225	226	230	231	232	235	241	255		
CA I				Ser			His						Gln							Thr		
CA II	Ser														Asn	Phe						
CA III	Lys	Glu	Cys		Arg	Phe		Ile	Val	Leu	Leu	Asp		Leu			Ser	Asn	Leu			

[a] These are unique and invariant residues in mammalian CA isozymes based on seven sequences for CA I (as for TABLE 1), eight sequences for CA II (as for TABLE 1 plus mouse CA II[27]) and four sequences for CA III (as for TABLE 1 plus horse CA III of H. F. Deutsch, personal communication). * indicates active-site residues. - indicates homologous residue not present in CA II or CA III. Sequence numbering based on CA I.

[b] Position 126 is included because of its presence only in CA I. It is not, however, unique and invariant as horse CA I is Ser.

substitution of an aliphatic residue in the hydrophilic side of the active site, while Arg-91 represents the substitution of a hydrophilic residue in the hydrophobic side of the active site. These substitutions may account for the lowered sulfonamide binding of CA III.[9,19,33] Of these five residues, only Val-69 has been fully characterized in chicken CA III, but there are strong indications that 67 and 91 are Arg or Lys based on tryptic cleavage at these sites (D. Hewett-Emmett, unpublished data).

Residues 67 and 69 are also unique and invariant in CA II (Asn and Glu, respectively), and until the horse CA I sequence was determined,[34] were also unique and invariant in CA I (His and Asn, respectively). It is worth noting that the replacements in horse (Gln-67, Lys-69) may be responsible for the markedly lower CO_2 hydratase activity of the horse isozyme relative to other mammalian CA Is.[35]

BUILDING AND ROOTING EVOLUTIONARY TREES

Building evolutionary trees from amino acid and nucleic acid sequence data has been carried out with increasing sophistication over the 16 years or so since Fitch and Margoliash[36] made their pioneering efforts. An attractive approach is the maximum parsimony method, which is based on the assumptions that evolution is mainly a process of divergence rather than convergence (parallelism), and that evolution has taken the most economical course.[37] This method does not, however, require the assumption that evolutionary rates are equal in all lineages. In recent years, certain complications have become apparent with the realization that gene conversion (correction of one gene sequence by its neighbor) and the expansion and contraction of gene copy number within clusters of similar genes may not be uncommon events.[21] Until we obtain strong evidence to indicate otherwise, we will assume that these complications are not occurring.

The computer algorithms used in the maximum parsimony approach are well described elsewhere[37] and they have already been used in studying the carbonic anhydrase isozymes.[4,17,20] The computer algorithms provide only the *network* of lowest nucleotide replacement length; to convert this network into a *tree* with a time dimension, a "root" must be subjectively placed on one of the branches. This problem is illustrated in FIGURE 1. In the case of the carbonic anhydrase isozymes, we are presently dealing with three isozyme lineages and have no "primitive" carbonic anhydrase with which to root the tree. It was for this purpose that we chose to characterize spinach carbonic anhydrase. However, as

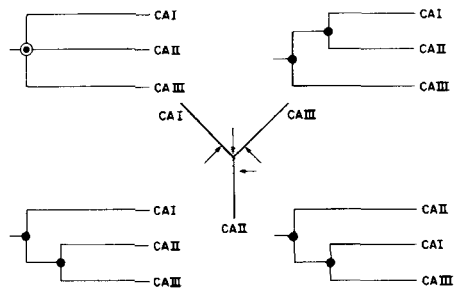


FIGURE 1. Four alternative trees generated from a single network (*center*). Placing the "root" on the three CA linkage branches generates trees in which the gene duplication occur at different times (dichotomies). The fourth placement results in a trichotomy (*top, left*) which represents the situation where the *two gene duplications* occurred very close together in time.

TABLE 4. Residues Located, or Postulated to Occur, in the Active Site Regions of the Carbonic Anhydrase Isozymes, CA I, CA II, and CA III[a]

Isozymes	Residue Number														
	7	29	61	64	65	67	69	91	92	94	96	106	107	117	119
Human I	Tyr	Ser	Asn	His	Ser	His	Asn	Phe	Gln	His	His	Glu	His	Glu	His
Rhesus I	Tyr	Ser	Asn	His	Ser	His	Asn	Phe	Gln	His	His	Glu	His	Glu	His
Ox I	Tyr	Ser	Asn	His	Ser	His	Asn	Phe	Gln	His	His	Glu	His	Glu	His
Horse I	Tyr	Ser	Asn	His	Ser	Gln	Lys	Val	Gln	His	His	Glu	His	Glu	His
Turtle I	Tyr	Ser	Asn	His	Ser	His	Asn	His	Gln	His	His	Glu	His	Glu	His
Human II	Tyr	Ser	Asn	His	Ala	Asn	Glu	Ile	Gln	His	His	Glu	His	Glu	His
Ox II	Tyr	Ser	Asn	His	Ser	Asn	Glu	Val	Gln	His	His	Glu	His	Glu	His
Horse II	Tyr	Ser	Asx	His	Ser	Asn	Glu	Ile	Gln	His	His	Glu	His	Glu	His
Rabbit II	Tyr	Ser	Asn	His	Ser	Asn	Glu	Ile	Gln	His	His	Glu	His	Glu	His
Mouse II	Tyr	Ser	Asn	His	Ser	Asn	Glu	Ile	Gln	His	His	Glu	His	Glu	His
Chicken II		Ser	Asn	His	Ser	Asn	Glu								
Human III	Tyr	Ser	Asn	Lys	Thr	Arg	Val	Arg	Gln	His	His	Glu		Glu	His
Ox III	Tyr	Ser	Asn	Lys	Thr	Arg	Val	Arg	Gln	His	His	Glu	His	Glu	His
Horse III	Tyr	Ser	Asn				Val		Gln	His	His	Glu	His	Glu	His
Chicken III		Ser				X	Val	X	Gln	His	His				

