Motif-weighted Structure Alignment for Classification and Evolutionary Studies of Carbonic Anhydrase

Shi, Hongyi [1] *

[1] West Valley College, USA

*Correspondence to: hshi11@mywvm.wvm.edu

Abstract

Carbonic anhydrases (CAs) attract interest for their critical roles in various physiological processes and potential application in CO_2 sequestration to combat global warming. Despite being an important enzyme family, the classification and evolution of CAs remain elusive due to their high sequence diversity and long evolutionary history. In this paper, the *in-silico* strategy, Motif-weighted Alignment for Structure-based Protein Classification (MASPC) was developed, which uses OmegaFold simulated CA structures combined with weighted structural motif alignment, TM-weighted, to facilitate more precise polymorphic analysis of large enzyme datasets in a robust manner. The MASPC strategy was first validated by 74 ground-truth CA structures extracted from PDB, showing improved performance compared to sequence-based polymorphic analysis (ClustalO-RAxML). Subsequently, MASPC was applied to analyze a representative database, which contains 1603 CAs from 117 model organisms, with focus on α-, β-, and- γ- CA classes, to cover organisms from across life evolution history. The results indicated that α-, β-, and γ-CAs were well grouped in their own classes, with clearer clustering associated with the CA's organism. The structural differences among the α-, β-, and γ-CAs revealed by MASPC supported the current understanding that CA classes are the results of convergent evolution. The sub-clusters in α- and β-CAs are highly associated with organisms according to their appearance in evolutionary history, demonstrating a close correlation between CA evolution and life evolution. Furthermore, the MASPC method was also applied to identify 27 potential α-CAs from the NCBI database with less than 40% sequence similarity to a template human carbonic anhydrase II (HCA-II) sequence, demonstrating possible applications in enzyme identification studies.

Introduction

CAs are metalloenzymes with the enzyme designation EC 4.2.1.1 that catalyze the reversible reaction between carbon dioxide and water to form bicarbonate. As a highly diverse enzyme family with 7 classes described in literature, including α-, β-, γ-, δ-, ζ-, η-, θ-, and ι-classes, the most thoroughly explored are the α-, β-, and γ-CAs **(Hewett-Emmett & Tashian, 1996)**. The classification of CA is based on sequence and structure differences with Zn^{2+} ion as the cofactor in most CAs. The Zn^{2+} ion is typically coordinated by multiple His residues in the enzymes (**Lindskog, 1997**) but has been shown to be replaced with other di-cations such as Cd^{2+} or Fe^{2+}. Because of its CO_2 sequestration abilities, CAs attract attention for their potential to combat global warming (**Boone et al., 2013**). Moreover, recent studies of CAs have been reported that they are involved in many important physiological processes in animals, including the control of neuronal excitability (**Ruusuvuori & Kaila, 2014**), neuroactive alkaloid biosynthesis (**Nett et al., 2023)**.

CA structures from different classes are largely unique but share some important characteristics. The α-, and γ - CAs use three His residues, along with water, to coordinate the Zn^{2+} ion, while β-CAs differ slightly, using at least two Cys residues and one His residue (**Supuran, 2016**). Additionally, α-CAs are active as a monomer and γ-CAs as trimers. It has been found that CAs are ubiquitous, appearing in archaea, bacteria, plants, animals, and humans (**Smith et al., 1999**). As an ancient enzyme, it is theorized that CAs likely co-evolved with their host organisms, leading to the high diversity in the over 61,000 EC 4.2.1.1 annotated CAs in UniProtKB. It is widely accepted that the CA classes evolved from different ancestors and its evolution is often used as a model example of convergent evolution due to the fundamentally different protein structures between CA classes but similar catalytic pathways (**Liljas & Laurberg, 2000**). Despite their ubiquity, there is a non-uniform distribution of CA classes across organism domains and even kingdoms (**Smith & Ferry, 2000)**. For example, vertebrates only contain α-CAs, while plants contain α, β, and γ-CAs. In fact, the high sequence diversity and long evolution history further complicates CA classification, bringing a need of more powerful methods for its classification and evolution studies. Moreover, with the rapid advancement of genome sequencing technologies, the UniProtKB database now contains approximately 250 million sequences, though only a small fraction of them have been experimentally characterized. This adds to the increasing need for scalable and effective new solutions for large-scale protein classification, especially for proteins showing low sequence similarity with well characterized proteins.

Conventional protein classification was built upon conserved motifs, mainly depending on conserved catalytic residues. In addition, sequence similarity and phylogenetic relationships were also utilized to help improve accurate classification. While being largely successful, this strategy is less effective when applied to proteins with high diversity and different ancestors, such as CAs. For example, the η-CAs were originally identified as α-CAs due to sharing similar His motifs in the active site, but later were separated as a new class because of their much longer sequence lengths, among other differences (**Del Prete et al., 2014**). Besides the sequence-based methods, Gene3D, a domain-based method was developed for protein annotation and classification in 2008, utilizing Hidden Markov Models to help identify proteins (**Yeats et al., 2008)**. Other sequence and motif-based methods, such as InterPro (**Hunter et al., 2009**), PANTHER (**Mi et al., 2005**), and MobiDB (**Piovesan et al., 2023**), have also been developed. However, these methods can be susceptible to misaligning highly diverse proteins due to their highly sophisticated and specialized approaches, including inaccurate assumptions about the motifs, as seen in η-CAs.

Compared to the sequence- or motif- based methods, protein structure alignment could provide more accurate classification due to the close correlation between protein structures and functions. Different experimental technologies, including Xray, NMR, and Cryo-EM can resolve high resolution protein structures. These methods provide useful insight into protein functionality, therefore facilitating more accurate protein classification. The major drawback of these experimental methods is that they currently cannot be scaled up due to their high costs and are only limited to *in vitro* purified proteins under laboratory conditions. For example, after removing similar entries, it is found that there are less than 100 unique CA structures in the PDB, covering a low number of organisms. Recently, with the expansion of the PDB and machine learning developments in protein science, protein structure simulation models based on proteins' amino acid sequences were developed and exhibited strong structure simulation capability. Among these methods, AlphaFold 2 stands out for its use of multiple sequence alignment (MSA) to generate highly accurate protein structure predictions (**Jumper et al., 2021**). In comparison, the Large Language Model based ESMFold (**Lin et al., 2023**) and OmegaFold (**Wu et al., 2022**) are faster, exhibiting the ability to predict protein structures without clear homologs. Compared to experimental structure characterization, structure simulation is a trade-off of accuracy for speed and scale, but still allows for structure-based protein analysis. Using these models, several attempts at protein classification, identification, and prediction using simulated protein structures have been performed, including FoldTree **(Moi et al., 2023)**, which is built upon structural alphabets to classify proteins and generate polymorphic trees.

Here, an *in silico* method called MASPC was developed, which utilized OmegaFold simulated structures to explore the classification and evolution of CAs (**Figure 1)**. MASPC starts by creating an initial database containing only unique CAs, called CA-DB-I, from which a representative model dataset designed to be as diverse as possible was generated to capture the breadth of CA variability and used as the basis for further analysis. Next, the structure simulation using OmegaFold was performed to convert the sequence dataset to a simulated structure dataset. Finally, polymorphic studies of CAs based on the simulated structures were performed using TM-weighted and Neighborhood Joining (NJ) to incorporate the significant motifs into the analysis. Using the MASPC strategy, 1603 CAs were analyzed to improve the current understanding of CA classification, which provided higher fidelity extraction of protein relationships, demonstrated by the resulting phylogenetic trees. Moreover, using human carbonic anhydrase II (HCA-II) as the target template, MASPC was also successfully applied to identify 27 potentially active CA candidates in the hypothetical and uncharacterized protein database extracted from NCBI.

Methods

Comprehensive CA Sequence Database Creation

A data cleaning setup was built into the MASPC framework's Database Creation pipeline, which performed boilerplate data filtering and name convention conversion. A protein is decided to be a CA based on its GenBank file. To remove redundant CA sequences, the protein similarity was handled within each organism instead of against the whole database. This assumes that similar CAs within the same organism are likely to be redundant, while similar CAs appearing in different organisms are likely to be the result of convergent evolution. To account for the diversity of CAs, a similarity of 70% was used as a maximum threshold for the pairwise Needleman-Wunsch sequence similarity comparisons. The database was further paired down to proteins between 100 to 500 amino acids to facilitate the structure prediction. In the process, near-identical sequences were filtered, which resulted in a complete database (CA-DB-I, **Supplement Information File 1**) of all unique CA sequences of a given organism. There are a total of 16890 unique organisms in CA-DB-I.

Carbonic Anhydrase Structure Database

Protein sequences used in this study were sourced from UniProt (**www.uniprot.org**) and NCBI (**https://www.ncbi.nlm.nih.gov/**). All downloads occurred before March 17th, 2024. BLAST is used to create a custom CA database with all found CAs in processed databases.

To obtain a high-quality structure dataset for evaluating the structure prediction accuracy of OmegaFold and the MASPC strategy, a dataset containing 74 CA structures was extracted from PDB (**www.rcsb.org)**. Firstly, a search was performed to find all likely CA entries, resulting in a collection approximately 1000 PDB entries that fit the constraints. A BLAST database was then initialized upon CA-DB-I. Subsequently, each of the collected PDB entries were searched using BLAST across CA-DB-I using a high sequence similarity threshold of 98% similarity as determined by BLAST, verifying the PDB entry as a CA. A further structural comparison was performed using a TM-score threshold of 0.95 to remove identical structures and obtain the candidate set of PDB structures. A final manual check of each structure resulted in a representative dataset of structures that was used in further analysis as unique representatives of all PDB CAs. The PDB structure database is provided in the **Supplement Information File 1**.

Structure Simulation

OmegaFold was used for CA structure prediction in this study. Structural similarity between protein structures was evaluated using TM-score (**Zhang & Skolnick, 2004**) calculated with TM-align, and TM-weighted, while Biopython was used for handling biological data (**Cock et al., 2009**). All the structure simulations were performed on Linux 22.04 LTS.

