Defining Lists of Conserved Amino Acid Residues in Alpha Carbonic Anhydrases

Master's in Bioinformatics Thesis
Rawnak Jahan Hoque
Institute of Biosciences and Medical Technology
(BioMediTech)
University of Tampere, Finland
December 2014

Acknowledgements

This research work was carried out at the Tissue Biology research group of the School of medicine in the Institute of Biosciences and Medical Technology (BioMediTech), University of Tampere. I express my sincere thanks to Professor Seppo Parkkila for warmly welcoming me in his research group. His valuable thoughts and guidelines during the group meetings provided me critical insights about the research processes on carbonic anhydrase.

I render my deepest gratitude and thank to my supervisor Dr. Martti Tolvanen for supervising the work and reviewing my thesis. Without his tremendous support and efficient supervision, this work would have been very difficult. His appropriate guidance, constructive criticism and cordial advice helped me to reach my goal.

I would like to thank Harlan Barker for providing me his software tool for the assistance of my analysis and giving valuable suggestions on the language of this thesis. I would also like to thank Professor Matti Nykter for reviewing my thesis. I express my profound gratitude to all my course teachers who taught me during this MSc. degree.

My exceptional thank goes to my husband Munir Hossain for his inspiration, guidance and moral support during the entire period of my study. I would like to thank my sisters for their support during the difficult times. I like to express my gratefulness to my parents for their endless support and motivation during the journey of my study. Their precious advice and love are the key strength of my life.

Master's Thesis

Place	UNIVERSITY OF TAMPERE Tissue Biology group, School of Medicine Institute of Biosciences and Medical Technology
Author	HOQUE, RAWNAK JAHAN
Title	Defining Lists of Conserved Amino Acid Residues in Alpha Carbonic Anhydrases
Pages	72
Supervisors	Dr. Martti Tolvanen; Professor Seppo Parkkila
Reviewers	Professor Matti Nykter; Dr. Martti Tolvanen
Date	December 2014

Abstract

Background and Aims

Carbonic anhydrases comprise a large enzyme family that catalyzes the reversible conversion of carbon dioxide to bicarbonate for controlling the acid-base balance in blood and other types of tissues in almost all types of living bodies. Conservation study is an indispensable approach to identify the functional elements in proteins. This can help with invention of inhibitors of diseases, determination of the structure, protein-protein interfaces etc. The goal of this research is to trace the conserved residues that are shared among all alpha carbonic anhydrase isoforms in vertebrates, most notably those CAs containing Histidines in the active center. The study of sequence conservation of all α-CA isozymes is important to do comparative analysis among different isozymes and define their functional significance. There are a few other conserved residues that have been recorded in previous literature, but the conservation profiles have not been studied exhaustively.

Methods

To facilitate the study a Python-based pipeline was used that can automatically retrieve a maximal number of orthologous sequences from the Ensembl database, do quality checks, and quantify conservation at each residue based on the K_a/K_s approach of an automatically generated codon-based alignments. A comparison made for the conservation profiles of individual isozyme results from previous output, and also comparing these results to conservation profiles of two largest groups that is conservation in all cytoplasmic and extracellular isozymes.

Results

I have produced a complete and definitive list of absolutely and highly conserved residues in the alpha CAs of tetrapods. Ninety percent of the conserved residues were shown to be buried in the protein core. Structural and functional roles of the individual residues were identified by literature review and inspection of structures, and high-quality visualizations were produced in the human CA-II 3D crystallographic structure. Complete list of residues conserved exclusively in cytoplasmic and extracellular CAs were made and compared to reveal that the cytoplasmic isozymes might share common binding sites on the surface for interacting with other molecules whereas the extracellular isoforms have unique surfaces. Finally, N-linked glycosylated sites of CA-VI, IX, XII, and XIV were studied. It was seen that these extracellular isoforms did not share any precise glycosylated positions. However, many glycosylation sites were observed positioned at the entrance of the active cavity, which may facilitate the protein not to interact with other proteins that might block the active site.

Conclusions

This thesis constitutes the most extensive structural interpretation of the roles of conserved residues in alpha CAs thus far. I have discovered previously undocumented structural features and interpretations for several universally conserved residues (Trp-16, Gln-28, Pro-30, Asn-61, Leu-44, Ser-105, His-122, Ala-134, Ala-142, Pro-186, Tyr-194, Ser-197, Pro-201, Gln-222, Asn-244, Arg-246, and Arg-254). The comparison for the conservation profile of cytoplasmic and extracellular isozymes revealed a possible common protein-binding interface in the cytoplasmic isoforms. Finally, it was speculated from the visual comparison of conserved N-glycosylation sites that the glycosylation sites around the passage of the catalytic cavity may inhibit interactions with other proteins, and keep a clear passage to the active site.

Abbreviations

CA	Carbonic anhydrase
hCA	Human Carbonic anhydrase
RSA	Relative Solvent Accessibility
GPI	Glycosylphosphatidylinisitol
MSA	Multiple Sequence Alignment
DSSP	Dictionary of Secondary Structure Prediction
CDS	Coding DNA Sequence
DNA	Deoxyribonucleic Acid
cDNA	Complementary Deoxyribonucleic Acid
3D	Three Dimensional
BLAST	Basic Local Alignment Search Tool
PDB	Protein Data Bank
POV-ray	Persistence of Vision Raytracer
SA	Solvent Accessibility
POOL	Partial Order Optimal Likelihood
aa	Amino acid
Xaa	Unknown amino acid
IP_3	Inositol 1,4,5-trisphosphate
IPTR1	Inositol 1,4,5-trisphosphate receptor type 1

Amino Acid Codes

Ala	A	Alanine
Cys	C	Cysteine
Asp	D	Aspartic acid
Glu	E	Glutamic acid
Phe	F	Phenylalanine
Gly	G	Glycine
His	H	Histidine
Ile	I	Isoleucine
Lys	K	Lysine
Leu	L	Leucine
Met	M	Methionine
Asn	N	Asparagine
Pro	P	Proline
Gln	Q	Glutamine
Arg	R	Arginine
Ser	S	Serine
Thr	T	Threonine
Val	V	Valine
Trp	W	Tryptophan
Tyr	Y	Tyrosine
	X	Unspecified or unknown

Table of Contents

1 Introduction 1
2 Literature Review 3
2.1 The Structure of Alpha Carbonic Anhydrases 3
2.2 Catalytic Mechanism 5
2.3 Theories of the Most Essential Methods Used for Conservation Analysis 7
2.3.1 K_a/K_s Ratio 7
2.3.2 BioPython 8
2.3.3 DSSP 8
2.3.4 PAL2NAL 9
2.3.5 Selecton 9
2.3.6 Chimera 9
2.4 Physicochemical Properties Used to Investigate and Define the Role of Universally Conserved Residues 10
2.4.1 Hydrogen Bonds 10
2.4.2 Polarity and Hydrophilicity 11
2.4.3 Aromaticity 12
2.4.4 Relative Solvent Accessibility 13
2.4.5 Hydrophobic Interaction 14
2.5 N-glycosylation Site 14
3 Aim of the Study 15
4 Material and Methods 16
4.1 Conservation Analysis for Universal, Cytoplasmic and Extracellular Group 16
4.1.1 Species and Isoform Selection 16
4.1.2 Sequence Retrieval, MSA and K_a/K_s Scoring 16
4.1.3 Manual Alignment and Universal Conserved Group 17
4.1.4 Pool Rank 19
4.1.5 Cytoplasmic and Extracellular Conserved Residues 19
4.2 Conserved N-glycosylation Site Prediction 20
4.2.1 MSA and N-glycosylation Site Identification 20
4.2.2 Categorization of the N-glycosylated Sites 20

4.2.3 Modelling of the Missing Part of CA12 Structure 21
5 Results 22
5.1 List of "Universally Conserved" Residues 22
5.2 Roles of the "Universally Conserved" Residues 24
5.3 Roles of Conserved Residues in the Active Site 25
5.3.1 His-94, His-96, and His-119 25
5.3.2 Thr-199, Thr-200, and Glu-106 26
5.3.3 Gln-92 27
5.3.4 Val-121, Leu-198, Val-207, and Val-143 27
5.3.5 His-64 28
5.4 Roles of structurally important conserved residues 29
5.4.1 Trp-16 29
5.4.2 Gln-222 29
5.4.3 Gln-249 30
5.4.4 Asn-61 30
5.4.5 Asn-244 31
5.4.6 Ser-105 31
5.4.7 Ser-29 32
5.4.8 Ser-197 32
5.4.9 Pro-201 33
5.4.10 Pro-30 33
5.4.11 Pro-186 34
5.4.12 Arg-246 and Arg-254 34
5.4.13 Gln-28 35
5.4.14 His-107 and Glu-117 35
5.4.15 Trp-97 36
5.4.16 Gly-63, Gly-197, and Gly-104 36
5.4.17 Tyr-194 and Trp-209 37
5.4.18 Leu-44 37
5.4.19 His-122 38
5.4.20 Ala-134 38
5.4.21 Ala-142 39
5.5 Statistical Analysis 39

5.6 List of Residues Conserved Only in Cytoplasmic or Extracellular CA Isozymes 41
5.7 Cytoplasmic and Extracellular Conserved Surface Visualization 42
5.8 Visualization and Comparison of N-glycosylation Sites on Structures 44
6 Discussion 45
6.1 "Universally Conserved" Residues 45
6.2 Cytoplasmic and Extracellular α-CAs 47
6.3 N-glycosylation Sites 48
Conclusions 49
References 50
Appendix 1 58
Appendix 2 63
Appendix 3 65
Appendix 4 66
Appendix 5 68

1 Introduction

Carbonic anhydrases (CA, EC 4.2.1.1) form a large protein group consisting of a number of distinct families: α, β, γ, δ and ζ. They are often called metalloenzymes as they bind a metal ion, mostly zinc, at the active center that is an essential component for the catalytic reaction. Carbonic anhydrases actively participate in the catalysis of a CO_2 (de)hydration reaction, that is crucial for the maintenance of various physiological and biochemical processes in almost all the living bodies (Dodgson 1991). They mostly control the respiration and acid-base balance in blood and other tissues throughout the rapid interconversion of carbon dioxide and bicarbonate as the carbon dioxide molecules react with waters to form bicarbonates and protons:

$$CO_2 + H_2O \rightleftharpoons HCO_3^- + H^+$$

To date the α-CA family is the most studied consisting of 16 different isozymes contributing in a wide variety of cellular functions (Esbaugh 2006).

The α-CA family is further divided into the following distinguished subfamilies according to their subcellular locations: cytosolic isozymes (CA-I, II, III, VII and XIII), mitochondrial (CA-V), transmembrane (CA-IX, XII, and XIV), secreted (CA-VI) and the GPI-linked (CA-IV and XV, XVII) (Leggat 2005) (Esbaugh 2006) (Tolvanen 2012). There is also another distinct subfamily called carbonic anhydrase related proteins (CARPs) that consists of CA-VIII, X and XI. Despite lacking important Histidines (the key catalytic elements) in the active site, CARPs are included in the α-CA family due to their highly conserved motifs across the α-CAs (Lovejoy 1998). The role of CA-VIII has already been discovered in the regulation of the calcium channel in the endoplasmic reticulum (ITPR1) and in the interaction with the IP_3 receptor (Aspatwar 2012) (Hirota 2003) but the specific roles of CA-X and XI are still unknown.

Study of the conservation profile of amino acids in proteins is an essential tool for identifying the structural and functional properties. Conserved areas can be considered to be the

important functional elements of the proteins. The regions that are conserved in a 3D or tertiary protein structure provide insights to determine protein-protein or protein-ligand interaction sites, area of the dimer interfaces, and most importantly potential inhibitor binding sites. Apart from that, conservation analysis is a powerful approach to explore the phylogenetic relationship among species, their habitat, function and evolution.

This thesis work particularly concerns conservation analysis and identification of the most important common functional elements across the α-CA family, focusing on non-ray-finned-fish jawed vertebrates. To date, there are various highly conserved residues identified that are shared between all alpha carbonic anhydrases, most notably the Histidines in the active center. There are a few other important residues that have been recorded in the literature, but the conservation profile study has not been performed thoroughly. This prompted me to study the conservation profile of the most important species group, vertebrates, and make a complete list of highly conserved residues that are functionally active. Secondly, a manual/visual comparison was made to study the conservation profiles of individual amino acids in a crystallographic structure, and define their specific structural and functional roles. For the analysis, human carbonic anhydrase II was considered as a standard reference sequence and structure due to its high catalytic rate up to $k_{cat} = 1.4*10^6\ s^{-1}$ or a million times a second (Berg 2010), availability of high quality crystallographic structures, and as it is the most well studied CA isozyme to date. Thirdly, a comparison of these results to conservation profiles of two largest groups, cytoplasmic and extracellular isozymes, were done to understand their structural and functional importance in the individual sub groups. Finally, an application of the conservation profile study was applied to predict functional and non-functional N-linked glycosylation sites in the extracellular domain of four isoforms, CA-VI, IX, XII, and XIV.

2 Literature Review

2.1 The Structure of Alpha Carbonic Anhydrases

The catalytically active alpha carbonic anhydrases are similar in structure with their conserved motifs of the active site cavity. To date, the crystallographic structure of human CA-I, II, III, IV, VI, VII, VIII, IX, XII, XIII, and XIV have been determined and are available in the protein data bank (www.PDB.org). All the alpha CAs have similar tertiary structure and centrally bind a divalent metal ion, most often a zinc (Zn^{2+}), held as a prosthetic group. The zinc ion is coordinated with three imidazole rings of histidine residues and a water molecule forming a distorted tetrahedral geometry at the cone shaped active cavity (Liljas 1972). This geometric figure is essential for accelerating the rapid reaction of CO_2 hydration (Silverman 1988). There were several studies done to understand whether all divalent metals show the same coordination geometry or not. The Zn(II) was replaced by the divalent Co(II), Ni(II), Mn(II) and Cu(II) and the result revealed that only zinc and cobalt show the tetrahedral coordination geometry at about pH-8 (Liljas 1994).

The dominating structure of the protein core is composed of ten-twisted beta sheets, where two of them are parallel and rest are antiparallel. There are seven right-handed alpha helices positioned on the surface of the molecule that are connected through some short length coils including hairpin-bends and type-I and type-II reverse turns distributed in the different points of the structure (Venkatachalam 1968) (Crawford 1973).

The active cavity of the structure is cone shaped and strictly separated into two distinct parts, one of which contains hydrophobic residues and other one contains hydrophilic residues (Chegwidden 2000). The conserved hydrophilic part contains His-94, His-96, His-119, Tyr-7, His-102, Asn-62, His-64, Asn-67, Thr-199, and Thr-200, and the hydrophobic part consists of Val-121, Val-143, Leu-198, Val-207, and Trp-209. To be mentioned, all the amino acid positions in this paragraph are according to the human carbonic anhydrase II crystallographic structure PDB: 3KS3.

The key feature of the catalytic cavity is that, a number of ordered water molecules are positioned connecting themselves through hydrogen bonds and form a water chain or network. A water called the “deep water”, or DW, molecule is placed in the deepest end of the cavity forming hydrogen bond with zinc bound water (ZW) which is further connected to the Oγ1 of Thr-199 (Figure 1) (Liljas 1994) (Fisher 2010). Another water molecule, W1 is oriented to the ZW and Glu-106 by forming two hydrogen bonds separately. It was assumed that another water molecule W2 connects the third coordination site of the W1 molecule, which in turn form a cascade of waters (W2, W3a, W3b). It was also assumed that W3a forms hydrogen bond with Tyr-7 and W3b forms hydrogen bond with Asn-62 and Asn-67. The H atom of W2 is oriented towards the carbonyl oxygen of His-64 to trigger the shuttle of protons by His-64 side chain, which is in a continuous transformation to the inward and outward conformation. Several studies found that the imidazole side chain of the His-64 predominantly oriented in the inward position (Figure 1). (Fisher 2010) (Merz 1990) (Nair 1991) (Fisher 2005)

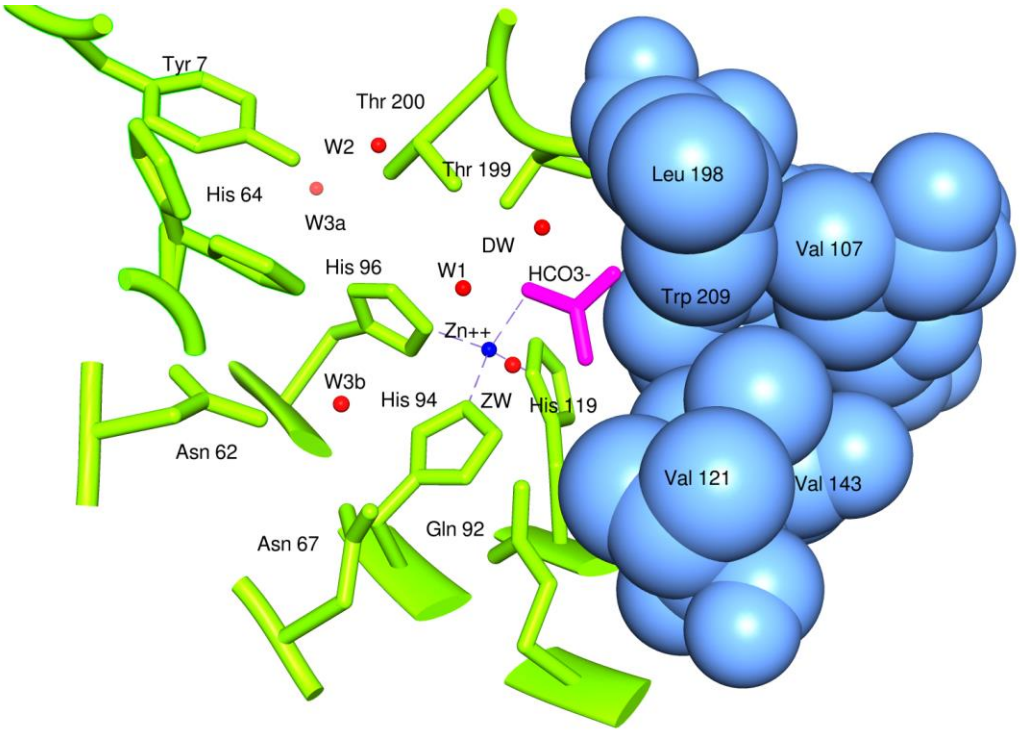


Figure 1 Active site of hCA-11 (PDB: 2VVB), showing the hydrophobic part in blue color spheres, the hydrophilic residues are in green, bicarbonate ion is in pink, red dots are representing the active site waters and the blue dot is the zinc ion. The figure was created in Chimera (Pettersen 2004) to show the typical active site composition of human CAs. The idea of the water chain and water numbering was adapted from Fisher, 2010.

2.2 Catalytic Mechanism

The most important usage of the catalytic reaction performed by the carbonic anhydrase is maintenance of pH balance of the blood, and other tissues, during aerobic metabolism. There have been several reaction mechanisms proposed. The general catalytic mechanism that was proposed by Le Chatelier is described as follows:

The reaction starts at the position of the zinc bound water molecule, where the zinc held as a metal cofactor and polarizes the water molecule. The zinc releases a proton from the bound water to create a hydroxyl ion and the reaction moves towards a de-protonation state while the pKa of the water changes from its usual value of 15.7 to 7. It has been proved by several studies that the released protons are accepted by the His-64 (Tu 1989). The zinc bound hydroxide ($ZnOH^-$) donates the H to the nearby Oγ1 atom of Thr-199 forming a hydrogen bond and simultaneously one of the lone pairs of the zinc bound O^- is ready to accept a CO_2 molecule. The Hydroxyl ion conducts a nucleophilic attack on the positively charged carbon to convert it to the reaction intermediate bicarbonate ion (HCO_3^-). At the same time, the O^- in HCO_3^- forms an intermediate van der Waals interaction with the Zn. At this stage, the HCO_3^- and the proton of His-64 is released and subsequently the enzyme repeats the reaction (Figure 2).

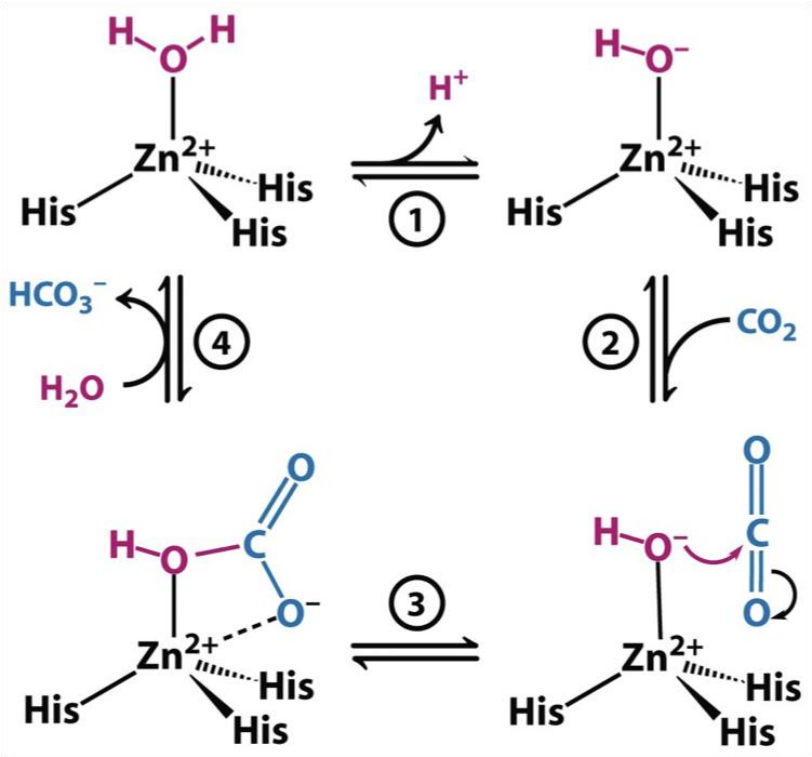


Figure 2 The overall catalytic mechanism of carbonic anhydrases (Berg 2010).

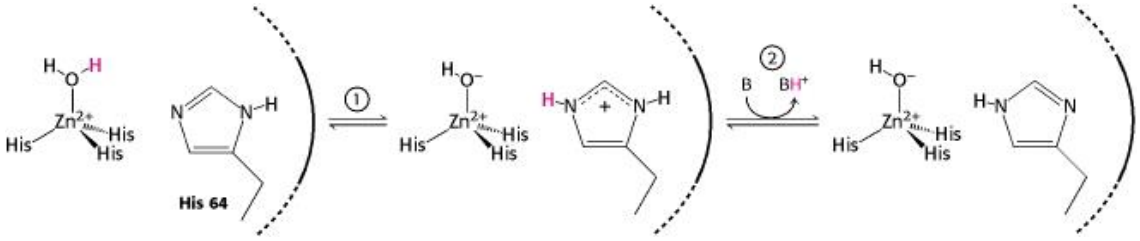


Figure 3 The proton shuttle mechanism by His-64 (Domsic 2008)

The following two different mechanisms (Figure 4) for the rapid interconversion of CO_2 and HCO_3^- have been proposed by Lipscomb (Liang 1987) and Lindskog (Lindskog 1983). According to Lipscomb, the Zn in the Zn-HCO_3^- intermediate is in monodentate form where the proton is influxed by the original Zn-OH^- ion. On the other hand, Lindskog proposed that the reaction intermediate forms a bidentate ion Zn-Zn-HCO_3^- that receives the O^- from the original CO_2 molecule and directly interact with zinc. The former mechanism creates a tetrahedral geometry in contrast to the later one forming trigonal bi-pyramidal geometry at the zinc binding site (Figure 4).

