Homology Modeling and docking study of *Danio rerio* Carbonic Anhydrase VI - Pentraxin protein and bioinformatics analysis of extra-cellular CAs

Prajwol Manandhar

Master's Thesis of M.Sc. Bioinformatics

BioMediTech

University of Tampere

Finland

Acknowledgements

The two and half years I spent in Finland was the most valuable moment of my life. I would like to express my deepest gratitude to this wonderful nation for providing me with such a great opportunity to carry out my higher education in one of the best universities of the world. The quality of education I have acquired during this tenure have brought a self-confidence in me to aim high in order to attain greatest endeavors of my life in the future.

My sincere thanks to University of Tampere for granting me a summer stipend in 2013 and Professor Seppo Parkkila for warmly welcoming me to his research group to carry out the summer research, a journey which led me into a research world of bioinformatics, and later providing me the opportunity to conduct my master's thesis research in his group. And most of all, the greatest source of my success and inspiration is Dr. Martti Tolvanen who have guided me since the beginning of my studies first as our Program coordinator, Lecturer, then as my summer research Supervisor and finally my master's thesis Supervisor. He was also the most amazing Finnish friend I have had during my stay in Finland. His guidance and supervision have hugely helped me accomplish one of the important goals of my life. I think I am very much grateful to have been his student. Similarly, I express my genuine thanks to Professor Matti Nykter who reviewed my thesis and for always being flexible during some of my difficult circumstances. Lastly, I thank all the teachers, lecturers, staffs of the University of Tampere and Turku and my colleagues in the Tissue Biology research group because of whom I have been able to gain this level of education.

My friends who have made me feel like a home away from home in this foreign country are the most significant parts that ever happened to me here. Harlan, who has helped me in many of my hurdles be it studies related or any other matter, I think he is the most genius guy I have ever met. Nirmal, who have always been by my side, never let me feel that we are from different nations. His friendship is something that will bond me with my neighboring country for a much eternal time. Praveen, who is more of an elder brother to me, have always made me feel like a family in this land thousands of miles away from my home. And all other amazing friends I have met in Finland, I thank you for your wonderful friendship. While lastly and most prominently, my mom, my dad and my sister are my ever-lasting medium of encouragement and the reason for my success in life. Without their support, I would have never been able to be the person I am today, my utmost respect and love to my family.

Master's thesis

Place	University of Tampere Tissue Biology group, School of Medicine Institute of Biosciences and Medical Technology (BioMediTech)
Author	MANANDHAR, PRAJWOL
Title	Homology Modeling and docking study of *Danio rerio* Carbonic Anhydrase VI - Pentraxin protein and bioinformatics analysis of extra-cellular CAs
Pages	83
Supervisor	Dr. Martti Tolvanen
Reviewers	Professor Matti Nykter Dr. Martti Tolvanen
Date	August 2015

Abstract

Background and Aims

Computational prediction and protein structure modeling are the marvelous inventions of computer sciences that have come to the rescue of various biological problems. The technology has revolutionized the biological world of research and helped scientists and researchers to gain insights into their biological questions much efficiently to design experimental research. Carbonic anhydrase (CA) is ubiquitous enzyme existing in all living beings and most importantly serves in catalyzing the reversible reaction of carbon dioxide and bicarbonate interconversion. There are at least 16 different isozymic forms of CAs in higher vertebrates which are mainly categorized on the basis of their sub-cellular localizations, broadly extracellular and intracellular. And recently, certain sub-population of transmembrane isoform CA IX, which is an extracellular CA, has been reported to also exist in nucleus i.e. in the intracellular environment. Likewise, it had been discovered that CA VI, another extracellular isoform, of non-mammalian vertebrates have an additional novel domain related to Pentraxins.

The main goal of this research was to look for computational prediction of the nuclear-cytoplasmic signals in the sequences of all three transmembrane CAs: CA IX, CA XII and CA XIV. And, another goal was to model the complete structure of the complex of CA VI and Pentraxin domains of zebrafish *Danio rerio*. While additionally, some preliminary sequence analyses of the extracellular CAs and Pentraxin proteins were also targeted.