Polymorphic Analysis and Evolution Tree Creation

Protein sequence alignments were performed using ClustalO (**Sievers et al., 2011**) and sequence similarity percentages are reported using a normalized Needleman-Wunsch method (**Needleman & Wunsch, 1970**). The phylogenetic trees built from distance matrices were constructed using Neighbor-joining (NJ) **(Saitou & Nei, 1987)**. Sequence based trees were constructed using ClustalO and RAxML (**Stamatakis, 2006**). iTOL was used to display the phylogenetic trees (**Letunic & Bork, 2024**), and the coloring is based on existing CA annotations in NCBI and UniProt databases.

The sequence-based trees were constructed using the best 100 iterations of RAxML. For structure-based trees, structural distances between protein structures were calculated using the structure comparison methods, TM-score or TM-weighted. PDB files were compared pairwise, converting the resulting TM-score or TM-weighted scores to distance measures. Structure based phylogenetic trees were generated using NJ.

TM-weighted

TM-weighted is a Python implementation of TM-score. A heuristic weight is applied when aligning His residues in proteins is introduced to consider the active residues of the potential CA sequences (see below). TM-weighted uses a deterministic Kabsch algorithm in line with the original TM-score implementation. In the initial alignment step, a weighting factor is applied to each residue, allowing specific residues to have a greater influence on the alignment. The final alignment performed using TM-weighted considers important local motifs as well as global alignment. The comparison of TM-weighted and TM-score on CA structure analysis is presented in **Figure S1**.

$$TM_{\text{heuristic}} = \frac{1}{L}\sum_{i=1}^{L} w_i \cdot \frac{1}{1+(d_i/d_0)^2}$$

where

$$w_i = \begin{cases} w_{\text{His}} & \text{if residue } i \text{ is histidine in both structures} \\ 1 & \text{otherwise} \end{cases}$$

Low sequence similarity CA confirmation in NCBI database

The structure-based analysis was also used to confirm CAs with low sequence similarity in the NCBI database. A loose regex was used to further filter sequences using well-known motifs, keeping the sequence search relevant and controllable. Sequences were then categorized into similarity tiers based on their sequence's similarity to the target protein, in this case chosen to be HCA-II. First, the whole NCBI database was searched using the target HCA-II sequence as a template followed by α-CA motif filtering. The sequences with similarities between 30-39% to the HCA-II were collected, followed by structure simulation using OmegaFold. This specific range was chosen for containing most of the low sequence similarity proteins (**Figure S4**). Among these, sequences with a structural similarity of a TM-weighted score greater than 0.5 as compared to the HCA-II structure (PDB: 1BIC) were considered similar (**Xu & Zhang, 2010**) and selected for further analysis. Proteins with any reference to CAs in their GenBank file were considered unrelated and were removed to focus the search on unique edge-case proteins. The remaining CAs were verified by manual structure alignment and checked for correctly oriented active residues. The structures of the final sorted sequences were then referenced with their corresponding UniProtKB entries.

Results

Validation of structure-based polymorphic analysis with PDB CA structures

Firstly, the MASPC strategy was validated by published CA structures in PDB. Despite the presence of thousands of CA entries in the PDB, only 74 unique CA structures have been resolved in the PDB after being cross-referenced with CA-DB-I and further filtered. Among these structures, 10, 17, 5, and 1 CAs are annotated as α-, β-, γ-, and ι-CAs, respectively, with the rest being unclassified within their PDB entry files (**Table 1**).

To evaluate the CA classification strategy, a comparative analysis of sequence-based and structure-based polymorphism was conducted using these 74 structures (**Figure 2**). An optimized sequence-based phylogenetic tree generated using the best result after 100 iterations of RAxML (**Figure 2A**) still failed to group all the CAs as expected, creating two clades of unclassified CAs. In comparison, structure-based trees clearly separated α-, β-, γ-CA clades (**Figure 2B**). Though the trees have shown similar clustering, their pair-wise normalized Robinson-Foulds distances are all greater than 0.5 suggesting though the clustering is similar, the trees show variation within each clade. These results suggested that the structure-based polymorphism was a promising strategy for clustering highly diverse CAs, which could provide more accurate large-scale analysis of CA classification and evolution. To focus on the analysis of active CAs, TM-weighted, a method based on TM-score was developed to be incorporated into the MASPC strategy for better enzyme classification. Compared to TM-score, TM-weighted aligns more closely with human intuition by emphasizing critical His residues, which form key motifs in active CAs, over other residues. This reduces the impact of superfluous protein structures that add unnecessary noise, resulting in a more relevant alignment metric tailored specifically for comparisons within a particular protein class. Similarly, this created a mechanism for weeding out improperly predicted structures or non-CAs into their own category, due to their lack of the target structural motif. TM-weighted created accurate CA class-based clades similar to TM-score (**Figure 2B, 2C**). In the TM-weighted polymorphic tree, γ-CAs and β-CAs were clustered well with no splitting in the clades. Notably, the ι-CA was grouped closer to α-CA clade in TM-weighted-based analysis, while it was closer to γ-CAs in the TM-score-based analysis.

Confirmation of the OmegaFold simulated CA structures

Next, OmegaFold was evaluated for later use in the large-scale model CA structure simulation. The LLM-based OmegaFold was chosen in this study because it has been shown to be effective in predicting orphan proteins, a favorable capability given the diversity of protein sequences like CAs. Besides, its low computing resource demands meet the speed and cost-efficiency required for large-scale structure simulation. To evaluate the applicability of OmegaFold for CA structure simulation specifically for phylogenetic analysis, the OmegaFold simulated structures based on the sequences of 74 PDB CA entries were compared to the ground-truth PDB structures. Our results showed that the OmegaFold-generated CA structures agreed to their ground truth PDB counterparts, with 67 out of 74 showing TM-scores greater than 0.9, and only one with a TM-score below 0.7 of 0.69 (**Figure S2**). Not surprisingly, polymorphic studies based on OmegaFold's simulated structures closely matched that of the ground truth PDB structures, further confirming its effectiveness on CA structure simulation for classification (**Figure 2C, 2D**).

TM-weighted High diversity and fast evolution of CAs

With the observation that OmegaFold structures could provide good CA classification, a more extensive analysis was performed on a broader range of model organisms to further understand CA's classification and evolution. 117 representative model organisms were chosen from across all walks of life, representing major landmarks of evolution, including archaea, bacteria, plants, fish, and mammals. The 117 selected model organisms (**Table 2**) were searched against the CA-DB-I database to obtain annotated CAs with focus on α-, β-, γ-CAs, allowing for a non-repeating database of model organism sequences. To ensure enzyme diversity regardless of the model organisms, further sets of α-, β-, γ-, δ-, ζ-, η-CAs were collected (**Del Prete et al., 2014**). A total of 1544 CAs from these representative model organisms with the addition of 59 high-variety CAs from literature and the PDB were finally selected for a systematic study of CA's classification and evolution (**Table 2, Supplement Information File 1**). Among these 1603 CAs, there were potentially 869 α-CAs, 481 β-CAs, 200 γ-CAs and 53 indeterminate CAs from other classes according to previous UniProt sequence annotation and subsequent phylogenetic analysis. Most organisms contain multiple CA isoforms, ranging from a single CA isoform to 29 isoforms in *Ustilago maydis* and *Oryza sativa subsp. japonica*, respectively (**Figure 3**). There are 169 unique CAs in *E. coli*, which is mainly due to the high diversity of individual *E. coli* strains. In addition, some organisms, mainly plants, contain all α-, β-, and γ-CAs, such as *Arabidopsis thaliana*, while the other organisms contain only one single class CAs, such as *Mus musculus* and *Methanosarcina thermophila*, which contain only α- or γ-CAs, respectively. An example of high sequence variability in contrast to structure conservation are the 12 CA isoforms that have been identified in humans. These CA isoforms all belong to α-CAs, with sequence similarity between 28% and 64%. However, their structures are very similar, with TM-scores all greater than 0.71 (**Figure S3**).

Model CA sequence-based phylogenetic analysis

The collected CAs were subsequently applied to both sequence- and structure-based analysis. Similar to the 74 CAs extracted from the PDB, the 1603 CAs from different classes were not grouped well in the sequence-based phylogenetic tree (**Figure 4A**), providing limited information for their classification. Using a maximum likelihood tree construction, RAxML, for sequence-based polymorphic tree construction, CAs were largely grouped by organisms, as shown by the outermost color rings of both trees. It was observed that although β- and γ-CAs were differentiated into their own branches, the clades were not clearly defined. The sequence-based tree separated type-1 β-CAs and type-2 β-CAs **(Figure 4A)** into different subclades, which have been reported to possess distinct difference of the ligation state and the orientation of amino acid residues around the active site.

Additionally, the sequence-based analysis encountered challenges in classifying α-CAs. Unlike the β- or γ-CAs, the α-CA clade did not cleanly contain all α-CA proteins under a single branch, instead forming staircase-like branches, which could not be categorized as a single clade without encompassing the rest of the tree as well. The main α-CA was split into 5 subclades, but the remaining stair-case clades could not be easily put into a single clade. This was paired with three identifiable sub-clades corresponding to plant, invertebrate, and vertebrate CAs. The sequence-based approach also had difficulties in classifying **ζ**-CAs, with a subgroup between the β- and γ-CA clades, and another subgroup appearing to be mixed in with η-CAs and unclassified CAs between the α- and γ-CA clades **(Figure 4A**). The η-CAs were grouped well but showed up clustered alongside unclassified structures and a portion of the **ζ**-CAs (**Figure 4A**), suggesting that the sequence-based approach was not able to truly classify them. Further issues arose with the sequence-based approach when considering the human-verified proteins. Notably a verified β-CA XP_001699151.1 **(Del Prete et al., 2019**) appeared grouped under the α-CA clade (**Figure 4A**).

