Using Evolutionary Rates to Investigate Protein Functional Divergence and Conservation: A Case Study of the Carbonic Anhydrases

Bjarne Knudsen,*,[1] Michael M. Miyamoto,† Philip J. Laipis‡ and David N. Silverman§

**Bioinformatics Research Center, University of Aarhus, 8000 Århus C, Denmark,* †*Department of Zoology, University of Florida, Gainesville, Florida 32611-8525,* ‡*Department of Biochemistry and Molecular Biology, College of Medicine, University of Florida, Gainesville, Florida 32610-0245 and* §*Department of Pharmacology and Therapeutics, College of Medicine, University of Florida, Gainesville, Florida 32610-0267*

Manuscript received November 19, 2002
Accepted for publication April 9, 2003

ABSTRACT

Functional constraints on proteins limit their evolutionary rates at specific sites. These constraints allow for the interpretation of conserved residues and sites with a rate change as those most likely underlying the functional similarities and differences among protein subfamilies, respectively. This study describes new likelihood-ratio tests (LRTs) that complement existing ones for the identification of both conserved and rate change sites. These identifications are validated by the recovery of residues that are known from existing biochemical and structural information to be critical for the functional similarities and differences among carbonic anhydrases (CAs). In combination with this other information, these LRTs also support a unique antioxidant defense role for the puzzling CA III. As illustrated by the CAs, these LRTs, in combination with other biological evidence, offer a powerful and cost-effective approach for testing hypotheses, making predictions, and designing experiments in protein functional studies.

FUNCTIONALLY important sites and regions of biological sequences are under strong purifying selection and therefore evolve slowly according to the rule of functional constraint in molecular evolution (Kimura 1983; Li 1997). This widely acknowledged rule forms the foundation of many comparative approaches for the functional analysis of protein and nucleic acid sequences (Hughes 1999; Nei and Kumar 2000; Landgraf *et al.* 2001; Gaucher *et al.* 2002). For example, conserved amino acids are routinely interpreted as those that are most likely critical for an enzyme's function. In turn, those homologous sites with varying evolutionary rates among protein subfamilies are often interpreted as those that most likely underlie the functional differences among their groups. When integrated with biochemical, structural, and other biological information, these rate tests of functionally important sites offer a powerful and cost-effective way to generate new hypotheses and experiments for testing protein function (Golding and Dean 1998).

Protein functional divergence is related to gene duplications and major speciations (Ohno 1970; Nei *et al.* 1997; Hughes 1999; Lynch and Force 2000; Gaucher *et al.* 2002). In particular, gene duplications provide the additional coding and regulatory sequences for the origins of new protein functions and subspecializations of their ancestral roles. Correspondingly, most rate tests of functional divergence focus on the subfamilies from duplications (Gu 1999, 2001; Knudsen and Miyamoto 2001). For relatively recent events, these tests usually rely on comparisons of the nonsynonymous (replacement) to synonymous (silent) substitution rates for coding DNAs (Hughes 1999; Nei and Kumar 2000; Yang and Bielawski 2000). However, this approach is limited by the relatively rapid saturation of the synonymous substitutions by multiple hits. Thus, studies of older protein subfamilies usually rely on the replacement rates alone to identify sites that are most likely responsible for their divergent, as well as conserved, functions (Gaucher *et al.* 2002).

In these studies of protein functional divergence, replacement rates are most often evaluated on a site-by-site basis and according to whether they differ between subfamilies or are accelerated in their stems (*i.e.*, in the direct ancestral or basal-most lineage that leads to the most recent common ancestor of the group; Knudsen and Miyamoto 2001; Gaucher *et al.* 2002; Pupko and Galtier 2002). The two latter patterns of rate change, as evidenced by their subfamily differences or stem accelerations, have been referred to as type I and type II divergences, respectively (Gu 1999, 2001). The most obvious example of a type I site is a homologous position that is conserved for a particular amino acid in one subfamily, but highly variable in another. Such a site can be interpreted as functionally important in the first subfamily, but less so in the second. In contrast, the best example of a type II site is one that is fixed for radically different amino acids between the subfamilies. Here, the functional interpretation is that this site fulfills different, but equally important roles in the two subfamilies.

[1] *Corresponding author:* Department of Zoology, Box 118525, University of Florida, Gainesville, FL 32611-8525.
E-mail: knudsen@zoo.ufl.edu

Type I and II divergences belong to a series of five nested hypotheses for rate change and conserved sites (Figure 1). These related hypotheses are sequentially interconnected from the simpler to more complex by three rate parameters. New rigorous likelihood-ratio tests (LRTs) have been recently described for type I and conserved sites (H_{1a}, H_2, and H_3; Knudsen and Miyamoto 2001; Pupko and Galtier 2002) and compared to other current methods (Gu 1999, 2001; Dermitzakis and Clark 2001; Gaucher *et al.* 2001). This study complements these LRTs for type I and conserved sites by describing new ones for type II positions (H_0 and H_{1b}). As an illustration of its utility, this comprehensive series of LRTs is applied to a set of carbonic anhydrases (CAs). These LRTs recover known sites of functional importance to CAs and support a distinct biological role for their puzzling CA III.

LIKELIHOOD-RATIO TESTS FOR RATE CHANGE AND CONSERVED SITES

Type II model: In type II divergence, the evolutionary rate for a specific site is accelerated somewhere along the basal internode that connects the two subfamilies (Figure 1). This basal acceleration can be modeled by multiplying the overall rate for this internode, as estimated for the entire protein, by a factor of $a > 1$. Alternatively, this acceleration can be modeled by increasing the length of the basal internode by a positive amount. These two approaches are identical when there are no prior constraints on the basal acceleration.

In this study, this acceleration is modeled with the new factor, thereby yielding three parameters for the type I and II tests (a for the basal increase and r_I *vs.* r_{II} for the site-specific rates in subfamily I *vs.* II, respectively). These parameters are included in the likelihood calculations by extending the relevant branches in the tree by corresponding amounts (Felsenstein 1981). As Figure 1 shows, the significance of these parameters is tested in a hierarchical fashion, first for a basal acceleration and/or rate shift (H_0, H_{1a}, or H_{1b} to H_2). If there is no rate shift or basal acceleration, the site is then tested for whether it is evolving slower or faster than the overall average for the protein (H_2 to H_3).