		121	131	141	143	145	192	194	199	200	201	202	204	206	207	211
Human	I	Ala	Leu	Leu	Val	Gly	Trp	Tyr	Thr	His	Pro	Pro	Tyr	Ser	Val	Ile
Rhesus	I	Val	Leu	Leu	Val	Gly	Trp	Tyr	Thr	His	Pro	Pro	Tyr	Ser	Val	Ile
Ox	I	Val	Phe	Leu	Leu	Gly	Trp	Tyr	Thr	His	Pro	Pro	Leu	Ser	Val	Ile
Horse	I	Val	Phe	Leu	Ile	Gly	Trp	Tyr	Thr	His	Pro	Pro	Tyr	Ser	Val	Val
Turtle	I	Val	Ile	Leu	Leu			Tyr	Thr	His	Pro	Pro	Ser	Ser	Val	Ile
Human	II	Val	Phe	Leu	Val	Gly	Trp	Tyr	Thr	Thr	Pro	Pro	Leu	Cys	Val	Val
Ox	II	Val	Phe	Leu	Val	Gly	Trp	Tyr	Thr	Thr	Pro	Pro	Leu	Ser	Val	Val
Horse	II	Val	Phe	Leu	Val	Gly	Trp	Tyr	Thr	Thr	Pro	Pro	Leu	Cys	Val	Val
Rabbit	II	Val	Phe	Leu	Val	Gly	Trp	Tyr	Thr	Thr	Pro	Pro	Leu	Cys	Val	Val
Mouse	II	Val	Phe	Leu	Val	Gly	Trp	Tyr	Thr	Thr	Pro	Pro	Leu	Cys	Val	Val
Human	III	Val	Phe				*TRP*	Tyr	*THR*	*THR*	Pro	Pro				
Ox	III	Val	Ile	Ile	Val	Gly	Trp	Tyr	Thr	Thr	Pro	Pro	Cys	Cys	Ile	Leu
Horse	III	Val	Tyr				Trp	Tyr	Thr	Thr	Pro	Pro	Glu	Cys	Ile	Leu

[a] Residues are numbered according to CA I. Boxed residues are either invariant in all three isozymes or are unique to and invariant in a particular isozyme (e.g., Val-69 in CA III or His-200 in CA I). Sequences are referenced in Tashian *et al.*[19] except ox CA III,[20] mouse CA II,[27] horse CA II,[42] chicken CA II (Yoshihara *et al.*,[29] D. Hewett-Emmett, unpublished) and horse CA III (H. F. Deutsch, personal communication). Capitalized and italicized residues are tentative. X indicates residue is either Lys or Arg based on tryptic cleavage at this position (D. Hewett-Emmett, unpublished).

described below, it is apparently not suitable for rooting the tree as it has diverged so radically from the animal carbonic anhydrases that it is difficult to align.

Therefore, we will assume in the absence of any other strong evidence for any one of the three dicotomous arrangements in FIGURE 1, that the gene duplications that led to the CA I, II, and III lineages occurred very close together in time and that the tricotomy is correct.

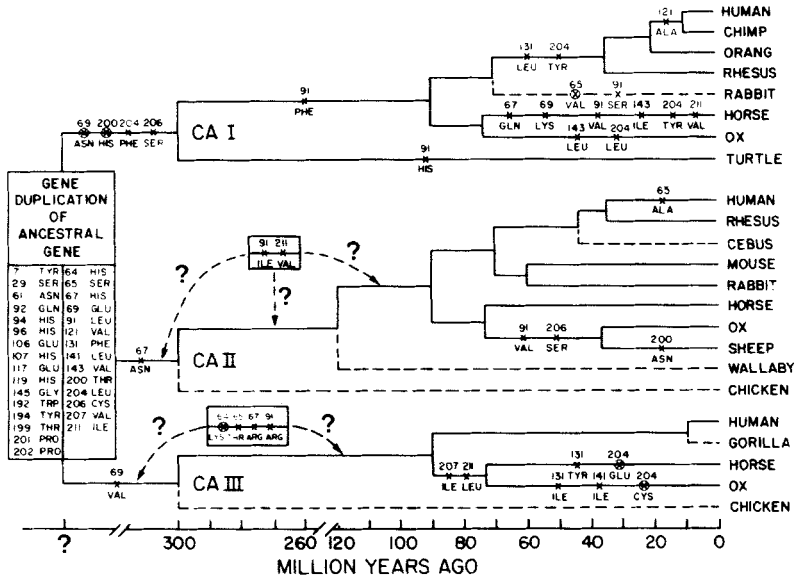


FIGURE 2. Biological phylogenetic tree constructed using the available complete and partial CA sequences listed in TABLE 5. It has a nucleotide replacement length of 742. Changes at the active-site residues shown in TABLE 4 are displayed on the branches. X = single nucleotide substitution; ⊗ = double mutation; ? = exact location of these substitutions are ambiguous owing to the partial sequences (branches with broken lines). The box represents the ancestral node at which the gene duplications occurred. Sequences in this box are those of the active-site of the ancestral CA; those on the left are invariant; those on the right are subsequently substituted in some lineages. Time-scale is based on the fossil record.[36] Dashed lines are sequences less than 70% complete.