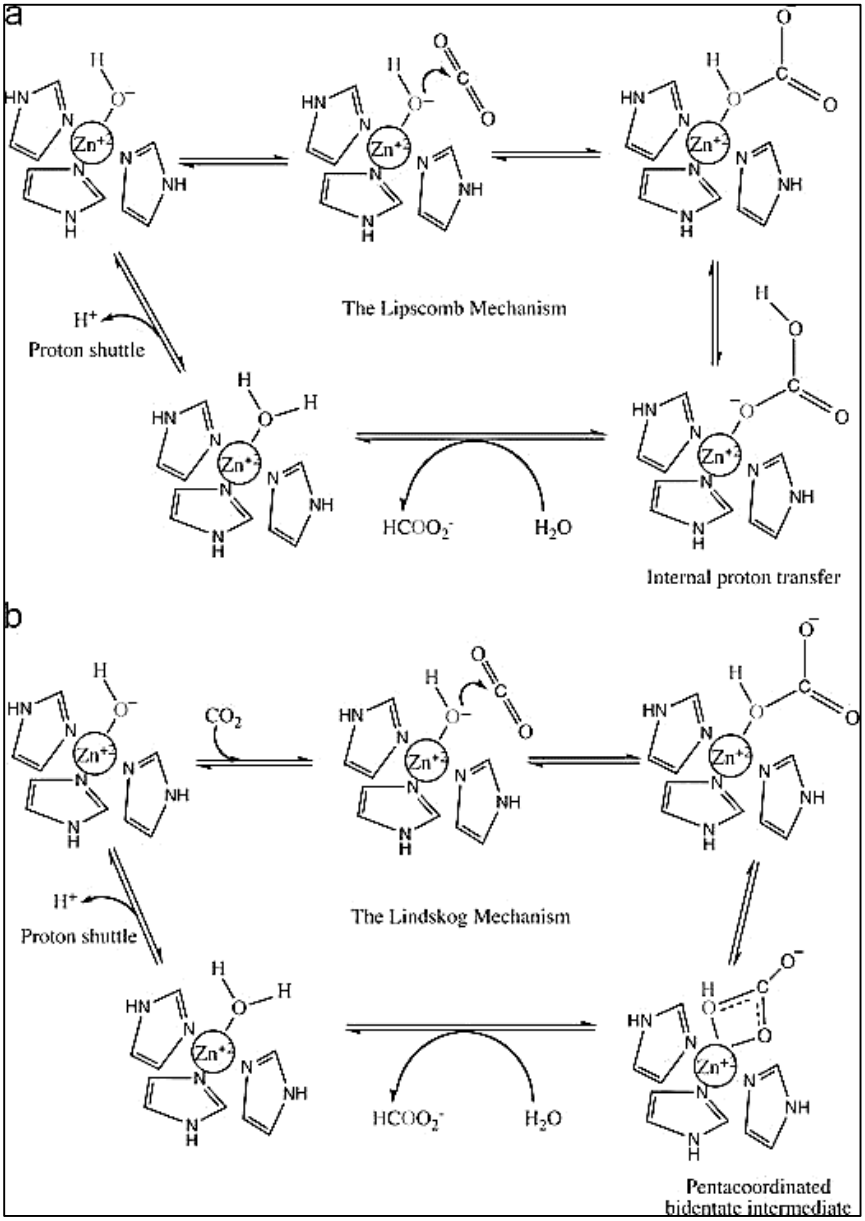


Figure 4 The (de)hydration mechanism of hCA-II proposed by Lipscomb (Liang 1987) (a) and Lindskog (Lindskog 1983) (b).

2.3 Theories of the Most Essential Methods Used for Conservation Analysis

2.3.1 K_a/K_s Ratio

The ratio of the number of non-synonymous substitutions per non-synonymous site (K_a) and the number of synonymous substitutions per synonymous site (K_s) is called the K_a/K_s ratio.

(Miyata 1980) (Ina 1995) (Comeron 1995). This substitution rate is used to calculate the evolutionary pressure on protein coding sequences.

In conservation analysis, the K_a/K_s ratio method is a very effective way for determining the conserved residues within a group of homologous protein sequences. While the typical multiple sequence alignment method for a group of homologous protein only can show the conserved residues, the K_a/K_s ratio analysis of the codon aligned nucleotide sequences is a more sophisticated way to detect which residues are under evolutionary conservation pressure. The output of the K_a/K_s analysis are numeric values assigned for each amino acid, where the higher value, $K_a/K_s \geq 1$ indicates less conserved or positive selection occurred as opposed to the $K_a/K_s < 1$, which means highly conserved (Stern 2007).

2.3.2 BioPython

Python is an open source programming language (python.org) widely used in several application domains. The scripts that were used to analyze the conservation profile were written in Python 2.7 version. Biopython (biopython.org) is an open source python tool specially made for computational biology and bioinformatics analysis. The downloadable version of the software is compatible for LINUX, WINDOWS and MAC operating systems, available for both 32 GB and 64 GB machine. The Biopython tutorial and cookbook is available online and freely accessible by the users.

2.3.3 DSSP

DSSP (**D**efine **S**econdary **S**tructure of **P**roteins) is a dictionary where secondary structure information for each of the protein residues of a given protein structure is kept (Kabsch 1983). The dictionary was created by Wolfgang Kabsch and Cristian Sander in 1983. The algorithm that is used in DSSP for assigning protein secondary structure for each amino acid is based on the atomic coordinate data obtained from each of the X-ray crystallographic structures. The main function of the DSSP algorithm is to analyze the hydrogen-bonding pattern and related geometric features to identify secondary structure information. The DSSP program (http://swift.cmbi.ru.nl/gv/dssp/) takes PDB files as an input and automatically

creates the output DSSP formatted files. DSSP also can determine the solvent exposure values of the protein residues from a given protein structure.

2.3.4 PAL2NAL

PAL2NAL is a program used to compare protein sequence alignment with corresponding coding DNA sequence (CDS) (Suyama 2006). The program takes amino acid and CDS sequence alignment files as input, then matches the corresponding codon and finally produces their respective CDS alignment file. This codon alignment is required for the proper computation of K_a/K_s values for identifying conserved residues. PAL2NAL is available as both web server (http://www.bork.embl.de/pal2nal) and downloadable version.

2.3.5 Selecton

The Selecton (Stern 2007) is a freely available web based tool located at (http://selecton.tau.ac.il/). The tool is also available as a downloadable version. This tool identifies conserved amino acids in the 3D structure of a protein. The program takes codon aligned CDS sequences as an input, performs K_a/K_s analysis, categorizes the result as numeric values from 1-7 (where 1 means the least conserved and 7 stands for most conserved) and marks them according to pre-specified color grid for each numeric value in the 3D structure.

2.3.6 Chimera

Chimera is a molecular visualization software for visualizing and interactive analysis of 3D molecular structures, and their properties such as: electron density, molecular self-assembly, conformational changes, sequence-structure alignment, investigating molecular docking results etc. (Pettersen 2004). The program was developed by Resource for Biocomputing, Visualization, and Informatics at the University of California, San Francisco (UCSF) and is freely available as a downloadable version located at (https://www.cgl.ucsf.edu/chimera/). The software has both command line and manual operation interfaces. Chimera has interfaces for MODELLER (for comparative modeling of the 3D structure), protein BLAST, POV-Ray

and Amber tools. In this thesis chimera version 1.9 for Windows was used for structural analysis.

2.4 Physicochemical Properties Used to Investigate and Define the Role of Universally Conserved Residues

2.4.1 Hydrogen Bonds

A hydrogen bond is formed between two polar molecules (one donor and one acceptor) when an electromagnetic attraction occurs between them. The hydrogen bond can be both intramolecular and intermolecular. The hydrogen bonds in protein are mainly intramolecular which stabilizes the secondary and tertiary structure of the proteins. The amino acids in the proteins are interconnected by the hydrogen bonds and form specific secondary or tertiary shapes of the proteins. There are mainly three types of hydrogen bonds which occur in proteins. The first type is a hydrogen bond between the side chains of two separate amino acids; second type is formed between the backbone of beta sheets and third type is formed between the turns of alpha helices. A hydrogen bond is formed by one of the lone pair of electrons in oxygen attaching to an electronegative atom (such as a nitrogen atom) (Figure 5). The oxygen in -OH (as in Ser, Thr, and Tyr) or HOH, and the nitrogen in $-NH_3^+$ (as in Lys and Arg) or -NH- (as in the main chain peptide bond, Trp, His, Arg, and nucleotide bases), are typical donors. In 1997, Jeffrey categorized distances of H bond 2.2-2.5 Å as "strong, mostly covalent", 2.5-3.2 Å as "moderate, mostly electrostatic", 3.2-4.0 Å as "weak, electrostatic" (Jeffrey 1997). The hydrogen bonds that are found in proteins are mostly in the moderate category. Proper hydrogen bonding patterns are exceptionally important for stabilizing the protein folding followed by protein structures.

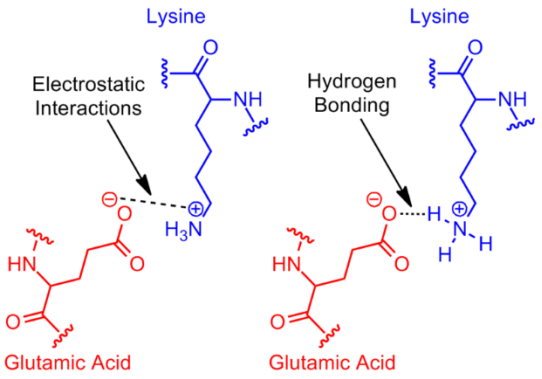


Figure 5 The electrostatic interaction formation between amino acid residues. Image borrowed from http://www.chemguide.co.uk/organicprops/aminoacids/proteinstruct.html.

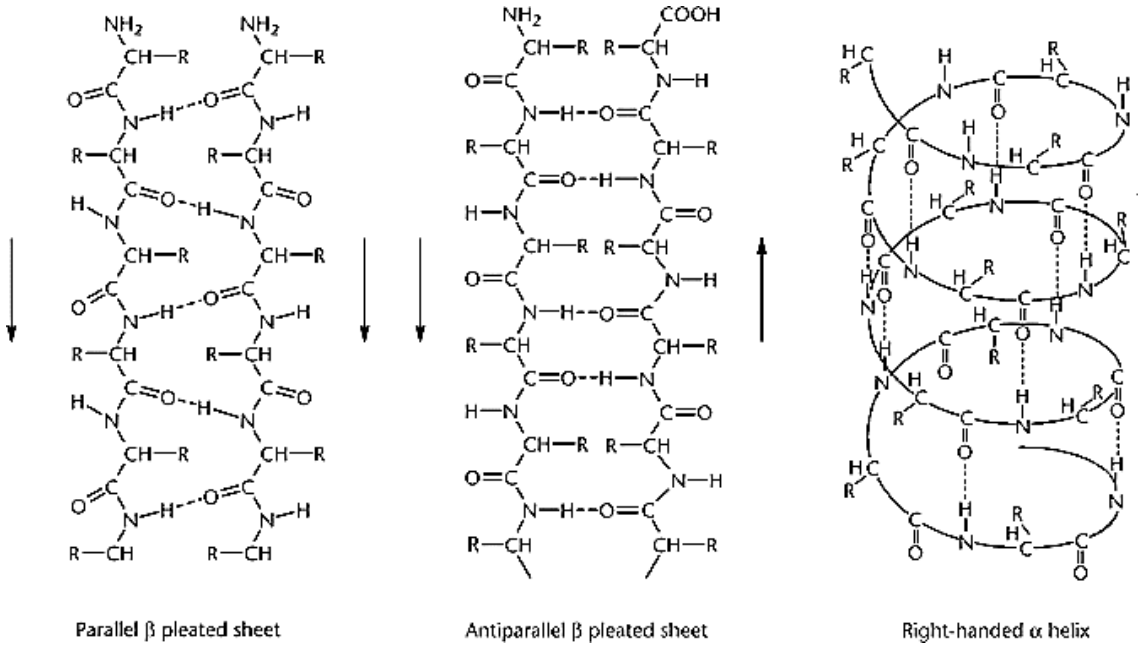


Figure 6 Hydrogen bonds in parallel and antiparallel beta sheets in left figure. Hydrogen bonds in right handed alpha helices in right figure. Image borrowed from http://humanbiology2011.wordpress.com/proteins/.

2.4.2 Polarity and Hydrophilicity

Each of the 20 amino acids fall in either the hydrophilic category, having hydrophilic side chains or the hydrophobic category, having hydrophobic side chains. Polar amino acids include Glutamine, Asparagine, Histidine, Serine, Threonine, Tyrosine, Cysteine, Methionine and Tryptophan; hydrophobic amino acids are Alanine, Isoleucine, Leucine, Phenylalanine, Valine, Proline and Glycine (Figure 7). Glycines are yet to be considered as hydrophilic as they have functional groups (-NH_2 and -COOH) that can form hydrogen bond

with solvents. Polar amino acids are more prone to form hydrogen bonds in tertiary structures. Some amino acids, that are called amphipathic, show both hydrophilic and hydrophobic properties due to the presence of both polar and non-polar groups in the side chains. Threonine, Lysine, Tyrosine, Methionine and Tryptophan are amino acids that fall into this category (Creighton 1992). If the amphipathic amino acids are located in the protein surface, they predominantly interact with other protein molecules (Creighton 1992).

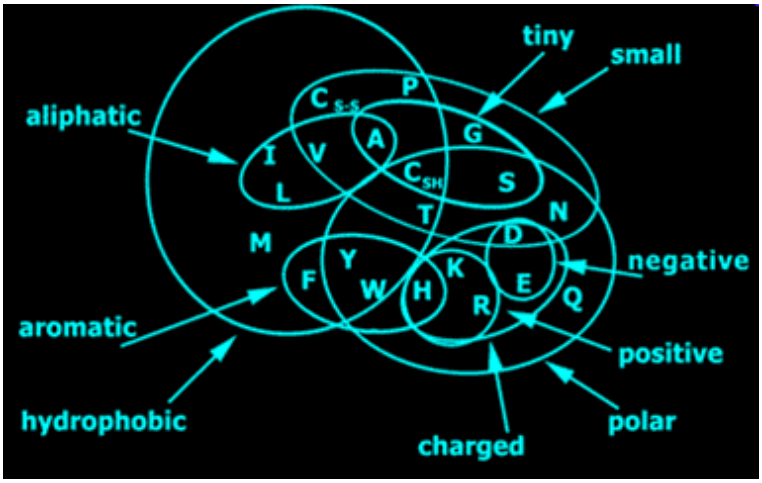


Figure 7 Classification of the amino acids based on their chemical properties. Modified image from Livingstone & Barton, 1993 (Livingstone 1993). Image courtesy of Jukka Lehtiniemi.

2.4.3 Aromaticity

Aromaticity is a chemical property of a compound with a conjugated ring of unsaturated bonds. This property arises due to the delocalization of the electrons in such conjugated systems. (Holfmann 1855). Among 20 amino acids, only Phe, Pro, His, Tyr, Trp have the aromatic side chains. Like other type of interactions, aromatic side chains also show some specific interactions between them (Burley 1986). In the tertiary structure of proteins, such interaction is seen as non-covalent π-π stacking interaction where two closely positioned aromatic systems form a weak electrostatic interaction between them to stabilize the structure. Such π-π stacking interactions are named: sandwich, parallel displaced, and T-shaped (or edge-to-face configuration) (McGaughey 1998) (Figure 8). These interactions are important in protein folding as well as stabilizing the protein structure. The most important

property of such interaction is that they can form a stable interaction at a distance larger than the average van der Waals radius (McGaughey 1998).

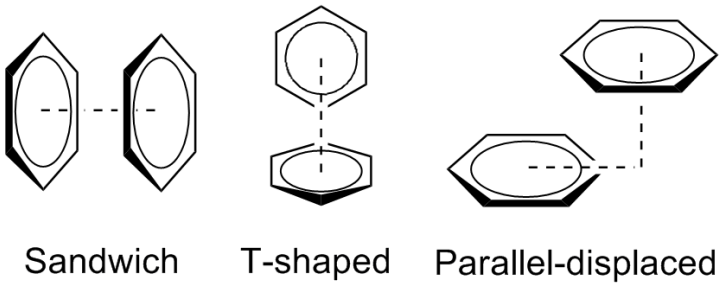


Figure 8 Different types π-π stacking interactions between benzene rings. Image source http://en.wikipedia.org/wiki/Stacking_%28chemistry%29

2.4.4 Relative Solvent Accessibility

The RSA, or relative solvent accessibility value determines the solvent exposure level of an amino acid residue in a given protein structure. Amino acids that are located on a protein surface are more prone to react with solvents whereas the buried residues participate in stabilizing the structure by forming hydrogen bonds or other non-covalent interactions. A cutoff value of the RSA scores is determined to distinguish between buried and surfaced residues. The solvent accessibility (SA) values can be determined by the DSSP server (Kabsch 1983). The equation that is used to determine the SA values was described by Kabsch and Sander, is as follows:

$$\mathbf{W = Area/Volume(water\ molecule)^{2/3}}$$

Where, W = number of water molecules interacting with the surface of the residue
Area = Total surface area of the amino acid
Volume = The total volume of the amino acid residue

If the solvent exposure value is divided by the total surface area of the single amino acid residue, it returns the RSA value of the specific residue (Miller 1987). A RSA value around 0.25 can be considered as the boundary line for the exposed versus buried residues (Adamczak 2005) where any value lower than 0.25 is buried and greater than 0.25 will be exposed.

2.4.5 Hydrophobic Interaction

Amino acids having hydrophobic side chains can interact with each other. The interaction takes place when two or more hydrophobic molecules are present in a water medium. An American chemist Walter Kauzmann described this interaction, as the hydrophobic molecules form a clump in the water medium aggregating themselves in a cluster because in such a way they can be in a minimal contact with the solvent molecules. These properties are often seen in protein tertiary structures, where amino acid residues with hydrophobic side chains interact and are buried in the protein core, away from the solvent exposure.

2.5 N-glycosylation Site

Glycosylation is a process in which a carbohydrate molecule or glycan (glycosyl donor) is attached to a protein, lipid or other organic molecule (glycosyl acceptor) to form a glycosydic bond (one kind of covalent bond). In proteins, glycosylation occurs during the co-translational and post-translational stage in protein biosynthesis. These kinds of modifications are essential for protein folding, which give stability, and participate in different types of cellular functions (Freeze 2009). There are different types of glycosylation, such as N-linked glycosylation, O-linked glycosylation, phospho-serine glycosylation, C-mannosylation, and glypiation (GPI anchors). Among them N-linked glycosylation is the most common type of modification that occurred in the proteins and the sites are easy to trace from the proteins primary structures. In N-linked glycosylation, glycans are attached to the N atom of Asparagine (Asn) side chains. The pattern of the N-linked glycosylation sites includes consensus amino acid residues, Asn-Xaa-Ser or Thr, where Xaa can be any other protein except proline, as the side-chain of proline can hinder/impair the N-glycosylation process (Schwarz 2011) (Gavel 1990). The Asn-X-Cys motif is also found to be glycosylated, however it is very rare. The N-linked glycosylation sites can be determined by the "NetNGlyc" (Gupta 2004), a freely accessible tool for predicting N-glycosylation sites, located at (http://www.cbs.dtu.dk/services/NetNGlyc/).

3 Aim of the Study

The main aim of this study is to construct a perfect list of important conserved residues in alpha carbonic anhydrases, with a focus on vertebrates, and define their roles in 3D protein structure, as well as for catalysis. Additionally, the conservation study was done for the two largest α-CA subfamilies, cytoplasmic and extracellular. At the end, the conservation profile study was made on the analysis of the N-glycosylated sites in the extracellular domains of the CA-VI, IX, XII, and XIV. The whole process was divided in the following steps:

1. Select the appropriate number of species and isozymes for each group (Universal, Cytoplasmic and Extracellular) and retrieve best quality sequences from Ensembl.
2. Use an automated method for K_a/K_s scoring to rank the conserved residues.
3. Produce multiple sequence alignment using Clustal Omega software.
4. Manual alignment of the K_a/K_s derived top ranked conserved residues with MSA derived 100%-conserved residues.
5. Make a complete list of important conserved residues from the manual alignment.
6. Define the role of the conserved residues by investigating 3D structure based on physicochemical properties and literature survey.
7. Following the same K_a/K_s based approach, make a list of unique residues for the bigger subgroups (cytoplasmic and extracellular), comparing them with universally conserved group and visualizing in the 3D structures.
8. Finally, a separate conservation study for mapping the conserved N-linked glycosylated sites in the 3D protein structures for the extracellular isozymes (CA-VI, IX, XII, and XIV) and structural visualization.

To date, several studies have sought to identify the functional and structural importance of conserved residues in α-CAs. However, no through study has been completed for each of the highly conserved amino acid residues, for instance, the recent study done by Aggarwal (Aggarwal 2013). This analysis created a list of the most important conserved α-CA residues, reviewed previous literature on their functions, and made intelligent guesses for each of the highly conserved residues. The procedure that was used to identify conserved residues can be applied for any group of the homologous species.

4 Material and Methods

4.1 Conservation Analysis for Universal, Cytoplasmic and Extracellular Group

4.1.1 Species and Isoform Selection

Species selection is the most crucial part in case of conservation analysis for a specific gene family. Here, non-ray-finned-fish jawed vertebrates (tetrapods plus the lobe-finned fish *Latimeria*) were considered for the analysis of the conservation profile since ray-finned fishes have a different set of cytoplasmic CAs (Esbaugh 2006). For the "universal group", human, mouse and chicken (or turkey if chicken was unavailable) were chosen for the analysis. The selection was made due to the diversity and high coverage genome sequences (at least 6X) available in genomic databases. Although, a more diverged choice would have been human, frog, and chicken but the choice of mouse was justified by the most certain sequences available for all of the isozymes. For the cytoplasmic and extracellular group, frog (or lizard if frog was unavailable) and giant panda were added to the previous selection for the universal group. The number of species was increased for keeping the consistency with the decreased number of isozymes in each sub group. Further, the choice of panda was based on availability of good-quality genome and the evolutionary distance of panda from human and mouse. All active α-CA isozymes except CA-XVII, which is a novel isozyme and restricted to non-mammalian species, were chosen for the analysis. The cytoplasmic isozymes CA-I, II, III, VII, and XIII were included in the cytoplasmic group and extracellular isozymes CA-VI, IX, XII, and XIV were included the extracellular group. CA-V was excluded from the cytoplasmic group as it is located in the mitochondria and serves different purposes than the other cytoplasmic isozymes and CA-VI was included in the extracellular group as being secreted its ultimate location is extracellular.

4.1.2 Sequence Retrieval, MSA and K_a/K_s Scoring

The DNA and protein sequences were retrieved with automated methods *"Orthologer"* and *"SEQs2Categories"* (Barker 2013), which retrieve the maximal number of orthologous sequences (protein and CDS) from the Ensembl database (Flicek 2013), do quality checks (if

there is any bad sequence lacking Methionine at the first position or any missing residues), and produces separate FASTA files for each of the isozymes. The targeted protein and CDS sequences (not to be confused with cDNA) were selected and put in two separate document files manually for the later approaches. Then, the script "*Unaligned2KaKs"* (Barker 2013) was used to produce protein alignment and quantify conservation at each residue based on the K_a/K_s approach (of automatically generated codon-based alignments). For the execution of the codon-based alignment, *Clustal Omega* (McWilliam 2013) and *PAL2NAL* (Suyama 2006) were called in the *"Unaligned2KaKs*" program. Clustal omega was used to create the protein alignment file whereas *PAL2NAL* created the codon alignment file using the protein alignment and respective unaligned CDS sequences. At this stage, the program generated two separate alignment files (one for protein and another one for CDS sequences) and the K_a/K_s output file, containing the K_a/K_s values for each of the amino acid in the human CA-II sequence (see Appendix 1). The same program was run for all three groups, universal, cytoplasmic and extracellular. Finally, Selecton was called by the program, which analyzed the codon aligned file, created K_a/K_s score for each of the amino acids to categorized them based on predefined parameters by Selecton and generated an output file containing most to the least conserved residues.

During the generation of K_a/K_s values, human CA-II was selected as the template sequence for the "universal group" and the "cytoplasmic group" due to available good quality structure in the protein data bank. For the "extracellular group" human CA-XII was used as template sequence for the same reason.

4.1.3 Manual Alignment and Universal Conserved Group

To obtain the most important conserved residues, the K_a/K_s score table was aligned with the MSA of protein sequences. At first, the amino acids were arranged in ascending order so that the lower K_a/K_s scores are shown at the top of the table (lower K_a/K_s score means higher conservation). Then the protein alignment file was investigated for 100% conserved residues which were noted with 'XX' (beside the K_a/K_s score column) and named as "perfectly conserved" (see Appendix 1) (*Table 2*). It was observed that Selecton detected most conserved residues were found to be 100% conserved in the universal group. However, there

were some highly conserved residues that are not 100% conserved. This was often due to a mismatch at the single isozyme or single species level, and these residues had K_a/K_s scores quite close to those of the perfectly conserved residues. For example, Val-206 has variant in human, mouse and chicken but only in the CA-III isozyme so it was marked as a single exception "X". Ala-133 has a single variant at CA-IV mouse so it was also marked as "X". On the other hand, Leu-163 has a variant at Chicken CA-III and CA-XII so it was not considered as highly conserved due to the occurrence of the exception at two different isozyme positions. Thus, there were nine residues found for "highly conserved" type and marked as "X". Further, three other residues Thr-199, Val-142 and His-64 were also included in the highly conserved group due to their strong conservation score that are quite close to the highly conserved residues and most notably, all are important for the catalysis of the CO_2 (de)hydration reaction and location in the active site of the enzyme (see Appendix 1). So, the list of the "highly conserved" residues was finalized with eleven residues (*Table 3*). The roles of the conserved residues, both "perfectly conserved" and "highly conserved" will be discussed in the result section.