Model CA simulated structure-based phylogenetic analysis

TM-weighted structure-based phylogenetic tree is different from the sequence based phylogenetic tree, with the Robinson-Foulds distance being calculated as 0.87 (**Figure 4**). These results were in line with the analysis of the CA PDB structures (Table 3). Compared to the sequence-based analysis, CAs were better clustered in the structure-based phylogenetic tree (**Figure 4B**), with the annotated α-CAs forming the largest cluster and the annotated γ-CAs forming the most conservative cluster. The γ-CAs were more highly conserved than both α-CAs and β-CAs, which formed two main sub-clusters. One sub-cluster contained CAs predominantly from archaea, and a few from bacteria. The other sub-cluster contains CAs from archaea, bacteria, plants and protists. None of γ-CAs came from invertebrates, vertebrates, or fungi. The structure-based analysis gave a better clustering of γ-CAs (**Figure 4A, 5B**), and their high structural similarity suggested a conserved evolutionary pathway for γ-CAs, potentially originating from a common ancestor. Similarly, β-CAs also largely formed several distinct sub-clusters (**Figure 4B**). One main sub-cluster containing CAs mainly from bacteria and plants, while CAs of the rest sub-clusters came from a broader range of organisms, including archaea, bacteria, protists, plants, invertebrates, and a few vertebrates. The structure and organism differences between those sub-clusters suggested that there may have been a divergence in their evolution, indicating possible sub-classifications for β-CAs. It was noticed that the currently identified type-1 and type-2 β-CAs were not closely grouped, indicating that more studies are needed for a better sub-classification.

Although α- and γ-CAs share similar active sites, in which they utilize three His residues to coordinate the Zn^{2+} ion, the polymorphic tree showed a significant evolutionary distance between the α-CA and γ-CA clusters (**Figure 4B**). Different from and γ-CAs, α-CAs were primarily found in invertebrates, vertebrates, and plants, with no representatives from archaea. Similar to the sequence-based analysis, the structure-based analysis still could not deliver clear clustering of the α-CAs, but it was clear that the α-CAs were grouped better in the structure-based polymorphic tree, forming several sub-groups with close evolutionary distance (**Figure 4B**). In the structure-based polymorphic tree, CAs from invertebrates and vertebrates clearly separated from those of plant origin, which seemed to have closer evolutionary relationships of CAs from protists, Archaeplastida, etc. It was also noticed that CAs from invertebrates and vertebrates could also be separated by two distinct sub-clusters. These distinct sub-clusters indicated the very high diversity of α-CAs and their possible differences in origins. More detailed analysis requested further sub-classification of the α-CA group.

In addition to the new observations of α-, β-, and γ-CAs, TM-weighted's classification of η-CAs reflected their highly similar active site geometry which originally caused η-CAs to be misclassified as α-CAs **(Figure 4B)**. Based on the structure similarity analysis, although η-CAs were still grouped with the α-

CAs, they were isolated clearly from the majority of α-CAs, showing an evolutionary position between α- and γ-CAs. Moreover, a miscellaneous zone made up of 10 α-CAs, 5 β-CAs, 1 γ-CA, 13 unknown CAs, and all 9 ζ-CAs were presented in the structure-based polymorphic tree, sitting between the β-, and γ-CAs. These structures were loosely folded, and did not contain any geometric motif, suggesting that the structures of these CAs had low prediction quality or even that those proteins were possibly misannotated as carbonic anhydrases. These proteins were likely grouped together for having low similarity with CAs in other clades due to not sharing any expected structural or His motifs with those CA clades.

Identification of low sequence similarity CAs from database

Based on the studies presented, it was speculated that the OmegaFold simulated structure-based strategy could also be applied for protein classification without heavily relying on sequence similarity. To explore its applicability, HCA-II was used as a target protein for a sequence search against the complete NCBI database. Proteins with sequence similarity lower than 40%, numbering 14048 sequences, were chosen (**Figure S4**). Subsequently, structures for the proteins were predicted, followed by structure similarity comparisons using TM-weighted against the experimentally determined structure of HCA-II. This process discovered 27 new α-CAs not currently identified as CAs by their GenBank files. Although these 27 proteins have a <40% sequence similarity with HCA-II, their TM-scores are in the range of 0.87 to 0.56 (**Figure 5A**).

Next, a structure alignment and comparison with the HCA-II structure (PDB: 1BIC) was performed. The identified CAs possessed the HCA-II-like active site pocket, with three characteristic His residues in the correct orientation, suggesting they might be active CAs. As a representative example, HBH53009.1, which was labeled as a hypothetical protein in the NCBI database with a low sequence similarity of 38.2% to HCA-II yet showed high structure similarity with a TM-weighted score of 0.56 (**Figure 5A**). The predicted structure of HBH53009.1 showed different structural features compared to HCA-II's structure at the periphery but maintained the highly overlapped β-sheets and coordinating His residues in the catalytic core that is critical to its function as an α-CA (**Figure 4B**). This result further demonstrates the applications of the structure-based CA classification method in low sequence similarity CA identification.

Discussion

In this study, a structure-based *in silico* method, MASPC, which weighted key motifs in target proteins, was developed to improve the classification of CAs, a group of highly sequence-diverse ancient enzymes with carbonic dioxide capturing capability. In MASPC, the structure-based polymorphic analysis, as applied through TM-weighted, provided superior accuracy in classifying CAs compared to sequence-based methods. Based on an analysis of 1603 CAs from 117 model organisms, it was observed that α-, β-, and γ-class CAs possessed distinct structural features, suggesting they might have evolved from different ancestral proteins. Further analysis indicated that CA evolution was highly related to their host organisms, revealing CA's correlations to life evolution. Other than for CA classification, this *in silico* method was also applied to identify CAs with low sequence similarity. Using HCA-II as a reference, 27 active CAs with sequence similarity under 40% in the NCBI database were identified, further validating the power of structure-based approaches.

Sequence-based methods face significant challenges when dealing with proteins that have evolved from different evolutionary ancestors. Both the findings from this research (**Figures 2, 5**) and other recent studies (**Moi et al., 2023**) indicated that simulated structures offer more reliable clustering for diverse proteins. The predicted-structure-based polymorphic tree correctly grouped CAs from different classes, maintaining detailed relationships such as evolutionary distances like the experimentally determined structure-based polymorphic trees (**Figure 2B-D**). In contrast, the sequence-based tree was unable to capture these fine details and highlighted the extensive sequence variability among CAs (**Figure 2A**). The better accuracy of structure-based analysis was further confirmed in a large-scale dataset analysis (**Figure 4**). These results underscore that while CA sequences are highly variable, their structures remain conserved, and structure-based polymorphic analysis offers more precise CA classification, which is less reliant on the organism from which the protein originates. To enhance this structure-based approach for CA classification, TM-weighted was developed, building on the TM-score measure to emphasize key motifs during polymorphic alignment, ensuring more accurate representations of catalytic features. This reasoning is in line with protein evolution theory (**Ribeiro et al., 2023**), in which enzyme active sites are strictly conserved through evolution. TM-weighted differentiates from TM-score, which emphasizes the overall topology and the larger structural features. TM-weighted is thus making it more suitable for CA analysis, where diverse sequences can complicate protein classification. TM-weighted outperformed TM-score in sub-clustering accuracy (**Figures 2B-C, Figure S3**), providing a clearer distinction between enzyme subclasses. In the large and more diverse database analysis, TM-weighted also provided a better clustering of different CAs from various organisms. ζ-CAs were not grouped, falling into a misc-zone grouped with hard-to-be predicted structures.

Structure provides more accurate clustering guidance; however, it is largely limited by structure simulation models. Different structure-based methods can emphasize more complex characteristics of proteins. The Robinson-Foulds distance between the structure-based tree using TM-weighted and TM-score is 0.801. The TM-score approach groups a single ζ-CA with γ-CAs and groups the other 8 ζ-CAs within the misc clade. Though the η-CAs are grouped near each other, they belong to separate clades. The lack of clustering suggests that the TM-score metric can be improved by a weighting mechanism.

Although AlphaFold2 provides the state-of-the-art structure prediction, it requires large amounts of computing resources for large scale analysis. Comparably, LLM-based models, such as OmegaFold, achieve similar performance but with a much lower resource requirement. OmegaFold is also known for orphan protein structure prediction as it does not utilize Multiple Sequence Alignment. Still, there are evident issues with transitioning from experimental structures to predicted structures between trees generated using experimental methods and trees generated through prediction technologies, exemplified in the ground truth vs predicted structure trees **(Figure 2, Table 3)**. Although CAs are far from being considered orphan proteins, their evolution spans great periods of time and is subject to vast (**Figure 3, 5A**). In the research here, OmegaFold simulated CA structures showed good agreement with the PDB ground truth CA structures in both alignment metric and resulting trees (**Figure S2**). This suggests at its applicability to much larger datasets. Due to the computing resource limitations, the research performed here was limited to proteins with length under 500 aa. Most proteins, especially annotated CAs in the UniProt database, are shorter than that. Importantly, this *in silico* approach will continue to improve alongside the development of protein prediction technologies and further fine-tuning of the models used to predict the structures in the dataset.

The CAs in the tree demonstrate many different classes of CAs belonging to single types of organisms, such as plants, which can contain all 3 main CA classes. Furthermore, multiple isoforms of the same CA class in a single organism might be correlated to their functional diversity in different tissues and organs. Studies increasingly show that CO2 serves various physiological roles, such as a signaling molecule (**Phelan et al., 2021**), and specific CA isoforms predominate in different cell types and tissues (**Agarwal et al., 2019**).

The structure-based polymorphic analysis of the CAs from model organisms sheds new light on their evolution and correlation with life evolution (**Figure 3B**). The structure and sequence analysis of the model CAs performed in this study support that different CA classes evolved from different ancestors. Furthermore, the research here also supports that α-, β-, and γ-CA classes are dominant in different organisms. For example, two main sub-clusters exist in α-CAs, of which one sub-cluster is composed of invertebrates and vertebrates CAs, and the other sub-cluster consists of α-CAs mainly from invertebrates and plants. The branches in these two sub-clusters are separated by different organisms. The coexistence of invertebrate CAs in these two sub-clusters, along with the organism-associated CA clades, suggests a correlation between α-CA evolution and the broader process of life evolution. It is believed that vertebrates evolved from invertebrates, which in turn, had already undergone significant evolutionary diversification before plants emerged. It is speculated that invertebrate α-CAs evolved into two different forms, with one pathway leading to similarities with the α-CAs found in plants and the other leading to similarities with those found in vertebrates. In addition, the structure-based analysis also showed that several bacterial α-CAs have high structural similarities with vertebrate CAs (**Figure 3B**). The bacteria-containing α-CA clade is made up of 6 CAs from *Aliivibrio fisceri*, *Pseudomonas fluorescens*, a Curvibacter symbiont subsp. *Hydra magnipapillata*, and *E. Coli*. Their closest vertebrate neighbor is a CA from *Danio rerio*. Notably, H. *magnipapillata* is an aquatic organism and *A. fisceri* is a symbiote with aquatic organisms, showing a living environment which overlaps with *D. rerio*.