Methods

For the first goal, the orthologous sequences of all transmembrane CAs, CA VI and Pentraxin proteins CRP and SAP were retrieved from Ensembl database, and was addressed to analyses to identify some key features through certain bioinformatics tools. The nuclear localization signal was predicted from NucPred webserver tool while the nuclear export signal was predicted from NetNES webserver tool for

transmembrane CAs. While for other sequence analyses, sub-cellular localization prediction was done from TargetP webserver, transmembrane helix prediction was done from TMHMM webserver.

As for the second goal, the structures of both CA domain and Pentraxin domain of zebrafish was modeled first using homology modeling technique from their respective template structures analyzed from the PDB database. The homology modeling was done in MODELLER interface of Chimera visualization software. And subsequently, these two generated comparative models of each of the domains were docked together computationally using HADDOCK docking suite available in the webserver.

Results

Almost all analyzed transmembrane CA sequences were predicted to have N-terminal signal peptide, with few exception of some sequences that have missing N-terminal regions in their sequence reads. The NetNES webserver tool predicted the NES sequence motifs mostly in the starting region of the transmembrane helical domain of the transmembrane CAs. In addition, the NucPred webserver tool predicted NLS sequence motifs at the cytoplasmic domains of transmembrane CAs, right at the region where the transmembrane domain ends and the cytoplasmic domain starts. Most of the analyzed sequences of transmembrane CAs were predicted to have these nuclear-cytoplasmic signal motifs with just a few exceptions. Sequence analyses of transmembrane CAs revealed there were dimerization signal motifs in the transmembrane regions of CA XII and CA XIV that could drive the dimerization in the tertiary structure of the proteins. Moreover, there were two extra Cysteine residues conserved among the Pentraxin domain of non-mammalian CA VI which are not present in any of classical Pentraxin CRP and SAP.

The comparative models of zebrafish CA VI domain was generated using human CA VI structure as the template and its RMSD was calculated to be 0.254 Å with reference to the template structure. Similarly, the comparative models of zebrafish Pentraxin domain was generated using human SAP structure as the template and its RMSD was calculated to be 0.288 Å with reference to the template structure. Successively, these comparative models of each domain were computationally docked using HADDOCK webserver software, and a docked complex of complete model of zebrafish CA VI with Pentraxin was generated having Haddock score of -115.9 +/- 5.2 and Z-score of -2.5.

Conclusion

The transmembrane CAs are predicted to have NLS and NES sequence motifs in their transmembrane and cytoplasmic domains distinct to these isozyme groups of CAs, which could reflect on their secondary role in the nucleus apart from the normal CA role in extracellular region. Similarly, computational modeling and/or docking study could be very useful for generating models of such biomolecular complexes whose structure would be otherwise difficult to determine through experimental procedures. A good quality model of the zebrafish CA VI with Pentraxin domain was generated through computational modeling and docking procedures that could be useful for researchers for concluding various interpretations.

Abbreviations

AIR	Ambiguous Interaction Restraints
ANN	Artificial Neural Network
AP	Amphipathic
API	Application Programming Interface
BLAST	Basic Local Alignment Search Tool
CA	Carbonic Anhydrase
CARP	Carbonic Anhydrase Related Protein
CAS	Cellular Apoptosis Susceptibility Gene
CPORT	Consensus Prediction of Interface Residues in Transient complexes
CRP	C - reactive protein
CSP	Chemical Shift Perturbation
EBI	European Bioinformatics Institute
GPI	Glycosyl-phosphatidyl-inositol
GUI	Graphical User Interface
HMM	Hidden Markov Models
IC	Intracellular
MAV	Multi Align Viewer
MS	Mass Spectrometry
MSA	Multiple Sequence Alignment
NCBI	National Center for Biotechnology Information
NES	Nuclear Export Signal
NLS	Nuclear Localization Signal
NMR	Nuclear Magnetic Resonance
NPC	Nuclear Pore Complex
NPR	Neuronal Pentraxin Receptor
PDB	Protein Data Bank
PG	Proteoglycan
PRR	Pattern Recognition Receptor
PTX	Pentraxin
RCSB	The Research Collaboratory for Structural Bioinformatics
REST	Representational State Transfer
RMSD	Root Mean Square Deviation
SAP	Serum Amyloid P Component
TM	Transmembrane
UCSF	University of California, San Francisco