N.B. All the amino acid positions that were used in this paragraph are based on the protein primary structure information.

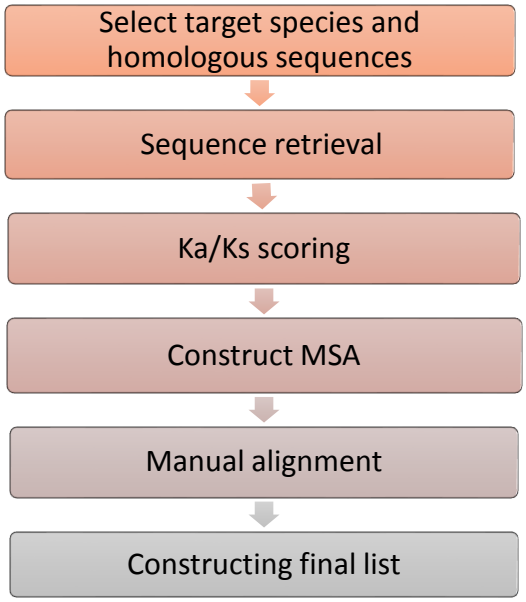


Figure 9 The work flow of the conservation analysis

4.1.4 Pool Rank

In this analysis the pool scores were used to cross check the functional importance of the conserved residues that were listed in the "universally conserved" group. POOL stands for **P**artial **O**rder **O**ptimal **L**ikelihood. This is a machine learning method used to predict proteins' functional elements (Somarowthu 2011). The program was developed by "Ondrechen research group" at Northeastern University, Boston, MA, USA. POOL estimates the probability that a residue is functional according to the results achieved by the following three programs: THEMATICS, a computational program for identifying the active sites of the enzymes based on the electrostatic data) (Wei 2007) (Ko 2005) (Ondrechen 2001); INTREPID, a program for identifying functional residues based on conservation and phylogeny analysis (Sankararaman 2008); and ConCavity, a computational approach for identifying binding cavity. Together, these programs rank the functionally important residues in the active site of a protein 3D structure. The program takes a PDB id as input and returns the rank of the residues present in the whole protein structure with their corresponding pool values. The residues that obtain top positions in the pool rank are considered to be functionally important in the structure.

4.1.5 Cytoplasmic and Extracellular Conserved Residues

The same procedure that was used to identify the absolutely conserved residues in universal group, was applied to identify the absolutely conserved residues in the cytoplasmic and extracellular group. The idea underlying this analysis is to create a list of conserved residues that are unique in these subgroups. For example, the conserved residues that are not present in the universal group but in the cytoplasmic group are likely important residues for that group and assumed to have a specific role. Therefore, the cytoplasmic and extracellular conserved residue lists were constructed according to the residues, those are not present in the universal group but in the cytoplasmic group (CA-I, II, III, VII, and XIII) or extracellular group (CA-VI, IX, XII, and XIV) respectively. The resulting two distinct conserved residue groups were further visualized in the protein structure of human CA-II (PDB: 3KS3) for cytoplasmic group and human CA-XII (PDB: 4HT2) for extracellular group, and analyzed

for understanding the significance of those residues for being conserved in their subcellular part.

4.2 Conserved N-glycosylation Site Prediction

4.2.1 MSA and N-glycosylation Site Identification

In this section, the conservation analysis was applied to predict functional N-glycosylation sites in the extracellular domain of CA-VI, IX, XII, and XIV as only secretory and transmembrane proteins have post translational modification. Though CA-VI was previously analyzed by Patrikainen (Patrikainen 2012) but here CA-VI was analyzed again only with good quality protein sequences and a comparison was done for the conserved N-glycosylation site among all the extracellular isozymes. As this analysis did not concern about a straightforward ranking of the residues according to conservation, likewise in the universal or the subgroups study, the method was kept simple. All the good quality sequences of the extracellular group were selected from the previously downloaded sequences from Ensembl (Flicek 2013). The multiple sequence alignments were done using Clustal Omega (clustalomega.org) and the alignment file was analyzed in the GeneDoc software (http://www.psc.edu/biomed/genedoc) (Nicholas 1997). The N-Glycosylation sites were predicted from the NetNGlyc 1.0 server (Gupta 2004). The conserved N-glycosylation pattern (Asn-Xaa-Ser/Thr) (where Xaa is not Pro) and Asn-X-Cys (Taylor 2006), were identified and colored in the alignment file (see Appendix 4). Four different colors were used for four different isozyme groups.

4.2.2 Categorization of the N-glycosylated Sites

N-glycosylation sites for all the available good quality sequences of the extracellular α-CA isozymes (CA-VI, IX, XII and XIV) were detected and manually colored in the MSA. The conserved sites were identified and categorized according to the frequency of the sites at each of the conserved positions (*Table 1*). The total number of sites that were detected in CA6, CA9, CA12, and CA14 are 20, 22, 18, and 26 respectively. The 50% cutoff was chosen for the highly conserved glycosylation sites. The conserved sites having frequency 50% or more

were considered to be "conserved glycosylated" sites, in which the glycan part might be functional. The second cutoff was chosen at 25%, so the sites that were 25% or more frequent were considered as "frequently glycosylated" sites. Consequently, the sites having less than 25% frequency were called as "occasionally glycosylated" sites.

Table 1 Percent frequency of the glycosylated sites in each position of the individual isoforms.

Positions	CA6/20	%freq	CA9/22	%freq	CA12/18	%freq	CA14/26	%freq
pos1	10	50	1	4.55	16	88.89	1	3.85
pos2	1	5	22	100	2	11.11	1	3.85
pos3	3	15	17	77.27	1	5.56	3	11.54
pos4	18	90	2	9.09	2	11.11	2	7.69
pos5	1	5	0	0	17	94.44	1	3.85
pos6					3	16.67	25	96.15
pos7					1	5.56		
pos8					1	5.56		
pos9					2	11.11		
pos10					3	16.67		
pos11					10	55.56		
pos12					1	5.56		

4.2.3 Modelling of the Missing Part of CA12 Structure

The next step was to visualize the N-glycosylation sites on the 3D protein structure. The structural investigation found that the first N-glycosylation site was missing in the available crystallographic structure for CA-XII. Therefore, the N terminal part of CA-XII protein was modelled using the "MODELLER" (Eswar 2006) program, which is also available through chimera (http://www.cgl.ucsf.edu/chimera/). The missing region is a short segment of three residues, "NGS" from residue number 1 to 3. The missing segment was modelled using hCA12 and hCA13 as template. The best model was chosen according to the high structural similarity with both of the templates and a rational guess wad made so as the part is quite available to be glycosylated. The modelled region was then spliced and added to the original CA-XII structure to make the structure prepared for the later analysis.

5 Results

5.1 List of "Universally Conserved" Residues

The conservation analysis for the universal group included ten alpha CA isozymes (CA-I, CA-II, CA-III, CA-IV, CA-VA, CA-VB, CA-VI, CAVII, CA-IX, CA-XII, CA-XIII, CA-XIV, and CA-XV) for three species human, mouse and chicken. The K_a/K_s analysis was performed with the human CA-II sequence (Ensembl transcript id: ENST00000285379) as the target. Along with the K_a/K_s values, an output table includes Ensembl ids and positions, PDB ids and positions, their chemical properties, RSA values and locations based on solvent exposure, secondary structure information and pKa values (Appendix 4).

***Table 2. List of "perfectly conserved" residues in vertebrate CAs for universal group. Ens_Pos = Ensembl position for human CA-II, Ens_res = Ensembl residue, PDB_pos = PDB position, PDB_res = PDB residue, RSA = relative solvent accesibility, LOC = Location, Sec_struc = Secondary structure, Chem_prop = Chemical properties. In secondary structure information, G =Helix-3, E = Strand, S = Bend, T = Turn, H = Alpha helix, B = Beta bridge. In chemical properties information, NP = Non-polar, P = Polar, A = Amphipathic, L = Hydrophilic, B = Hydrophobic. pKa values (see Appendix 4) were derived from DEPTH server* (Tan 2013).**

Serial	Ens _Pos	PDB_res	PDB_pos	K_a/K_s	RSA	LOC	pKa	Structure	Chem _prop
1	16	W	16	0.021	0.02	Buried	-	G	NP/A
2	28	Q	28	0.0061	0.02	Buried	-	-	P/L
3	29	S	29	0.0059	0	Buried	-	S	P/L
4	30	P	30	0.0092	0	Buried	-	-	NP/B
5	44	L	44	0.011	0.18	Buried	-	-	NP/B
6	61	N	61	0.0055	0	Buried	-	E	P/L
7	94	H	94	0.0061	0.12	Buried	5.1	E	P/L
8	96	H	96	0.0059	0.01	Buried	2.91	E	P/L
9	97	W	97	0.021	0	Buried	-	E	NP/A
10	104	G	104	0.0091	0	Buried	-	-	NP/L
11	106	E	106	0.0077	0.01	Buried	7.53	S	P/L
12	107	H	107	0.006	0	Buried	1.67	S	P/L
13	117	E	117	0.0076	0	Buried	8.37	E	P/L
14	119	H	119	0.0062	0.02	Buried	2.6	E	P/L
15	122	H	122	0.0061	0	Buried	1.67	E	P/L
16	141	A	142	0.0066	0	Buried	-	E	NP/B
16	185	P	186	0.0099	0.09	Buried	-	-	NP/L
17	193	Y	194	0.0097	0.03	Buried	-	E	P/A
18	195	G	196	0.0093	0	Buried	-	E	NP/L
19	196	S	197	0.006	0	Buried	-	-	P/L
20	198	T	199	0.0055	0.04	Buried	-	-	P/A
21	200	P	201	0.01	0.08	Buried	-	S	NP/L
22	208	W	209	0.021	0.02	Buried	-	T	NP/A
23	221	Q	222	0.006	0.02	Buried	-	E	P/L
24	245	R	246	0.0086	0.01	Buried	-	H	P/L
25	248	Q	249	0.0061	0.21	Surface	-	-	P/L
26	253	R	254	0.0083	0.11	Buried	-	-	P/L

Table 3. List of "highly conserved" residues. Ens_Pos = Ensembl position for human CA-II, Ens_res = Ensembl residue, PDB_pos = PDB position, PDB_res = PDB residue, RSA = relative solvent accesibility, LOC = Location, Sec_struc = Secondary structure, Chem_prop = Chemical properties. In secondary structure information, G =Helix-3, E = Strand, S = Bend, T = Turn, H = Alpha helix, B = Beta bridge. In chemical properties information, NP = Non-polar, P = Polar, A = Amphipathic, L = Hydrophilic, B = Hydrophobic. pKa values (see Appendix 4) were derived from DEPTH server (Tan 2013)***.***

Serial	Ens_pos	PDB_res	PDB_pos	Ks/Ks	RSA	LOC	pKa	Structure	Chem_prop
1	63	G	63	0.031	0.18	Buried	-	S	NP/L
2	64	H	64	0.046	0.28	Surface	5.76	S	P/L
3	92	Q	92	0.029	0.15	Buried	-	-	P/L
4	105	S	105	0.019	0	Buried	-	-	P/L
5	121	V	121	0.018	0.06	Buried	-	E	NP/B
6	133	A	134	0.017	0	Buried	-	H	NP/B
7	142	V	143	0.045	0.04	Buried	-	-	NP/B
8	197	L	198	0.03	0.18	Buried	-	-	NP/B
9	199	T	200	0.029	0.21	Surface	-	S	P/A
10	206	V	207	0.016	0	Buried	-	E	NP/B
11	243	N	244	0.027	0	Buried	-	-	P/L

5.2 Roles of the "Universally Conserved" Residues

Different factors were considered while analyzing the role of the universally conserved residues. Mostly, the physicochemical properties were considered for the analysis such as, hydrogen bond, hydrophobic interaction and pKa values. For each of the hydrogen bonds the distances were measured and checked whether the values are in their usual range or not. Literature surveys along with structural investigations were conducted to figure out most possible structural and functional role of those conserved residues that are discussed in the following sections. The protein structure, PDB accession id: 3KS3 (Avvaru 2010) of human CA2 was used due to the high-resolution (0.9Å) crystallographic structure.

5.3 Roles of Conserved Residues in the Active Site

This section describes the structural and functional roles of the active site conserved amino acid residues of human carbonic anhydrase II, PDB id: 3KS3 listed in *Table 2* and *Table 3*, which are important for catalysis. Residues, which are positioned at the active site and directly or indirectly involved in substrate binding or proton donation, and therefore assist the catalytic mechanism, are also discussed. The conserved residues of our concern are colored in yellow in the figures.

5.3.1 His-94, His-96, and His-119

The invariant Histidine triad (His-94, His-96, and His-119) forms a hydrophilic cluster that is essential for coordinating the zinc metal ion Zn^{2+} at the active site of this metalloenzyme. Being polar and having an imidazole ring, their major contribution is in binding a metal ion at the catalytic core. They can react with solvents and polar substrates, due to their high reactive nature, and form a distorted tetrahedral geometry (Liljas 1994), which is an essential coordination geometry for the CO_2 (de)hydration reaction mechanism. Besides this specific structural property, His-94, His-96, and His-119 show unusual pKa values 5.1, 2.91, and 2.6 respectively, that are much lower than their usual pKa value of 6.5 (see Appendix 4) which indicates their functional importance for catalysis. These three hydrophilic residues also take part in formation of the hydrophilic half of the catalytic core (Figure 10).

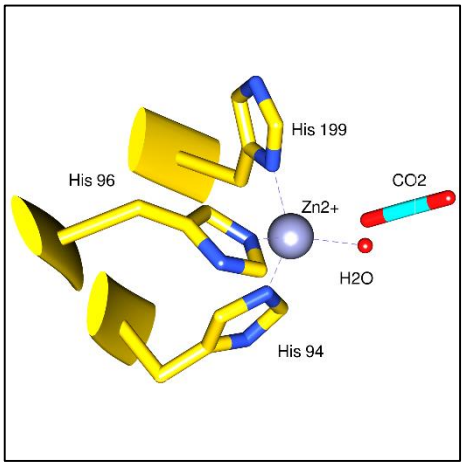


Figure 10 His-94, His-96 and His-119 coordinating with zinc ion. Zinc is bound with water (red dot) and a CO2 molecule is interacting at the active site.

5.3.2 Thr-199, Thr-200, and Glu-106

Thr-199 plays a very important role in the catalysis of CO_2. The catalytic cavity of CAs reaches to its deepest position at Thr-199, which binds a water molecule called the "deep water" or DW. Thr-199 forms a hydrogen-bonded network with Glu-106, DW that is further hydrogen bonded to the zinc bound hydroxide, forming an optimal coordination geometry which facilitates the solvents for the optimal nucleophilic attack on CO_2 (Xue 1993) (Merz 1990) (Figure 11). Due to this special phenomenon, Thr-199 is called the "door-keeper" residue (Liljas 1994). A previous site specific mutation study also revealed that such a hydrogen bonding pattern stabilizes the (de)hydration reaction transition state (E-HCO_3) and the zinc-hydroxide (Zn-OH^-) (Krebs 1993).

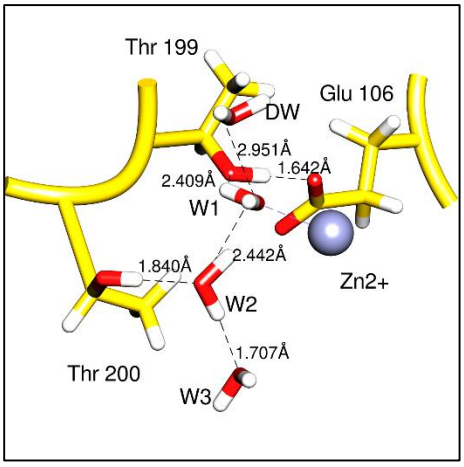


Figure 11 Hydrogen bonded network along with active Thr-199, Thr-200 and Glu-106. hCA-II structure, PDB:3TMJ was used for constructing the figure.

Like Thr-199, Thr-200 is also a catalytically active residue and takes part in the CO_2 hydration reaction (Krebs 1991). Being polar, Threonines have high affinity to water molecules and it was found that they stabilize the W1 in the hydrogen bonded water network at the active site (Fisher 2011) (Figure 11). A Thr-200-Ser site-specific mutation study has also been done to understand the hydration activity, and the result was that Ser-200 stabilizes the E-HCO_3 complex two fold greater than the wild type one. So in a reverse idea it is proved that Thr-200 stabilizes the reaction transition state even though to lesser extent (Krebs 1991).

5.3.3 Gln-92

Gln-92 is hydrogen bonded to His-94, and their position in the hydrophilic half of the catalytic core clearly indicating that they have a distinct catalytic role for the CO_2 hydration (Figure 12). A molecular dynamic study revealed that Gln-92 acts as a CO_2 binding site (Liang 1990). In an another study, Turkoglu et al. performed a mutation for Gln-92 to Ala-92 and found that the hydration activity was 30% lower in the variant than that of the wild type (Turkoglu 2012), which clearly demonstrates its role for catalyzing the (de)hydration reaction.

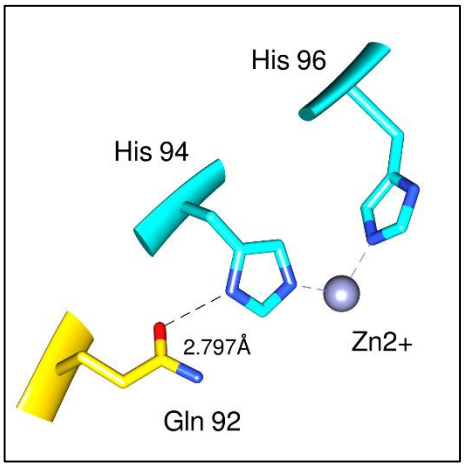


Figure 12 Gln-92 forming hydrogen bond with His-92

5.3.4 Val-121, Leu-198, Val-207, and Val-143

Valines and Leucines are hydrophobic in nature due to absence of polar side chain. Here, Val-121, Leu-198, Val-207, and Val-143 are forming the mouth of the hydrophobic pocket at the active site (Figure 13). The hydrophobic environment thus possibly facilitates the water molecules to be repulsed by the hydrophobic site to the hydrophobic site for (de)hydration. However Nail et al., 1991, stated that all of these residues participate in CO_2 hydratase activity (Nair 1991). Further investigation found Val-143 to be highly efficient at the position as the mutations (Val-143-Ile, Val-143-Leu) caused 20-fold lower efficiency the catalysis (West 2012).

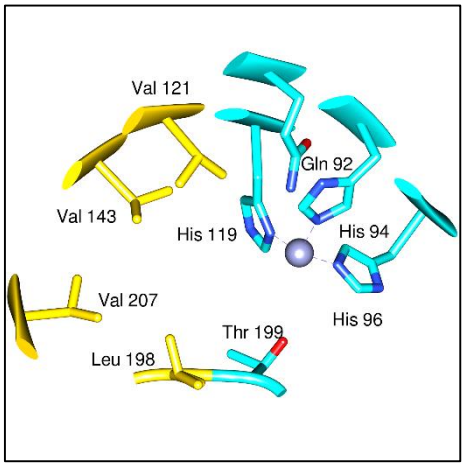


Figure 13 Hydrophobic residues Val 121, Val-143, Val-207 and Leu-198

5.3.5 His-64

His-64 is believed to be involved in proton transfer during catalysis as proved by several investigations (Tu 1989). The site-specific mutation (His-64-Ala) study proved that the catalytic efficiency decreased 20-fold in the modified enzyme, from than that of wild type (Tu 1989). It was also found that the imidazole side chain of the residue tends to be in both inward and outward directions (Figure 14), though several studies found it to be in the inward conformation as the delta nitrogen reaches the closest to the zinc in this position (Maupin 2007). On the other hand, the outward conformation was found to decrease proton shuttle activity (Zheng 2008) (Maupin 2009). The structural investigation shows the reason His-64 tends to be in outward conformation, despite being less favourable for the catalytic role, may be due to the tendency of forming π-π stacking interaction with Trp-5. Further, the pKa value also proved that despite of being basic they have obtained lower pKa value of 5.76 (see Appendix 4) from their usual value of 6.5 and stood 9th position in the pool rank (see Appendix 3). This denotes they are in protonated form in aqueous solution. To be mentioned, His-64 is neither absolutely nor highly conserved as isozymes CA-III and CA-V lack this residue for all the species (human, mouse and chicken). This was still included in the conserved residue list due its high specificity for proton influx mechanism.

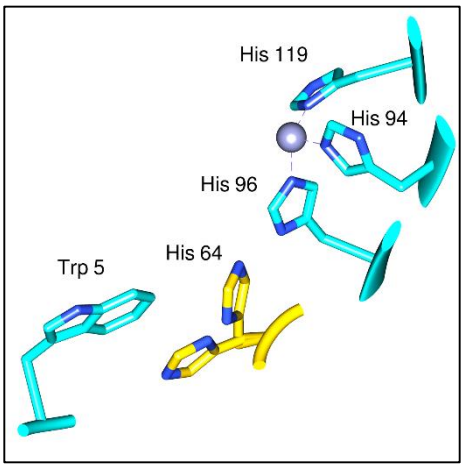


Figure 14 His-64 in inward and outward conforamtion along with Trp-5, and Histidine triad at the active site

5.4 Roles of structurally important conserved residues

5.4.1 Trp-16

Trp-16 is located in the terminal end of the protein structure and the side chain is positioned inwards to the center. Further, as being closely located to the Trp-5, Trp-16 is highly likely to form a T-shaped π-π stacking interaction (Figure 15) therefore, stabilizes the corresponding alpha helix. Trp-5 was left under the conservation thresholds in the study, but is conserved in most CA isoforms except for CA-VA and CA-VB.

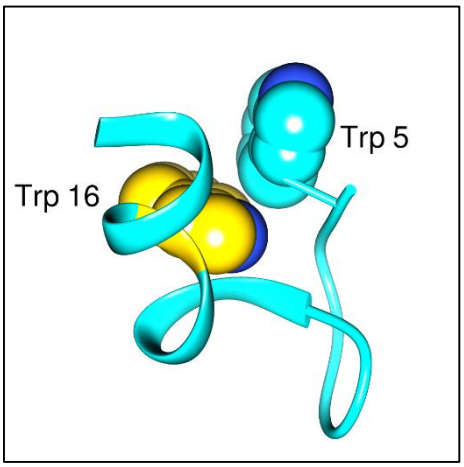


Figure 15 Trp-16 and Trp-5 π-π stacking interaction

5.4.2 Gln-222

Glutamines are polar in nature and usually tend to be in the surface of the protein but here, Gln-222 is buried and forms four hydrogen bonds with four different neighboring residues (Figure 16). Most notably, the hydrogen bonds are the only bonds that interconnect two parallel alpha helices and hence, stabilize the structure. Moreover, the fact that there are no other conserved residues present at closer distances highlights the importance of this specific structural role.

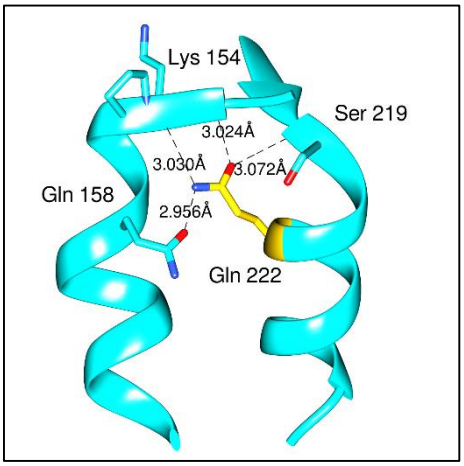


Figure 16 Gln-222 and the corresponding hydrogen bonds.

5.4.3 Gln-249

Gln-249 is located on the surface, so the residue is involved in interacting with the solvents or ligands from the outer environment (Figure 17). The study conducted by Whittington et al. proved the need of this highly conserved residue for the stabilization of dimer interaction in CA-XII. They have identified 19 hydrogen bonds in the interface of the CA-XII dimer, where two of them were highly conserved and formed by Gln-249 (Whittington 2001).