γ-CAs are the most conservative CAs mainly existing in bacteria and archaea, with a few in protists and plants. Although the active site of γ-CAs are also composed three His residues similar to α-CA, its functional unit is a trimer, with three active sites spanning the monomer-monomer interfaces. In this active site, the Zn^{2+} ion is coordinated by each the His residues all three subunits (**Ferry, 2010**). γ-CAs may be the most ancient CAs, considering they mainly exist in prokaryotes and no γ-CA has been identified in animals. It is believed that γ-CAs are well conserved in photosynthetic organisms, and a recent study showed that γ-CAs are subunits of the mitochondrial complex I of diatoms (**Cainzos et al., 2021**). Compared to α- and γ-CAs, β-CAs exist in more diverse organisms, covering all organisms except vertebrates and being dominant in bacteria. β-CAs also exhibited broad structural diversity in this study, and it was reported recently that some β-CAs even possess a non-catalytic CO_2 binding site (**Cronk et al., 2006**) which can be explored for further weighted motif studies.

Understanding the diverse β-CA classification and its evolution, including the correlation between type-1 and type 2 β-CAs, is still challenging, but it is worth noting that bacterial CAs ubiquitously exist in the different branches of the β-CA cluster (**Figure 4B**). This may be correlated to bacterial horizontal gene transfer (HGT), a key factor in bacterial evolution (**Zolfaghari Emameh et al., 2016**). In comparison, α-CA clades are dominated by vertebrates, invertebrates, and plants. The α-CA also shows more variety in sub-clusters, which may be a result of less extreme HGT events allowing for differentiation (**Crisp et al., 2015**).

The structure-based method used for CA classification was also applied to facilitate CA identification from a low sequence similarity pool (**Figure 5**). Combined with TM-weighted, the analysis was further focused on the potentially active CA candidates. Surprisingly, despite low sequence similarities and His weighting in alignment, some proteins still showed high structure similarity with well aligned active sites (**Figure 5B**). These results indicated that the simulated structure-based strategy developed has possible applications in more accurate methods for protein annotation. No doubt structure-based analyses will become more and more effective for protein annotation and classification as more powerful structure prediction models are developed.

Data availability statement

All datasets generated for this study are included in the article/supplementary material.

Conflict of interest statement

The author claims no conflict of interest related to this paper.

References

Agarwal, T., Singla, R. K., & Garg, A. (2019). Carbonic Anhydrases and their Physiological Roles. *Proceedings of MOL2NET 2019, International Conference on Multidisciplinary Sciences, 5th Edition*, 6764. https://doi.org/10.3390/mol2net-05-06764

Boone, C. D., Habibzadegan, A., Gill, S., & McKenna, R. (2013). Carbonic Anhydrases and Their Biotechnological Applications. *Biomolecules*, *3*(3), 553–562. https://doi.org/10.3390/biom3030553

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & De Hoon, M. J. L. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Crisp, A., Boschetti, C., Perry, M., Tunnacliffe, A., & Micklem, G. (2015). Expression of multiple horizontally acquired genes is a hallmark of both vertebrate and invertebrate genomes. *Genome Biology*, *16*(1), 50. https://doi.org/10.1186/s13059-015-0607-3

Cronk, J. D., Rowlett, R. S., Zhang, K. Y. J., Tu, C., Endrizzi, J. A., Lee, J., Gareiss, P. C., & Preiss, J. R. (2006). Identification of a novel noncatalytic bicarbonate binding site in eubacterial beta-carbonic anhydrase. *Biochemistry*, *45*(14), 4351–4361. https://doi.org/10.1021/bi052272q

Del Prete, S., Vullo, D., Fisher, G. M., Andrews, K. T., Poulsen, S.-A., Capasso, C., & Supuran, C. T. (2014). Discovery of a new family of carbonic anhydrases in the malaria pathogen Plasmodium falciparum—The η-carbonic anhydrases. *Bioorganic & Medicinal Chemistry Letters*, *24*(18), 4389–4396. https://doi.org/10.1016/j.bmcl.2014.08.015

Del Prete, S., Vullo, D., Ghobril, C., Hitce, J., Clavaud, C., Marat, X., Capasso, C., & Supuran, C. T. (2019). Cloning, Purification, and Characterization of a β-Carbonic Anhydrase from Malassezia restricta, an Opportunistic Pathogen Involved in Dandruff and Seborrheic Dermatitis. *International Journal of Molecular Sciences*, *20*(10), 2447. https://doi.org/10.3390/ijms20102447

Ferry, J. G. (2010). The γ Class of Carbonic Anhydrases. *Biochimica et Biophysica Acta*, *1804*(2), 374. https://doi.org/10.1016/j.bbapap.2009.08.026

Hewett-Emmett, D., & Tashian, R. E. (1996). Functional diversity, conservation, and convergence in the evolution of the alpha-, beta-, and gamma-carbonic anhydrase gene families. *Molecular Phylogenetics and Evolution*, *5*(1), 50–77. https://doi.org/10.1006/mpev.1996.0006

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., … Yeats, C. (2009). InterPro: The integrative protein signature database. *Nucleic Acids Research*, *37*(Database issue), D211–D215. https://doi.org/10.1093/nar/gkn785

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

Letunic, I., & Bork, P. (2024). Interactive Tree of Life (iTOL) v6: Recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Research*, *52*(W1), W78–W82. https://doi.org/10.1093/nar/gkae268

Liljas, A., & Laurberg, M. (2000). A wheel invented three times. *EMBO Reports*, *1*(1), 16–17. https://doi.org/10.1093/embo-reports/kvd016

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., & Shmueli, Y. (2023). *Evolutionary-scale prediction of atomic level protein structure with a language model*.

Lindskog, S. (1997). Structure and mechanism of carbonic anhydrase. *Pharmacology & Therapeutics*, *74*(1), 1–20. https://doi.org/10.1016/s0163-7258(96)00198-2

Mi, H., Lazareva-Ulitsky, B., Loo, R., Kejariwal, A., Vandergriff, J., Rabkin, S., Guo, N., Muruganujan, A., Doremieux, O., Campbell, M. J., Kitano, H., & Thomas, P. D. (2005). The PANTHER database of protein families, subfamilies, functions and pathways. *Nucleic Acids Research*, *33*(suppl_1), D284–D288. https://doi.org/10.1093/nar/gki078

Moi, D., Bernard, C., Steinegger, M., Nevers, Y., Langleib, M., & Dessimoz, C. (2023). *Structural phylogenetics unravels the evolutionary diversification of communication systems in gram-positive bacteria and their viruses* (p. 2023.09.19.558401). bioRxiv. https://doi.org/10.1101/2023.09.19.558401

Needleman, S. B., & Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology*, *48*(3), 443–453. https://doi.org/10.1016/0022-2836(70)90057-4

Nett, R. S., Dho, Y., Tsai, C., Passow, D., Martinez Grundman, J., Low, Y.-Y., & Sattely, E. S. (2023). Plant carbonic anhydrase-like enzymes in neuroactive alkaloid biosynthesis. *Nature*, *624*(7990), 182–191. https://doi.org/10.1038/s41586-023-06716-y

Phelan, D. E., Mota, C., Lai, C., Kierans, S. J., & Cummins, E. P. (2021). Carbon dioxide-dependent signal transduction in mammalian systems. *Interface Focus*, *11*(2), 20200033. https://doi.org/10.1098/rsfs.2020.0033

Piovesan, D., Del Conte, A., Clementel, D., Monzon, A. M., Bevilacqua, M., Aspromonte, M. C., Iserte, J. A., Orti, F. E., Marino-Buslje, C., & Tosatto, S. C. E. (2023). MobiDB: 10 years of intrinsically disordered proteins. *Nucleic Acids Research*, *51*(D1), D438–D444. https://doi.org/10.1093/nar/gkac1065

Ribeiro, A. J. M., Riziotis, I. G., Borkakoti, N., & Thornton, J. M. (2023). Enzyme function and evolution through the lens of bioinformatics. *Biochemical Journal*, *480*(22), 1845–1863. https://doi.org/10.1042/BCJ20220405

Ruusuvuori, E., & Kaila, K. (2014). Carbonic anhydrases and brain pH in the control of neuronal excitability. *Sub-Cellular Biochemistry*, *75*, 271–290. https://doi.org/10.1007/978-94-007-7359-2_14

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., Lopez, R., McWilliam, H., Remmert, M., Söding, J., Thompson, J. D., & Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, *7*(1), 539. https://doi.org/10.1038/msb.2011.75

Smith, K. S., & Ferry, J. G. (2000). Prokaryotic carbonic anhydrases. *FEMS Microbiology Reviews*, *24*(4), 335–366. https://doi.org/10.1111/j.1574-6976.2000.tb00546.x

Smith, K. S., Jakubzick, C., Whittam, T. S., & Ferry, J. G. (1999). Carbonic anhydrase is an ancient enzyme widespread in prokaryotes. *Proceedings of the National Academy of Sciences*, *96*(26), 15184–15189. https://doi.org/10.1073/pnas.96.26.15184

Saitou, N., & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, *4*(4), 406–425. https://doi.org/10.1093/oxfordjournals.molbev.a040454

Stamatakis, A. (2006). RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, *22*(21), 2688–2690. https://doi.org/10.1093/bioinformatics/btl446

Supuran, C. T. (2016). Structure and function of carbonic anhydrases. *Biochemical Journal*, *473*(14), 2023–2032. https://doi.org/10.1042/BCJ20160115

Wu, R., Ding, F., Wang, R., Shen, R., Zhang, X., Luo, S., Su, C., Wu, Z., Xie, Q., Berger, B., Ma, J., & Peng, J. (2022). *High-resolution de novo structure prediction from primary sequence* (p. 2022.07.21.500999). bioRxiv. https://doi.org/10.1101/2022.07.21.500999