Testing the hypotheses: The likelihood values for each site are calculated using the method of Felsenstein (1981). Maximum-likelihood (ML) scores are obtained for all hypotheses by optimizing their free parameters given phylogeny of the sequences.

The ML scores for the three rate change hypotheses [$L(H_0)$, $L(H_{1a})$, and $L(H_{1b})$] are each tested against the ML score for the hypothesis with a single rate for the entire tree [$L(H_2)$]. These evaluations are quantified by the U values of their LRTs (Knudsen and Miyamoto 2001):

$$U_0 = -2\log\frac{L(H_2)}{L(H_0)}$$

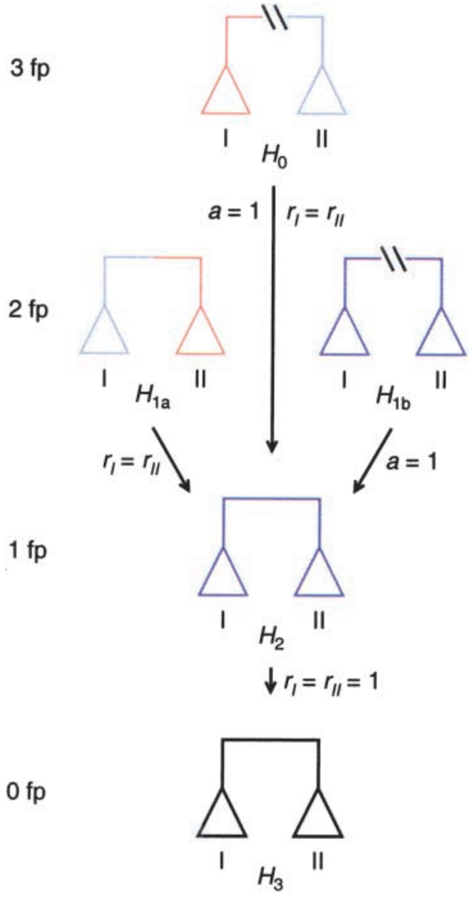

Figure 1.—Nested hypotheses and LRTs for rate change and conserved sites. Triangles represent the two protein subfamilies from a gene duplication or major speciation (for an example of subfamilial divergence due to the latter, see Gaucher *et al.* 2002). Red and blue denote a homologous position with different rates in the two subfamilies (*e.g.*, fast *vs.* slow, respectively) and therefore a type I site. A break along the basal internode, connecting the two subfamilies, signifies an accelerated rate for the position in one or both of the subfamily stems and therefore a type II site. Black and purple denote a site with a single rate that is equal or unequal to the overall average for the protein, respectively. The five hypotheses are for type I and II sites (H_0), type I or II positions (H_{1a} and H_{1b}, respectively), and those with a single rate that is unequal or equal to the overall average (H_2 *vs.* H_3). The numbers of free parameters (fp) for each hypothesis are listed to the left. These parameters include the basal acceleration factor (a) and separate rates for the two subfamilies (r_I *vs.* r_{II}). Arrows indicate which nested hypotheses are directly compared in the LRTs and which parameters are reduced from the more complex to simpler models.

$$U_{1a} = -2\log\frac{L(H_2)}{L(H_{1a})}$$

$$U_{1b} = -2\log\frac{L(H_2)}{L(H_{1b})}.$$

U_0 and U_{1b} are strongly influenced by an amino acid replacement along the basal internode. Thus, neither statistic closely follows a χ^2 or related distribution, since neither approximates a sum of squared normally distrib-

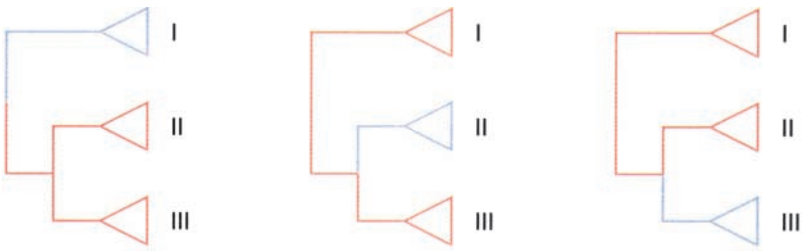

FIGURE 2.—Choosing which stem to test for a rate change when more than two subfamilies are represented. This selection is illustrated for three protein subfamilies and H_{1a} (Figure 1). There are three stems to test for a type I change given three subfamilies. The one with the greatest ML score is chosen for the LRT of H_{1a} to H_2. In the same way, single stems are selected for the LRTs of H_0 and H_{1b}.

uted values. Consequently, the 5% significance levels for U_0, U_{1a}, and U_{1b} ($U_0^{5\%}$, $U_{1a}^{5\%}$, and $U_{1b}^{5\%}$, respectively) are found with simulations (see below).

The 5% cutoffs from the simulations are compared to the observed U values (HUELSENBECK and RANNALA 1997):

$$\Delta U_0 = U_0 - U_0^{5\%}$$
$$\Delta U_{1a} = U_{1a} - U_{1a}^{5\%}$$
$$\Delta U_{1b} = U_{1b} - U_{1b}^{5\%}.$$

A positive ΔU indicates that the corresponding rate change hypothesis is a significantly better explanation of the data than is H_2. If ΔU is positive for more than one rate change hypothesis, then the one with the greatest difference is retained for the site in question. If no ΔU is positive, then the rate for this site is accepted as constant throughout the tree. The constant rate can then be tested against the average for the entire protein to determine whether this site is evolving significantly slow or fast. This test is done with the following LRT (KNUDSEN and MIYAMOTO 2001):

$$U_2 = -2 \log \frac{L(H_3)}{L(H_2)}.$$

Although U_2 approximately follows a χ^2 distribution, simulations are again recommended for the determination of its 5% cutoffs, since they are more reliable.