Table of Contents

1 Introduction

Carbonic anhydrases (CAs) are the enzymes that catalyze the reversible reactions involving hydration and dehydration of CO_2 and HCO_3^- respectively during active transport of CO_2 across the cells to eventually eliminate it from the body. These enzyme catalysts consist of a metal co-factor, mostly Zinc (Zn), in its active site that coordinate the dissociation of a proton from a water molecule during the reversible reaction. Such reaction would have been much slower without the presence of the enzyme (Lindskog and Coleman 1973). These enzymes have been invented convergently as well as divergently during the evolution of life on earth to be present in all the domains of life viz. Archaea, Bacteria, and Eukarya. Specifically, in higher animals such as vertebrates, the alpha gene family of CAs have been dominant throughout their evolution and the gene family is the most studied gene family.

Most of the studies on CAs have been based on the enzymes from higher organisms, while the prokaryotic CAs has a crucial role in shaping the ecology of earth's biosphere. The prokaryotic organisms from domains Bacteria and Archaea have key roles in earth's biogeochemical cycles, the CAs from these organisms help in procurement of CO_2 required for photosynthesis, while physiology of the prokaryotes help in decomposing the organic matter back to atmospheric CO_2 completing the global carbon cycle (Kumar and Ferry 2014). However, the alpha CAs present in the higher vertebrates, mostly being studied in mammals including humans, have their essential roles in the physiological functions in different cellular processes. There have been at least 16 isozyme forms of α-CAs identified so far in the vertebrates which have been classified further based on their sub-cellular localization (Hilvo et al. 2005; Supuran 2008).

Among the isozymes of the α-CAs, this thesis research is mostly based on the various bioinformatics studies related to a group of extracellularly localized isozymes which are CA VI, CA IX, CA XII and CAXIV. Of these extracellular isozymes as well, the last three are categorized as transmembrane-bound CAs while the former CA VI is the only secreted form of all CAs. The transmembrane isozyme CA IX with its general function in acid-base balance, intercellular communication, and cell proliferation, has been associated with most cancers. A 1998 study by *Saarnio et al* confirmed the unusual expression of CA IX in the areas with a high proliferative activity of colorectal tumor cells by immunohistochemical method (Saarnio et al. 1998). Further such studies done in various types of cancers have found out CA IX to be of great interest among all of the CAs in regard to their associations with cancers or tumors. Further studies that followed later showed that the CA9 gene expression and CA IX enzyme activity highly relates to regulating extracellular acidic pH and helping cancer cells in progression or metastasis, mostly under the hypoxic conditions of tumor cells (Ivanov et al. 2001; Robertson, Potter, and Harris 2004; Svastova et al. 2004; Thiry et al. 2006; Swietach, Vaughan-Jones, and Harris 2007). Another recent study found out the first evidence of CA IX interacting with the proteins of nuclear/cytoplasmic transport machinery in an interactome characterization study in hypoxic cells, a completely new finding for any alpha CAs (Buanne et al. 2013). This study suggested the existence of nuclear subpopulations of CA IX with its possible intracellular functions, distinct from their well-known role in the cell membrane.

These novel findings related to CA IX inspired us to research further on this with various bioinformatics approaches available to us. Hence, we did the assessment of sequence analysis for identifying any clues that would give insight about these proteins to be targeted to the nucleus using various prediction methods, which are discussed in detail later in this thesis. One of the previous studies also showed the overexpression of CA12 gene in cells under hypoxic conditions contributing in tumor microenvironment by sustaining extracellular acidic pH, and the cancer cells to grow and spread (Ivanov et al. 2001). Similar

to the analysis for CA IX, other two transmembrane CAs, CA XII and CA XIV, were also addressed to the sequence analysis for the predictions of nuclear/cytoplasmic transport.