Two sequence data sets were used. One represented the complete alignment of 261 amino acids comprising complete and partial sequences of eight CA Is, ten CA IIs, and five CA IIIs (TABLE 5). This data set was used to construct and test many trees of which two are shown in FIGURES 2 and 3. The second data set comprised the 177 sequence positions for which the human, horse, and ox CA IIIs were completely characterized. Also included were the five CA Is and six CA IIs completely sequenced for these 177 positions. These residues are boxed in TABLE 5. Two of the trees constructed from these data are shown in FIGURES 4 and 5.

Using the complete sequence data set, we show a "biological" tree (FIGURE 2) that has species arrangements that are consistent between the isozymes and that are in accord with current ideas of mammalian evolution. The substitutions shown

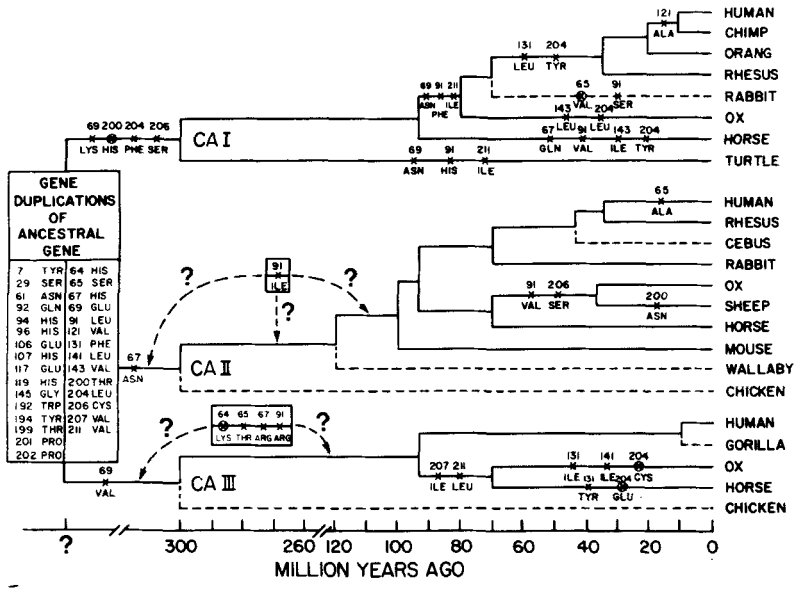


FIGURE 3. Phylogenetic tree of joint lowest nucleotide replacement length (733). Symbols are as in FIGURE 2.