In this series of LRTs, H_0 is directly compared to H_2, even though H_{1a} and H_{1b} are also nested in the former hypothesis (Figure 1). Thus, alternatively, H_0 could be directly evaluated against H_{1a} and H_{1b}, rather than against H_2. However, this alternative sequence is not preferred, since their 5% cutoffs are determined with simulations. Direct testing of H_0 against H_{1a} and H_{1b} requires the specification of r_{I} and r_{II} or a in their respective simulations. By comparing instead H_0 to H_2, these extra parameterizations are avoided.

Evaluating multiple subfamilies: The type I and II LRTs specifically test for rate changes in either or both of the subfamily stems (Figure 1). By analogy, these LRTs can be extended to the stems of multiple subfamilies (Figure 2). Given multiple subfamilies, ML scores are separately calculated under H_0, H_{1a}, and H_{1b} for a type I and/or II change along each stem. The one stem with the greatest ML score for H_0, H_{1a}, or H_{1b} is retained for further testing of that rate change hypothesis with U and ΔU. As before, if ΔU is positive for more than one rate change hypothesis, then the one with the greatest difference is accepted as the best explanation for the site in question.

The main reason that only a single rate shift or basal acceleration is allowed for each site is that the number of possible rate change configurations grows exponentially with the number of sequences. The introduction of even one extra rate change would lead to increased numbers of parameters, thereby making it much more difficult to estimate them reliably given the available information for a site.

The final selection of H_0, H_{1a}, or H_{1b} for a site with more than one significant ΔU does not inflate the overall significance of the accepted rate change hypothesis, since this decision is made after the LRTs are completed. In contrast, the selection of which stem to test given more than two subfamilies forms the basis of the LRTs themselves and is therefore vulnerable to the effects of multiple testing. This source of inflated significance can be readily corrected by establishing the 5% cutoffs in the simulations with only the best ML scores for the multiple subfamilies.

Type II LRTs—power analyses and phylogenetic errors: A site with a fixed amino acid difference between two subfamilies provides the clearest evidence of type II divergence (GU 2001). The probability of obtaining this fixed difference by chance (P) can be approximated under the Jukes-Cantor (JC) model that assumes equal replacement rates among all amino acids (JUKES and CANTOR 1969). A highly significant P will reflect a large difference in the probability of an amino acid change within *vs.* between the subfamilies. Thus, P can serve as a measure of the power available to type II LRTs.

Assuming that all replacements occur with equal frequency and that none are hidden due to multiple changes, P can be approximated for a given rate (r) under the JC model by the following equation (JUKES and CANTOR 1969):

$$P(r) \approx (1 - e^{-rl_0})e^{-r(l_t - l_0)} = e^{-r(l_t - l_0)} - e^{-rl_t}.$$

Here, l_0 and l_t refer to the branch length for the stem of the test subfamily *vs.* that for the total phylogeny. The gamma distribution can be incorporated to accommodate site-to-site variation in rates (YANG 1996):

$$P = \int_{r=0}^{\infty} \phi(r) P(r)\, dr \approx \int_{r=0}^{\infty} \phi(r)\,(e^{-r(l_t - l_0)} - e^{-rl_t})\, dr.$$

TABLE 1

Power analyses of the type II LRTs

A.	% $2l_t$ (CA)	% l_t (CA)	% $0.5l_t$ (CA)
$2l_0$(CA I)	0.72	2.51	7.32
l_0(CA I)	0.35	1.19	3.34
$0.5l_0$(CA I)	0.17	0.58	1.60

B.	l_0	% $\alpha = 1.15$	% $\alpha = \infty$
CA I	0.2140	1.19	0.82
CA II	0.1332	0.72	0.49
CA III	0.2953	1.67	1.18

Percentages are the probabilities of observing by chance (P) a fixed amino acid difference between one CA subfamily and the two others. (A) Power analysis for CA I *vs.* II and III, given varying lengths of the stem for the former subfamily [l_0(CA I) = 0.2140 replacements/site] *vs.* total tree [l_t(CA) = 3.3730 replacements/site]. (B) Power analysis for each CA subfamily, with and without a gamma distribution (α = 1.15 and ∞, respectively). The 1.15 estimate is the ML value for the CAs under the Jones, Taylor, and Thornton model with a gamma correction (Jones *et al.* 1992; Yang 1996).

The gamma density function, with parameter α, is calculated by

$$\phi(x) = \frac{\alpha^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\alpha x},$$

thereby leading to

$$P \approx \frac{\alpha^\alpha}{\Gamma(\alpha)} \int_{r=0}^{\infty} r^{\alpha-1} (e^{-r(l_t - l_0 + \alpha)} - e^{-r(l_t+\alpha)})\, dr.$$

The integral can be calculated by

$$\int_{x=0}^{\infty} x^a e^{-bx} dx = \Gamma(a+1)/b^{a+1},$$

thereby resulting in

$$P \approx \frac{\alpha^\alpha}{\Gamma(\alpha)} \left(\frac{\Gamma(\alpha)}{(l_t - l_0 + \alpha)^\alpha} - \frac{\Gamma(\alpha)}{(l_t + \alpha)^\alpha} \right)$$
$$= \frac{\alpha^\alpha}{(l_t - l_0 + \alpha)^\alpha} - \frac{\alpha^\alpha}{(l_t + \alpha)^\alpha}.$$

Varying the relative lengths of the stem for one CA subfamily *vs.* total phylogeny documents that power decreases with l_0 and increases with l_t (Table 1). Thus, power is maximized when the opportunity for a stem replacement is small, but that for a change within subfamilies is large. This conclusion becomes important when many sites of the protein have evolved as type II positions. In these cases, phylogenetic methods will overestimate the lengths of the stems and thereby lead to underestimates of the actual numbers of type II sites. In contrast, this conclusion also indicates that an obvious strategy to reduce l_0 and thereby increase power is to sample species that connect to the phylogeny along the stems of each subfamily.