Of the membrane-bound isozymes and all the α-CAs, the only one that has been characterized to be existing in secreted form is CA VI, which has been identified to be present in saliva and milk secretions (Henkin et al. 1975; Thatcher et al. 1998; Karhumaa et al. 2001). Its physiological function has been associated with growth-supporting role in taste buds, while as found to be one of the elementary factors in mammalian milk suggests its essential role in normal growth and development of the alimentary canal in infants (Karhumaa et al. 2001). Some preliminary observations during the course of Maarit Patrikainen's thesis in our research group discovered a different peptide sequence, found to be a Pentraxin, attached to the Carboxyl-terminal of the CA VI of certain species (Patrikainen 2012). Pentraxins are distinct families of protein which mainly consists of short Pentraxins and long Pentraxins, usually characterized by the presence of a 200 residue long Pentraxin domain in their Carboxyl-terminal with an 8 amino acid conserved Pentraxin signature, HxCxS/TWxS, where x is any amino acid residue (Garlanda et al. 2005). Short Pentraxins comprises of C-reactive protein (CRP) and Serum amyloid P component (SAP), while long Pentraxins include PTX3, neuronal Pentraxin 1 (NP1), neuronal Pentraxin 2 (NP2), neuronal Pentraxin receptor (NPR) and PTX4. The main structural difference was the presence of an amino-terminal domain in long Pentraxins coupled to the Pentraxin domains, which is not present in CRP and SAP (Garlanda et al. 2005). The novel type of Pentraxin domain discovered in the CA VI enzymes of non-mammalian vertebrates coupled to their carboxyl-terminal was found to be phylogenetically closely related to short Pentraxins [Tolvanen M., unpublished observation].

To study about the Pentraxin containing CA VI, a Zebrafish CA VI protein structure model was proposed which is to be achieved through various bioinformatics procedures. The Protein Data Bank (PDB) has in its database the x-ray structure model of Human CA VI, a closest homologous protein to the Zebrafish CA VI, and a couple of both short and long Pentraxin proteins. This second main focus of the study was designed to be accomplished using Homology Modelling followed by Protein-protein Docking approaches. The finalized model could give better insight into the idea of how this novel CA VI domain would serve to a potential new role of CA VI proteins in those groups of vertebrates.

2 Aims of the study

The aims of this research are to investigate more on the transmembrane CAs, to find out pieces of evidence about their localization into nucleus which is a new topic in any CAs so far. Various bioinformatics prediction methods are used to perform sequence analysis of these transmembrane CAs in multiple species in order to predict nuclear localization signal (NLS) and nuclear export signal (NES) in transmembrane CAs. Additionally few other sequence analyses are also performed for certain purposes.

The modeling part of this research aims to model a complete structure for Zebrafish CA VI with its Pentraxin domain attached at the carboxyl-terminus. Each of the CA VI catalytic and the Pentraxin domains is to be modeled separately by Homology Modelling method, using homologous template structures from the PDB database. These models are then to be addressed to protein-protein docking method for generating a complete CA VI with Pentraxin structure, from which it may be possible to propose an insight how Pentraxin domain might assist the non-mammalian CA VI in associating with the cell membrane or with other biomolecules.