Xu, J., & Zhang, Y. (2010). How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics*, *26*(7), 889–895. https://doi.org/10.1093/bioinformatics/btq066

Yeats, C., Lees, J., Reid, A., Kellam, P., Martin, N., Liu, X., & Orengo, C. (2008). Gene3D: Comprehensive structural and functional annotation of genomes. *Nucleic Acids Research*, *36*(Database issue), D414–D418. https://doi.org/10.1093/nar/gkm1019

Zhang, Y., & Skolnick, J. (2004). Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function, and Bioinformatics*, *57*(4), 702–710. https://doi.org/10.1002/prot.20264

Zolfaghari Emameh, R., Barker, H. R., Tolvanen, M. E. E., Parkkila, S., & Hytönen, V. P. (2016). Horizontal transfer of β-carbonic anhydrase genes from prokaryotes to protozoans, insects, and nematodes. *Parasites & Vectors*, *9*(1), 152. https://doi.org/10.1186/s13071-016-1415-7

Tables

Table 1. The summarized information of CAs from PDB

Class	Frequency	Percentage
α- CA	10	13.5%
β-CA	17	23.0%
γ-CA	5	6.8%
ι-CA	1	1.4%
Others	41	56.8%

Table 2. The summarized information of CAs in selected organisms

Class	Numbers	Percentage	Organisms (ordered by relative frequency)
α- CA	862	55.5%	Vertebrates, Invertebrates, Plants, Protists
β-CA	454	29.2%	Archaea, Bacteria, Plants,
γ-CA	195	12.6%	Archaea, Bacteria, Plants, Protists
Others	43	2.8%	Invertebrates, Plants, Protists

Table 3. Normalized RF distances of PDB trees

	A	B	C	D
A	0	0.78	0.96	0.72
B	0.78	0	0.93	0.80
C	0.96	0.93	0	0.96
D	0.72	0.80	0.96	0

Figures

Figure 1

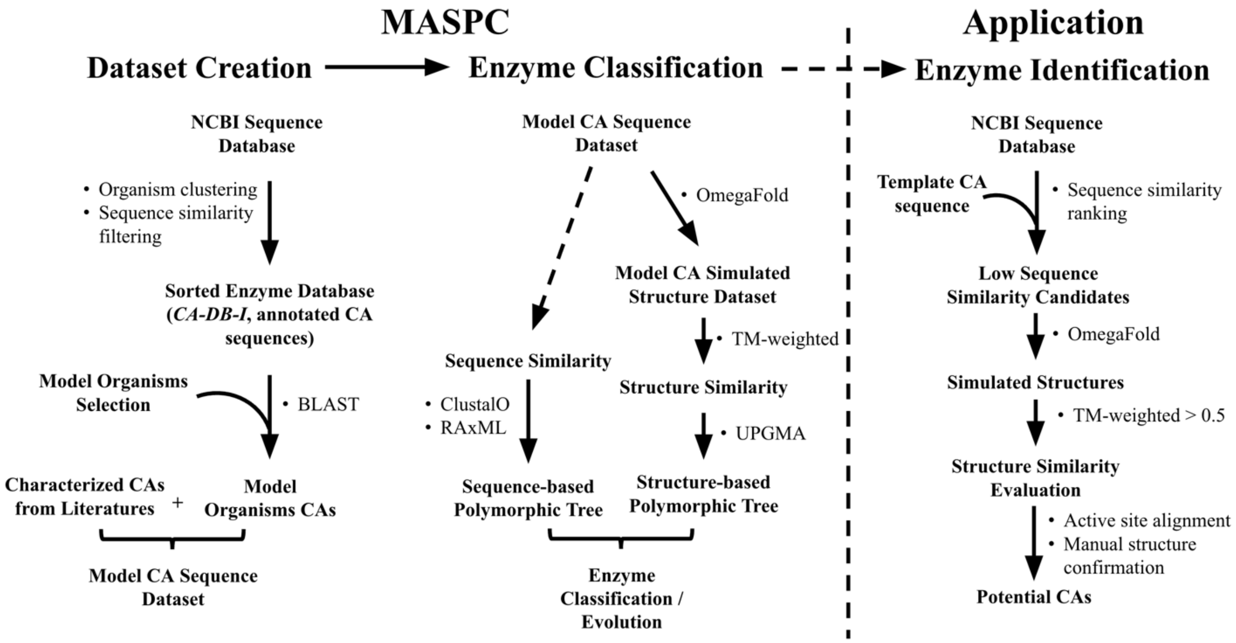

MASPC
Application
Dataset Creation
Enzyme Classification
Enzyme Identification
NCBI Sequence Database
• Organism clustering
• Sequence similarity filtering
Sorted Enzyme Database (CA-DB-I, annotated CA sequences)
Model Organisms Selection
• BLAST
Characterized CAs from Literatures
+
Model Organisms CAs
Model CA Sequence Dataset
Model CA Sequence Dataset
• OmegaFold
Model CA Simulated Structure Dataset
• TM-weighted
Sequence Similarity
Structure Similarity
• ClustalO
• RAxML
• UPGMA
Sequence-based Polymorphic Tree
Structure-based Polymorphic Tree
Enzyme Classification / Evolution
NCBI Sequence Database
Template CA sequence
• Sequence similarity ranking
Low Sequence Similarity Candidates
• OmegaFold
Simulated Structures
• TM-weighted > 0.5
Structure Similarity Evaluation
• Active site alignment
• Manual structure confirmation
Potential CAs

Figure 1. Motif-weighted Structure Alignment for Enzyme Classification (MASPC) strategy and its expanded application in CA identification