The power analyses further illustrate that rate heterogeneity among sites increases the chances of a type II position (Table 1). Under a gamma process with α = 1.15, a relatively large proportion of sites is slowly evolving (Yang 1996). For these slow sites, any chance replacement in the stems is less likely to be followed by a subsequent change within the subfamilies relative to a position evolving at or faster than the average rate. Thus, a stem replacement for a slow site is more likely to be preserved as a fixed difference among subfamilies. This conclusion reinforces the overall conservative nature of type II sites and their corresponding potential as indicators of protein function (Gu 2001).

An alternative strategy to increase power in the type II LRTs is to include an appropriate replacement matrix for unequal rates among amino acids [*e.g.*, the Jones, Taylor, and Thornton (JTT) model; Jones *et al.* 1992]. For CAs, P under the JC model is ~1.19% for a site with any fixed amino acid difference between CA I *vs.* II and III. In contrast, according to simulations, P under the JTT model varies from <0.01% for F *vs.* K to ~1.79% for H *vs.* Y of CA I *vs.* II and III, respectively. These extremes agree with the premise that radical amino acid differences are more informative than conservative ones about the functional divergence of proteins (Livingstone and Barton 1996; Gu 2001). By incorporating an appropriate unequal rate matrix in their LRTs, fixed radical differences can contribute even stronger evidence to the recognition of type II sites.

In general, phylogenetic errors are not expected to diminish greatly the power of the type II LRTs, since their strongest support is obtained from fixed amino acid differences among subfamilies. By definition, these fixed differences will remain, even if lineages are shifted within subfamilies and the latter are rearranged (Knudsen and Miyamoto 2001). However, the power analysis with l_0 and l_t serves as a reminder of the importance of accurate branch lengths, particularly for the stems (Table 1). In this regard, phylogenetic errors may indirectly affect the type II LRTs by influencing the branch length estimations.

Availability of a computer program: A computer program for the LRTs of types I and II and conserved sites is available as a web server at www.daimi.au.dk/~compbio/LRTs.

RATE AND FUNCTIONAL ANALYSES OF CARBONIC ANHYDRASES

CA I, II, and III: The CA family of ubiquitous enzymes catalyzes the reversible hydration of CO_2 to bicarbonate and protons in many fundamental biological processes (*e.g.*, respiration and photosynthesis; Lindskog 1997; Chegwidden and Carter 2000). At least 15 CAs are known in mammals, with each encoded by a different duplicate gene (Hewett-Emmett 2000). Their diverse biological significance, expression patterns, and catalytic efficiencies, coupled with the successful development of a CA glaucoma drug, ensures that this family will remain a primary target for biochemical, physiological, structural, and pharmacological research.

Phylogenetic and linkage analyses indicate that CA I, II, and III form a related group within their monophyletic CA family, even though their tissue expression patterns and CO_2 hydration rates vary almost as much as for all CAs (Hewett-Emmett and Tashian 1996; Lindskog 1997; Hewett-Emmett 2000). CA II remains the primary reference for the family, because of its high catalytic efficiency and broad tissue expression. Its high CO_2 hydration rate is related to its conserved H64 that functions as a highly effective intramolecular shuttle for proton transfer from the zinc catalytic center to the surrounding medium. CA II is also characterized by a set of five or six H and K residues at its N terminus for Cl^-/HCO_3^- anion exchanger (AE1) binding for bicarbonate channeling from inside to outside the cell (Vince *et al.* 2000). This set of basic residues may also contribute to the transfer of protons from H64 to the bulk solvent (Briganti *et al.* 1997).

CA I and III are more restricted in their tissue expressions and their catalytic rates are ~20% and <1% of that for CA II, respectively (Lindskog 1997). In CA I, H64 is also conserved, but its set of basic residues at the N terminus is greatly diminished (Briganti *et al.* 1997; Vince *et al.* 2000). Thus, CA I cannot bind to AE1 and its N terminus probably does not participate in proton shuttling. In light of its reduced, but significant CO_2 hydration rate, CA I is thought to be a backup to CA II. Conversely, the physiological role of CA III remains unresolved. Its active site shows several important changes (*e.g.*, K or R at position 64) and its N-terminal set of basic residues is reduced. In contrast, CA III evolves slowly and comprises ~8% and ~25% of the total soluble proteins in red skeletal muscle and adipose tissue, respectively. Collectively, these characteristics suggest a major biological role for CA III, which is distinct from the standard CA function of reversible CO_2 hydration.

CA sequences, phylogeny, and LRTs: To evaluate further their functional similarities and differences, all available sequences of CA I, II, and III were compiled, aligned, and analyzed with the LRTs for rate change and conserved sites (Figure 3; Tashian *et al.* 1980; Eriksson and Liljas 1993; Hewett-Emmett and Tashian 1996). The final alignment consisted of 260 positions for 11 CA I, 8 CA II, 6 CA III, and 5 CA Va and Vb (outgroup) sequences. The accepted phylogeny combined the CA gene tree from phylogenetic and linkage analyses with the eutherian mammal phylogeny from a recent molecular synthesis (Figure 4; Hewett-Emmett and Tashian 1996; Hewett-Emmett 2000; Murphy *et al.* 2001).

The standard approach in the LRTs is to measure the site-specific rates for the subfamilies and stems against the local averages for their regions of the phylogeny (Knudsen and Miyamoto 2001; Pupko and Galtier 2002). By relying on relative rates, changes in the local averages due to varying demographic (*e.g.*, population size) and mutation/repair factors are compensated, thereby allowing for functional interpretations of the significant sites (Gaucher *et al.* 2002). However, this reliance on relative rates also reduces the ability of the LRTs to detect large numbers of type I and II sites that may be involved in a major overall change in the function of a protein subfamily. In the case of duplicate genes, this reduction can be addressed by focusing on the common species and branch points of the different subfamilies. These shared species and nodes offer the common time points and uniform biological backgrounds to compare the site-specific rates among subfamilies on a more absolute basis and without the interference of variable demographic and mutation/repair factors. In this way, the sensitivity of the LRTs is enhanced, along with the functional interpretability of their statistically significant sites.