3 Review of literature

3.1 Nuclear-cytoplasmic transport mechanism

Eukaryotic cells consist of a separate nuclear compartment that is separated from the cytoplasmic environment with a double-layered membrane called nuclear envelope. Nucleus which houses the genetic material often has to transport its transcription products and other macromolecules into the ribosomes in cytoplasm for further processing while different proteins such as transcription factors, DNA and RNA polymerases, histones that are synthesized in the cytoplasm require an active transport into the nucleus. These mechanisms of nucleocytoplasmic transport of different macromolecules of molecular weight larger than ~40 kDa are carried out by family of proteins called as importins and exportins (Koepp and Silver 1998; Moroianu 1998; Chook and Blobel 2001; Goldfarb et al. 2004; Poon and Jans 2005; Kutay and Guttinger 2005). Importins are involved in actively transporting the cargo molecules from the cytoplasm into the nucleus, while exportins perform the transport from nucleus to the cytoplasm. These proteins specifically recognize signal sequences in their to-be cargo molecules following a metabolic process mediated by a small RAs-related Nuclear protein (Ran) or GTP-binding nuclear protein in order to transport the molecules actively. The proteins that need to be transported into the nucleus possess a nuclear localization signal (NLS), which act as a tag for importins. Likewise, those molecules requiring the transport from nucleus to cytoplasm possess a nuclear export signal (NES) which the exportins would recognize and thus bind with.

3.1.1 Importins and exportins

In classical nucleocytoplasmic transport pathway of macromolecules, importin-α forms a ternary complex with importin-β1 which then binds to NLS sequence in the cargo protein that is to be carried into the nucleus. These protein complexes after entering into the nucleus through nuclear pore complex (NPC), RanGTP binds with it which triggers the dissociation of the complex releasing the cargo protein, to ensure the active import of the cargo mediated by the energy dissociated from RanGTP in the form of GTP. Importin-α, after dissociation, is then recycled back to cytoplasm in another complex with an importin-α re-exporter called cellular apoptosis susceptibility gene (CAS) again in the presence of RanGTP (Koepp and Silver 1998; Lange et al. 2007). Alternatively, importin-β1 domain alone can also bind with some cargo proteins by recognizing the NLS sequence within them. The importin-β1 is recycled back to the cytoplasm in a complex with RanGTP (Okada et al. 2008).

As for the nuclear export, the cargo proteins possessing NES are bound by exportin-1 (XPO1), stimulated by RanGTP, which are exported into the cytoplasm also through NPC. In the cytoplasm, the hydrolysis of RanGTP to RanGDP occurs which is catalyzed by Ran GTPase-activating protein. This promotes the dissociation of the complex assembly and thus the cargo protein is released. And again, the XPO1 is recycled back to nucleus by binding with an NPC component called Nup358 (Kutay and Guttinger 2005).

3.1.2 NLS and NES

A nuclear localization signal (NLS) is a stretch of an amino-acid sequence tag present in certain proteins that are targeted to the cell nucleus through nucleocytoplasmic transport. A typical NLS sequence consists of one or more stretches of positively charged basic amino acids, usually lysines or arginines, exposed on the protein surface that are recognized by importins. The best-characterized NLS are the classical NLS that are further classified as monopartite or bipartite. Monopartite are such which have one stretch of basic amino acids such as PKKKRKV in the SV40 Large T-antigen (first NLS to be discovered) and EEKRKR in NF-κB p65 (Poon and Jans 2005). Bipartite signals usually have two clusters of basic amino acids, such as the

NLS of nucleoplasmin, KR[PAATKKAGQA]KKKK, has two basic amino acids clusters separated by a spacer of 10 amino acids (Dingwall et al. 1988). Both of these types of cNLSs are recognized by importin-α while some cNLSs are directly recognized by importin-β1 as typified by the sequence RKKRRQRRR in HIV-1 Tat (Truant and Cullen 1999). One of such is also importin-α which contains a bipartite NLS itself and hence is specifically recognized by importin-β.

Non-classical NLSs do not have basic amino acids clusters, and they bind directly to different importin-β homologues (Chook and Blobel 2001). Such signals in heterogeneous nuclear ribonucleoprotein A1 and other proteins is directly recognized by importin-β2/transportin-1/karyopherin-β2 (Lee et al. 2006). Additionally, importin-independent nuclear entry systems also exist, such as viral protein R (Vpr) of HIV-1 and β-catenin are known to directly interact with NPC components before passing through it (Jenkins et al. 1998; Yokoya et al. 1999).