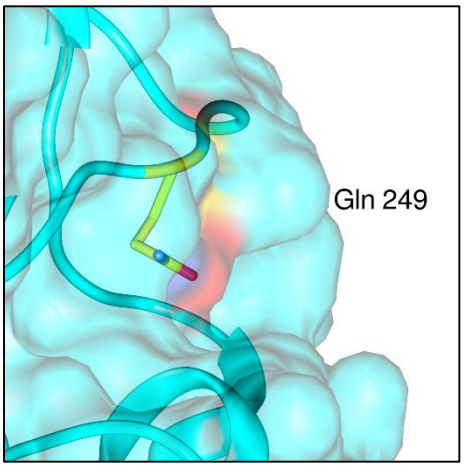


Figure 17 Gln-249 at the surface

5.4.4 Asn-61

Asn-61 is located outside of the active site cavity but as it is conserved, it likely has structural role. Asn-61 forms three hydrogen bonds with backbone atoms in Gly-63 (highly conserved), and Ile-167 and Asn-230. A closer look at the surroundings of Asn-61 suggests that the hydrogen bonds play a role for stabilizing the nearby bends and turns, most importantly the U-turn consisting of Asn-61 to Phe-66 and thus stabilizing the folding pattern of the structure (Figure 18).

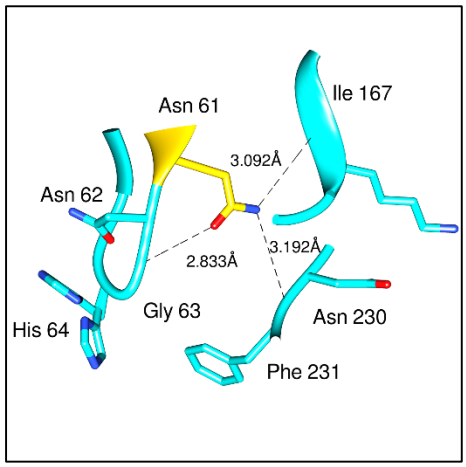


Figure 18 Asn-61 interacting with Gly-63, Ile-167 and Asn-230

5.4.5 Asn-244

Asn-244 forms two hydrogen bonds, one with the backbone of Trp-97 (another perfectly conserved residue) and the other one with the backbone of His-64 (Figure 19). The role of His-64 has already been described and it is highly probable that the Asn-244 provides the stability to the corresponding U-turn located towards the active center. The tip of this U-turn also forms part of the hydrophilic half of the active center.

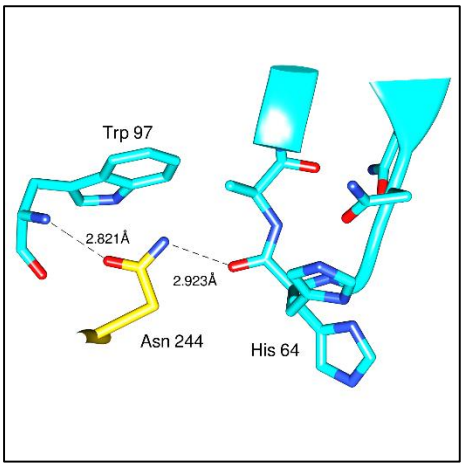


Figure 19 Asn-244 interacting with Trp-97 and His-64

5.4.6 Ser-105

Ser-105 forms two hydrogen bonds with the backbone of neighboring conserved residues His-107 and Tyr-114 (Figure 20). Ser-105 was found in a bend where three other residues His-107, Glu-106, and Gly-104 are also perfectly conserved and located in the same lining of the bend, might mean that the bend formation is necessary for the structural stability. The hydrogen bonding in the bend is probably essential for its conformational stability.

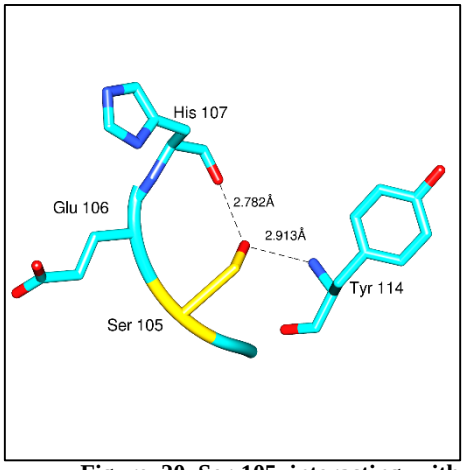


Figure 20 Ser-105 interacting with His-107 and Tyr-114 main chainoxygen and nitrogen respectively

5.4.7 Ser-29

Being small, Serines tend to be in turn or bend and due to the presence of hydroxyl groups, they are very reactive to the other polar residues and substrates (Betts and Russell, 2003). Ser-29 forms three hydrogen bonds with neighboring residues Ser-197 and Tyr-194, thus helping in formation of the bends Ser-29 to Pro-30 and Leu-198 to Gly-196 (Figure 21). As both Ser-197 and Tyr-194 are also perfectly conserved, it indicates that hydrogen bonds are required. Further, a site-specific mutation (Ser-29-Ala) study also revealed that the structure was quite unstable because necessary hydrogen bonds were lost when replaced with Ala-29 (Mårtensson 1992).

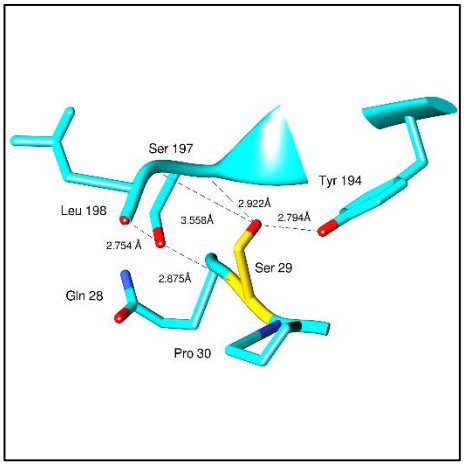


Figure 21 Ser-29 forming hydrogen bonds with Tyr-194 and Ser-197

5.4.8 Ser-197

The side chain of Ser-197 is connected with the main chain of Leu-198 and Ser-29 by forming two hydrogen bonds (Figure 22). The first hydrogen bond is formed with the next residue Leu-198. The second bond is formed with Ser-29, located in the opposite direction. It is interesting to see that both Leu-198 and Ser-29 are located in a lineup of other conserved residues, Thr-199, Ser-197, Pro-30 and Gln-28 and form bends. Therefore, the conservation of Ser-197, together with its environment, supports the idea that it might stabilize the bends and help in maintaining protein folding in this region devoid of regular secondary structure.

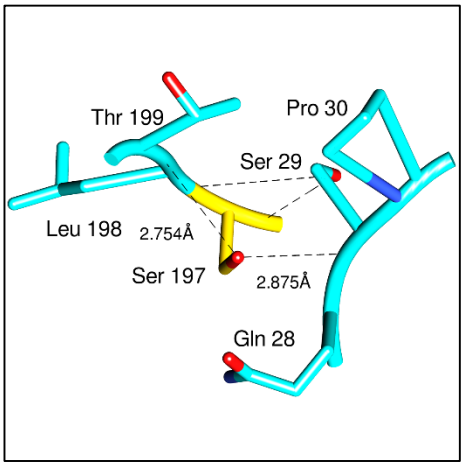


Figure 22 Ser-197 connection with Leu-198 and Ser-29

5.4.9 Pro-201

Being perfectly conserved but absent in the catalytic core indicates that Pro-201 might have a specific role for stabilizing the protein structure. “Prolines are the only amino acid which forms a unique structural feature where the side chain is connected to the protein backbone twice, forming a five-membered nitrogen-containing ring” (Betts 2003). This exceptional feature gives it a conformational rigidity to participate in formation of bend or tight turns (Betts 2003). In alpha CAs, the perfectly conserved residues (Gly-196, Ser-197, Leu-198, Thr-199, Thr-200, Pro-201), consecutive in the U-turn, clearly illustrate the importance for the turn in terms of structural stability (Figure 23) so the functionally important Thr-199 and Thr-200 (see 5.3.2) are kept in their exact positions.

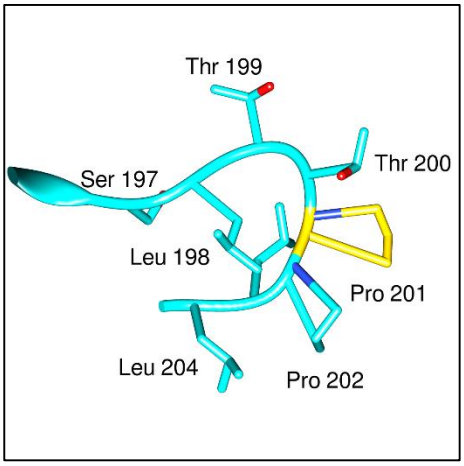


Figure 23 Pro-201 contributing in U-turn formation

5.4.10 Pro-30

If a peptide linkage is in *cis* form, it must have a definite structural role because most of the peptide linkages tend to be in the *trans* form (Donohue 1953). Besides that, the possibility of forming *cis* isomers by Prolines is high compared to the other amino acid residues (MacArthur 1991). The investigation done by Stewert et al. showed that *cis* prolines are highly likely to occur after serine, (Stewart 1990) and this is the case for Ser-29 and Pro-30. Here, the *cis* peptide of Pro-30 forms a slight bend where its main chain carbonyl oxygen forms a hydrogen bond with the amide nitrogen of neighboring conserved residue Gln-249 (Figure 24). The hydrogen bond is likely favoring the bend formation and therefore, giving the structural stability.

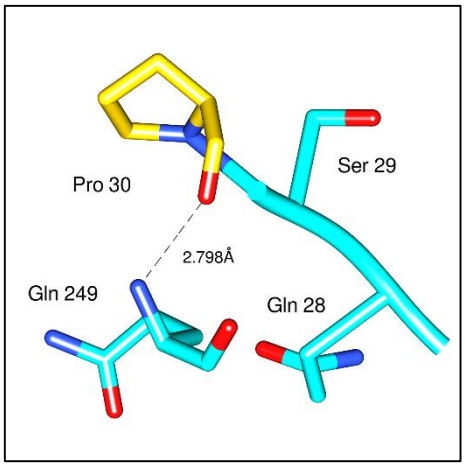


Figure 24 Pro-30 forming hydrogen bond with Gln-249

5.4.11 Pro-186

Prolines generally play special roles in protein structures like bend formation or fixing the dihedral angles to avoid the steric clashes (Woolfson 1990). Prolines are often found in the end of an alpha helix while acting as an alpha helix disruptor because their main chain angles are unable to obtain the normal helical conformation or formation of kinks (Barlow 1988). Pro-186 is also found in the end part of the alpha helix consisting of Asp-180 to Leu-185 (Figure 25). Therefore, it is obvious that Pro-186 is acting as an alpha helix breaker in this structure and thus helping in the protein folding. Further, the usual mean *phi* and *psi* angles for residues in alpha helices are -62, -41 respectively (Barlow 1988), here the *phi* and *psi* angles in Pro-186 show, -77.2 and 177.3 respectively which also show an unfavorable condition for continuing the helix.

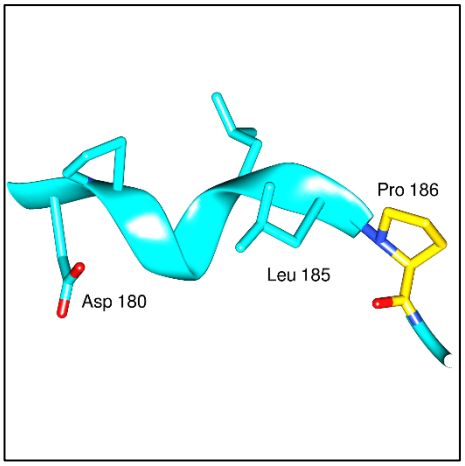


Figure 25 Pro-186 acting as a helix breaker

5.4.12 Arg-246 and Arg-254

Arginines are mainly polar but amphipathic in character. They naturally tend to be in the surface of the protein molecules (Betts 2003). In case of Arg-246 and Arg-254, both are surprisingly seen to be buried, (*Table 2*) indicating that they play exceptional roles for stabilizing the protein structure. Structural inspection revealed they form quite many hydrogen bonds compared to the other conserved residues in this protein molecule. Most notably Arg-254 forms six total hydrogen bonds with Leu-25, Pro-251, Pro-195 and Gln-28, whereas Arg-246 forms four hydrogen bonds with Ala-23, Gln-28 and Ser-29

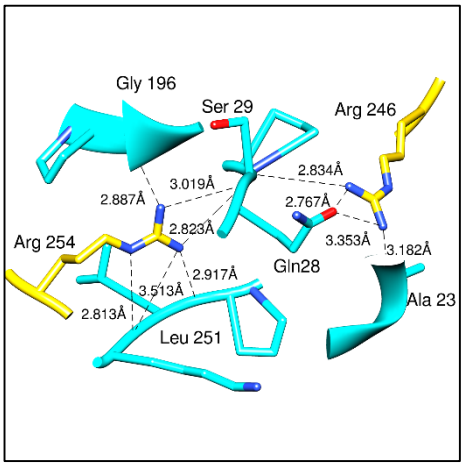


Figure 26 The important hydrogen bonds formed by Arg-246 and Arg-254

(Figure 26). By attaching to the backbone of Ser-29, these two Arginines lock the C-terminal part of the protein stably in the fold.

5.4.13 Gln-28

Gln-28 forms four hydrogen bonds where two of them are bonded with the Arg-246 side chain (Figure 27). A close observation of the structure found that, the hydrogen bond with Arg-246, along with hydrogen bond between Arg-246 and Ser-29, are crucial for the formation of the bend composed of consecutive Arg-27, Gln-28, Ser-29 and Pro-30.

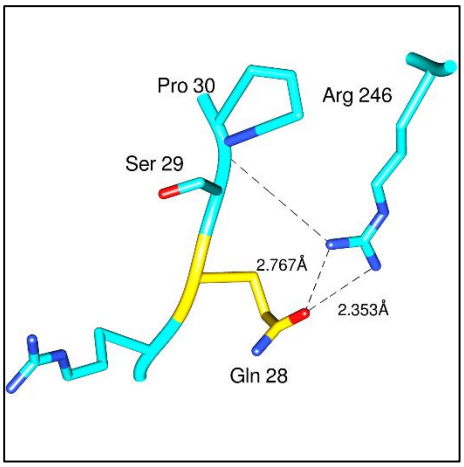


Figure 27 Gln-28 forming hydrogen bonds with Arg-246 and Ser-29

5.4.14 His-107 and Glu-117

His-107 forms hydrogen bonds with Glu-117 and Tyr-194 (Figure 28). These bonds are part of the conserved hydrogen bond network in the active site that promotes the catalytic potency (Kiefer 1995). A case study regarding CA-II deficiency syndrome done by Venta et al. also showed the importance of the invariant His-107, and its corresponding hydrogen bonds to other invariant residues Tyr-194 and Glu-117, for stabilizing the protein structure. Absence of His-107 was found to lack these hydrogen bonds and resulted in destabilization of the structure. (Venta 1991).

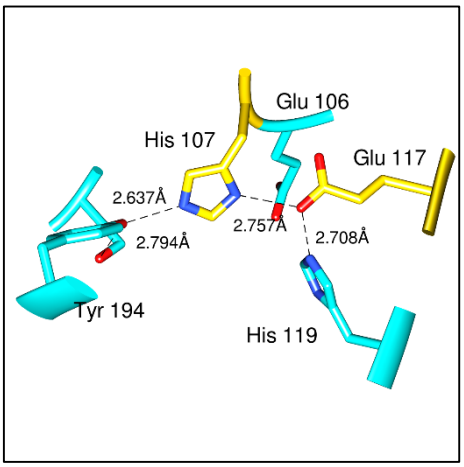


Figure 28 His-107 and Glu-117 participating in forming hydrogen bond network

5.4.15 Trp-97

Trp-97 is located just after His-96. It is bound with the main chain oxygen atom of Met-241 by forming a hydrogen bond in the opposite direction of the active center (Figure 29). The study done by Jennifer et al. showed that this structural orientation stabilizes the conformation of His-96. They have found that several variants at the Trp-97 position affected the zinc binding by the histidine triad, as binding affinity was several-fold slower than in the wild type (Hunt 1997). The data obtained from the study suggest that Trp-97 might anchor the beta strand in which it is located and restricts the conformational flexibility of His-94 and His-96 to inhibit the zinc dissociation process.

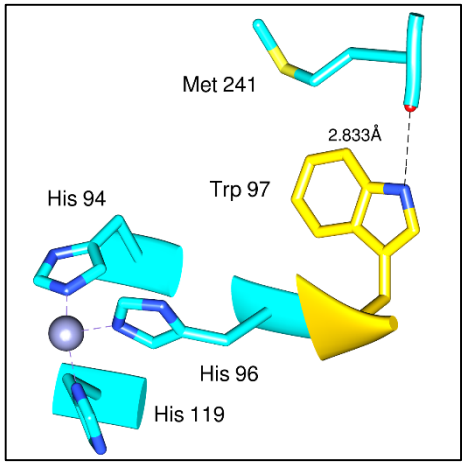


Figure 29 Trp-97 interacting with Met-241

5.4.16 Gly-63, Gly-197, and Gly-104

In hCA-II structure, Gly-63 is found in a tight turn (Figure 30). Glycine is the only amino acid having no side chain. Due to the absence of side chain, they easily form extreme bond angles compared to the other amino acids, therefore they are mostly found in tight turns or U-turns (Chou 1977). Further, Tamai et al. proved that Gly-63 ensures the rapid conformational changes by His-64 which plays a key role for efficient proton transfer. Moreover, a site-directed mutation (Gly-63-Gln) study showed 20-25% lower activity in catalysis than the wild form of the protein (Tamai 1996). Two other Glycines, Gly-196 and Gly-104, start slight turns in the structure and in both cases a number of consecutive conserved residues participating in the formation of bends clearly indicates their structural role.

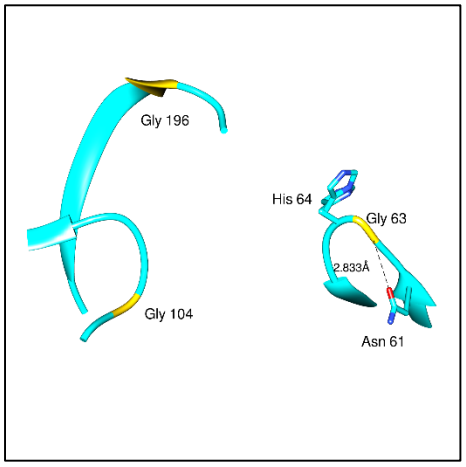


Figure 30 Gly-63, Gly-197 and Gly104 positioned in the turns and bend.

5.4.17 Tyr-194 and Trp-209

The hydrophobic interaction between Tyr-194 and Trp-209 probably necessary for the specific folding pattern in the protein core. Considering both are aromatic and their position in the same geometric plane, there is a parallel-displaced π-π stacking interaction (Dougherty 2007) between the two aromatic systems. Also, they form necessary hydrogen bonds with conserved residues Ser-29, Ser-197, and His-107 helping them fold into the protein core and maintain the typical beta sheet pattern of the alpha carbonic anhydrases (Figure 31).

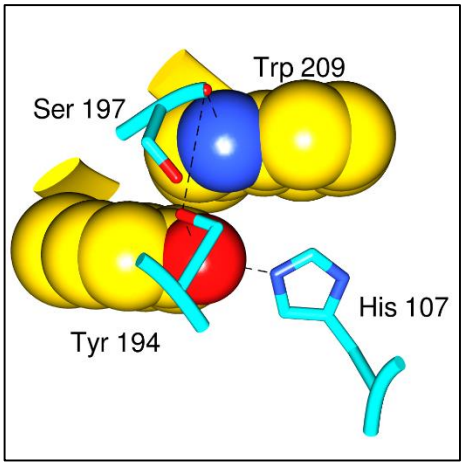


Figure 31 Tyr-194 and Trp-209 showing π-π stacking interaction

5.4.18 Leu-44

Leucines are nonpolar hydrophobic in nature and usually tend to be buried in the protein core. Due to the absence of a reactive group in the side-chain, they does not form any hydrogen bonds. Leu-44 is also buried in the protein core and are seen to be involved in hydrophobic packing with the nearby hydrophobic portion of the residues- Tyr-191, Asp-41, Ala-258 and Pro-83 (Figure 32). Although Leu-44 is positioned in the surface but the side-chain is directed towards the core. Moreover, corresponding main chain of the residue also bending towards the core and forming a specific folding pattern. Such orientation is clearly indicating the importance of the Leu-44 residue for hydrophobic packing as well as for the structural stability. The contacts with the residues in the proximity are showed in straight red line that were determined by the Chimera clash/contact tool.

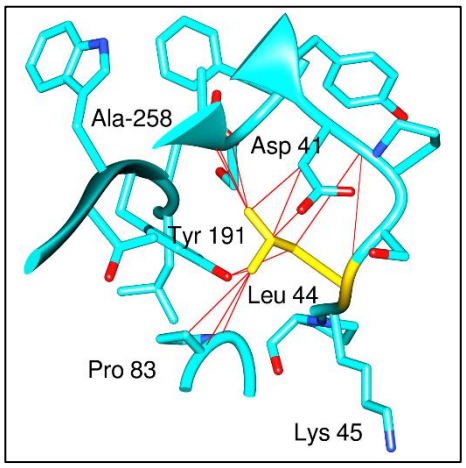


Figure 32 Leu-44 is participating in hydrophobic packing

5.4.19 His-122

The position of His-122 is just after the hydrophobic Val-121 where Val-121 is interacting with the other hydrophobic residues and forming the hydrophobic cluster in the active site. However, the side chain orientation of His-122 is opposite to the hydrophobic cluster as forming a hydrogen bond with the side chain of Tyr-51. The side chain orientation is justified by the fact that being hydrophilic, it is repulsed the hydrophobic residues- Val-121, Leu-141, Ile-91, Val-143 and Phe-131 from the active cavity. Further, having a calculated pKa value of 1.67, which is four times lower than their original pKa value 6.5 (see Appendix 4) shows that His-122 is quite acidic and needs to be in the deprotonated state most of the time.

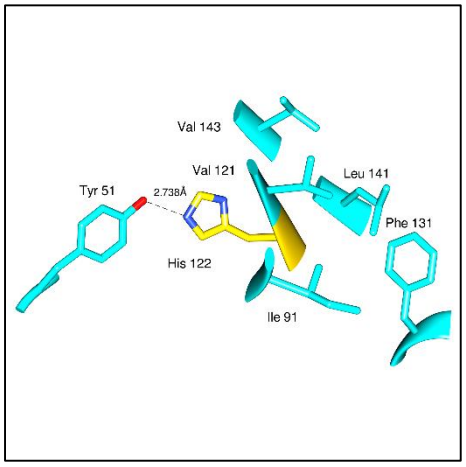


Figure 33 His-122 forming hydrogen bond with Tyr-51 positioned in the opposite side of the hydrophobic cluster

5.4.20 Ala-134

Alanines are tiny in size and hydrophobic in nature. Due to the absence of a reactive atom in side-chain, they do not form hydrogen bonds. Here, Ala-134 is buried but few hydrophobic contacts are seen with nearby hydrophobic residues, rather it is interacting with some of the nearby atoms. The residues located within 5Å distance of Ala-134 are visualized and the contacts, determined by Chimera clash/contact tool are shown in red straight lines (Figure 34). These interactions can be due to the act of van der Waals forces.

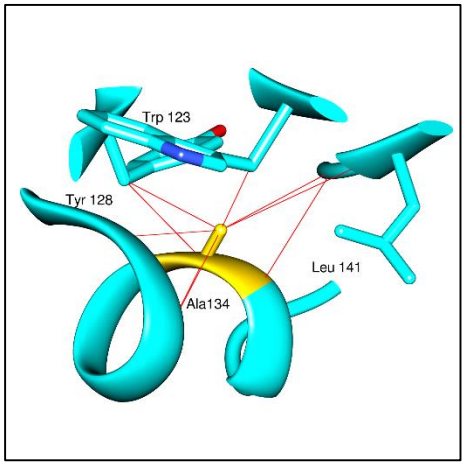


Figure 34 Ala-134 contacts with the hydrophobic residues in the proximity

5.4.21 Ala-142

Like Ala-134, Ala-142 also forms hydrophobic interactions but here the hydrophobic contacts are quite strong, formed by seven hydrophobic residues in close proximity. The residues residing within 5Å distance from Ala-142 were selected and all of them are hydrophobic in nature, as expected for a residue buried very deep in the structure. Therefore, Ala-142 certainly makes hydrophobic contacts with nearby hydrophobic residues Leu-120, Leu-144, Ile-210, Leu-79, Leu-84 and the aromatic ring of Tyr-88 thus stabilizing the structure (Figure 35).