Figure 2

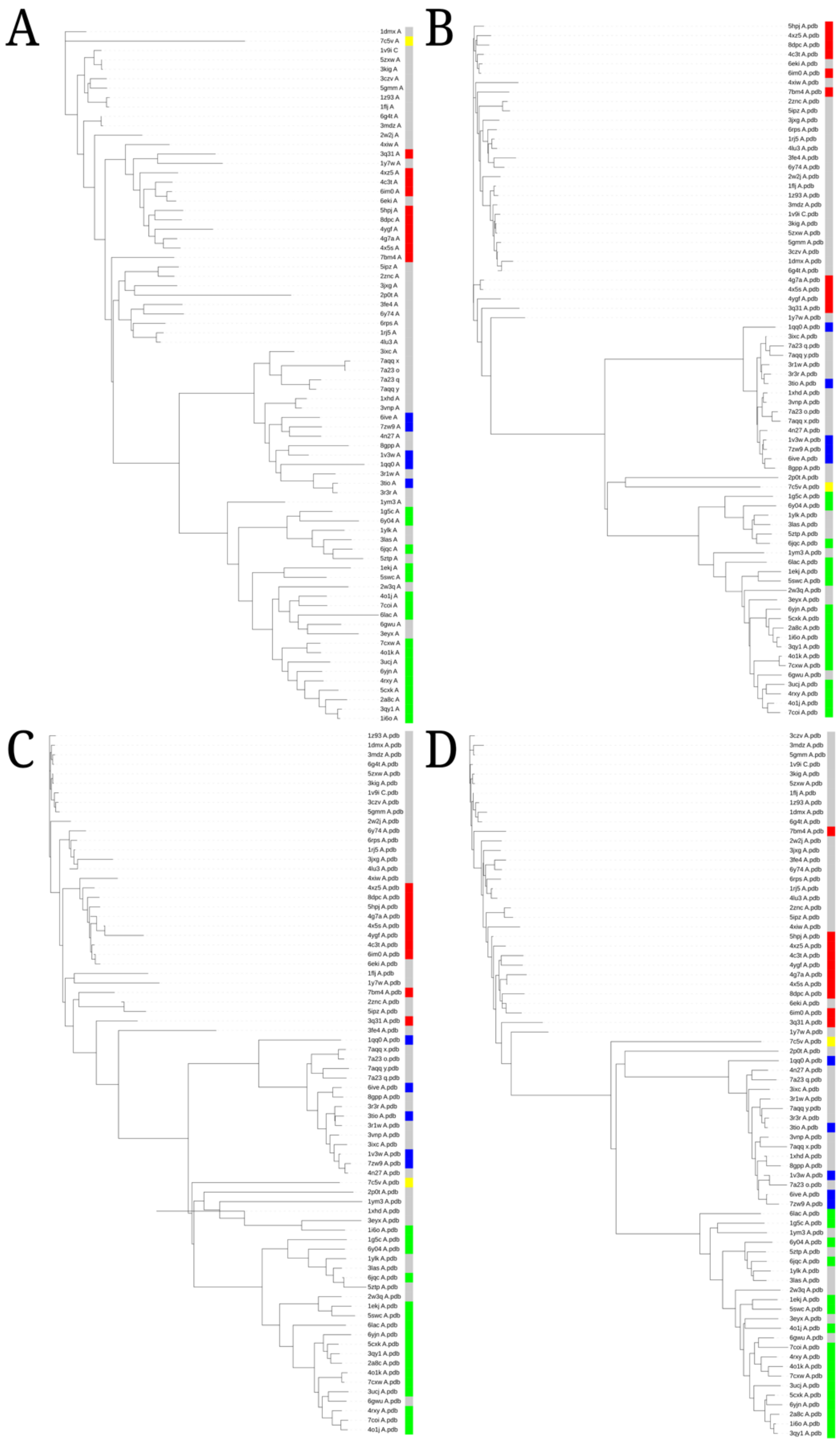

A
1dmx A
7c5v A
1v9i C
5zxw A
3kig A
3czv A
5gmm A
1z93 A
1flj A
6g4t A
3mdz A
2w2j A
4xiw A
3q31 A
1y7w A
4xz5 A
4c3t A
6im0 A
6eki A
5hpj A
8dpc A
4ygf A
4g7a A
4x5s A
7bm4 A
5ipz A
2znc A
3jxg A
2p0t A
3fe4 A
6y74 A
6rps A
1rj5 A
4lu3 A
3ixc A
7aqq x
7a23 o
7a23 q
7aqq y
1xhd A
3vnp A
6ive A
7zw9 A
4n27 A
8gpp A
1v3w A
1qq0 A
3r1w A
3tio A
3r3r A
1ym3 A
1g5c A
6y04 A
1ylk A
3las A
6jqc A
5ztp A
1ekj A
5swc A
2w3q A
4o1j A
7coi A
6lac A
6gwu A
3eyx A
7cxw A
4o1k A
3ucj A
6yjn A
4rxy A
5cxk A
2a8c A
3qy1 A
1i6o A
B
5hpj A.pdb
4xz5 A.pdb
8dpc A.pdb
4c3t A.pdb
6eki A.pdb
6im0 A.pdb
4xiw A.pdb
7bm4 A.pdb
2znc A.pdb
5ipz A.pdb
3jxg A.pdb
6rps A.pdb
1rj5 A.pdb
4lu3 A.pdb
3fe4 A.pdb
6y74 A.pdb
2w2j A.pdb
1flj A.pdb
1z93 A.pdb
3mdz A.pdb
1v9i C.pdb
3kig A.pdb
5zxw A.pdb
5gmm A.pdb
3czv A.pdb
1dmx A.pdb
6g4t A.pdb
4g7a A.pdb
4x5s A.pdb
4ygf A.pdb
3q31 A.pdb
1y7w A.pdb
1qq0 A.pdb
3ixc A.pdb
7a23 q.pdb
7aqq y.pdb
3r1w A.pdb
3r3r A.pdb
3tio A.pdb
1xhd A.pdb
3vnp A.pdb
7a23 o.pdb
7aqq x.pdb
4n27 A.pdb
1v3w A.pdb
7zw9 A.pdb
6ive A.pdb
8gpp A.pdb
2p0t A.pdb
7c5v A.pdb
1g5c A.pdb
6y04 A.pdb
1ylk A.pdb
3las A.pdb
5ztp A.pdb
6jqc A.pdb
1ym3 A.pdb
6lac A.pdb
1ekj A.pdb
5swc A.pdb
2w3q A.pdb
3eyx A.pdb
6yjn A.pdb
5cxk A.pdb
2a8c A.pdb
1i6o A.pdb
3qy1 A.pdb
4o1k A.pdb
7cxw A.pdb
6gwu A.pdb
3ucj A.pdb
4rxy A.pdb
4o1j A.pdb
7coi A.pdb
C
1z93 A.pdb
1dmx A.pdb
3mdz A.pdb
6g4t A.pdb
5zxw A.pdb
3kig A.pdb
1v9i C.pdb
3czv A.pdb
5gmm A.pdb
2w2j A.pdb
6y74 A.pdb
6rps A.pdb
1rj5 A.pdb
3jxg A.pdb
4lu3 A.pdb
4xiw A.pdb
4xz5 A.pdb
8dpc A.pdb
5hpj A.pdb
4g7a A.pdb
4x5s A.pdb
4ygf A.pdb
4c3t A.pdb
6im0 A.pdb
6eki A.pdb
1flj A.pdb
1y7w A.pdb
7bm4 A.pdb
2znc A.pdb
5ipz A.pdb
3q31 A.pdb
3fe4 A.pdb
1qq0 A.pdb
7aqq x.pdb
7a23 o.pdb
7aqq y.pdb
7a23 q.pdb
6ive A.pdb
8gpp A.pdb
3r3r A.pdb
3tio A.pdb
3r1w A.pdb
3vnp A.pdb
3ixc A.pdb
1v3w A.pdb
7zw9 A.pdb
4n27 A.pdb
7c5v A.pdb
2p0t A.pdb
1ym3 A.pdb
1xhd A.pdb
3eyx A.pdb
1i6o A.pdb
1g5c A.pdb
6y04 A.pdb
1ylk A.pdb
3las A.pdb
6jqc A.pdb
5ztp A.pdb
2w3q A.pdb
1ekj A.pdb
5swc A.pdb
6lac A.pdb
6yjn A.pdb
5cxk A.pdb
3qy1 A.pdb
2a8c A.pdb
4o1k A.pdb
7cxw A.pdb
3ucj A.pdb
6gwu A.pdb
4rxy A.pdb
7coi A.pdb
4o1j A.pdb
D
3czv A.pdb
3mdz A.pdb
5gmm A.pdb
1v9i C.pdb
3kig A.pdb
5zxw A.pdb
1flj A.pdb
1z93 A.pdb
1dmx A.pdb
6g4t A.pdb
7bm4 A.pdb
2w2j A.pdb
3jxg A.pdb
3fe4 A.pdb
6y74 A.pdb
6rps A.pdb
1rj5 A.pdb
4lu3 A.pdb
2znc A.pdb
5ipz A.pdb
4xiw A.pdb
5hpj A.pdb
4xz5 A.pdb
4c3t A.pdb
4ygf A.pdb
4g7a A.pdb
4x5s A.pdb
8dpc A.pdb
6eki A.pdb
6im0 A.pdb
3q31 A.pdb
1y7w A.pdb
7c5v A.pdb
2p0t A.pdb
1qq0 A.pdb
4n27 A.pdb
7a23 q.pdb
3ixc A.pdb
3r1w A.pdb
7aqq y.pdb
3r3r A.pdb
3tio A.pdb
3vnp A.pdb
7aqq x.pdb
1xhd A.pdb
8gpp A.pdb
1v3w A.pdb
7a23 o.pdb
6ive A.pdb
7zw9 A.pdb
6lac A.pdb
1g5c A.pdb
1ym3 A.pdb
6y04 A.pdb
5ztp A.pdb
6jqc A.pdb
1ylk A.pdb
3las A.pdb
2w3q A.pdb
1ekj A.pdb
5swc A.pdb
3eyx A.pdb
4o1j A.pdb
6gwu A.pdb
7coi A.pdb
4rxy A.pdb
4o1k A.pdb
7cxw A.pdb
3ucj A.pdb
5cxk A.pdb
6yjn A.pdb
2a8c A.pdb
1i6o A.pdb
3qy1 A.pdb

Figure 2. Polymorphic analysis of annotated CAs from the PDB.A: Sequence-based phylogenetic tree generated using RAxML and ClustalO with the best result from 100 iterations. B: Structure-based phylogenetic tree generated using TM-score and NJ based on PDB structures. C: Phylogenetic tree generated using TM-weighted and NJ based on PDB structures. D: Phylogenetic tree generated using TM-weighted and NJ based on predicted structures.

Figure 3

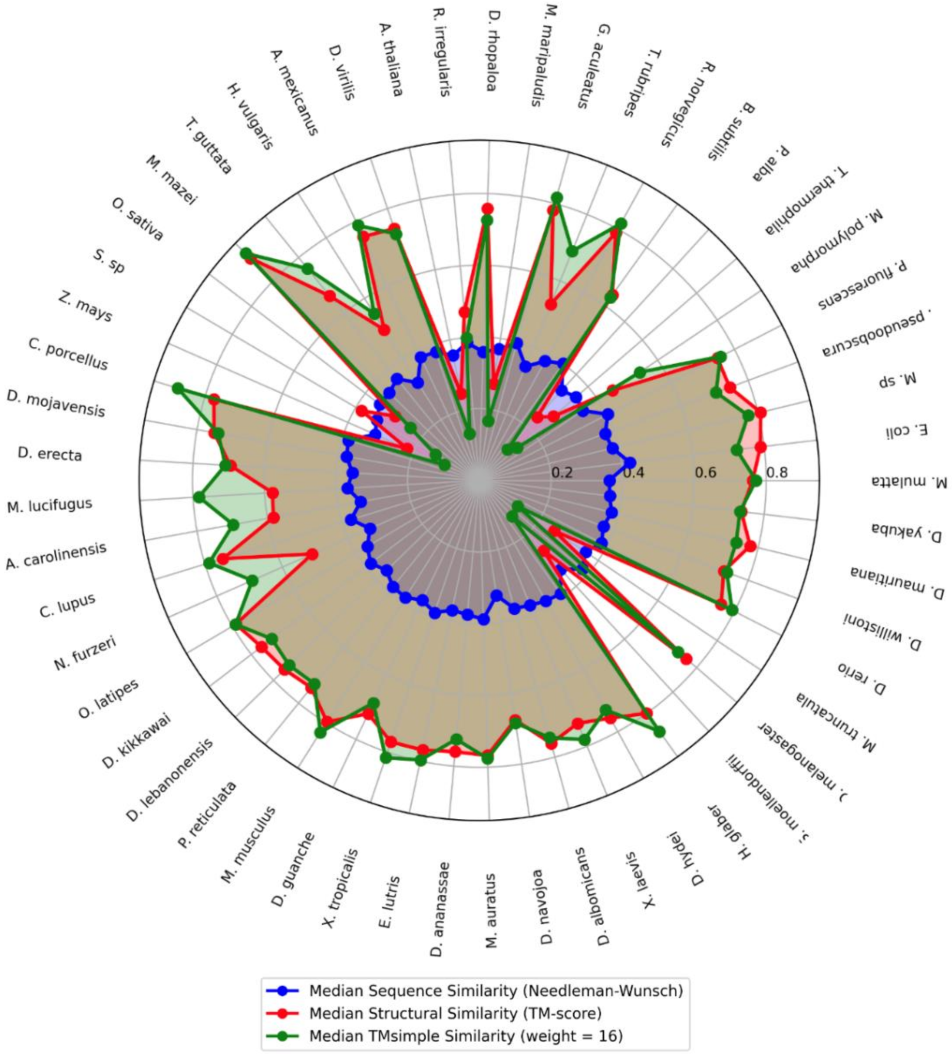


Figure 3. **Variance of CA sequences from selected model organisms**. The number of CAs in representative organisms are presented with the sequence similarities, TM-score-based structure similarity, and TM-weighted-based and structure similarity of CAs in the same organism are calculated.