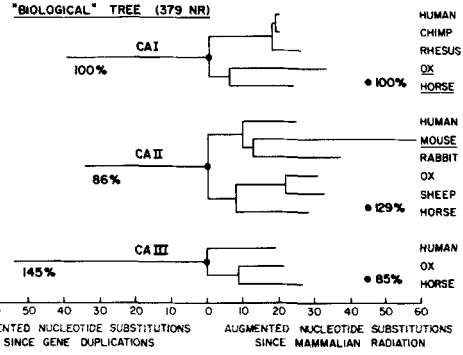
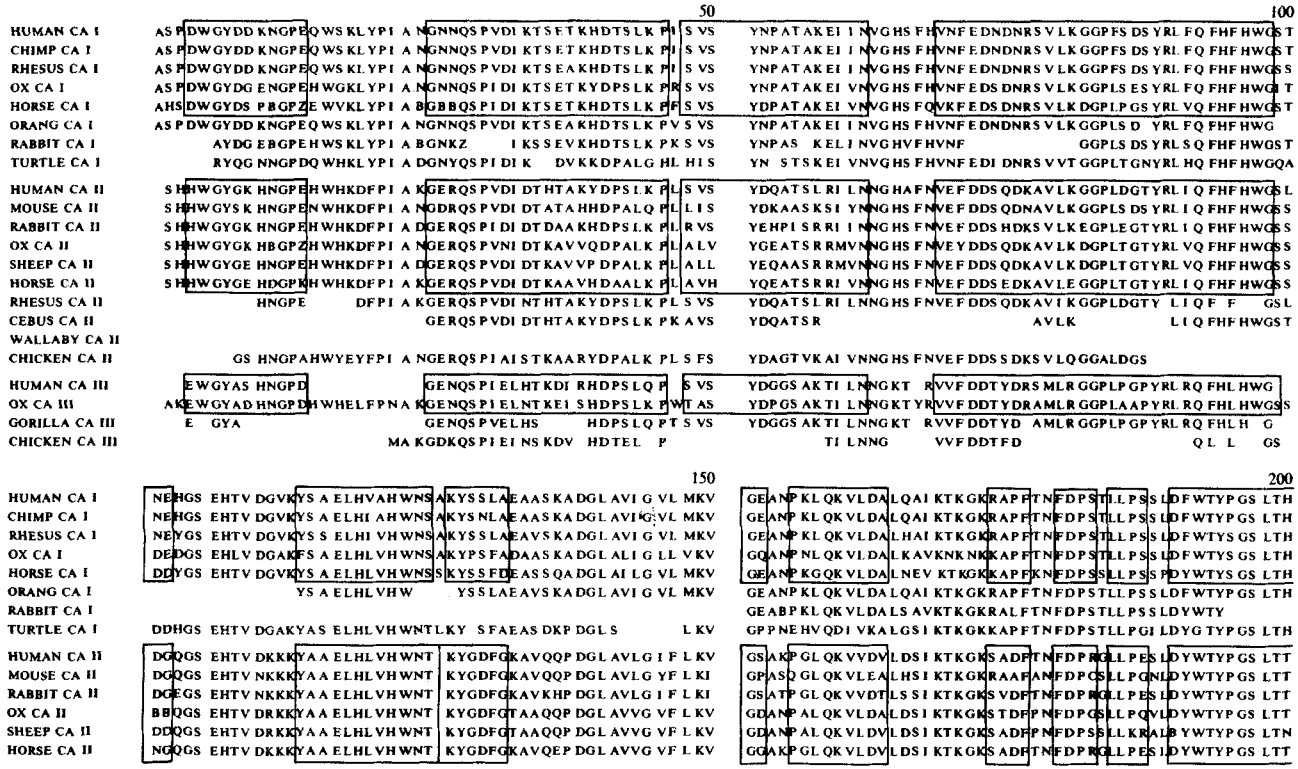
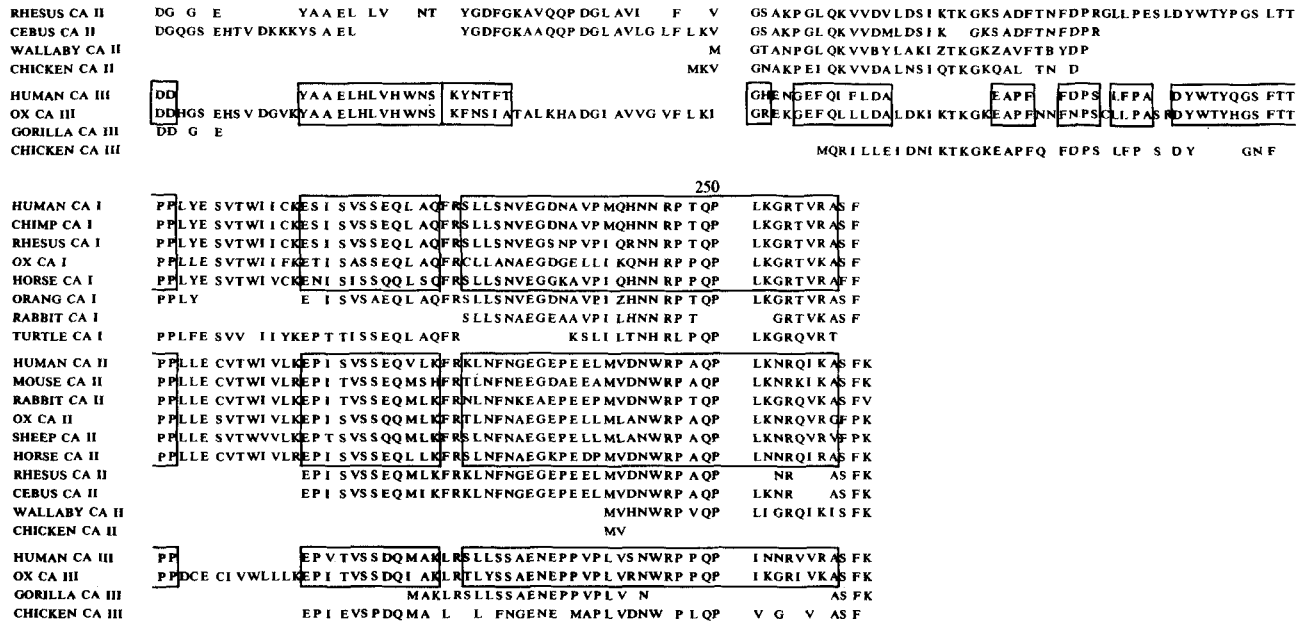


FIGURE 4. Biological phylogenetic tree constructed using the limited sequence data set boxed in TABLE 5. This tree has a nucleotide replacement length of 376. In this tree, branches are drawn to scale. ● represents the mammalian divergence. Scale to the left shows the relative evolutionary rates of the three isozymes between the gene duplications and the mammalian divergence. Scale to the right shows the relative evolutionary rates of the different isozymes within the mammals. Species underlined are those whose position in the tree differs in FIGURE 5. Different augments nucleotide substitution rates (unweighted average) are expressed as percentage of the rate for CA I.

TABLE 5. Alignment of Carbonic Anhydrase Sequences Used in the Construction of Phylogenetic Trees[a]





[a] Amino acid sequences are displayed in the one-letter code[41] and are referenced in Tashian *et al.*[19] except: ox CA III,[20] mouse CA II,[27] horse CA II,[42] wallaby CA II,[28] chicken CA II, residues 8–87 based on cDNA sequence[29] and the remainder on protein sequence.[19] Alignment numbering is based on human CA I as in Tashian *et al.*[20] Full sequence data set was used in constructing phylogenetic trees shown in FIGURES 2 and 3. Boxed residues constitute the limited data set used in constructing phylogenetic trees shown in FIGURES 4 and 5. Horse CA III (H. F. Deutsch, personal communication) was also used in constructing these trees, but the data are not shown.

on the branches are those at the 30 active-site residues. This particular tree required 742 nucleotide substitutions. Another of the many trees examined was generated by the branch-swapping algorithm[37] and had the joint lowest nucleotide replacement score (733) found. It differs from the biological tree in the arrangement of the horse and ox CA I and the mouse CA II. Our purpose in constructing the trees was not to test the species relationships themselves, which in the case of the placental mammalian radiations may well prove an impossible task, but to draw general conclusions about the way in which the isozymes have evolved with respect to the active-site and other regions of the isozymes. In this regard, the alternative trees do lead to the same conclusions: in the mammals, the active-site residues of CA II have been evolving the least, followed by CA III and CA I. Particularly notable in CA I are the similarities of the active-site residues of turtle CA I to those of the ancestral amniote CA I, in contrast to horse CA I, which has accepted several radical substitutions (e.g., Gln-67 and Lys-69) that, as noted