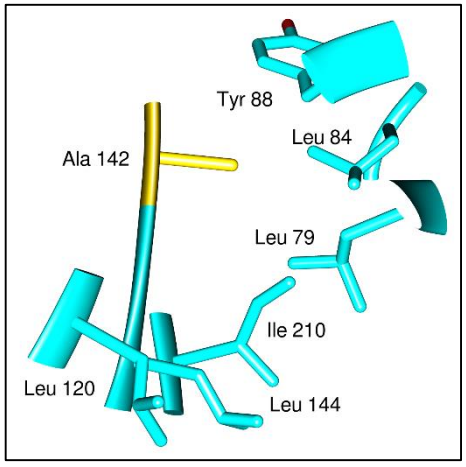


Figure 35 The hydrophobic interactions of Ala-142 with the neighbouring hydrophobic residues

5.5 Statistical Analysis

The statistical computing was done for visualizing different properties of the conserved residues in the universal group (Figure 36).

The first graph (Figure 36A) illustrates the frequency of the conserved polar residues vs. non-polar residues. It is clear from the graph that the number of conserved polar residues greater (by almost 15%) than the non-polar residues. The second graph (Figure 36B) shows that conserved residues are more than twice as likely to be hydrophilic than hydrophobic at 57.89% and 26.32% respectively. However, there is also a good number of residues which are amphipathic in character. In the third graph (Figure 36C), the proportion of buried residues is seen to be much greater than the surface residues (92.11% are buried compared to only 7.89% for surface). The fourth graph (Figure 36D) describes the overall distribution of the individual residues in the universally conserved group. Histidines are present at highest frequency, relative to the rest of the residues. The second highest position is occupied by Glutamine and the third most frequent residues are Glycine, Proline, Serine, Valine and Tryptophan, each seen in eight conserved positions. However, the lowest frequency residue

is Tyrosine. Nevertheless, amino acid residues Cysteine, Methionine, Isoleucine, Phenylalanine are completely absent in the universally conserved residue list.

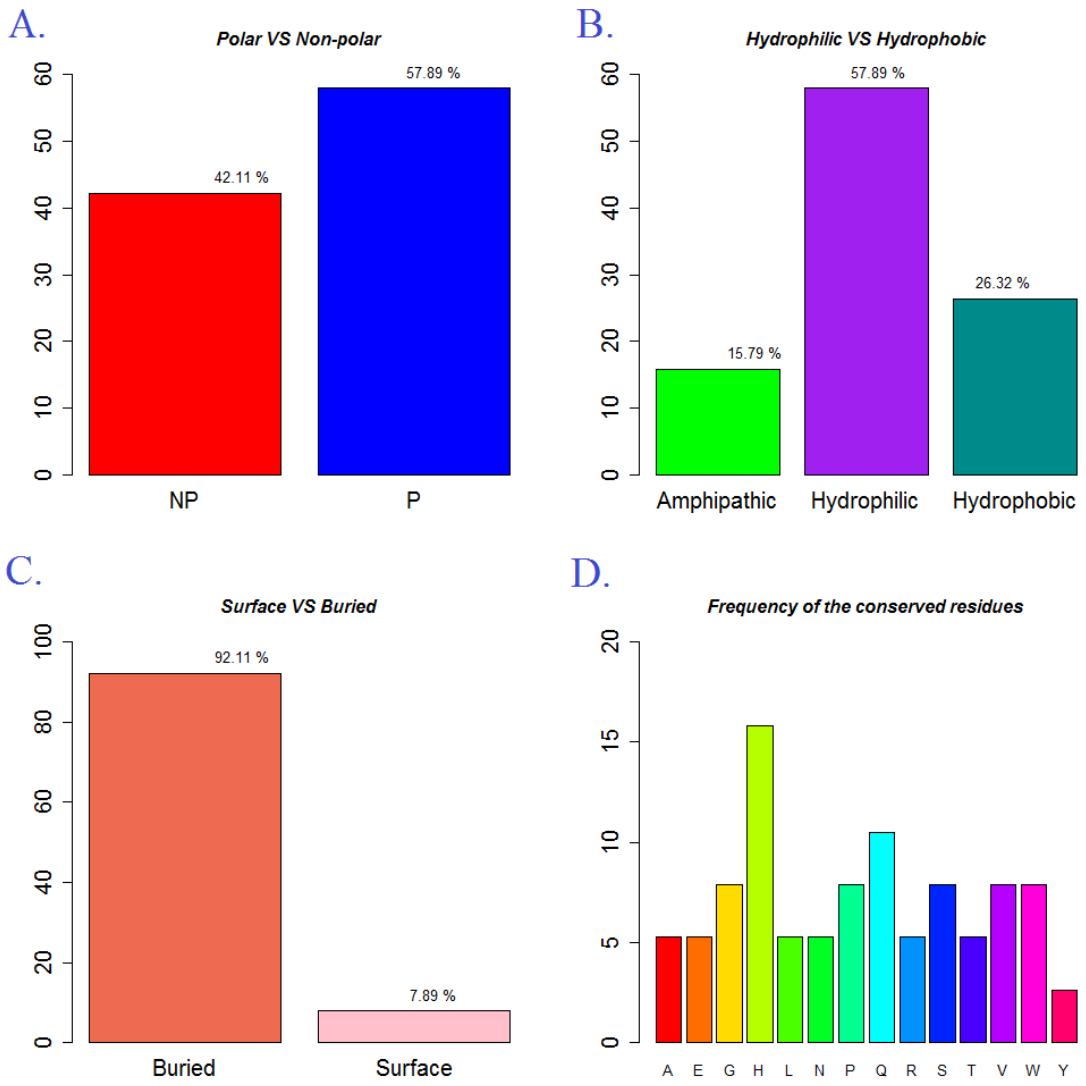


Figure 36 The frequency of distribution of the amino acid in the final universally conserved residue list and their frequency in terms of chemical properties. NP = Non-polar, P = Polar. The statistical analysis and the figures were made in the R (version 3.1.1) (http://www.r-project.org/) environment.

5.6 List of Residues Conserved Only in Cytoplasmic or Extracellular CA Isozymes

Table 4 was compiled of absolutely conserved residues that are only present in the cytoplasmic and extracellular isozymes. The most notable findings are- there are three Aspartate residues and many amphipathic residues that are present in the cytoplasmic group while none of them are found in the extracellular group. In addition to that, the Cysteines that are responsible for the formation of disulphide bridge in the extracellular domain are conserved.

Table 4 List of residues that are conserved only in cytoplasmic isozymes (CA-I, II, III, VII and XIII) or extracellular isozymes (CA-VI, IX, XII and XIV) and their corresponding K_a/K_s values

	Cytoplasmic residues- 3KS3			**Extracellular residues- 4HT2**		
Serial	PDB_res	PDB_pos	K_a/K_s	PDB_res	PDB_pos	K_a/K_s
1	W	5	0.027	G	8	0.029
2	Y	7	0.013	C	22	0.026
3	P	21	0.012	I	30	0.027
4	A	23	0.029	H	66	0.031
5	G	25	0.02	V	68	0.013
6	V	68	0.038	L	72	0.019
7	D	72	0.0095	A	87	0.03
8	G	82	0.014	G	95	0.02
9	R	89	0.0079	G	138	0.022
10	L	90	0.014	L	139	0.019
11	G	98	0.013	V	141	0.012
12	D	110	0.039	L	142	0.019
13	W	123	0.027	N	152	0.01
14	N	124	0.0069	P	201	0.016
15	D	139	0.0098	C	202	0.02
16	G	140	0.015	T	209	0.037
17	K	170	0.0089	V	210	0.031
18	F	176	0.014			
19	D	180	0.036			
20	L	184	0.014			
21	W	192	0.027			
22	T	193	0.026			
23	P	202	0.012			
24	E	205	0.039			
25	R	227	0.011			
26	P	250	0.012			

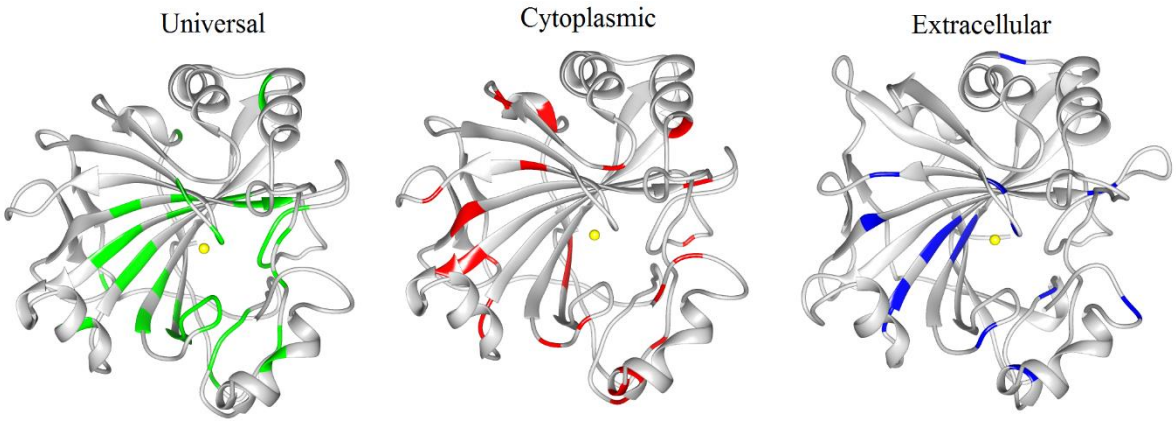


Figure 37 Comparison between only cytoplasmic and only extracellular conserved residues. The left figure shows the universally conserved residues in green color, middle figure shows unique cytoplasmic conserved residues in red color and right figure shows unique extracellular conserved residues in blue color. Left and middle figures were made from hCA-II structure (PDB: 3ks3) and right figure was made from hCA-XII structure (PDB: 4ht2)

5.7 Cytoplasmic and Extracellular Conserved Surface Visualization

The hydrophobic residues were visualized and marked in the protein structure PDB-3ks3 for cytoplasmic and PDB-4ht2 for extracellular conserved residue analysis (Figure 38). The figures denoting some specific information about the conserved surfaces, the conserved surface areas are scattered throughout the whole protein molecule in the cytoplasmic structure where in extracellular, the conserved surfaces are accumulated at the center, mostly around the active cavity. When coloring the surfaces based on the chemical properties, in the cytoplasmic isoforms, the quantity of the hydrophobic surfaces are less compared to the hydrophilic and amphipathic parts. However, in the extracellular isoforms, the amount of hydrophilic and hydrophobic surfaces are same but the amount of overall conserved surface area is very trivial.

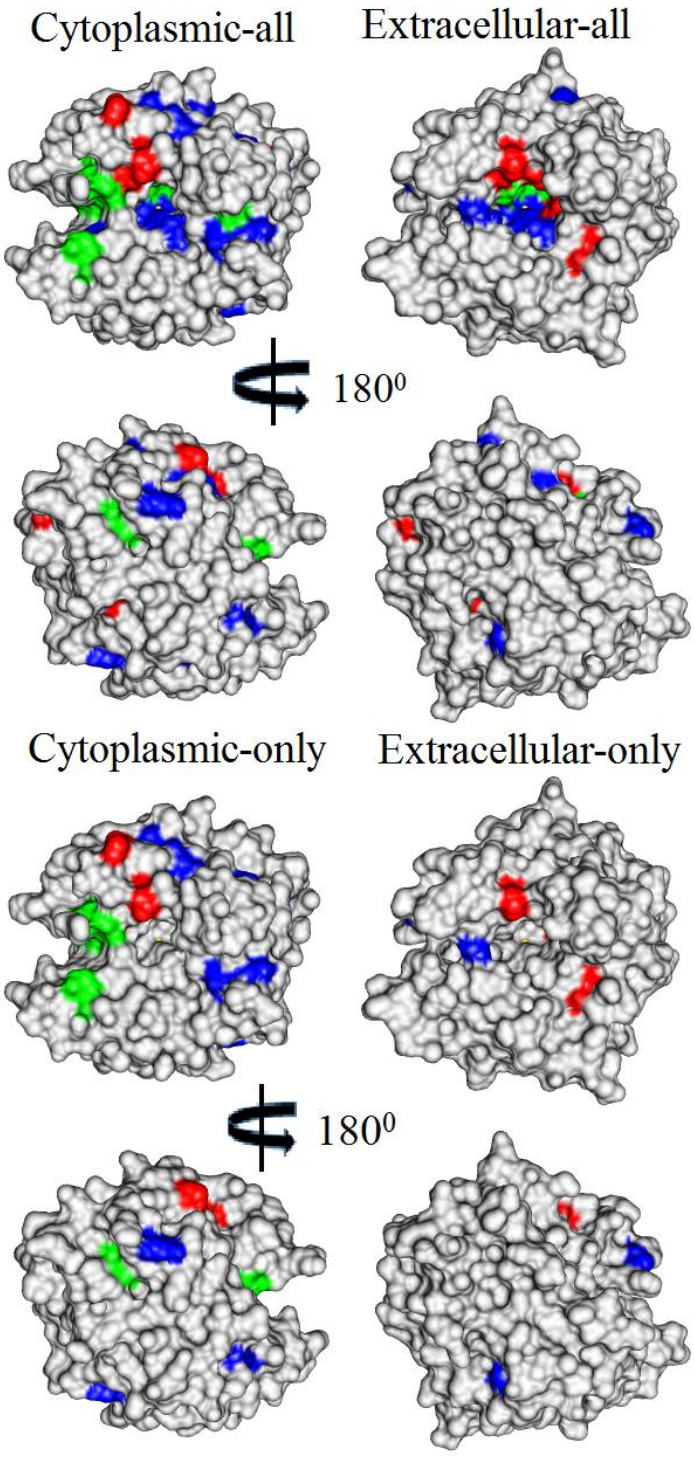


Figure 38 Cytoplasmic and extracellular conserved surfaces visualization and comparison, with hydrophobic-red, hydrophilic-blue and amphipathic-green. The first and third rows are showing straight view to the active site, representing all the absolutely conserved and residues conserved only in each group. The second and fourth rows, visualizing the surfaces rotating 180^0 vertically to the right. Protein structure PDB ids- 3ks3 for cytoplasmic and 4ht2 for extracellular.

5.8 Visualization and Comparison of N-glycosylation Sites on Structures

The categorized N-linked glycosylation sites were mapped and colored for each of the isozymes in their representative crystallographic structures. For the isozyme CA-VI, the human CA-VI (PDB: 3FE4); for isozyme CA-IX, human CA-IX (PDB: 4M2V); for isozyme CA-XII (PDB: 4HT2) and for isozyme CA-XIV (PDB: 4LU3) were selected from RCSB.org. Then a missing part of CA XII at N terminal end was modelled to accommodate one glycosylation site. The sites of different categories were colored differently from deep color to a lighter color describing the most conserved to the least conserved, respectively (Figure 39). Based on the alignment of Appendix 5, each of these isozymes has one to three conserved glycosylation sites (darkest colors in Figure 39), which are: Asn-256, Asn-67 (CA6); Asn-213 (CA9); Asn-1, Asn-52, and Asn-134 (CA12); Asn-195 (CA14).

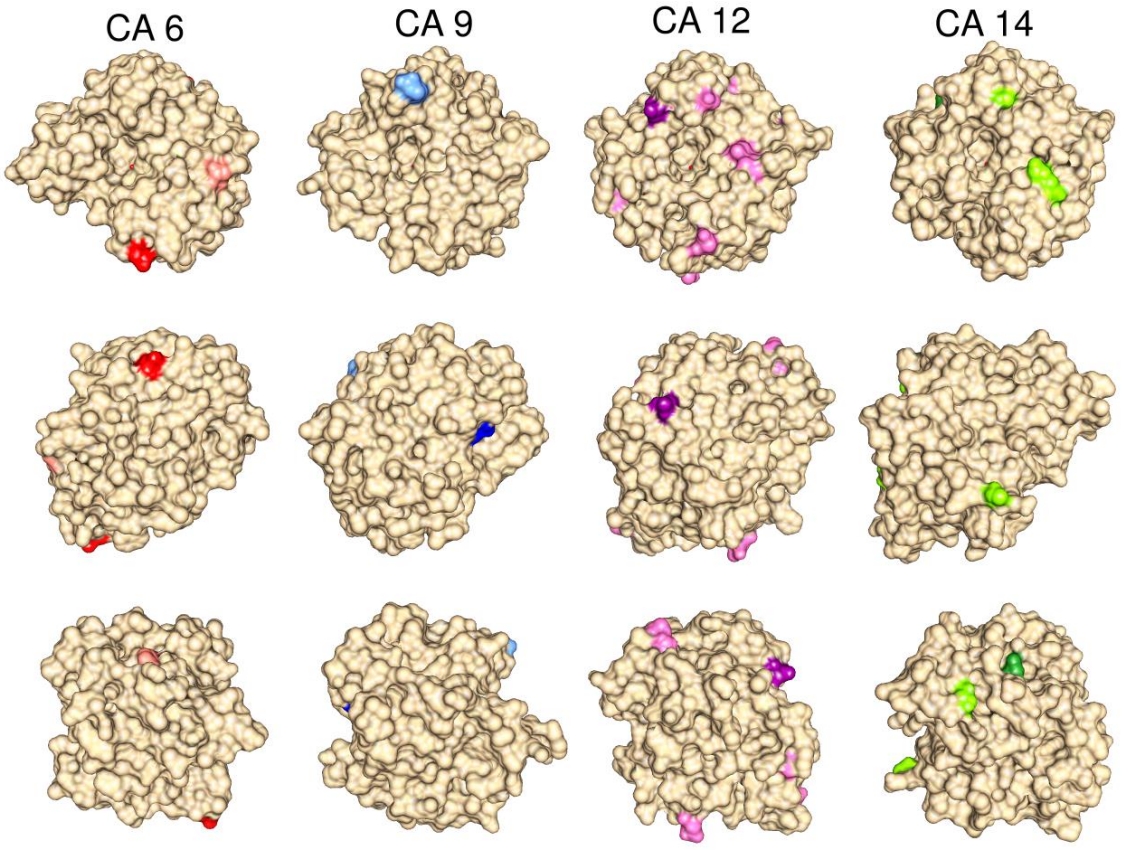


Figure 39 Comparative analyses of the N-linked glycosylation sites in CA-VI, IX, XII and XIV. The deeper colours denote "conserved glycosylated" sites while the lighter one shows "occasionally glycosylated" sites. First row shows the active site facing us; second row turned 120^0 vertically to the right and the third row turned 120^0 vertically to the left. The structures were edited in Chimera.

6 Discussion

6.1 "Universally Conserved" Residues

The universally conserved residue list, which includes both perfectly conserved and highly conserved residues, are predominantly participating in structural and functional activities. Residues that are conserved were found to be forming hydrogen bonds or hydrophobic interactions with the other conserved residues. This fact actually strengthens the logic that the residues are conserved across species and different isozymes. The invariant hydrogen bonds in the conserved areas mostly indicated that absence of any of the conserved residues in the respective regions could result in loss of bonding and become unstable. The information that was found from the literature survey specially described the evidences regarding the mutation studies that were done in various laboratories. In all of the cases, the mutated versions showed lower efficacy for both catalytic activity and/or structural stability. Therefore, the physicochemical study approach that was taken for each of the conserved universal amino acid residues was justified by the evidence from literature surveys in most of the cases. The residues that were not mentioned in any publications were intensively analyzed according to their chemical features and environment, and rational predictions of their roles were made for each.

In Figure 40, some of the coinciding structural features were portrayed with different colors. The most eye-catching figure is three conserved Histidines (94, 96, and 119) at the catalytic site contributing in stabilizing the zinc ion as well as proton flow during the reaction mechanism (showed in yellow). Besides that, many of the structural residues are highly likely to contribute in stabilizing the loops/turns (Figure 40). Nevertheless, turns and loops are the vital component of a globular protein. Figure 40 summarizes the stabilizing effects in various loops of the alpha CA folds. The loop that is seen in blue color is composed of conserved His-64, Gly-63 and Asn-61. Another loop containing conserved Leu-198, Thr-199, Thr-200, and Pro-201 are colored in green. The aforementioned two loops are the most vital loop in the alpha CA structures as they are actively participating in both catalysis and stabilizing the

structure. Furthermore, a group of consecutive conserved residues also stabilizes two major bends. One such bend is formed by Ser-105, Glu-106, and His-107 (red) and another one is with Gln-28, Ser-29, and Pro-30 (cyan).

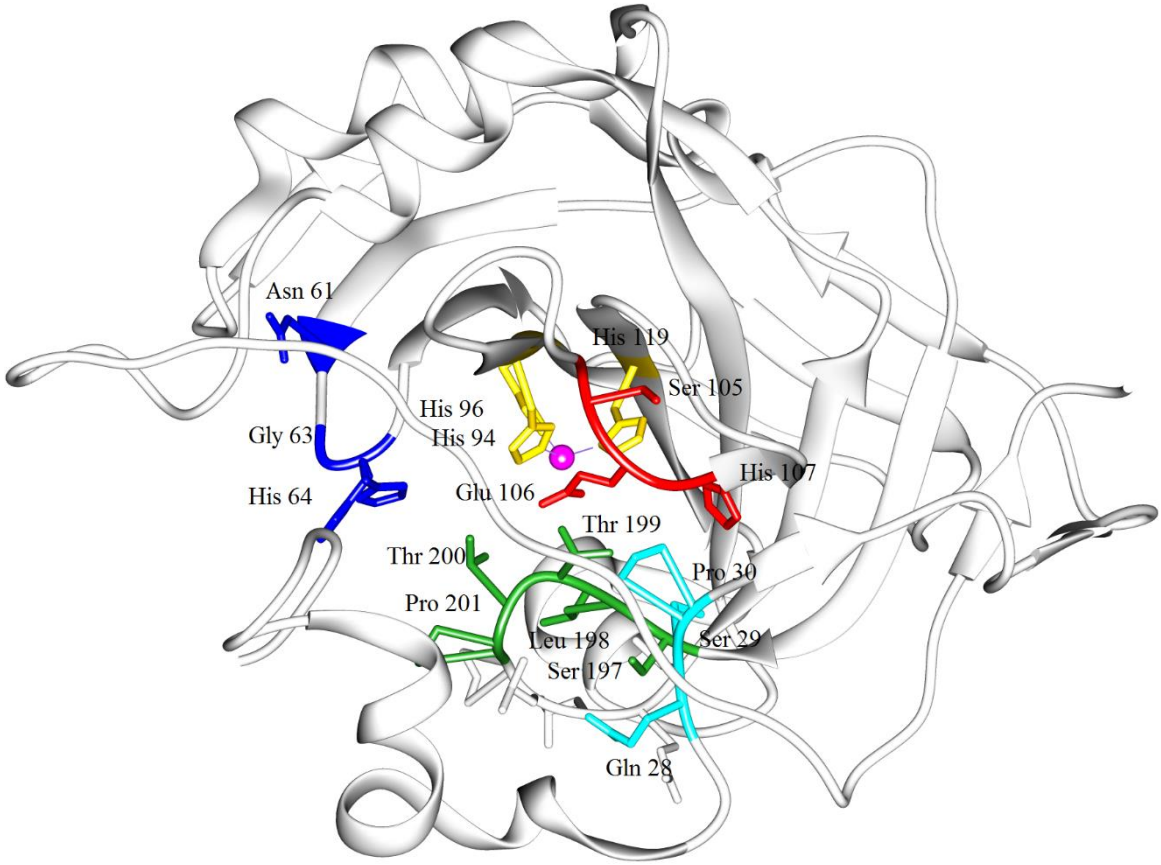


Figure 40 Two vital loops formed by Asn-61, Gly-63 and His-64 (blue), and Ser-197 to Pro-201 (green). Two major bends formed by Ser-105 to His-107 (red) and Gln-28 to Ser-29 (cyan). The catalytic Histidines are in yellow color.

The result of the statistical analysis showed that a vast majority of the universally conserved residues are buried, yet surprisingly most of them are hydrophilic in nature, due to extensive hydrogen bonding networks within the protein core. In addition to that, the Histidines are present in highest frequency; mainly because of the catalytic center of three zinc binding Histidines. The pKa analysis (Appendix 4) showed that Histidines are highly acidic in nature, which indicates that they are favoring proton flow, which is one of the key functional elements in catalysis.

6.2 Cytoplasmic and Extracellular α-CAs

This part of study was made to investigate the differences between cytoplasmic and extracellular conserved residues that might be worthy of attention in their respective group. The structural comparison between cytoplasmic and extracellular conserved residues disclosed that the areas that are conserved in both groups are unique; none of them shared the same conserved places (Figure 37). Therefore, it is probable that the conserved "only cytoplasmic" residues or "only extracellular" residues are functionally important for their respective roles in their subcellular parts.