In the case of the CAs, this advantage of shared lineages was accommodated by a constraint that required the distances from the root to each common species and node to be equal across subfamilies (Figure 4). The branch lengths of the phylogeny were then estimated with ML under the JTT model with the gamma distribution (JTT + Γ). As illustrated by the phylogeny, this constraint did not impose a molecular clock on the analysis in the classical sense, since rates remained free to vary across other lineages. The JTT + Γ model was significantly preferred over both the JTT and the JC + Γ models (log likelihood decreases of 149.48 and 677.17, respectively).

Ten thousand sites were simulated under H_2 to establish the 5% cutoffs for H_0, H_{1a}, and H_{1b}. These simulations relied on the JTT + Γ model with α set to its ML value of 1.15 for the CAs. Ten thousand sites were also simulated under H_3 (*i.e.*, with a single rate equal to the average for the entire protein) to determine the 5% significance for H_2. Finally, a set of 42 functionally important sites for CAs was defined according to the 36 positions of the active site and six basic residues of the N terminus for AE1 binding and/or proton shuttling (Hewett-Emmett and Tashian 1996; Briganti *et al.* 1997; Vince *et al.* 2000).

Significant sites and functional interpretations: The LRTs for conserved sites recovered 47 positions that were evolving significantly slower than the overall average for the entire protein (Figure 3). These 47 conserved sites were concentrated among the 42 functionally important positions (Table 2) and included the seven direct and indirect ligands to the zinc catalytic center of the active site (Q92, H94, H96, E117, H119, T199, and N244; Hewett-Emmett and Tashian 1996; Lindskog 1997). Collectively, these results reconfirmed the rule of functional constraint that the biologically

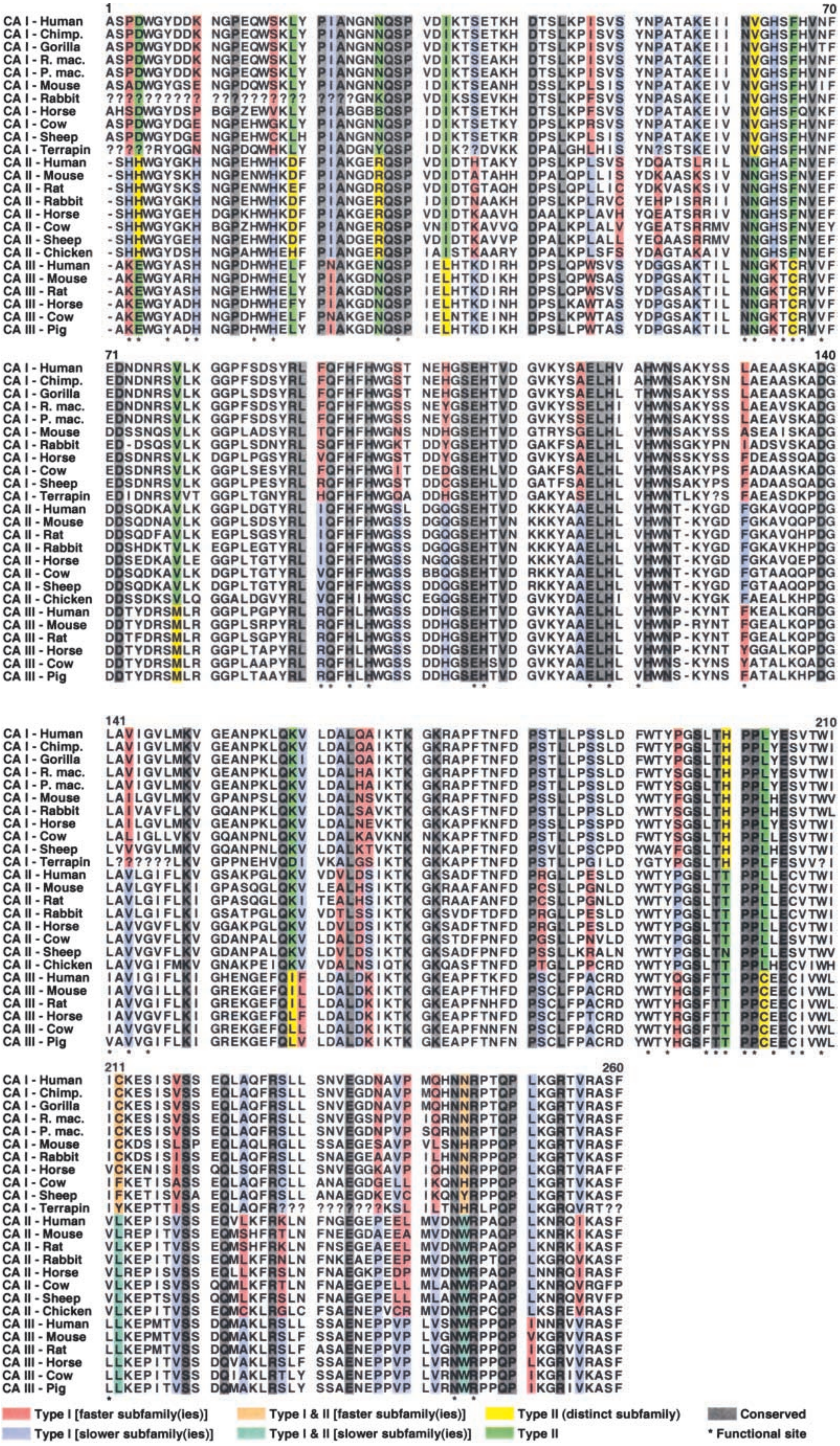
Type I [faster subfamily(ies)]
Type I [slower subfamily(ies)]
Type I & II [faster subfamily(ies)]
Type I & II [slower subfamily(ies)]
Type II (distinct subfamily)
Type II
Conserved
* Functional site

FIGURE 3.—Multiple sequence alignment for all 30 CAs. This alignment follows that of HEWETT-EMMETT and TASHIAN (1996). Sources for these sequences are given in Figure 4. The 42 functionally important positions are asterisked. The annotation of "faster" and "slower" subfamilies is relative to each specific site and not to the entire protein. Species abbreviations are: Chimp., chimpanzee; P. mac., pig-tailed macaque; and R. mac., rhesus macaque.