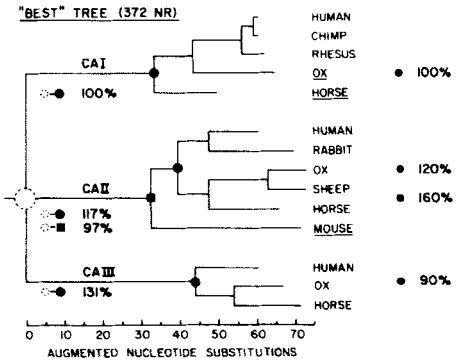


FIGURE 5. Phylogenetic tree of lowest nucleotide replacement length (369) using the limited sequence data set. Branches are drawn to scale. In this tree, the mammalian divergence is ambiguous owing to the position of mouse CA II. If this is an orthologous gene (i.e., represents the true mouse CA II and not a different but close relative), then the true mammalian radiation cannot be compared since comparable mouse data for CA I and CA III are not available. Thus rates before (○–●) and after (●) the node make the fairest comparisons. Once again augmented rates (unweighted average) are expressed as percentage of the rate for CA I. Species underlined are those whose position differs from FIGURE 4.

earlier, may account for its low CO_2 hydratase activity.[35] Since the gene duplications, the only major active-site change in the CA II lineage is the substitution of Asn for His at residue 67. This change presumably resulted in the unmasking of His-64 as discussed by Tashian *et al.*[19]

The limited sequence data set (177 amino acids and species boxed in TABLE 5) allows us to make more precise estimates of the evolutionary rates of the three isozymes. Once again, trees were constructed and the two that are comparable to those in FIGURES 2 and 3 are shown in FIGURES 4 and 5. This time, the branches of the trees are drawn to scale, the length of each branch being proportional to the augmented number of nucleotide substitutions occurring on it. These two trees and six others (not shown) differ only in the position of the ox, horse, and mouse sequences. The trees range in length from 369 to 382 nucleotide substitutions. The

rates of nucleotide substitution before and after the mammalian radiation show the same general pattern whichever of the eight trees is chosen. Notably, before the mammalian radiation, the substitution rate is CA III > CA I > CA II. After the mammalian radiation, the substitution rate is reversed, i.e., CA III < CA I < CA II. This pattern holds true whether the method of calculation involves averaging all branches radiating from the ancestral node equally (unweighted) or by weighted averaging of all the branches.

The pattern emerges, therefore, of an ancestral CA with an active site similar to that found in the CA II isozymes of all species. Subsequent to the gene duplications, both the CA I and CA III lineages fixed important substitutions in their active sites that are presumably responsible for their different properties. Since the mammalian radiation, CA I and, to a lesser extent CA III, have continued to fix substitutions in their active sites. Considering the molecule as a whole, however, a different pattern emerges. After the gene duplications, the CA III lineage in particular evolved quite rapidly. More recently, however, since the mammalian radiation, CA III has been the most conserved, and CA II the least conserved, of the three CA isozymes.

EXTERNAL CONSERVED REGIONS

One particularly intriguing aspect of the carbonic anhydrase isozymes is their true physiological role and the puzzling need for so many different isozymes. The relatively similar evolutionary conservatism of CA I, CA II, and CA III discussed earlier leaves little doubt that all three are being selected for and do play important physiological roles; but in the case of CA I and CA III, it remains to be determined what they are. In light of the fact, described earlier, that while the active site of CA II has remained relatively unchanged, and presumably represents a close approximation of the active site of the ancestral CA, the remainder of CA II has in fact been evolving more rapidly than CA I and considerably more rapidly than CA III. One possible explanation is that CA I and CA III interact with other molecules and that external (nonactive-site) regions are under selection. The mammalian CA isozyme sequences documented in TABLE 5 were therefore analyzed to seek candidates for such conserved regions, and two were found. Residues 18–37 of CA I are considerably more conserved in evolution than the homologous regions of CA II and CA III (TABLE 6). The only variation occurs at Asn-27 (Lys in rabbit), Val-31 (Ile in ox and rabbit) and Thr-35 (Ser in rabbit). In CA II, variation occurs at seven of these 20 residues and in CA III at eight residues. Residue 31 is the site of the polymorphic allelic variation in human CA III (Ile ↔ Val), and this may represent selectively neutral variation.[39] Residues 231–250 of CA III are remarkably conserved compared to the homologous residues in CA I and CA II (TABLE 7). The only variation is at Arg-243 (Ser in human), while CA I varies at 11 of the 20 residues and CA II at nine residues.

One common feature of these two regions is that they are external, and they wind around active-site regions containing residues specific to CA I and CA III. Thus, residues 18–37 (conserved in CA I) are close to His-200, which is thought to be responsible for some of the different kinetic properties of CA I. Residues 231–250 (conserved in CA III) wind around the region containing Lys-64 and behind residues 18–37. Whether or not the conservation of these regions proves to be connected to interactions of CA I and CA III with other molecules remains to be determined. An alternative explanation for such conserved regions is that selection is operating at the DNA or mRNA level.