The hydrophilic surfaces of the proteins are highly likely to bind with the surrounding solvent molecules whereas the hydrophobic surfaces are more prone to interact with ligands or other protein molecules. However, the amphipathic surfaces are ideal for making protein-protein interactions (Creighton 1992). Here, in the cytoplasmic isozymes, some conserved amphipathic surface residues are located close to each other and forming an amphipathic patch. This can be considered as a potential interaction region for other proteins. Besides, two small conserved hydrophobic regions are seen (cytoplasmic-only in Figure 38) close to the amphipathic patch. The conserved hydrophobic regions also possibly participate in protein-protein interactions. On the other hand, no amphipathic or even meaningful hydrophobic surfaces are seen conserved in the extracellular isoforms (the two hydrophobic residues that are conserved are located near the entrance of the cavity that probably would not participate in protein binding). This might mean that the cytoplasmic isozymes share a common binding function whereas the extracellular isozymes interact with different proteins.

Another eye-catching feature is that in the extracellular domain, Cys-22 and Cys-202 are conserved and forming a disulfide linkage. Because the disulfide is also quite near the important doorkeeper residues Thr-198 and Thr-199, it stabilizes that loop and the active center in general.

6.3 N-glycosylation Sites

The idea of this part of the analysis was to identify functional and non-functional glycosylation sites. The results from frequency calculation (*Table 1*) of the glycosylation sites explored conserved and rare glycosylation sites. The conserved glycosylation sites, which are found in the majority of the species, might have a functional purpose for the presence of the oligosaccharide (glycan). Further from the structural visualization (Figure 39), it is clear that none of the conserved glycosylated sites is shared among the isozymes. It is also seen that most of the glycosylation sites are present close to the entrance of the catalytic pocket (1st row of Figure 39). This is logical with the fact that other protein-protein interactions would not be likely to block the active site, whereas in the other two projections (2nd and 3rd row of Figure 39) they have more extensive clear regions that might be interaction regions with other proteins or dimerization regions.

Conclusions

The main purpose of this work was to identify the maximal number of highly conserved residues that are functionally important across the alpha carbonic anhydrases. The first list of the "universally conserved" residues were filled by the absolute conserved residues followed by the second list of "highly conserved" residues that were not perfectly conserved due to the presence of single exception at the species or isozyme level. The characteristics of the conserved residues showed interesting results. The statistics of the RSA values revealed that many of these conserved residues are buried in the protein core. Further, the structural investigation found quite distinct role for each of the conserved residues that justified their trend to be conserved. Besides, the surface visualization of the conserved cytoplasmic and extracellular residues gives a stronger motive about their specificity to be bound with the protein or ligand. It can be speculated that the cytoplasmic isoforms may have common binding modes as opposed to the extracellular isoforms that would seem to interact with different molecules in unique modes. Finally, the N-glycosylation site visualization presents that the densely aggregated sites are positioned at the entrance of the active site. This arrangement of the sites might be required so that any protein-protein interaction does not block the passage towards the active cavity.

The work procedure that was conducted to construct the universally conserved residues can be applied for determining the highly responsible functional elements in any protein family like beta CAs and gamma CAs or in different sub-groups like GPI-linked alpha CAs and CARPs etc. The alpha CA comparisons will be extended to further groups of organisms, such as invertebrates, protozoans, fungi, plants, and bacteria. Furthermore, the enigmatic, non-catalytic CA-related proteins are a good target for comparisons with the whole rest of the alpha family, using the conservation approach I have applied here. We expect that this would give meaningful insights in the functions of enzymatically active and inactive alpha CAs.

References

Adamczak, R., Porollo, A., Meller, J. 2005. "Combining prediction of secondary structure and solvent accessibility in proteins." *Proteins* 59 (3): 467-475.

Aggarwal, M., Boone, C. D., Kondeti, B., & McKenna, R. 2013. "Structural annotation of human carbonic anhydrases." *Journal of enzyme inhibition and medicinal chemistry* 28 (2): 267-277.

Aspatwar, A., Tolvanen, M. E., Jokitalo, E., Parikka, M., Ortutay, C., Harjula, S. K. E., ... & Parkkila, S. 2012. "Abnormal cerebellar development and ataxia in CARP VIII morphant zebrafish." *Human molecular genetics* dds438.

Avvaru, B. S., Kim, C. U., Sippel, K. H., Gruner, S. M., Agbandje-McKenna, M., Silverman, D. N., McKenna, R. 2010. "A short, strong hydrogen bond in the active site of human carbonic anhydrase II." *Biochemistry* 49 (2): 249-251.

Barker, H. R. 2013. *Development of a Protein Conservation Analysis Pipeline and Application to Carbonic Anhydrase IV.* Master's Thesis, University of Tampere.

Barlow, D. J., Thornton, J. M. 1988. "Helix geometry in proteins." *Journal of molecular biology* 201 (3): 601-619.

Berg, J. M., Tymoczko, J. L., Stryer, L. 2010. *Biochemistry; W. H.* 7th. New York: Freeman and Company.

Betts, M. J., Russell, R. B. 2003. "Amino acid properties and consequences of substitutions." *Bioinformatics for geneticists* 317: 289.

Burley, S. K., Petsko, G. A. 1986. "Amino-aromatic interactions in proteins." *Febs Letters* 203 (2): 139-143.

Chegwidden, W. R., Carter, N. D., & Edwards, Y. H. (Eds.). 2000. *The Carbonic Anhydrases: New Horizons.* Vol. 90. Springer.

Chou, P.Y., Fasman, G.D. 1977. "Beta-turns in proteins." *Journal of Molecular Biology* 115: 135-175.

Comeron, J. M. 1995. "A method for estimating the numbers of synonymous and nonsynonymous substitution per site." *Journal of Molecular Evolution* 41: 1152-1159.

Crawford, J. L., Lipscomb, W. N., & Schellman, C. G. 1973. "The reverse turns as a polypeptide conformation in globular proteins." *Proceedings of the National Academy of Sciences* 70: 538-542.

Creighton, T.E. 1992. *Proteins: Structure and Molecular Properties.* 2nd. New York: W. H. Freeman.

Dodgson, S. J., Tashian, R. E., Gross, G., Carter, N. D. 1991. *The Carbonic Anhydrases: Cellular Physiology and Molecular Genetics.* Springer.

Domsic, J. F., Avvaru, B. S., Kim, C. U., Gruner, S. M., Agbandje-McKenna, M., Silverman, D. N., & McKenna, R. 2008. "Entrapment of Carbon Dioxide in the Active Site of Carbonic Anhydrase II." 283 (45): 30766-30771.

Donohue, J. 1953. "Hydrogen bonded helical configurations of the polypeptide chain." *Proceedings of the National Academy of Sciences of the United States of America* 39 (6): 470.

Dougherty, D. A. 2007. "Cation-pi interactions involving aromatic amino acids." *The Journal of nutrition* 137 (6): 1504-1508.

Esbaugh, A. J. & Tufts, B. L. 2006. "The structure and function of carbonic anhydrase isozymes in the respiratory system of vertebrates." *Respiratory Physiology & Neurobiology* 154 (1-2): 185-198.

Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M. Y., ... & Sali, A. 2006. "Comparative Protein Structure Modeling With MODELLER." *Current protocols in bioinformatics* 5-6.

Fisher, S. Z., Kovalevsky, A. Y., Domsic, J. F., Mustyakimov, M., McKenna, R., Silverman, D. N., & Langan, P. A. 2010. "Neutron structure of human carbonic anhydrase II: implications for proton transfer." *Biochemistry* 49 (3): 415-421.

Fisher, Z., Hernandez Prada, J. A., Tu, C., Duda, D., Yoshioka, C., An, H., ... & McKenna, R. 2005. "Structural and kinetic characterization of active-site histidine as a proton shuttle in catalysis by human carbonic anhydrase II." *Biochemistry* 44 (4): 1097-1105.

Fisher, Z., Kovalevsky, A. Y., Mustyakimov, M., Silverman, D. N., McKenna, R., & Langan, P. 2011. "Neutron structure of human carbonic anhydrase ii: a hydrogen-bonded

water network switch is observed between ph 7.8 and 10.0." *Biochemistry* 50 (44): 9421-9423.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., ... & Yates, A. 2013. "Ensembl 2014." *Nucleic acids research.*

Freeze, H. H., & Elbein, A. D. 2009. *Essentials of glycobiology.* 2nd. Edited by A Varki. New York: Cold Spring Harbor Laboratory Press.

Gavel, Y., & von Heijne, G. 1990. "Sequence differences between glycosylated and non-glycosylated Asn-X-Thr/Ser acceptor sites: implications for protein engineering." *Protein engineering* 3 (5): 433-442.

Gupta, R., Jung, E., Brunak, S. 2004. "Prediction of N-glycosylation sites in human proteins." *In preparation.*

Hirota, J., Ando, H., Hamada, K., & Mikoshiba, K. 2003. "Carbonic anhydrase-related protein is a novel binding protein for inositol 1, 4, 5-trisphosphate receptor type 1." *Biochemical Journal* 372: 435-441.

Holfmann, A. W. 1855. "On Insolinic Acid." *Proceedings of the Royal Society* 8: 1-3.

Hunt, J. A., & Fierke, C. A. 1997. "Selection of Carbonic Anhydrase Variants Displayed on Phage. AROMATIC RESIDUES IN ZINC BINDING SITE ENHANCE METAL AFFINITY AND EQUILIBRATION KINETICS." *The Journal of Biological Chemistry* 272: 20364-20372.

Ina, Y. 1995. "New methods for estimating the numbers of synonymous and nonsynonymous substitutions." *Journal of molecular evolution* 40 (2): 190-226.

Jeffrey, George A. 1997. “An introduction to hydrogen bonding.” *Oxford University Press.*

Kabsch, W., Sander, C. 1983. "Dictionary of Protein Secondary Structure: Pattern Recognition of Hydrogen-Bonded and Geometrical Features." *Biopolymers* 22: 2577-2637.

Kiefer, L. L., Paterno, S. A., & Fierke, C. A. 1995. "Hydrogen bond network in the metal binding site of carbonic anhydrase enhances zinc affinity and catalytic efficiency." *Journal of the american chemical society* 117 (26): 6831–6837.

Ko, J., Murga, L. F., André, P., Yang, H., Ondrechen, M. J., Williams, R. J., ... & Budil, D. E. 2005. "Statistical criteria for the identification of protein active sites using

theoretical microscopic titration curves." *Proteins: Structure, Function, and Bioinformatics* 59 (2): 183-195.

Krebs, J. F., Fierke, C. A., Alexander, R. S., & Christianson, D. W. 1991. "Conformational mobility of his-64 in the thr-200. fwdarw. ser mutant of human carbonic anhydrase ii." *Biochemistry* 30 (38): 9153–9160.

Krebs, J. F., Ippolito, J. A., Christianson, D. W., & Fierke, C. A. 1993. "Structural and functional importance of a conserved hydrogen bond network in human carbonic anhydrase ii." *Journal of Biological Chemistry* 268 (36): 27458–27466.

Leggat, W., Dixon, R., Saleh, S., Yellowlees, D. 2005. "A novel carbonic anhydrase from the giant clam Tridacna gigas contains two carbonic anhydrase domains." *FEBS Journal* 272 (13): 3297-3305.

Liljas, A., Håkansson, K., Jonsson, B. H., Xue, Y. 1994. "Inhibition and catalysis of carbonic anhydrase Recent crystallographic analyses." *European journal of biochemistry* 219: 1-10.

Liljas, A., Kannan, K. K., Bergsten, P. C., Waara, I., Fridborg, K., Strandberg, B., ... & Petef, M. 1972. "Crystal Structure of Human Carbonic Anhydrase C." *Nature New Biology* 235: 131-137.

Lindskog, S. 1983. *Zinc Enzymes (Spiro, T. G., ed).* New York: John Wiley & Sons.

Livingstone, C. D., & Barton, G. J. 1993. "Protein sequence alignments: a strategy for the hierarchical analysis of residue conservation." *Computer applications in the biosciences: CABIOS* 745-756.

Lovejoy, D. A., Hewett-Emmett, D., Porter, C. A., Cepoi, D., Sheffield, A., Vale, W. W., & Tashian, R. E. 1998. "Evolutionarily conserved, "acatalytic" carbonic anhydrase-related protein XI contains a sequence motif present in the neuropeptide sauvagine: the human CA-RP XI gene (CA11) is embedded between the secretor gene cluster and the DBP gene at 19q13.3." *Genomics* 54 (3): 484-493.

MacArthur, M. W., Thornton, J. M. 1991. "Influence of proline residues on protein conformation." *Journal of molecular biology* 218 (2): 397–412.

Maupin, C. M., & Voth, G. A. 2007. "Preferred orientations of His64 in human carbonic anhydrase II." *Biochemistry* 46 (11): 2938-2947.

Maupin, C. M., Zheng, J., Tu, C., McKenna, R., Silverman, D. N., Voth, G. A. 2009. "Effect of Active-site Mutation at Asn67 on the Proton Transfer." *Biochemistry* 48 (33): 7996-8005.

McGaughey, G. B., Gagné, M., Rappé, A. K. 1998. "pi-Stacking interactions. Alive and well in proteins." *The journal of biological chemistry* 273 (25): 15458-15463.

McWilliam, H., Li, W., Uludag, M., Squizzato, S., Park, Y. M., Buso, N., ... & Lopez, R. 2013. "Analysis tool web services from the EMBL-EBI." *Nucleic acids research* 41 (W1): W597-W600.

Merz, Jr., Kenneth, M. 1990. "Insights into the function of the zinc hydroxide-thr199-glu106 hydrogen bonding network in carbonic anhydrases." *Journal of molecular biology* 214 (4): 799–802.

Merz, K. M. 1990. "Insights into the function of the zinc hydroxide-Thr199-Glu106 hydrogen bonding network in carbonic anhydrases." *Journal of molecular biology* 214 (4): 799-802.

Miller, S., Janin, J., Lesk, A. M., Chothia, C. 1987. "Interior and surface of monomeric proteins." *Journal of Molecular Biology* 196 (3): 641–656.

Miyata, T., & Yasunaga, T. 1980. "Molecular evolution of mRNA: a method for estimating evolutionary rates of synonymous and amino acid substitutions from homologous nucleotide sequences and its application." *Journal of Molecular Evolution* 16 (1): 23-36.

Mårtensson, L. G., Jonsson, B. H., Andersson, M., Kihlgren, A., Bergenhem, N., & Carlsson, U. 1992. "Role of an evolutionarily invariant serine for the stability of human carbonic anhydrase II." *Biochimica et Biophysica Acta (BBA)-Protein Structure and Molecular Enzymology* 1118 (2): 179–186.

Nair, S. K., & Christianson, D. W. 1991. "Unexpected pH-dependent conformation of His-64, the proton shuttle of carbonic anhydrase II." *Journal of the American Chemical Society* 113 (25): 9455–9458.

Nair, S. K., Calderone, T. L., Christianson, D. W., Fierke, C. A. 1991. "Altering the mouth of a hydrophobic pocket. Structure and kinetics of human carbonic anhydrase II

mutants at residue Val-121." *The Journal of biological chemistry* 266 (26): 17320-17325.

Nicholas, K. B., Nicholas, H. B. J., & Deerfield, D. W. 1997. "GeneDoc: Analysis and Visualization of Genetic Variation." *Embnet. news* 4: 14.

Ondrechen, M. J., Clifton, J. G., & Ringe, D. 2001. "THEMATICS: a simple computational predictor of enzyme function from structure." *Proceedings of the National Academy of Sciences* 98 (22): 12473-12478.

Patrikainen, M. 2012. *Pentraxin -Carbonic anhydrase CA VI: A novel multidomain protein.* MSc. Thesis, University of Tampere.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. 2004. "UCSF Chimera--a visualization system for exploratory research and analysis." *Journal of computational chemistry* 25 (13): 1605-1612.

Sankararaman, S., & Sjölander, K. 2008. "INTREPID—INformation-theoretic TREe traversal for Protein functional site IDentification." *Bioinformatics* 24 (21): 2445-2452.

Schwarz, F., & Aebi, M. 2011. "Mechanisms and principles of N-linked protein glycosylation." *Current opinion in structural biology* 21 (5): 576-582.

Silverman, D. N., & Lindskog, S. 1988. "The catalytic mechanism of carbonic anhydrase: implications of a rate-limiting protolysis of water." *Accounts of Chemical Research* 21 (1): 30-36.

Somarowthu, S., Yang, H., Hildebrand, D. G., & Ondrechen, M. J. 2011. "High-performance prediction of functional residues in proteins with machine learning and computed input features." *Biopolymers* 96 (6): 390-400.

Stern, A., Doron-Faigenboim, A., Erez, E., Martz, E., Bacharach, E., & Pupko, T. 2007. "Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach." *Nucleic acids research* 35: W506-W511.

Stewart, D. E., Sarkar, A., & Wampler, J. E. 1990. "Occurrence and role of cis peptide bonds in protein structures." *Journal of molecular biology* 214 (1): 253-260.

Suyama, M., Torrents, D., & Bork, P. 2006. “PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments.” *Nucleic acids research* 34 (suppl 2): W609-W612.

Tamai, S., Waheed, A., Cody, L.B. & Sly, W.S. 1996. "Gly-63-->Gln substitution adjacent to His-64 in rodent carbonic anhydrase IVs largely explains their reduced activity." *Proceedings of the National Academy of Sciences of the United States of America* 93 (24): 13647-13652.

Tan, K. P., Nguyen, T. B., Patel, S., Varadarajan, R., & Madhusudhan, M. S. 2013. "Depth: a web server to compute depth, cavity sizes, detect potential small-molecule ligand-binding cavities and predict the pKa of ionizable residues in proteins." *Nucleic acids research* gkt503.

Taylor, M. E., & Drickamer, K. 2006. *Introduction to glycobiology.* Oxford University Press, USA.

Tolvanen, M. E., Ortutay, C., Barker, H. R., Aspatwar, A., Patrikainen, M., & Parkkila, S. 2012. “Analysis of evolution of carbonic anhydrases IV and XV reveals a rich history of gene dupilcations and a new group of isozymes.” *Bioorganic & Medicinal Chemistry* 21 (6): 1503-1510.

Tu, C., Silverman, D. N., Forsman, C., Jonsson, B. H., & Lindskog, S. 1989. "Role of histidine 64 in the catalytic mechanism of human carbonic anhydrase II studied with a site-specific mutant." *Biochemistry* 28 (19): 7913-7918.

Turkoglu, S., Maresca, A., Alper, M., Kockar, F., Işık, S., Sinan, S., ... & Supuran, C. T. 2012. "Mutation of active site residues Asn67 to Ile, Gln92 to Val and Leu204 to Ser in human carbonic anhydrase II: influences on the catalytic activity and affinity for inhibitors." *Bioorganic & medicinal chemistry* 20 (7): 2208-2213.

Wei, Y., Ko, J., Murga, L. F., & Ondrechen, M. J. 2007. "Selective prediction of interaction sites in protein structures with THEMATICS." *BMC bioinformatics* 8 (1): 119.

Venkatachalam, C. M. 1968. "Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units." *Biopolymers* 6: 1425-1436.

Venta, P. J., Welty, R. J., Johnson, T. M., Sly, W. S., & Tashian, R. E. 1991. "Carbonic anhydrase II deficiency syndrome in a Belgian family is caused by a point mutation

at an invariant histidine residue (107 His----Tyr): complete structure of the normal human CA II gene." *American journal of human genetics* 49 (5): 1082-1090.

West, D., Kim, C. U., Tu, C., Robbins, A. H., Gruner, S. M., Silverman, D. N., & McKenna, R. 2012. "Structural and Kinetic Effects on Changes in the CO2 Binding Pocket of Human Carbonic Anhydrase II." *Biochemistry* 51 (45): 9156-9163.

Whittington, D. A., Waheed, A., Ulmasov, B., Shah, G. N., Grubb, J. H., Sly, W. S., & Christianson, D. W. 2001. "Crystal structure of the dimeric extracellular domain of human carbonic anhydrase XII, a bitopic membrane protein overexpressed in certain cancer tumor cells." *Proceedings of the National Academy of Sciences of the United States of America* 98 (17): 9545-9550.

Woolfson, D. N. & Williams, D. H. 1990. "The influence of proline residues in alpha helical structure." *FEBS Letters* 277: 185-188.

Xue, Y., Liljas, A., Jonsson, B. H., & Lindskog, S. 1993. "Structural analysis of the zinc hydroxide–thr-199–glu-106 hydrogen-bond network in human carbonic anhydrase ii." *Proteins: Structure, Function, and Bioinformatics* 17 (1): 93–106.

Zheng, J., Avvaru, B. S., Tu, C., McKenna, R., & Silverman, D. N. 2008. "Role of Hydrophilic Residues in Proton Transfer." *Biochemistry* 47: 12028-12036.

Appendix 1

The result of the K_a/K_s analysis and RSA values arranged in ascending order along with the manual alignment showing in the right most columns. 100% or perfectly conserved residues are noted with 'XX' and highly conserved residues having single exceptions are noted with 'X'. Conserved residues having exceptional roles are noted with '***'.

ENS_res	PDB_res	PDB_pos	RSA	LOC	k_a/k_s	Alignment	Exceptions/special roles
N	N	61	0	Buried	0.0055	XX	
T	T	199	0.04	Buried	0.0055	XX	
S	S	29	0	Buried	0.0059	XX	
H	H	96	0.01	Buried	0.0059	XX	
H	H	107	0	Buried	0.006	XX	
S	S	197	0	Buried	0.006	XX	
Q	Q	222	0.02	Buried	0.006	XX	
Q	Q	28	0.02	Buried	0.0061	XX	
H	H	94	0.12	Buried	0.0061	XX	
H	H	122	0	Buried	0.0061	XX	
Q	Q	249	0.21	Surface	0.0061	XX	
H	H	119	0.02	Buried	0.0062	XX	
A	A	142	0	Buried	0.0066	XX	
E	E	117	0	Buried	0.0076	XX	
E	E	106	0.01	Buried	0.0077	XX	
R	R	254	0.11	Buried	0.0083	XX	
R	R	246	0.01	Buried	0.0086	XX	
G	G	104	0	Buried	0.0091	XX	
P	P	30	0	Buried	0.0092	XX	
G	G	196	0	Buried	0.0093	XX	
Y	Y	194	0.03	Buried	0.0097	XX	
P	P	186	0.09	Buried	0.0099	XX	
P	P	201	0.08	Buried	0.01	XX	
L	L	44	0.18	Buried	0.011	XX	
V	V	207	0	Buried	0.016	X	CA3
A	A	134	0	Buried	0.017	X	CA4_mus_musculus
V	V	121	0.06	Buried	0.018	X	CA1_homo_sepiens
S	S	105	0	Buried	0.019	X	CA14_gallus_gallus
W	W	16	0.02	Buried	0.021	XX	
W	W	97	0	Buried	0.021	XX	
W	W	209	0.02	Buried	0.021	XX	
N	N	244	0	Buried	0.027	X	CA6
Q	Q	92	0.15	Buried	0.029	X	CA6_mus_musculus
T	T	200	0.21	Surface	0.029	***	Stabilize Water chain
L	L	198	0.18	Buried	0.03	X	CA3
G	G	63	0.18	Buried	0.031	X	CA4_mus_musculus
V	V	143	0.04	Buried	0.045	***	Hydrophobic pocket formation
H	H	64	0.28	Surface	0.046	***	Proton shuttle
L	L	164	0.04	Buried	0.053		
G	G	140	0	Buried	0.054		
F	F	70	0.05	Buried	0.056		
G	G	98	0	Buried	0.056		
N	N	124	0.01	Buried	0.057		
L	L	90	0.01	Buried	0.058		
L	L	203	0	Buried	0.059		
W	W	5	0.14	Buried	0.061		