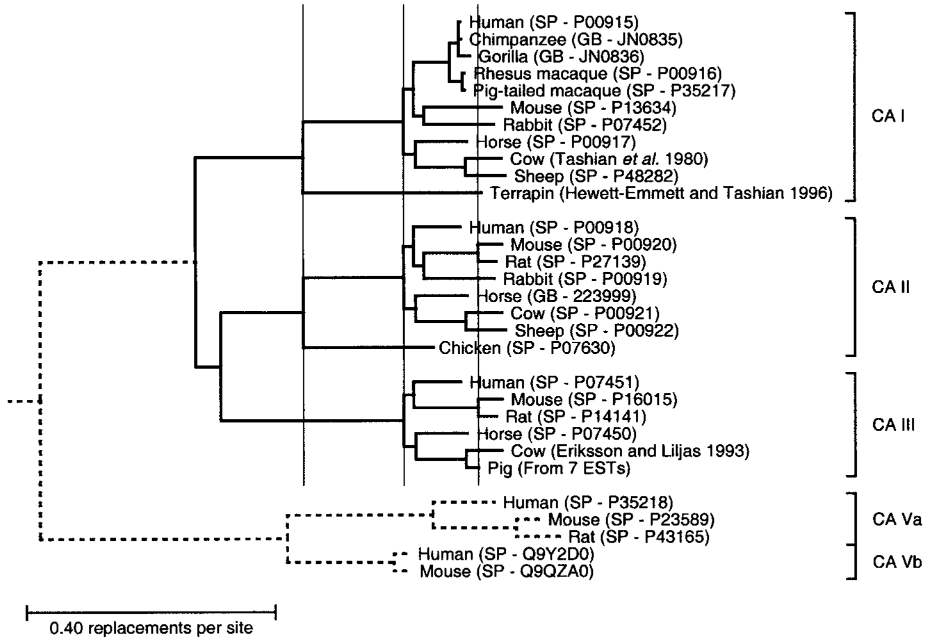

Figure 4.—Accepted phylogeny for the CAs. The distances from the root to identical species and ancestors are fixed across subfamilies, as illustrated by the thin vertical lines. This constraint allows for the more direct interpretation of the site-specific rates among subfamilies in terms of their absolute rather than relative differences (see text). Sites 1 and 121 are excluded from the branch length estimations, because of their gaps in >25% of the sequences. Furthermore, the CA Va and Vb outgroups are not constrained in these estimations, since they are included only to root the phylogeny. Sources for the 30 CAs are given in parentheses and include their GenBank (GB) and SWISS-PROT (SP) accession numbers or original references for those sequences that are not available in the databases (Benson *et al.* 1999; Bairoch and Apweiler 2000). In the case of the pig CA III, this sequence is derived from an analysis of seven overlapping ESTs (GB AJ301094, AJ301207, AJ301290, AJ301337, AU059476, BF074991, and BI360558). BF074991 varied from AJ301337 and AJ301094 by one silent difference. In turn, the two terminal nucleotides of AJ301290 were ignored, since they differed from the corresponding identical bases of BI360558 and AJ301207.

important sites of proteins are under the strongest purifying selection and thereby evolve the slowest.

The LRTs for rate change sites identified 32, 10, and 2 type I, II, and I/II positions, respectively (Figure 3). The expected numbers of type I, II, and I/II sites were 11.8, 11.7, and 2.9 according to the simulations, respectively. Thus, almost three times as many type I sites were recovered as expected by chance. The 32 type I sites included position 64, with its fixed H in CA I and II *vs.* variable R and K in CA III (Hewett-Emmett and Tashian 1996; Lindskog 1997).

Despite their near equal observed to expected frequencies, further analyses validated the importance of the type II sites to the greater understanding of CA functional divergence. The 44 type I and/or II sites were concentrated among the 42 functionally important positions (Table 2). However, this significance depended on the recognition of both divergences, since *P* became ~0.20 when the 10 type II sites were instead counted among the "other positions." Thus, type II divergence complements type I change and both processes must be considered in evolutionary studies of protein function (Gu 2001).

Of the 10 type II sites, 4 mapped to functionally important positions (Figure 3). These 4 type II sites emphasized fixed radical differences among the subfamilies within two primary functional regions of CAs. For example, type II site 4 highlighted the fixed radical difference of H in CA II against the acidic D and E in CA I and III. H4 of CA II is one of the five or six basic residues at its N terminus for AE1 binding. A truncation mutant of CA II, which is missing its first five residues (and therefore H4), shows a measurable decrease in AE1 binding (Vince *et al.* 2000). One obvious follow-up experiment is to retest the AE1 binding of a site-directed CA II mutant after the replacement of its H4 with acidic D or E (Golding and Dean 1998).

DISCUSSION

Functional predictions for CA III: Available biochemical, mutagenic, and structural information defines a series of sites that are of known importance to the common and unique functions of CAs. The ability of the LRTs to detect these known sites, as demonstrated both collectively (Table 2) and individually (*e.g.*, H4, H64, and the seven direct and indirect ligands to the zinc catalytic center), validates their utility for both testing existing hypotheses and generating new ones. In the case of CA III, these LRTs, in combination with biochemical, structural, and other bioinformatic information, support a distinct role for this enigmatic isozyme.

In CA III, C183 and C188 are unique surface residues that are known binding targets for glutathione (GSH; Figures 3 and 5). CA III is among the first proteins to be glutathiolated during oxidative stress and a mutant cell line that is deficient for this isozyme is particularly sensitive to oxyradical insults (Chai *et al.* 1994; Räisänen *et al.* 1999). Thus, CA III is hypothesized to function as an oxyradical scavenger, whereby glutathiolation protects its C183 and C188 from irreversible oxidation. In support of an antioxidant defense role, the LRTs recover three rate change sites that lie next to or directly

TABLE 2

Distributions of conserved and type I and/or II sites among the CAs

	Functionally important positions	Other positions	Totals
A.			
Conserved sites	13 (6.7)	34 (40.3)	47
Other sites	18 (24.3)	151 (144.7)	169
Totals	31	185	216
B.			
Type I and/or II sites	11 (6.0)	33 (38.0)	44
Other sites	18 (23.0)	151 (146.0)	169
Totals	29	184	213

Type I and/or II sites are not included in A, whereas conserved positions are conversely excluded from B. Expected counts are given in parentheses. Functionally important positions refer to the 42 residues of the active site and N terminus for AE1 binding and/or proton shuttling. The chi-square tests for both contingency tables are significant ($P = 0.005$ and 0.017, respectively).