TABLE 6. Conservative Region of Mammalian CA I[a]

	Residue Number																		
	18		20					25				30					35		37
CA I	Lys	Leu	Tyr	Pro	Ile	Ala	Asn	Gly	Asn	Asn Lys	Gln	X Ser	Pro	* Val Ile	Asp	Ile	Lys	Thr Ser	Ser	Glu
CA II	Lys	Asp	Phe	Pro	Ile	Ala	Lys Asn Asp	Gly	Glu Asp	Arg	Gln	X Ser	Pro	* Val Ile	Asp Asn	Ile	Asp Asn	Thr	Lys His Ala Asp	Thr Ala
CA III	Glu	Leu Phe	Phe Tyr	Pro	Lys Ile	Ala	Lys	Gly	Glu Asp	Asn	Gln	X Ser	Pro	* Ile Val	Glu	Leu	His Asn	Thr Ser	Lys	Asp Glu

[a] Based on mammalian CA sequences used in TABLE 1 and also mouse CA II[27] and horse CA III (H. F. Deutsch, personal communication). Residues are numbered according to CA I. * = site of Ile/Val silent polymorphism in Human CA III.[39] X = active-site residue.

TABLE 7. Conservative Region of Mammalian CA III[a]

	231				235					240						245				250
CA III:	Ser	Ser	Ala	Glu	Asn	Glu	Pro	Pro	Val	Pro	Leu	Val	Arg Ser	Asn	Trp	Arg	Pro	Pro	Gln	Pro
CA I:	Ser Ala	Asn	Val Ala	Glu	Gly	Asp Ser Gly Glu	Asn Gly Lys Ala	Ala Pro Glu	Val Leu	Pro Leu	Met Ile	Gln Lys Leu	His Arg Gln	Asn	Asn	Arg	Pro	Thr Pro	Gln	Pro
CA II:	Phe	Asn	Gly Ala Glu Lys	Glu	Gly Ala	Glu Asp Lys	Pro Ala	Glu	Glu Leu Asp	Leu Pro Ala	Met	Val Leu	Asp Ala	Asn	Trp	Arg	Pro	Ala Thr	Gln	Pro

[a] Based on mammalian CA sequences used in TABLE 1 and also mouse CA II[27] and horse CA III (H. F. Deutsch, personal communication). Numbering of residues according to CA I.

SPINACH CARBONIC ANHYDRASE

During the discussion of the evolutionary trees, it was assumed that CA I, CA II, and CA III resulted from two gene duplications occurring close together in time. However, the determination of the structure of a carbonic anhydrase in a lineage that diverged before these gene duplications could provide some evidence about these events and enable us to "root" the trees. Spinach carbonic anhydrase might provide such information. It was purified according to the method of Kandel *et al.*,[38] and one of the fragments obtained by chemically cleaving the enzyme at methionine residues with cyanogen bromide was sequenced. The sequence obtained showed no clear homology with the CA I, CA II, or CA III isozymes; however, it was decided to test this more rigorously by aligning this 20-residue segment of sequence (with two unidentified residues) with human CA I by sliding it along the 260-residue sequence and determining the number of sequence matches for each of these 241 comparisons. The result is shown in TABLE 8. The best match was five identities, and this occurred three times as shown in TABLE 9. Of these alignments, only that between spinach residues 1–15 and human CA I residues 77–91 shows significant evidence of homology ($p = 0.1\%$) using the Moore and Goodman test.[40] Clearly, more sequence data on spinach CA need to be determined to align it correctly with the vertebrate CA isozyme sequences, but it appears to be too distantly related to be useful in "rooting" the trees. Perhaps carbonic anhydrase from a primitive chordate (e.g. tunicate) or an invertebrate (e.g., sea urchin or *Drosophila*) might prove more suitable for this purpose.

SUMMARY

1. Work on membrane-bound and subcellular forms of CA at the protein level, and the possibility of multiple forms of the mouse CA II gene at the DNA level, indicate that CA may represent an extensive multigene family.
2. A method for classifying newly sequenced CA molecules, or genes encoding them, is discussed.
3. Phylogenetic trees based on the existing sequence data are presented and discussed in terms of gene evolution.

TABLE 8. Matches of the 241 Different 20 Residue Segments of Human CA I with a 20 Residue Spinach CA Sequence[a]

Number of Matching Amino Acids	Number of 20 Residue Segments
0	72
1	102
2	39
3	19
4	6
5	3
6 or higher	0

[a] This sequence is from a fragment derived from cyanogen bromide cleavage of spinach CA purified according to Kandel *et al.*[38] It is shown in TABLE 9.

TABLE 9. Best Homologies of Spinach CA with Human CA I[a]

Spinach CA	Gly 6	Leu	Ala	Asp	Gly 10	Gly	Thr	Pro	Ser	Ala 15	Ser	Tyr	Pro	Val	Gln 20	X	Ile	X	Glu	Gly 25
Human CA I	*GLY* 77	Tyr	Asp	*ASP* 80	Lys	Asn	Gly	*PRO*	Glu 85	Gln	Trp	Ser	Lys	Leu 90	Tyr	Pro	*ILE*	Ala	Asn	*GLY* 96
Human CA I	Ser 107	Val	Leu	Lys 110	*GLY*	*GLY*	Pro	Phe	*SER* 115	Asp	*SER*	*TYR*	Arg	Leu 120	Phe	Gln	Phe	His	Phe	His 126
Human CA I	His	Thr	Val	*ASP*	*GLY*	Val	Lys	Tyr	*SER*	*ALA*	Glu	Leu	His	*VAL*	Ala	His	Trp	Asn	Ser	Ala

[a] Human CA I sequences are taken from Tashian *et al.*[20] Residues that match the spinach CA sequence are italicized and in caps.