ENS_res	PDB_res	PDB_pos	RSA	LOC	k_a/k_s	Alignment	Exceptions/special roles
V	V	31	0.01	Buried	0.063		
I	I	33	0	Buried	0.063		
A	A	23	0.07	Buried	0.066		
Y	Y	88	0.02	Buried	0.068		
Y	Y	191	0.01	Buried	0.068		
K	K	154	0.19	Buried	0.069		
S	S	259	0.22	Surface	0.069		
T	T	193	0.07	Buried	0.073		
L	L	184	0	Buried	0.074		
E	E	205	0.24	Surface	0.074		
V	V	211	0	Buried	0.075		
I	I	167	0	Buried	0.076		
L	L	185	0.19	Buried	0.077		
G	G	25	0.19	Buried	0.079		
F	F	66	0	Buried	0.079		
N	N	62	0.24	Surface	0.082		
Y	Y	7	0.15	Buried	0.086		
I	I	210	0	Buried	0.089		
I	I	256	0.08	Buried	0.089		
D	D	110	0.44	Surface	0.094		
V	V	68	0	Buried	0.096		
I	I	59	0	Buried	0.097		
D	D	41	0.24	Surface	0.098		
V	V	218	0.02	Buried	0.1		
G	G	145	0	Buried	0.11		
I	I	216	0.02	Buried	0.11		
L	L	251	0.38	Surface	0.11		
P	P	13	0.19	Buried	0.12		
T	T	108	0.06	Buried	0.12		
P	P	202	0.43	Surface	0.12		
L	L	141	0.01	Buried	0.13		
A	A	65	0.02	Buried	0.14		
K	K	111	0.75	Surface	0.14		
Y	Y	128	0.25	Surface	0.14		
W	W	192	0.04	Buried	0.14		
M	M	241	0	Buried	0.14		
P	P	46	0.84	Surface	0.15		
F	F	179	0.03	Buried	0.15		
R	R	227	0.12	Buried	0.15		
P	P	21	0.73	Surface	0.16		
D	D	32	0.18	Buried	0.16		
V	V	109	0.13	Buried	0.16		
A	A	116	0.01	Buried	0.16		
K	K	170	0.43	Surface	0.16		
L	L	212	0	Buried	0.16		
A	A	77	0.06	Buried	0.17		
R	R	89	0.17	Buried	0.17		
I	I	146	0	Buried	0.17		
K	K	149	0.5	Surface	0.17		
P	P	215	0.18	Buried	0.17		
F	F	226	0.01	Buried	0.17		
A	A	258	0.06	Buried	0.17		
G	G	6	0.13	Buried	0.18		
N	N	11	0.36	Surface	0.18		
T	T	35	0.31	Surface	0.18		
L	L	84	0.07	Buried	0.18		
A	A	115	0.04	Buried	0.18		
L	L	118	0	Buried	0.18		
L	L	120	0	Buried	0.18		

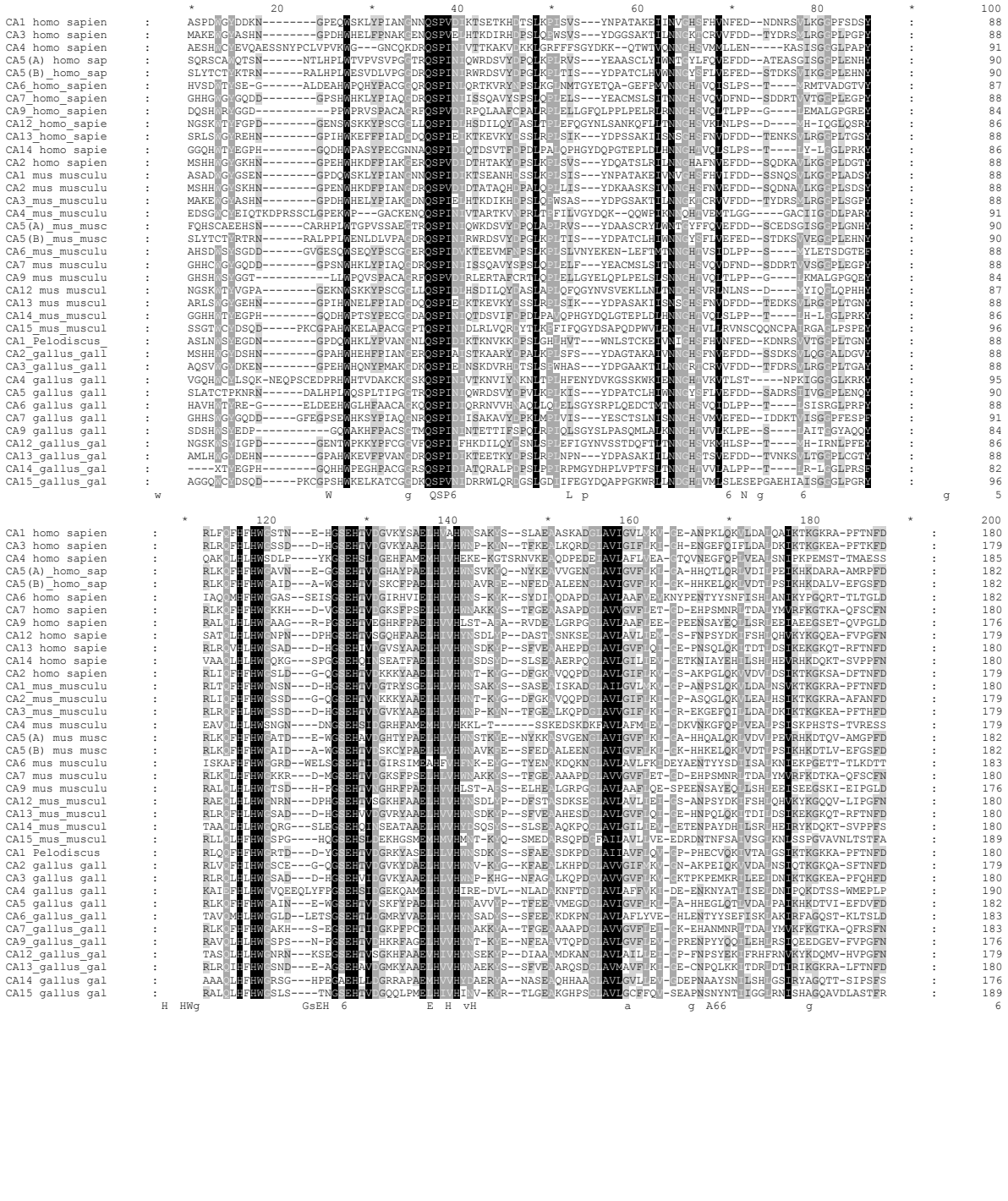
ENS_res	PDB_res	PDB_pos	RSA	LOC	k_a/k_s	Alignment	Exceptions/special roles
L	L	144	0	Buried	0.18		
F	F	147	0	Buried	0.18		
D	D	180	0.39	Surface	0.18		
C	C	206	0	Buried	0.18		
T	T	208	0.18	Buried	0.18		
F	F	260	0.2	Buried	0.18		
V	V	160	0	Buried	0.19		
S	S	219	0.08	Buried	0.19		
L	L	229	0.01	Buried	0.19		
E	E	236	0.52	Surface	0.19		
P	P	250	0.59	Surface	0.19		
K	K	252	0.44	Surface	0.19		
L	L	47	0.1	Buried	0.2		
G	G	82	0.08	Buried	0.2		
F	F	93	0	Buried	0.2		
F	F	95	0	Buried	0.2		
L	L	157	0	Buried	0.2		
G	G	171	0.69	Surface	0.2		
A	A	174	0.19	Buried	0.2		
D	D	190	0.35	Surface	0.2		
W	W	245	0.19	Buried	0.2		
G	G	12	0	Buried	0.21		
H	H	15	0.29	Surface	0.21		
D	D	75	0.55	Surface	0.21		
G	G	151	0.54	Surface	0.21		
F	F	176	0.1	Buried	0.21		
P	P	83	0.31	Surface	0.22		
D	D	139	0.33	Surface	0.22		
K	K	172	0.38	Surface	0.22		
D	D	130	0.5	Surface	0.23		
Q	Q	158	0.31	Surface	0.23		
V	V	161	0.03	Buried	0.23		
P	P	181	0	Buried	0.23		
V	V	223	0.01	Buried	0.23		
D	D	52	0.56	Surface	0.25		
D	D	162	0.58	Surface	0.25		
D	D	72	0.26	Surface	0.26		
V	V	78	0.09	Buried	0.26		
G	G	81	0.21	Surface	0.26		
D	D	243	0.26	Surface	0.26		
N	N	253	1.01	Surface	0.26		
I	I	22	0.33	Surface	0.27		
T	T	125	0.29	Surface	0.27		
K	K	133	0.53	Surface	0.27		
K	K	228	0.54	Surface	0.27		
E	E	234	0.75	Surface	0.27		
H	H	10	0.84	Surface	0.28		
A	A	38	0.16	Buried	0.28		
S	S	73	0.65	Surface	0.28		
S	S	99	0.33	Surface	0.28		
L	L	148	0.01	Buried	0.28		
N	N	230	0.03	Buried	0.28		
P	P	247	0.4	Surface	0.28		
L	L	57	0.48	Surface	0.29		
K	K	127	0.55	Surface	0.29		
S	S	166	0.46	Surface	0.29		
P	P	42	0.71	Surface	0.3		
V	V	150	0.33	Surface	0.3		
H	H	4	1.01	Surface	0.31		

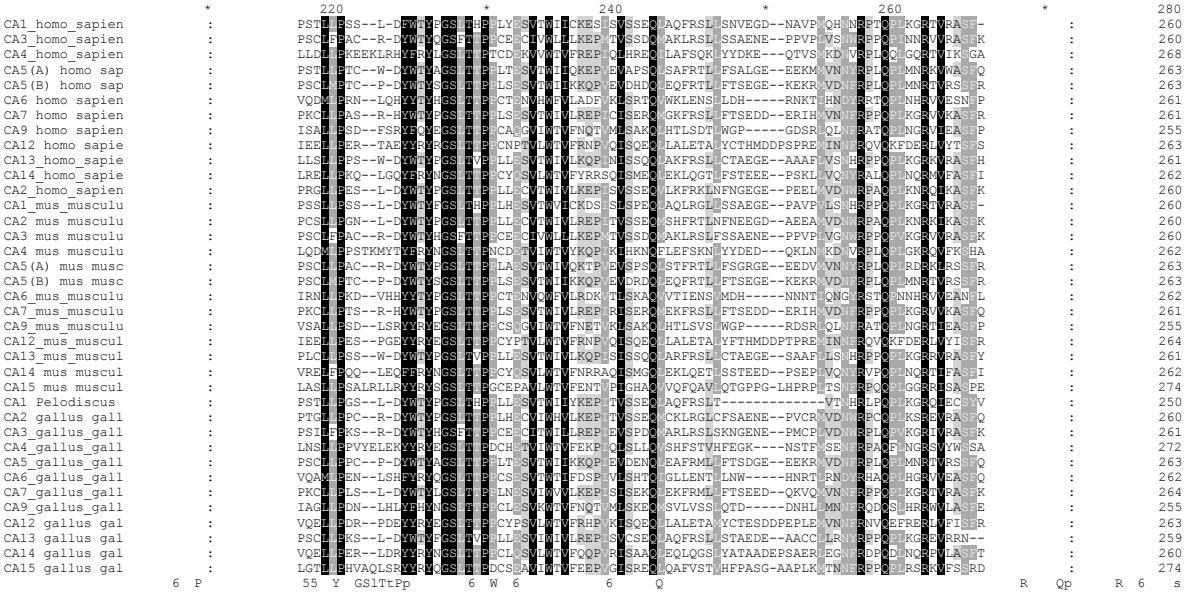
ENS_res	PDB_res	PDB_pos	RSA	LOC	k_a/k_s	Alignment	Exceptions/special roles
H	H	17	0.36	Surface	0.31		
K	K	24	0.67	Surface	0.31		
D	D	34	0.41	Surface	0.31		
W	W	123	0.04	Buried	0.31		
T	T	169	0.08	Buried	0.31		
E	E	238	0.59	Surface	0.31		
Y	Y	40	0.39	Surface	0.32		
K	K	168	0.45	Surface	0.32		
V	V	242	0.21	Surface	0.32		
V	V	49	0.23	Surface	0.33		
Y	Y	51	0.01	Buried	0.33		
S	S	56	0.01	Buried	0.33		
K	K	113	0.44	Surface	0.33		
K	K	159	0.56	Surface	0.33		
T	T	177	0.57	Surface	0.33		
E	E	214	0.31	Surface	0.33		
E	E	221	0.62	Surface	0.33		
S	S	188	0.22	Surface	0.34		
K	K	213	0.34	Surface	0.34		
K	K	257	0.2	Surface	0.34		
D	D	71	0.21	Surface	0.35		
N	N	232	0.03	Buried	0.35		
N	N	67	0.15	Buried	0.36		
Y	Y	114	0.14	Buried	0.36		
Q	Q	137	0.27	Surface	0.36		
V	V	163	0.18	Buried	0.37		
L	L	189	0.37	Surface	0.37		
N	N	178	0.85	Surface	0.38		
F	F	231	0.18	Buried	0.38		
A	A	248	0.43	Surface	0.38		
P	P	138	0.6	Surface	0.39		
T	T	55	0.32	Surface	0.4		
K	K	18	0.66	Surface	0.42		
F	F	20	0.28	Surface	0.42		
S	S	43	0.69	Surface	0.42		
D	D	101	0.47	Surface	0.42		
M	-	-	NA	NA	0.43		
K	K	39	0.64	Surface	0.43		
G	G	102	0.68	Surface	0.43		
A	A	153	0.51	Surface	0.43		
L	L	224	0.55	Surface	0.43		
E	E	69	0.24	Surface	0.44		
K	K	112	0.55	Surface	0.44		
F	F	131	0.33	Surface	0.44		
R	R	27	0.2	Buried	0.45		
S	-	-	NA	NA	0.46		
K	K	9	0.77	Surface	0.46		
D	D	85	0.83	Surface	0.46		
A	A	54	0.23	Surface	0.47		
V	V	135	0.14	Buried	0.47		
G	G	183	0.37	Surface	0.47		
H	H	36	0.93	Surface	0.48		
G	G	156	0.32	Surface	0.48		
T	T	37	0.8	Surface	0.49		
Q	Q	136	0.67	Surface	0.49		
P	P	155	0.83	Surface	0.49		
L	L	79	0	Buried	0.5		
I	I	91	0.21	Surface	0.5		
R	R	182	0.48	Surface	0.5		

ENS_res	PDB_res	PDB_pos	RSA	LOC	k_a/k_s	Alignment	Exceptions/special roles
G	G	86	0.56	Surface	0.51		
S	S	220	0.37	Surface	0.51		
G	G	235	1	Surface	0.52		
G	G	132	0.44	Surface	0.53		
E	E	26	0.63	Surface	0.54		
H	-	-	NA	NA	0.56		
K	K	225	0.4	Surface	0.56		
G	G	233	0.06	Buried	0.56		
Q	Q	53	0.64	Surface	0.58		
R	R	58	0.34	Surface	0.58		
D	D	165	0.76	Surface	0.58		
S	S	217	0.27	Surface	0.58		
Q	Q	103	0.26	Surface	0.59		
G	G	8	0.44	Surface	0.64		
K	K	45	0.66	Surface	0.65		
K	K	76	0.23	Surface	0.66		
P	P	195	0.39	Surface	0.67		
P	P	237	0.82	Surface	0.68		
K	K	261	1.09	Surface	0.69		
T	T	87	0.2	Surface	0.7		
S	S	152	0.71	Surface	0.7		
S	S	50	0.32	Surface	0.71		
L	L	60	0.2	Buried	0.72		
G	G	129	0.67	Surface	0.73		
L	L	204	0.33	Surface	0.74		
E	E	14	0.6	Surface	0.77		
S	S	48	0.31	Surface	0.78		
Q	Q	74	0.6	Surface	0.8		
E	E	239	0.57	Surface	0.8		
K	K	80	0.38	Surface	0.84		
S	S	173	0.42	Surface	0.85		
Q	Q	255	0.67	Surface	0.87		
E	E	187	0.93	Surface	0.9		
D	D	19	0.76	Surface	0.93		
L	L	100	0.65	Surface	0.95		
D	D	175	0.9	Surface	0.96		
L	L	240	0.52	Surface	0.98		

Appendix 2

MSA of the “universal group” containing human, mouse and chicken/turkey CA sequences.




CA1_homo_sapien
CA3_homo_sapien
CA4_homo_sapien
CA5(A) homo sap
CA5(B) homo sap
CA6 homo sapien
CA7 homo sapien
CA9 homo sapien
CA12 homo sapie
CA13_homo_sapie
CA14_homo_sapie
CA2_homo_sapien
CA1_mus_musculu
CA2 mus musculu
CA3 mus musculu
CA4 mus musculu
CA5(A) mus musc
CA5(B) mus musc
CA6_mus_musculu
CA7_mus_musculu
CA9_mus_musculu
CA12_mus_muscul
CA13_mus_muscul
CA14 mus muscul
CA15 mus muscul
CA1 Pelodiscus
CA2 gallus gall
CA3_gallus_gall
CA4_gallus_gall
CA5_gallus_gall
CA6_gallus_gall
CA7_gallus_gall
CA9_gallus_gall
CA12 gallus gal
CA13 gallus gal
CA14 gallus gal
CA15 gallus gal

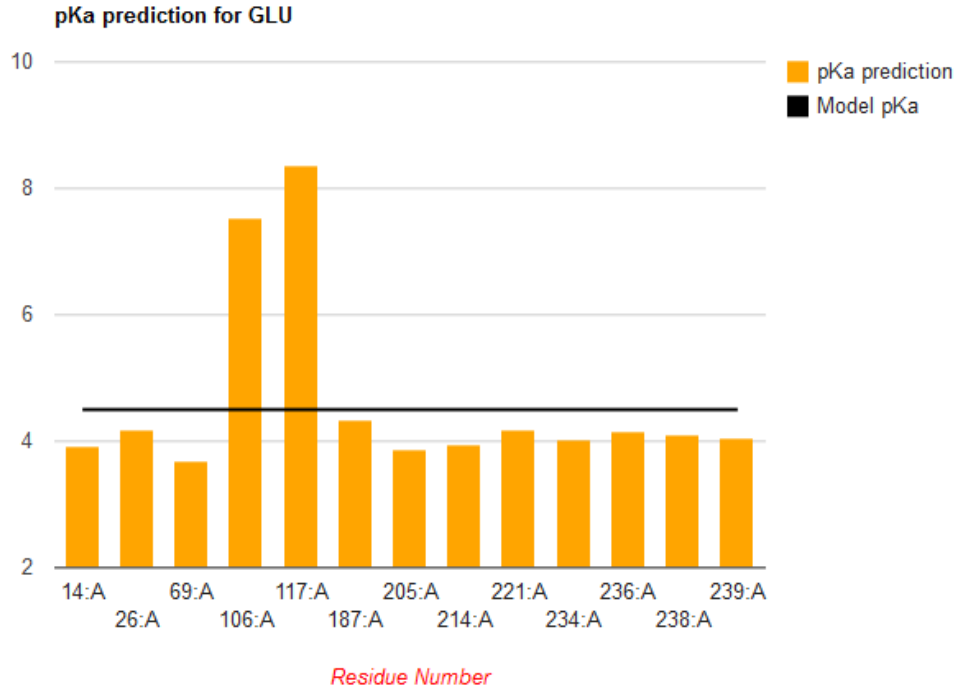
Appendix 3

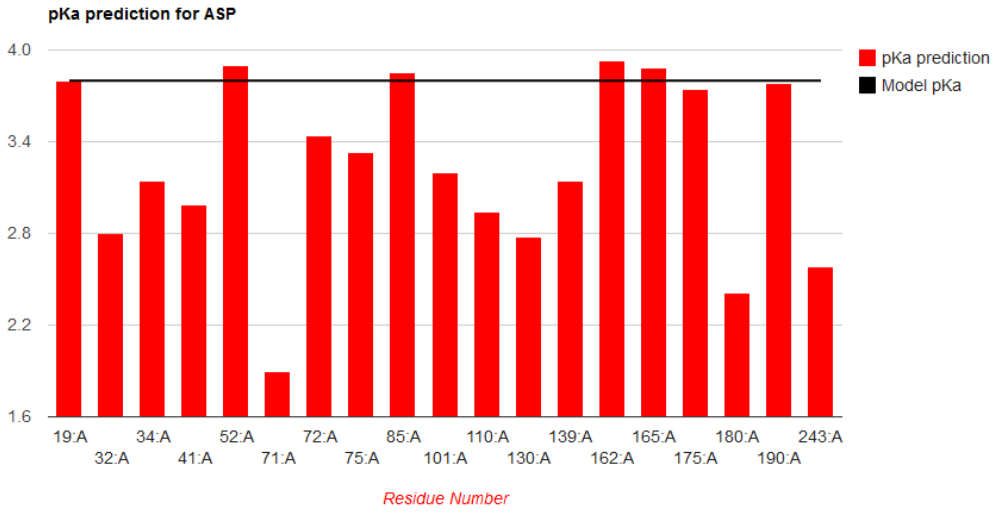
The POOL rank scores which denote the functionally important residues of a given protein structure identified by a machine learning approach. The POOL server is located at http://www.pool.neu.edu/wPOOL/. It has been seen that all the highly conserved residues that ranked top positions in the K_a/K_s score were also found within the top 50 positions of the POOL rank among 257 residues in the given protein structure, PDB: 3KS3

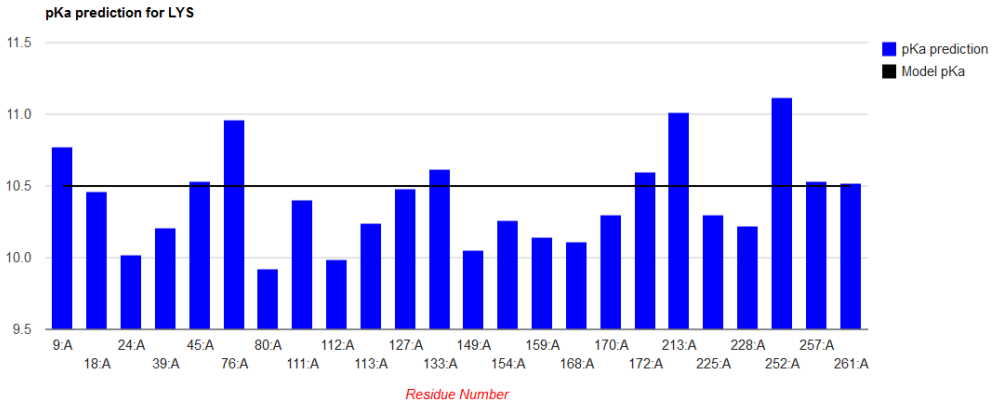
Rank	POOL score	Residue Number
1	0.006893208250403	HIS:A 94
2	0.006618639454246	HIS:A 119
3	0.001615040004253	HIS:A 96
4	0.000763847958297	GLU:A 106
5	0.000144612000440	GLU:A 117
6	0.000073695999163	HIS:A 107
7	0.000033696000173	ARG:A 246
8	0.000026879999496	TYR:A 7
9	0.000014742000531	HIS:A 64
10	0.000013860000763	TYR:A 194
11	0.000012149999748	THR:A 199
12	0.000009072000466	THR:A 200
13	0.000009072000466	LEU:A 198
14	0.000006480000593	VAL:A 121
15	0.000004752000223	PHE:A 95
16	0.000004049999916	ASN:A 67
17	0.000003888000265	VAL:A 143
18	0.000002964000032	ASP:A 32
19	0.000001890000021	GLN:A 92
20	0.000001700999746	TRP:A 209
21	0.000001248000103	LEU:A 118
22	0.000001214999998	ALA:A 65
23	0.000001200000042	PHE:A 93
24	0.000001200000042	LEU:A 120
25	0.000001120000093	CYS:A 206
26	0.000000864000128	PRO:A 30
27	0.000000864000128	GLY:A 145
28	0.000000850000106	ASP:A 243
29	0.000000800000066	ARG:A 254
30	0.000000720000116	HIS:A 122
31	0.000000480000040	SER:A 105
32	0.000000480000040	PHE:A 147
33	0.000000480000040	LEU:A 144
34	0.000000480000040	ILE:A 146
35	0.000000480000040	ALA:A 116
36	0.000000352000029	ASN:A 244
37	0.000000351000011	TYR:A 114
38	0.000000350000022	TYR:A 128
39	0.000000270000015	TRP:A 97
40	0.000000270000015	TRP:A 245
41	0.000000256000050	ASP:A 72
42	0.000000243000017	HIS:A 4
43	0.000000224000033	TYR:A 51
44	0.000000224000033	ARG:A 27
45	0.000000210000010	GLN:A 249
46	0.000000180000001	VAL:A 207
47	0.000000161999992	GLY:A 104
48	0.000000161999992	ALA:A 248
49	0.000000160000013	TYR:A 88
50	0.000000160000013	SER:A 29

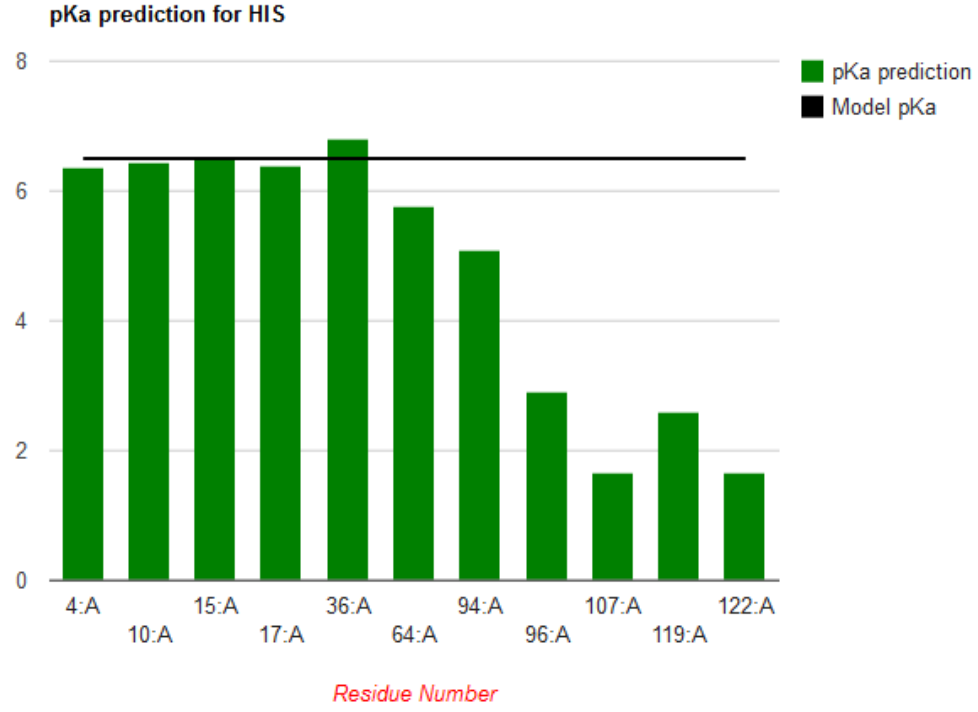
Appendix 4

The pKa values were predicted from DEPTH server (Tan 2013). It calculates the pKa values for acidic and basic amino acid residues from a given protein structure (PDB:3KS3). Each of the vertical bars in the following graphs are showing the predicted pKa values of the residues and the horizontal black lines are representing the standard pKa value of the residues.