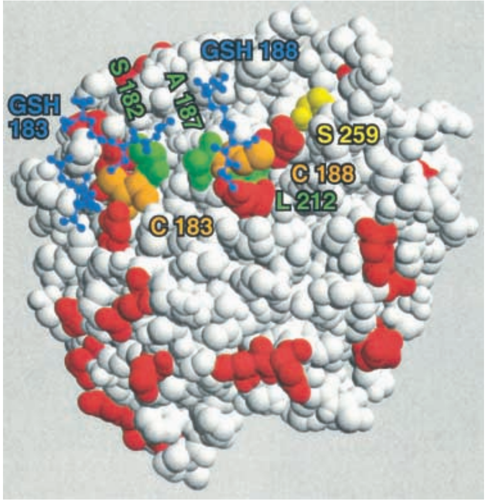

FIGURE 5.—Spacefill model of the tertiary structure for rat CA III with bound GSH (Protein Data Bank accession no. 1FLJ), as rendered with RasMol (SAYLE and MILNER-WHITE 1995; BERMAN *et al.* 2000; MALLIS *et al.* 2000). This view focuses on the key residues around C183 and C188, with both alternative conformations of GSH183 shown. Acidic D and E residues are red.

underneath C183 and C188 (positions 182/187 and 212, respectively). At these three type I and I/II sites, CA III is slowly evolving in contrast to the more variable CA I or II. The conserved or nearly conserved residues of CA III at these rate change sites may contribute to the greater surface exposure and weaker acidic surroundings that enhance GSH binding at C188 over that at C183 (MALLIS *et al.* 2000).

Interestingly, S259, which lies close to C188 at the surface (Figure 5), is a potential phosphorylation site according to NetPhos (an artificial neural network algorithm; BLOM *et al.* 1999). The score for conserved S259 being a phosphorylation site is 0.995 (out of 1.000) for every CA III, except for that of cow (0.928). Thus, S259 phosphorylation/dephosphorylation may affect C188 glutathiolation or vice versa. Through these reversible covalent modifications, CA III may then function as a sensor of oxidative stress, whose activity is tied to the signaling pathways for antioxidant defense (RÄISÄNEN *et al.* 1999; CHEGWIDDEN and CARTER 2000).

Future studies: The only other likelihood-based procedure for type II sites is the Bayesian method (GU 2001). This method calculates separate gamma-distributed rates for the subfamily stems *vs.* crown groups (*i.e.*, most recent common ancestors and their descendants of the subfamilies) and then tests for the independence of these estimates by their site-specific posterior probabilities. This separate treatment of the subfamily stems *vs.* crown groups and incorporation of the gamma process contrasts with the use of whole phylogenies and distribution-free rate estimates by the type II LRTs. In turn, the current and Bayesian methods share the assumption of a fixed point for potential functional divergence in the phylogeny. Although problems with multiple testing and reduced power are thereby avoided, this fixation limits the analyses to predefined groups (see below). The performance of the type II Bayesian method has not yet been studied with real or simulated data nor is it currently implemented in a computer program for general distribution (*e.g.*, DIVERGE; GU and VANDER VELDEN 2002). Further comparisons of the type II LRTs and Bayesian method await the implementation, testing, and application of a generally available computer program for the latter.

As for their Bayesian counterparts, the type I and II LRTs are designed for the study of functional divergence among protein subfamilies that are clearly distinct according to available biochemical, structural, and phylogenetic information (*e.g.*, CA I, II, and III; GU 1999, 2001; KNUDSEN and MIYAMOTO 2001; PUPKO and GALTIER 2002). In turn, nonsynonymous-to-synonymous rate tests using coding DNAs allow for the detection of functional change among the more closely related members of each subfamily (HUGHES 1999; NEI and KUMAR 2000; YANG and BIELAWSKI 2000). Thus, as the two approaches are complementary, both are recommended for more comprehensive studies of functional divergence and conservation between and within protein subfamilies.

We thank A. C. Harmon, R. L. Levine, and M. R. Tennant for their comments about our research. This study was supported by National Institutes of Health grant GM25154 (P.J.L. and D.N.S.) and by funds from the Department of Zoology, University of Florida.

LITERATURE CITED

BAIROCH, A., and R. APWEILER, 2000 The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res. **28:** 45–48.

Benson, D. A., M. S. Boguski, D. J. Lipman, J. Ostell, B. F. Ouellette *et al.*, 1999 GenBank. Nucleic Acids Res. **27:** 12–17.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat *et al.*, 2000 The protein data bank. Nucleic Acids Res. **28:** 235–242.

Blom, N., S. Gammeltoft and S. Brunak, 1999 Sequence- and structure-based prediction of eukaryotic protein phosphorylation sites. J. Mol. Biol. **294:** 1351–1362.

Briganti, F., S. Mangani, P. Orioli, A. Scozzafava, G. Vernaglione *et al.*, 1997 Carbonic anhydrase activators: X-ray crystallographic and spectroscopic investigations for the interaction of isozymes I and II with histamine. Biochemistry **36:** 10384–10392.

Chai, Y. C., S. Hendrich and J. A. Thomas, 1994 Protein S-thiolation in hepatocytes stimulated by t-butyl hydroperoxide, menadione, and neutrophils. Arch. Biochem. Biophys. **310:** 264–272.

Chegwidden, W. R., and N. D. Carter, 2000 Introduction to the carbonic anhydrases, pp. 13–28 in *The Carbonic Anhydrases: New Horizons*, edited by W. R. Chegwidden, N. D. Carter and Y. H. Edwards. Birkhäuser Verlag, Basel, Switzerland.