4. The active-site residues of CA II have been more conserved in evolution than those of CA I or CA III.
5. After the gene duplications, CA III and CA I initially evolved more rapidly than CA II.
6. Since the mammalian radiation, the CA II molecule as a whole has been accepting substitutions more frequently than CA I, which in turn is evolving more rapidly than CA III.
7. These findings can be explained if external regions of CA I and CA III have been conserved in evolution owing to interactions with other molecules. Two such regions appear to be residues 18–37 in CA I and 231–250 in CA III.
8. Spinach CA was purified and a small amount of sequence data collected. The difficulty in aligning it with animal CAs suggests that a plant CA may not be suitable to shed light on the active site and character of the ancestral eukaryote CA.

ACKNOWLEDGMENTS

We thank Corinne Yoshihara, Jerry Dodgson, and Harold Deutsch for communicating and allowing us to use their sequence data before publication. We are grateful to Morris Goodman for the use of computer algorithms developed in his laboratory. We thank Ya-Shiou Yu for her help in purifying spinach CA.

REFERENCES

1. MELDRUM, N. V. & F. J. W. ROUGHTON. 1933. Carbonic anhydrase: Its preparation and properties. J. Physiol. (London) **80:** 113–142.
2. DAVENPORT, H. W. 1984. The early days of research on carbonic anhydrase. Ann. N.Y. Acad. Sci. **429:** 4–9. This volume.
3. EDSALL, J. T. 1984. Some historical perspectives on carbonic anhydrase since 1960. Ann. N.Y. Acad. Sci. **429:** 18–25. This volume.
4. TASHIAN, R. E., M. GOODMAN, R. J. TANIS, R. E. FERRELL & W. R. A. OSBORNE. 1975. Evolution of carbonic anhydrase isozymes. *In* Isozymes. Vol. **4:** 207–223. C. L. Markert, Ed. Academic Press. New York.
5. MAREN, T. H., C. S. RAYBURN & N. E. LIDDELL. 1976. Inhibition by anions of human red cell carbonic anhydrase B: Physiological and biochemical considerations. Science **191:** 469–472.
6. TASHIAN, R. E., A. G. KENDALL & N. D. CARTER. 1980. Inherited variants of human red cell carbonic anhydrases. Hemoglobin **4:** 635–651.
7. HOLMES, R. S. 1976. Mammalian carbonic anhydrase isozymes: Evidence for a third locus. J. Exp. Zool. **197:** 289–295.
8. HOLMES, R. S. 1977. Purification, molecular properties, and ontogeny of carbonic anhydrase isozymes: Evidence for A, B, and C isozymes in avian and mammalian tissues. Eur. J. Biochem. **78:** 511–520.
9. REGISTER, A. M., M. K. KOESTER & E. A. NOLTMANN. 1978. Discovery of carbonic anhydrase in rabbit skeletal muscle and evidence for its identity with "basic muscle protein." J. Biol. Chem. **253:** 4143–4152.
10. TASHIAN, R. E. 1978. Evolution of carbonic anhydrases. *In* Evolution of Protein Molecules. H. Matsubara & T. Yamanaka, Eds.: 343–360. Japan Scientific Societies Press. Tokyo.
11. SCOPES, R. K. 1966. Isolation and properties of a basic protein from skeletal muscle sarcoplasm. Biochem. J. **98:** 193–197.
12. BLACKBURN, M. N., J. M. CHIRGWIN, G. T. JAMES, T. D. KEMPE, T. F. PARSONS,

A. M. REGISTER, K. D. SCHNACKERZ & E. A. NOLTMANN. 1972. Pseudoisoenzymes of rabbit muscle phosphoglucose isomerase. J. Biol. Chem. **247:** 1170–1179.