Figure 4

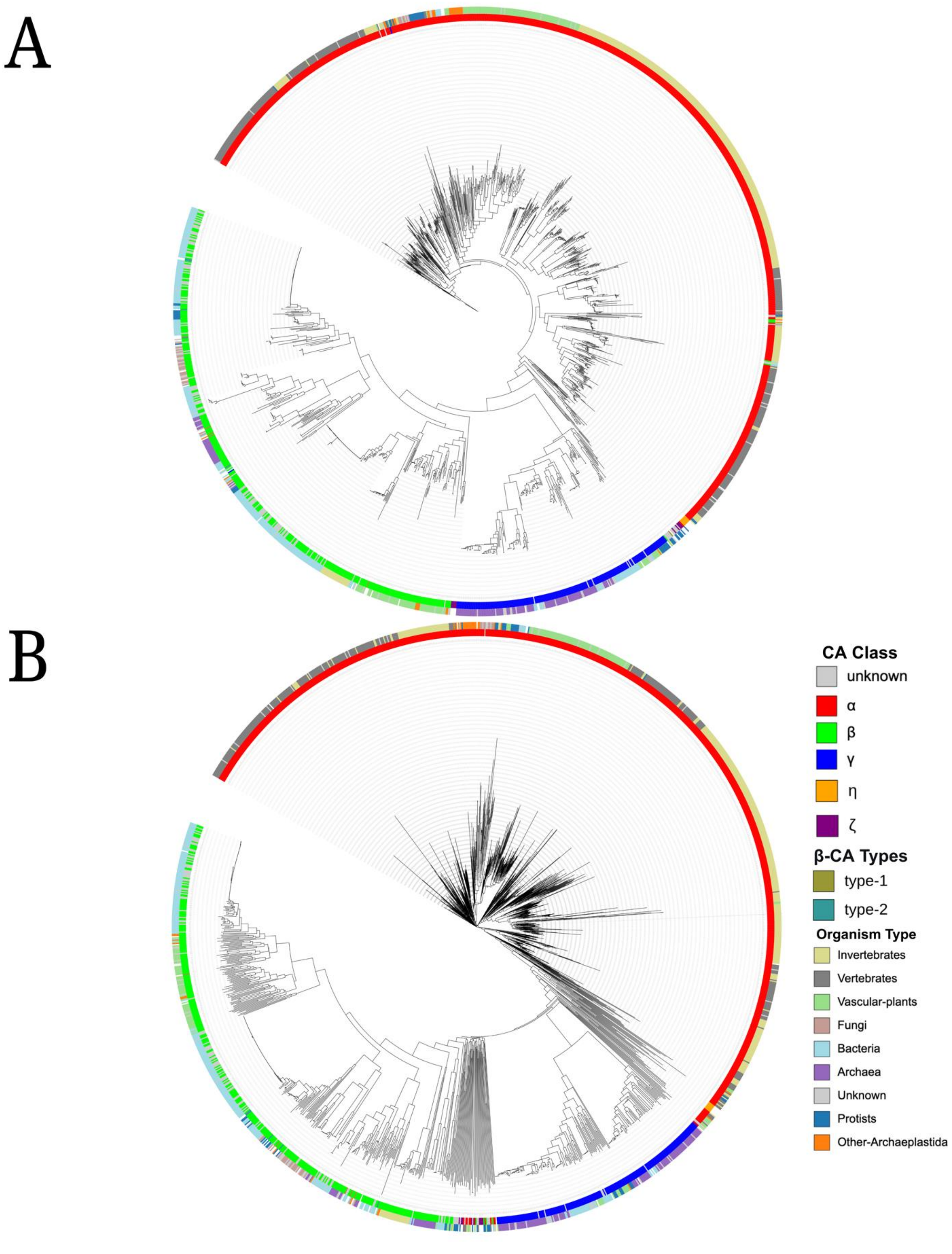
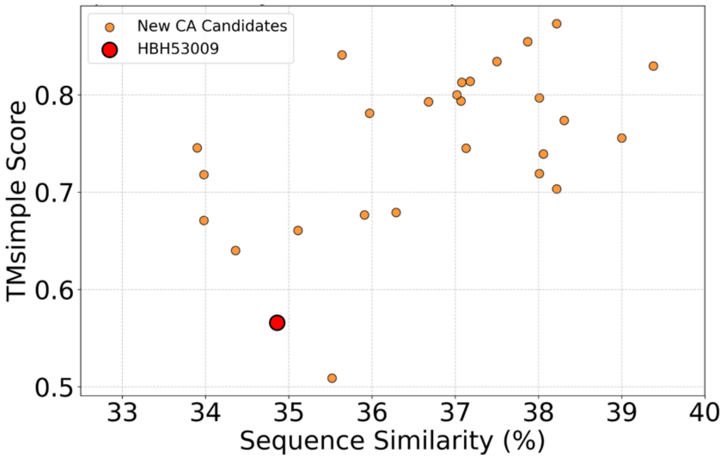

A
B
CA Class
unknown
α
β
γ
η
ζ
β-CA Types
type-1
type-2
Organism Type
Invertebrates
Vertebrates
Vascular-plants
Fungi
Bacteria
Archaea
Unknown
Protists
Other-Archaeplastida

Figure 4. Polymorphic analysis of CAs selected from the model organisms.

A: Sequence-based phylogenetic tree generated using RAxML and ClustalO with the best result from 100 iterations. B: Structure-based phylogenetic tree generated using TM-weighted and NJ based on predicted structures.

Figure 5

A



B

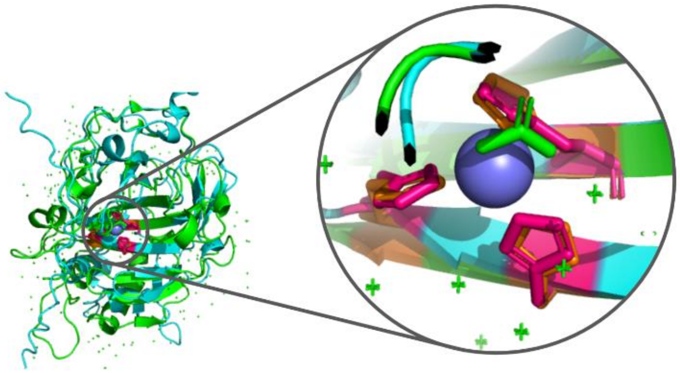


Figure 5. Application of structure-based CA classification for identifying low sequence similarity CAs from NCBI database. A) Analysis of NCBI database using HCA-II as template sequence. After the sequence blasting, the sequences with similarity in 30-39% bin (**Figure S3**) were selected for OmegaFold-mediated protein structure simulation, followed by TM-score-based structure similarity evaluation. B) Protein structure of the predicted structure HBH53009.1 aligned with HCA-II, 1BIC. Close up of His-Zn^{2+} active site.

Supplementary Information

Motif-weighted Structure Alignment for Classification and Evolutionary Studies of Carbonic Anhydrase

Hongyi Shi

Henry M Gunn High School, Palo Alto, CA 94306, USA

Email: hongyi.z.shi@gmail.com

Supplement Information File 1

File 1: CA-DB-I

Supplement Information File 2

Sheet 1: β-CAs

Sheet 2: Selected CAs

Sheet 3: Model Proteins

Sheet 4: Model Organisms

Sheet 5: PDB CAs

Supplemental Figures

Figure S1

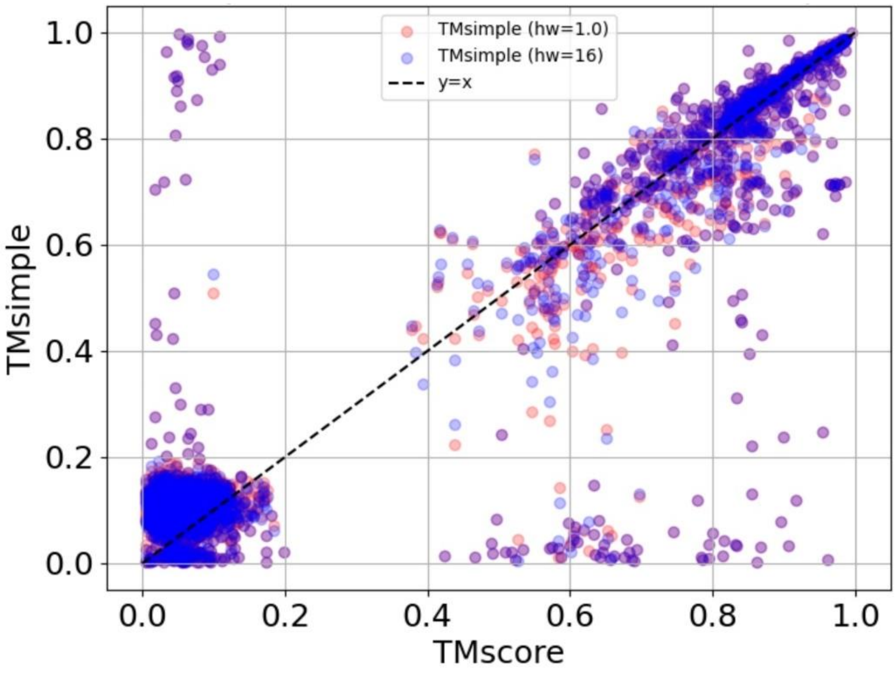


Figure S1. Comparison of TM-score with TM-weighted under different motif weights. Alignments are performed on all pairs of the 74 PDB CA structures.

Figure S2

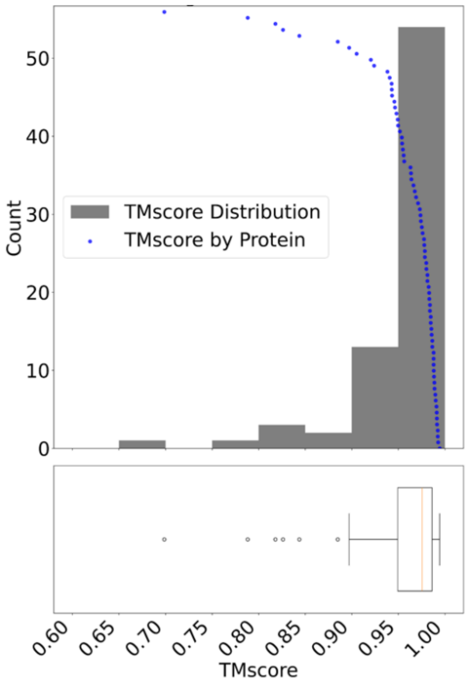


Figure S2. PDB vs OmegaFold TM-score distribution OmegaFold is shown to be an accurate protein prediction method which can be used for further CA prediction applications performed in this study.