Dermitzakis, E. T., and A. G. Clark, 2001 Differential selection after duplication in mammalian developmental genes. Mol. Biol. Evol. **18:** 557–562.

Eriksson, A. E., and A. Liljas, 1993 Refined structure of human carbonic anhydrase II at 2.0 Å resolution. Proteins **16:** 29–42.

Felsenstein, J., 1981 Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. **17:** 368–376.

Gaucher, E. A., M. M. Miyamoto and S. A. Benner, 2001 Function-structure analysis of proteins using covarion-based evolutionary approaches: elongation factors. Proc. Natl. Acad. Sci. USA **98:** 548–552.

Gaucher, E. A., X. Gu, M. M. Miyamoto and S. A. Benner, 2002 Detecting functional divergence in protein evolution by site-specific rate shifts. Trends Biochem. Sci. **27:** 315–321.

Golding, G. B., and A. M. Dean, 1998 The structural basis of molecular adaptation. Mol. Biol. Evol. **15:** 355–369.

Gu, X., 1999 Statistical methods for testing functional divergence after gene duplication. Mol. Biol. Evol. **16:** 1664–1674.

Gu, X., 2001 Maximum-likelihood approach for gene family evolution under functional divergence. Mol. Biol. Evol. **18:** 453–464.

Gu, X., and K. Vander Velden, 2002 DIVERGE: phylogeny-based analysis for functional-structural divergence of a protein family. Bioinformatics **18:** 500–501.

Hewett-Emmett, D., 2000 Evolution and distribution of the carbonic anhydrase gene families, pp. 29–76 in *The Carbonic Anhydrases: New Horizons*, edited by W. R. Chegwidden, N. D. Carter and Y. H. Edwards. Birkhäuser Verlag, Basel, Switzerland.

Hewett-Emmett, D., and R. E. Tashian, 1996 Functional diversity, conservation, and convergence in the evolution of the alpha-, beta-, and gamma-carbonic anhydrase gene families. Mol. Phylogenet. Evol. **5:** 50–77.

Huelsenbeck, J. P., and B. Rannala, 1997 Phylogenetic methods come of age: testing hypotheses in an evolutionary context. Science **276:** 227–232.

Hughes, A. L., 1999 *Adaptive Evolution of Genes and Genomes.* Oxford University Press, New York.

Jones, D. T., W. R. Taylor and J. M. Thornton, 1992 The rapid generation of mutation data matrices from protein sequences. Comput. Appl. Biosci. **8:** 275–282.

Jukes, T. H., and C. R. Cantor, 1969 Evolution of protein molecules, pp. 21–123 in *Mammalian Protein Metabolism*, edited by H. N. Munro. Academic Press, New York.

Kimura, M., 1983 *The Neutral Theory of Molecular Evolution.* Cambridge University Press, Cambridge, UK.

Knudsen, B., and M. M. Miyamoto, 2001 A likelihood ratio test for evolutionary rate shifts and functional divergence among proteins. Proc. Natl. Acad. Sci. USA **98:** 14512–14517.

Landgraf, R., I. Xenarios and D. Eisenberg, 2001 Three-dimensional cluster analysis identifies interfaces and functional residue clusters in proteins. J. Mol. Biol. **307:** 1487–1502.

Li, W.-H., 1997 *Molecular Evolution.* Sinauer, Sunderland, MA.

Lindskog, S., 1997 Structure and mechanism of carbonic anhydrases. Pharmacol. Ther. **74:** 1–20.

Livingstone, C. D., and G. J. Barton, 1996 Identification of functional residues and secondary structure from protein multiple sequence alignment. Methods Enzymol. **266:** 497–512.

Lynch, M., and A. Force, 2000 The probability of duplicate gene preservation by subfunctionalization. Genetics **154:** 459–473.

Mallis, R. J., B. W. Poland, T. K. Chatterjee, R. A. Fisher, S. Darmawan *et al.*, 2000 Crystal structure of S-glutathiolated carbonic anhydrase III. FEBS Lett. **482:** 237–241.

Murphy, W. J., E. Eizirik, S. J. O'Brien, O. Madsen, M. Scally *et al.*, 2001 Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science **294:** 2348–2351.

Nei, M., and S. Kumar, 2000 *Molecular Evolution and Phylogenetics.* Oxford University Press, New York.

Nei, M., X. Gu and T. Sitnikova, 1997 Evolution by the birth-and-death process in multigene families of the vertebrate immune systems. Proc. Natl. Acad. Sci. USA **94:** 7799–7806.

Ohno, S., 1970 *Evolution by Gene Duplication.* Springer-Verlag, Berlin.

Pupko, T., and N. Galtier, 2002 A covarion-based method for detecting molecular adaptation: application to the evolution of primate mitochondrial genomes. Proc. R. Soc. Lond. Ser. B **269:** 1313–1316.

Räisänen, S. R., P. Lehenkari, M. Tasanen, P. Rahkila, P. L. Härkönen *et al.*, 1999 Carbonic anhydrase III protects cells from hydrogen peroxide-induced apoptosis. FASEB J. **13:** 513–522.

Sayle, R., and E. J. Milner-White, 1995 RasMol: biomolecular graphics for all. Trends Biochem. Sci. **20:** 374.

Tashian, R. E., D. Hewett-Emmett, S. K. Stoup, M. Goodman and Y.-S. L. Yu, 1980 Evolution of structure and function in the carbonic anhydrase isozymes of mammals, pp. 165–176 in *Biophysics and Physiology of Carbon Dioxide*, edited by C. Bauer, G. Gros and H. Bartels. Springer-Verlag, Berlin.

Vince, J. W., U. Carlsson and R. A. Reithmeier, 2000 Localization of the Cl^-/HCO_3^- anion exchanger binding site to the amino-terminal region of carbonic anhydrase II. Biochemistry **39:** 13344–13349.

Yang, Z., 1996 Among-site rate variation and its impact on phylogenetic analyses. Trends Ecol. Evol. **11:** 367–372.

Yang, Z., and J. P. Bielawski, 2000 Statistical methods for detecting molecular adaptation. Trends Ecol. Evol. **15:** 496–503.

Communicating editor: J. Hein