13. CARTER, N. D., D. HEWETT-EMMETT, S. JEFFERY & R. E. TASHIAN. 1981. Testosterone-induced, sulfonamide-resistant carbonic anhydrase of rat liver is indistinguishable from skeletal muscle carbonic anhydrase III. FEBS Lett. **128:** 114–118.
14. GARG, L. C. 1974. The effect of sex hormones on rat liver carbonic anhydrase. J. Pharmacol. Exp. Ther. **189:** 557–562.
15. KING, R. W., L. C. GARG, J. HUCKSON & T. H. MAREN. 1974. The isolation and partial characterization of sulfonamide-resistant carbonic anhydrases from the liver of the male rat. Mol. Pharmacol. **10:** 335–343.
16. HALL, G. E. & R. SCHRAER. 1979. Purification and partial characterization of high- and low-activity carbonic anhydrase isoenzymes from *Malaclemys terrapin centrata*. Comp. Biochem. Physiol. **63B:** 561–567.
17. TASHIAN, R. E., D. HEWETT-EMMETT & M. GOODMAN. 1980. Evolutionary diversity in the structure and activity of carbonic anhydrase. *In* Protides of the Biological Fluids, Colloq. 28. H. Peters, Ed.: 153–156. Pergamon Press, Ltd. Oxford.
18. TASHIAN, R. E., S. K. STROUP & G. E. HALL. 1981. Primary sequence of turtle low-activity red cell carbonic anhydrase: Homology with mammalian low-activity CA I isozymes. Fed. Proc. **40:** 1677.
19. TASHIAN, R. E., D. HEWETT-EMMETT & M. GOODMAN. 1983. On the evolution and genetics of carbonic anhydrases I, II and III. *In* Isozymes: Current Topics in Biological and Medical Research. M. C. Rattazzi, J. G. Scandalios & G. S. Whitt, Eds. Vol. **7:** 79–100. Alan R. Liss, Inc. New York.
20. TASHIAN, R. E., D. HEWETT-EMMETT, S. K. STROUP, M. GOODMAN & Y.-S. L. YU. 1980. Evolution of structure and function in the carbonic anhydrase isozymes of mammals. *In* Biophysics and Physiology of CO_2. C. Bauer, G. Gros & H. Bartels, Eds.: 165–176. Springer-Verlag. Berlin, Heidelberg, New York.
21. HEWETT-EMMETT, D., P. J. VENTA & R. E. TASHIAN. 1982. Features of gene structure, organization, and expression that are providing unique insights into molecular evolution and systematics. *In* Macromolecular Sequences in Systematics and Evolutionary Biology. M. Goodman, Ed.: 357–405. Plenum Press. New York.
22. FERNLEY, R. T., R. D. WRIGHT & J. P. COGHLAN. 1979. A novel carbonic anhydrase from the ovine parotid gland. FEBS Lett. **105:** 299–302.
23. WHITNEY, P. L. & T. V. BRIGGLE. 1982. Membrane-associated carbonic anhydrase purified from bovine lung. J. Biol. Chem. **257:** 12056–12059.
24. SANYAL, G., N. I. PESSAH & T. H. MAREN. 1981. Kinetics and inhibition of membrane-bound carbonic anhydrase from canine renal cortex. Biochim. Biophys. Acta **657:** 128–137.
25. DODGSON, S. J., R. E. FØRSTER, B. T. STOREY & L. MELA. 1980. Mitochondrial carbonic anhydrase. Proc. Natl. Acad. Sci. USA **77:** 5562–5566.
26. VENTA, P. J., J. C. MONTGOMERY, K. WIEBAUER, D. HEWETT-EMMETT & R. E. TASHIAN. 1984. Organization of the mouse and human carbonic anhydrase II genes. Ann. N.Y. Acad. Sci. **429:** 309–323. This volume.
27. CURTIS, P. J., E. WITHERS, D. DEMUTH, R. WATT, P. J. VENTA & R. E. TASHIAN. 1983. The nucleotide sequence and derived amino acid sequence of cDNA coding for mouse carbonic anhydrase II. Gene **25:** 325–332.
28. JONES, G. L. & D. C. SHAW. 1982. Purification, properties, partial sequence, and evolutionary relationships of marsupial erythrocyte carbonic anhydrase. Biochim. Biophys. Acta **709:** 284–303.
29. YOSHIHARA, C. M., M. FEDERSPIEL & J. B. DODGSON. 1984. Isolation of the chicken carbonic anhydrase II gene. Ann. N.Y. Acad. Sci. **429:** 332–334. This volume.
30. NOTSTRAND, B., I. VAARA & K. K. KANNAN. 1975. Structural relationship of human erythrocyte carbonic anhydrase isozymes B and C. *In* Isozymes: Molecular Structure. C. L. Markert, Ed. Vol. **1:** 575–599. Academic Press. New York.
31. KANNAN, K. K., M. RAMANADHAM & T. A. JONES. 1984. Structure, refinement, and function of carbonic anhydrase isozymes. Refinement of human carbonic anhydrase I. Ann. N.Y. Acad. Sci. **429:** 49–60. This volume.

32. KOESTER, M. K., L. M. PULLAN & E. A. NOLTMANN. 1981. The *p*-nitrophenyl phosphatase activity of muscle carbonic anhydrase. Arch. Biochem. Biophys. **211:** 632–642.
33. SANYAL, G., E. R. SWENSON, N. I. PESSAH & T. H. MAREN. 1982. The carbon dioxide hydration activity of skeletal muscle carbonic anhydrase: Inhibition by sulfonamides and anions. Mol. Pharmacol. **22:** 211–220.
34. JABUSCH, J. R., R. P. BRAY & H. F. DEUTSCH. 1980. Sequence of the low activity equine erythrocyte carbonic anhydrase and delineation of the amino acid substitutions in various polymorphic forms. J. Biol. Chem. **255:** 9196–9204.
35. TASHIAN, R. E., R. J. TANIS & R. E. FERRELL. 1972. Comparative aspects of the primary structures and activities of mammalian carbonic anhydrases. Alfred Benzon Symp. IV. pp. 353–362.
36. FITCH, W. M. & E. MARGOLIASH. 1967. Construction of phylogenetic trees. Science **155:** 279–284.
37. GOODMAN, M. 1981. Decoding the pattern of protein evolution. Prog. Biophys. Mol. Biol. **37:** 105–164.
38. KANDEL, M., A. G. GORNALL, D. L. CYBULSKY & S. I. KANDEL. 1978. Carbonic anhydrase from spinach leaves. J. Biol. Chem. **253:** 679–685.
39. HEWETT-EMMETT, D., R. J. WELTY & R. E. TASHIAN. 1983. A widespread silent polymorphism of human carbonic anhydrase III (31 Ile ↔ Val): Implications for evolutionary genetics. Genetics **105:** 409–420.
40. MOORE, G. W. & M. GOODMAN. 1977. Alignment statistic for identifying related protein sequences. J. Mol. Evol. **9:** 121–130.
41. DAYHOFF, M. O. 1972. Atlas of protein sequence and structure, Vol. V, Suppl. 2. National Biomedical Research Foundation. Washington, D.C.
42. JABUSCH, J. R. & H. F. DEUTSCH. 1984. Sequence of the high activity equine erythrocyte carbonic anhydrase: N-terminal polymorphism (acetyl-Ser/acetyl-Thr) and homologies to similar mammalian isozymes. Biochem. Genet. **22:** 357–367.