Figure S3

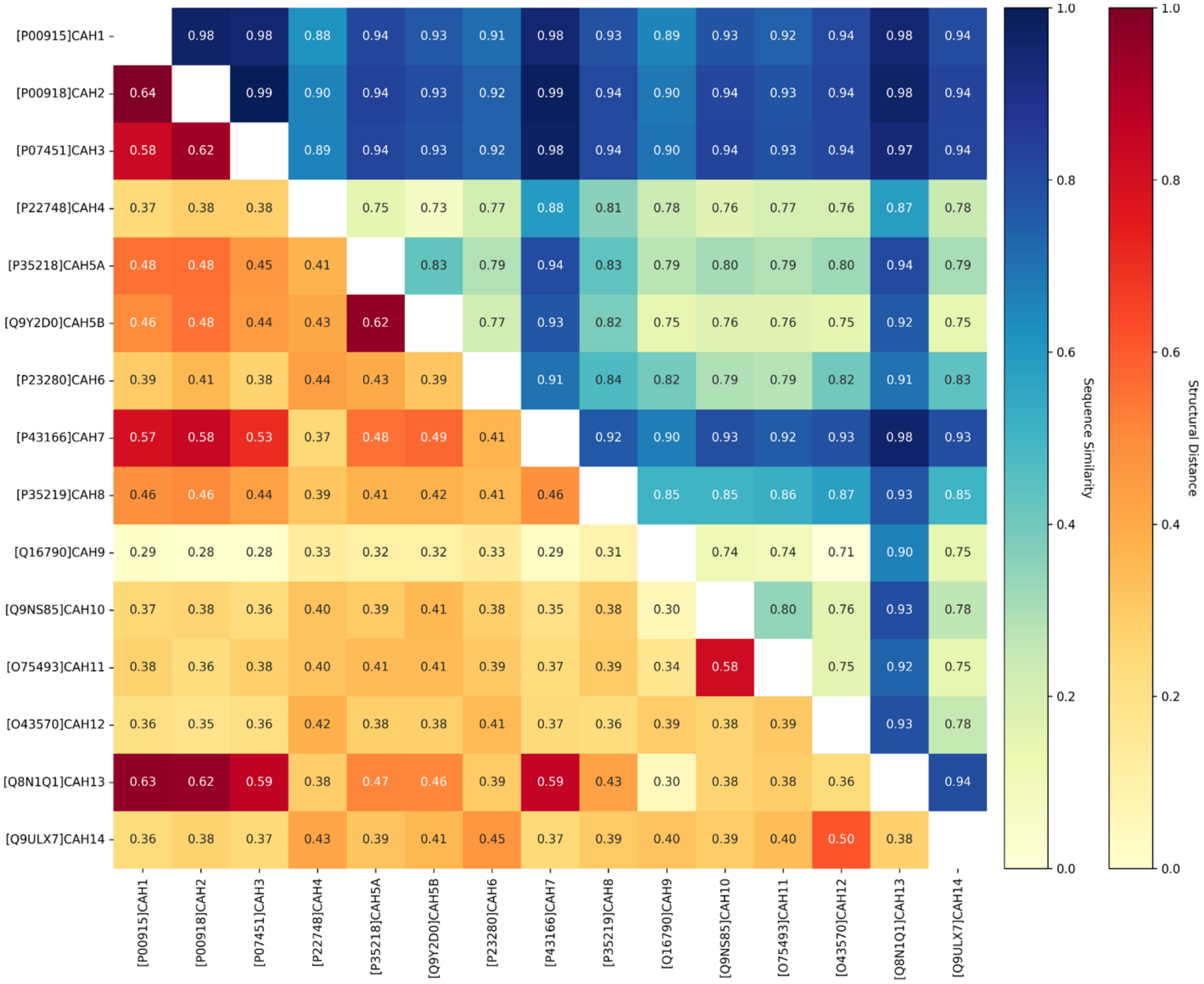


Figure S3. Comparison of structure and sequence similarity for human carbonic anhydrase isoforms. In contrast with low sequence similarities, the high structural similarity between structures show the diversity of CA sequences, yet relative uniformity of same-class CAs.

Figure S4

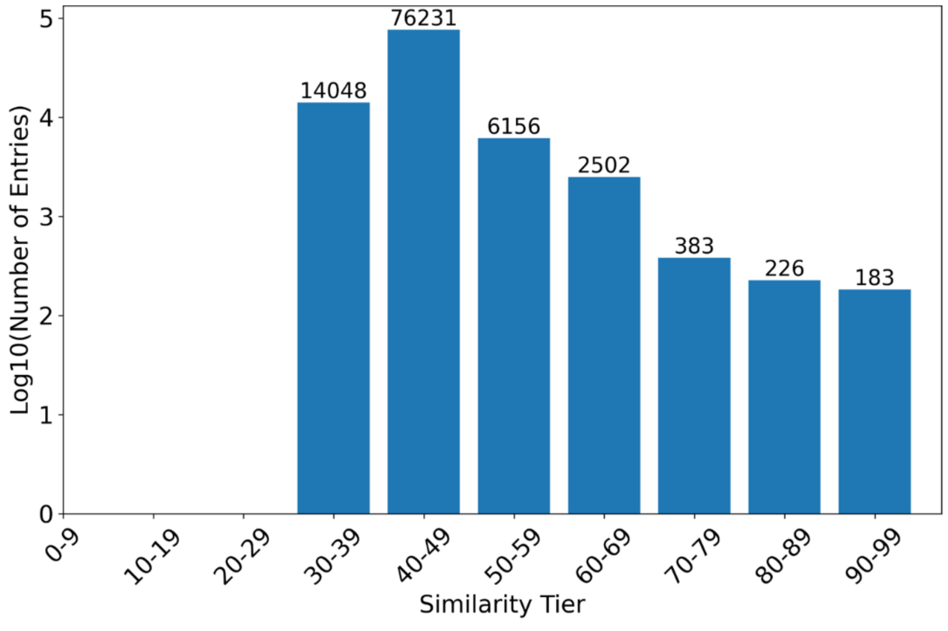


Figure S4. Numbers of CAs in each similarity tier after sequence blasting with HCAII in NCBI database. The 30-39% similarity bin is the most promisi/ng range to search in due to high numbers of potentials CAs and low sequence similarities.

Figure S5

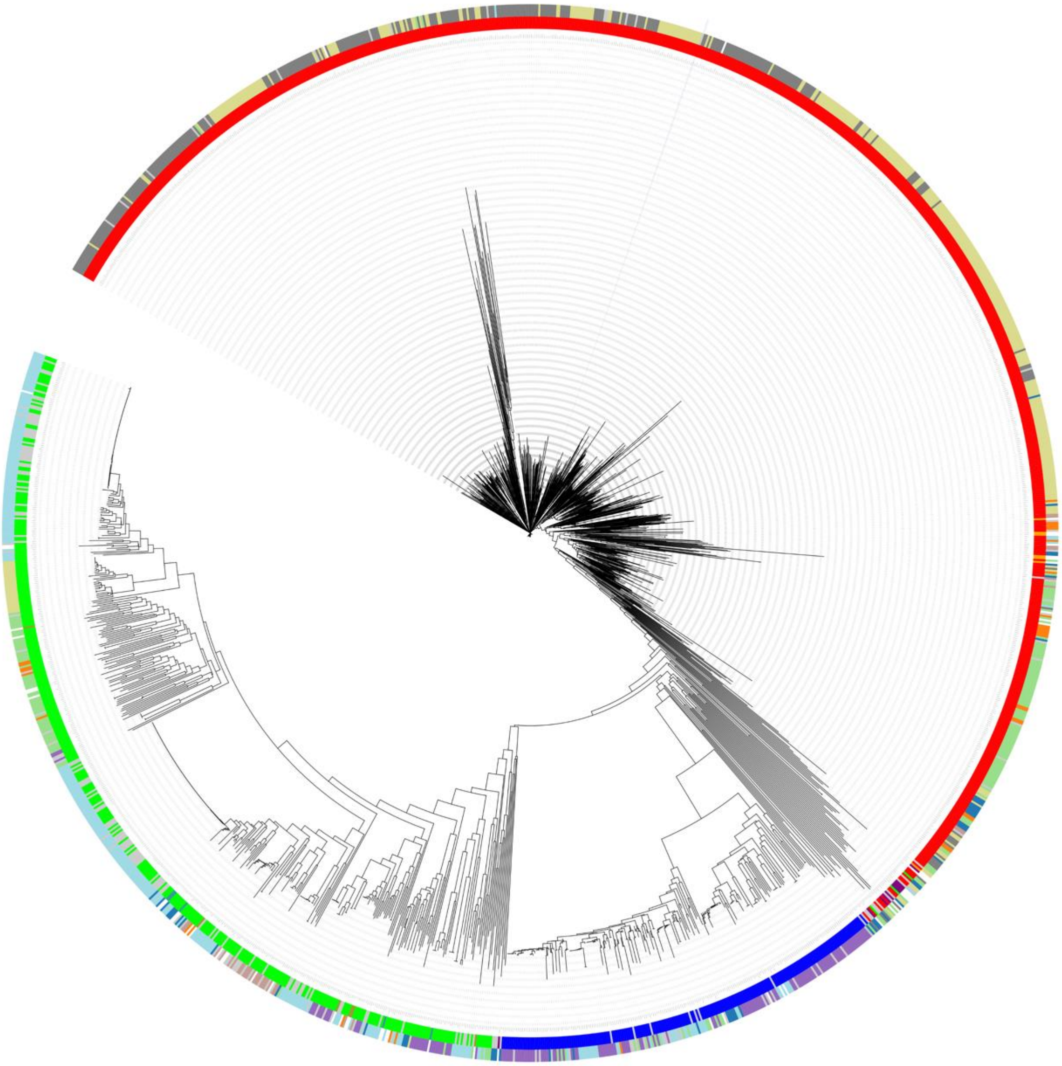


Figure S5. Structure-based phylogenetic tree generated using TM-score and NJ