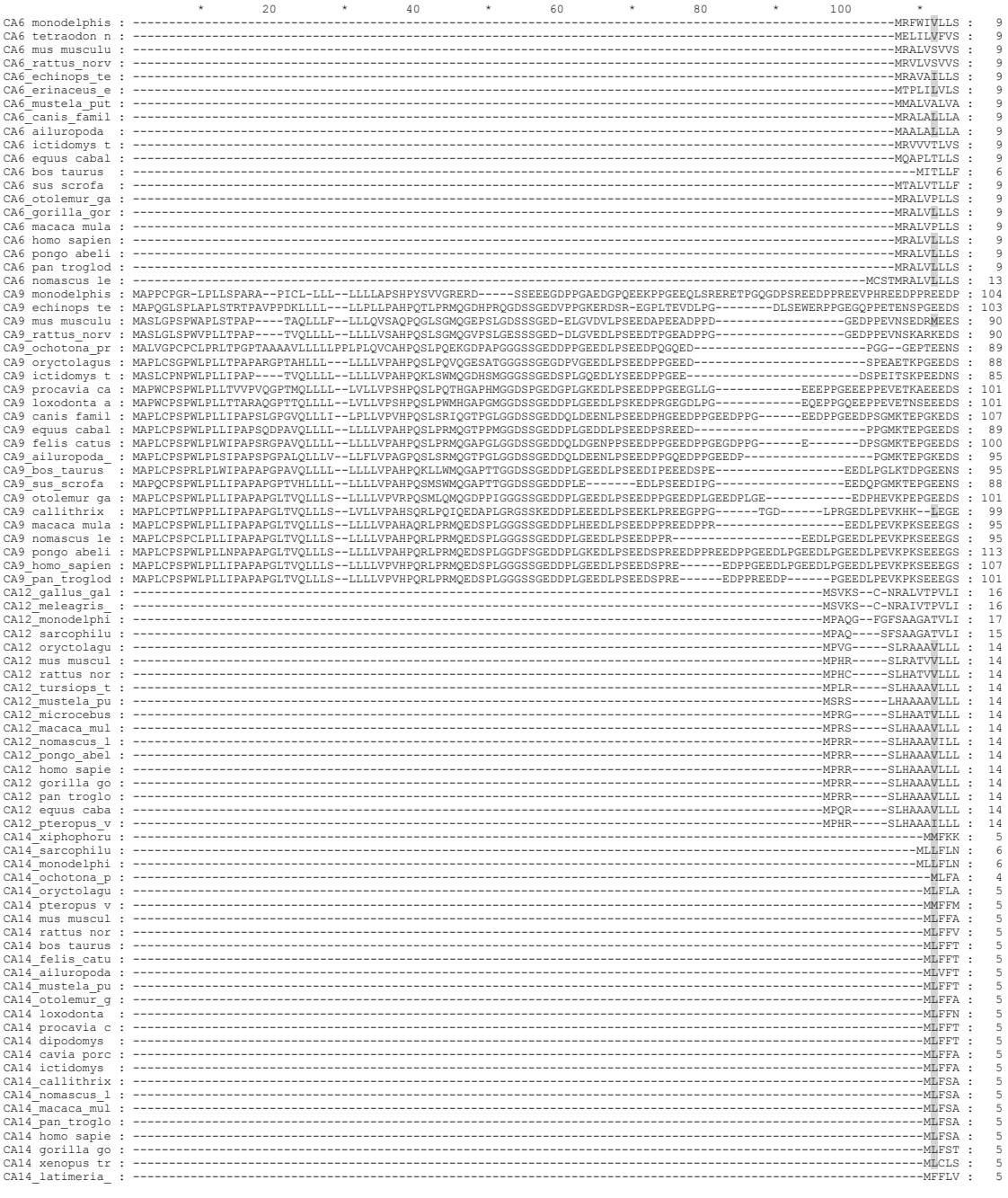


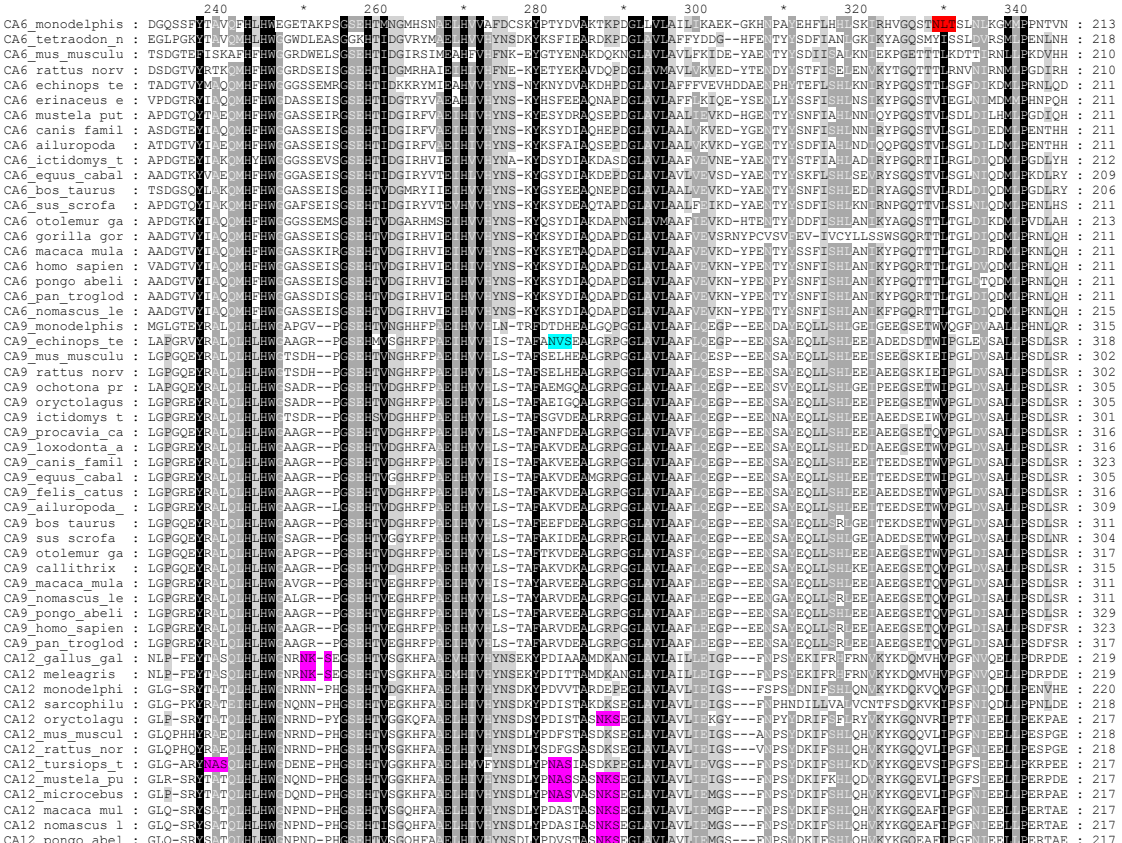

pKa prediction for LYS
11.5
11.0
10.5
10.0
9.5
pKa prediction
Model pKa
9:A
18:A
24:A
39:A
45:A
76:A
80:A
111:A
112:A
113:A
127:A
133:A
149:A
154:A
159:A
168:A
170:A
172:A
213:A
225:A
228:A
252:A
257:A
261:A
Residue Number

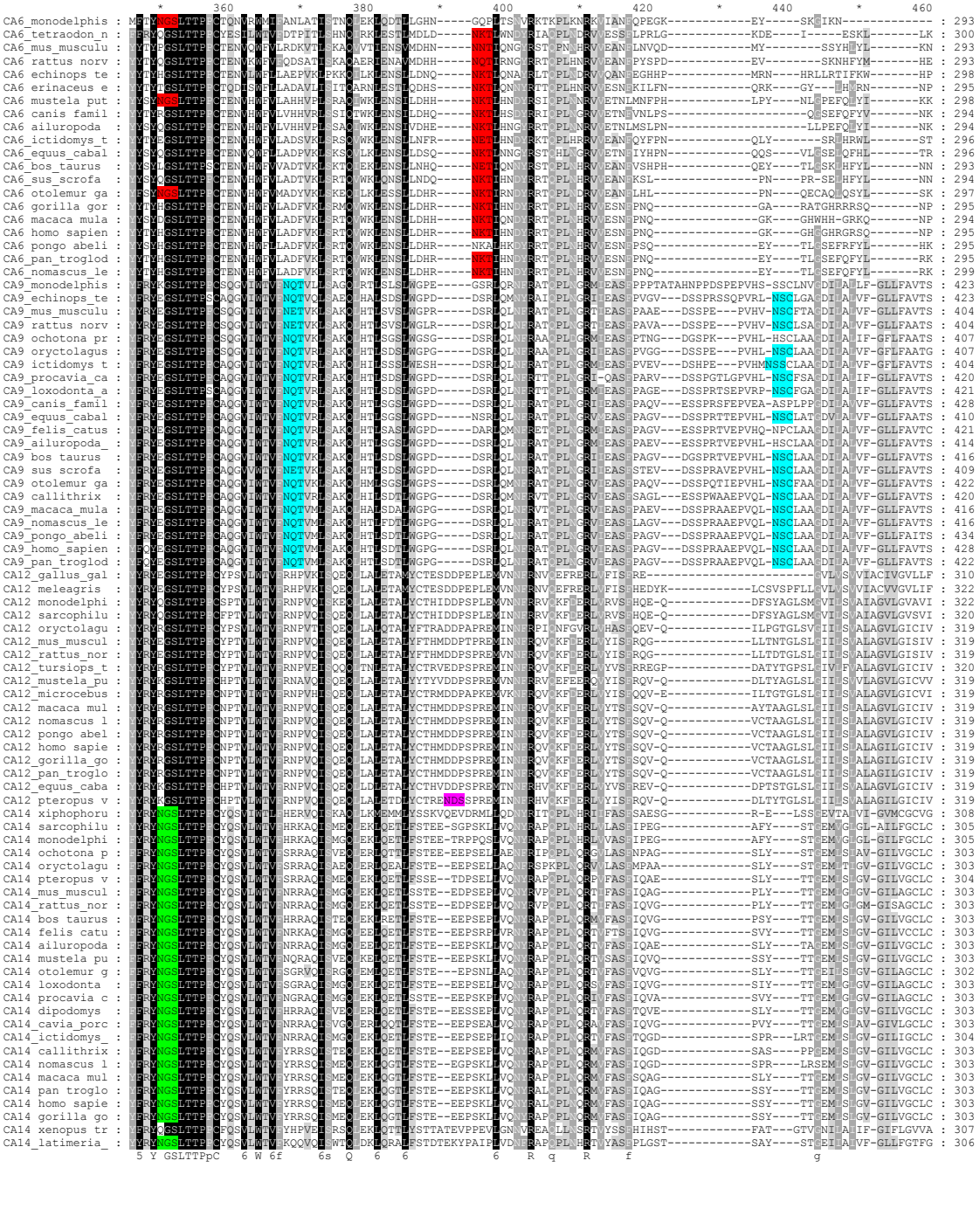

pKa prediction for HIS
8
6
4
2
0
pKa prediction
Model pKa
4:A
10:A
15:A
17:A
36:A
64:A
94:A
96:A
107:A
119:A
122:A
Residue Number

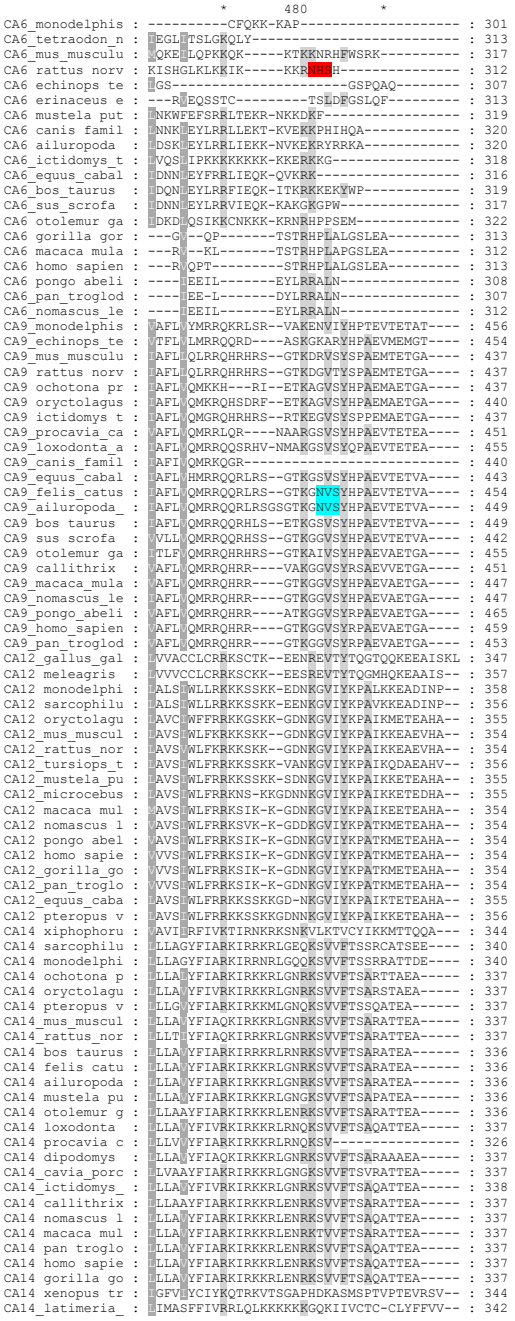
Appendix 5

The MSA of CA-VI, CA-IX, CA-XII, and CA-XIV with identified N-linked glycosylation sites with colors, red for CA-VI, cyan for CA-IX, purple for CA-XII and green for CA-XIV.








* 480 *
CA6_monodelphis : ----------CFQKK-KAP---------- : 301
CA6_tetraodon_n : EGLTSLGKQLY---------- : 313
CA6_mus_musculu : QKELLQPKKQK-----KTKKNRHFWSRK---------- : 317
CA6 rattus norv : KISHGLKLKKIK-----KKRNHSH---------- : 312
CA6 echinops te : IGS----------GSPQAQ-------- : 307
CA6 erinaceus e : ---RLEQSSTC---------TSLDFGSLQF--------- : 313
CA6 mustela put : NKWFEFSRRLTEKR-NKKDKF---------- : 319
CA6 canis famil : NNKEYLRRLLEKT-KVEKKPHIHQA---------- : 320
CA6 ailuropoda : DSKEYLRRLIEKK-NVKEKRYRRKA---------- : 320
CA6_ictidomys_t : VQSIPKKKKKKK-KKERKKG---------- : 318
CA6_equus_cabal : DNNEYFRRLIEQK-QVKRK---------- : 316
CA6_bos_taurus : DQNEYLRRFIEQK-ITKRKKEKYWP---------- : 319
CA6_sus_scrofa : DNNEYLRRVIEQK-KAKGKGPW---------- : 317
CA6 otolemur ga : DKDQSIKKCNKKK-KRNRHPPSEM---------- : 322
CA6 gorilla gor : ---GLE--QP-------TSTRHPLALGSLEA--------- : 313
CA6 macaca mula : ---RLE--KL-------TSTRHPLAPGSLEA--------- : 312
CA6 homo sapien : ---RLQPT---------STRHPLALGSLEA--------- : 313
CA6 pongo abeli : -----EEIL-------EYLRRALN---------- : 308
CA6_pan_troglod : -----EE-L-------DYLRRALN---------- : 307
CA6_nomascus_le : -----EEIL-------EYLRRALN---------- : 312
CA9_monodelphis : VAFLYMRRQKRLSR--VAKENVIYHPTEVTETAT---- : 456
CA9_echinops_te : VTFLLMRRQQRD----ASKGKARYHPAEVMEMGT---- : 454
CA9_mus_musculu : HAFLQLRRQHRHRS--GTKDRVSYSPAEMTETGA---- : 437
CA9 rattus norv : HAFLQLRRQHRHRS--GTKDGVTYSPAEMTETGA---- : 437
CA9 ochotona pr : HAFLQMKKH---RI--ETKAGVSYHPAEMAETGA---- : 437
CA9 oryctolagus : HAFLQMKRQHSDRF--ETKAGVSYHPAEMAETGA---- : 440
CA9 ictidomys t : HAFLQMGRQHRHRS--RTKEGVSYSPPEMAETGA---- : 437
CA9_procavia_ca : VAFLQMRRLQR----NAARGSVSYHPAEVTETEA---- : 451
CA9_loxodonta_a : HAFLQMRRQQSRHV-NMAKGSVSYQPAEVTETEA---- : 455
CA9_canis_famil : HAFIQMRKQGR---------- : 440
CA9_equus_cabal : HAFLHMRRQQRLRS--GTKGNVSYHPAEVTETVA---- : 443
CA9_felis_catus : HAFLQMRRQQRLRS--GTKGNVSYHPAEVTETVA---- : 454
CA9_ailuropoda_ : HAFLQMRRQQRLRSGSGTKGNVSYHPAEVTETVA---- : 449
CA9 bos taurus : HAFLQMRRQQRHLS--ETKGSVSYHPAEVTETVA---- : 449
CA9 sus scrofa : VLLLQMRRQQRHSS--GTKGGVSYHPAEVTETVA---- : 442
CA9 otolemur ga : HTLFQMRRQHRHRS--GTKAIVSYHPAEVAETGA---- : 455
CA9 callithrix : VAFLQMRRQHRR----VAKGGVSYRSAEVVETGA---- : 451
CA9_macaca_mula : VAFLQMRRQHRR----GTKGGVSYHPAEVAETGA---- : 447
CA9_nomascus_le : HAFLQMRRQHRR----GTKGGVSYHPAEVAETGA---- : 447
CA9_pongo_abeli : VAFLQMRRQHRR----ATKGGVSYRPAEVAETGA---- : 465
CA9_homo_sapien : VAFLQMRRQHRR----GTKGGVSYRPAEVAETGA---- : 459
CA9_pan_troglod : VAFLQMRRQHRR----GTKGGVSYRPAEVAETGA---- : 453
CA12_gallus_gal : VVACCLCRRKSCTK--EENREVTYTQGTQQKEEAISKL : 347
CA12 meleagris : VVVCCLCRRKSCKK--EESREVTYTQGMHQKEAAIS-- : 357
CA12 monodelphi : HALSWLLRKKKSSKK-EDNKGVIYKPALKKEADINP-- : 358
CA12 sarcophilu : HALSWLLRKKSSKK-EENKGVIYKPAVKKEADINP-- : 356
CA12 oryctolagu : HAVCWFFRKKGSKK-GDNKGVIYKPAIKMETEAHA-- : 355
CA12_mus_muscul : HAVSWLFKRKKSKK--GDNKGVIYKPAIKKEAEVHA-- : 354
CA12_rattus_nor : HAVSWLFKRKKSKK--GDNKGVIYKPAIKKEAEVHA-- : 354
CA12_tursiops_t : HAVSWLFRRKKSSKK-VANKGVIYKPAIKQDAEAHV-- : 356
CA12_mustela_pu : HAVSWLFRRKKSSKK-SDNKGVIYKPAIKKETEAHA-- : 355
CA12_microcebus : HAVSWLFRRKNS-KKGDNNKGVIYKPAIKKETEDHA-- : 355
CA12 macaca mul : VAVSWLFRRKSIK-K-GDNKGVIYKPAIKEETEAHA-- : 354
CA12 nomascus l : VAVSWLFRRKSVK-K-GDDKGVIYKPATKMETEAHA-- : 354
CA12 pongo abel : VAVSWLFRRKSIK-K-GDNKGVIYKPATKMETEAHA-- : 354
CA12 homo sapie : VVVSWLFRRKSIK-K-GDNKGVIYKPATKMETEAHA-- : 354
CA12_gorilla_go : VVVSWLFRRKSIK-K-GDNKGVIYKPATKMETEAHA-- : 354
CA12_pan_troglo : VVVSWLFRRKSIK-K-GDNKGVIYKPATKMETEAHA-- : 354
CA12_equus_caba : HAVSWLFRRKKSSKKGD-NKGVIYKPAIKTETEAHA-- : 355
CA12 pteropus v : HAVSWLFRRKKSSKKGDNNKGVIYKPAIKKETEAHA-- : 356
CA14 xiphophoru : VAVIRFIVKTIRNKRKSNKVLKTVCYIKKMTTQQA--- : 344
CA14 sarcophilu : HLLAGYFIARKIRRKRLGEQKSVVFTSSRCATSEE---- : 340
CA14 monodelphi : HLLAGYFIARKIRRNRLGQQKSVVFTSSRRATTDE---- : 340
CA14 ochotona p : HLLAYFIARKIRKKRLGNRKSVVFTSARTTAEA----- : 337
CA14 oryctolagu : HLLAYFIVRKIRKKRLGNRKSVVFTSARSTAEA----- : 337
CA14 pteropus v : HLLGYFIARKIRKKMLGNQKSVVFTSSQATEA------ : 337
CA14_mus_muscul : HLLAYFIAQKIRKKRLGNRKSVVFTSARATTEA----- : 337
CA14_rattus_nor : HLLTYFIAQKIRKKRLGNRKSVVFTSARATTEA----- : 337
CA14 bos taurus : HLLAYFIARKIRRKRLRNRKSVVFTSARATEA------ : 336
CA14 felis catu : HLLAYFIARKIRKKRLGNRKSVVFTSARATEA------ : 336
CA14 ailuropoda : HLLAYFIARKIRKKRLGNRKSVVFTSARATEA------ : 336
CA14 mustela pu : HLLAYFIARKIRKKRLGNGKSVVFTSAPATEA------ : 336
CA14 otolemur g : HLLAAYFIARKIRKKRLENRKSVVFTSARATTEA----- : 336
CA14 loxodonta : HLLAYFIVRKIRKKRLRNQKSVVFTSAQATTEA----- : 337
CA14 procavia c : HLLVYFIARKIRKKRLRNQKSV---------------- : 326
CA14 dipodomys : HLLAYFIAQKIRKKRLGNRKSVVFTSARAAAEA----- : 337
CA14_cavia_porc : HLVAAYFIARKIRKKRLGNGKSVVFTSVRATTEA----- : 337
CA14_ictidomys_ : HLLAYFIVRKIRKKRLGNRKSVVFTSAQATTEA----- : 338
CA14 callithrix : HLLAAYFIARKIRKKRLENRKSVVFTSARATTEA----- : 337
CA14 nomascus l : HLLAYFIARKIRKKRLENRKSVVFTSAQATTEA----- : 337
CA14 macaca mul : HLLAYFIARKIRKKRLENRKTVVFTSARATTEA----- : 337
CA14 pan troglo : HLLAYFIARKIRKKRLENRKSVVFTSAQATTEA----- : 337
CA14 homo sapie : HLLAYFIARKIRKKRLENRKSVVFTSAQATTEA----- : 337
CA14 gorilla go : HLLAYFIARKIRKKRLENRKSVVFTSAQATTEA----- : 337
CA14 xenopus tr : HGFVLYCIYKQTRKVTSGAPHDKASMSPTVPTEVRSV-- : 344
CA14_latimeria_ : HIMASFFIVRRLQLKKKKKKGQKIIVCTC-CLYFFVV-- : 342