Likewise, nuclear export signal (NES) is short amino acid sequence of hydrophobic residues which has an opposite function to that of the NLS, i.e. it targets the protein for export from the cell nucleus out into the cytoplasm through the NPC. The NES on the protein surface is recognized and bound by the exportins that transport the cargo actively. These signals recognized by exportins usually have short sequences stretch with several clusters of hydrophobic amino acids (often leucine), exemplified as RFLSLEPL and TPTDVRDVDI in cyclin D and LQKKLEELEL in mitogen-activated protein kinase (Poon and Jans 2005; Kutay and Guttinger 2005). The occurrence of the hydrophobic residues (L or D) with certain spacing may be explained by evaluating the protein structures which contain an NES, these crucial residues would usually orient at the same face of the adjacent secondary structures that they are associated to, enabling them to interact notably with the exportins (la Cour et al. 2004). RNA, which is synthesized in the nucleus, has to be exported into the cytoplasm but as it is composed of nucleotides and hence lacks NES, so most RNAs bind with protein to form ribonucleoprotein complex before getting exported to the cytoplasm.

3.1.3 Carbonic anhydrase aspect

The alpha gene family of carbonic anhydrases are classified into several isozymes classes mainly based on their sub-cellular localization, such as cytoplasmic, mitochondrial, secreted, transmembrane CAs. For decades of CA research, it has never been known about the functional role of any α-CAs in the cell nucleus, although there have been several suspicion about the same in some several experiments. It is still a mystery about the possible functionality of any CAs in the nucleus, however, it is also not unexpected of the existence of an enzyme in the nucleus with CA activity.

3.2 Alpha Carbonic Anhydrases

With no any sequence or structural similarity but having similar active site confirmation and no doubt the function, there have been three major inventions of different families of CAs, viz. Alpha, Beta and Gamma. While more expansive classification also includes two additional minor families, viz. Delta and Zeta. The previously thought separate family of CAs, Epsilon, was later found out to be a special type included within Beta CA family. The β-CAs occur in most prokaryotes like bacteria, phototrophic organisms such as plants, and fungi (Hewett-Emmett and Tashian 1996). Likewise, the CAs from archaea and eubacteria are identified as γ-CAs, later also discovered in mitochondria of plants (Alber and Ferry 1994; Parisi et al. 2004; Smith et al. 1999). And the δ- and ζ-classes which have cadmium as the metal co-factor in their active sites have been discovered in marine phytoplankton and diatoms respectively (McGinn and Morel 2008; Xu et al. 2008). The α-CAs predominantly occur in higher eukaryotes from arthropods to all groups of vertebrates, but also been reported in some prokaryotes.

The α-CAs from mammalian species have been studied to greater extent so far than any other classes of CAs, there have been at least 16 different isoforms (CA I - CA Va, CA Vb - CA XV) of α-CAs identified and characterized in mammals (Hilvo et al. 2005; Supuran 2008). The maintenance of acid-base homeostasis in a living system is essential for the proper functioning of various metabolic reactions in the body. These metalloenzymes play a great role in regulating this balance in different cells and tissues of the body by catalyzing the reaction of reversible hydration of carbon dioxide and bicarbonate ions and maintaining the pH homeostasis. Different isoforms of the enzyme are expressed differentially in several groups of tissues of the body and are mainly grouped based on the specific sub-cellular localization. The broad groupings include mainly intracellular and extracellular forms, while more specifically in intracellular ones, cytosolic CAs are the group that include some of the first characterized CA isozymes CA I, II, III, VII and XIII, the latter two being discovered much recently than the rest which was in 70s. The other intracellular group includes the two mitochondrial localized isoforms CA Va and CA Vb. Likewise among the extracellular groups, CA VI is the only isoform to exist in secreted form in secretions such as saliva, milk. While, the membrane-associated forms include the isozymes CA IV, IX, XII, XIV and XV. Here, the CA IV and CA XV associate with the plasma membrane through a glycosyl-phosphatidyl-inositol (GPI) linkage, while the remaining isoforms CA IX, XII and XIV are transmembrane proteins. And lastly, the three remaining isoforms are often called CA-Related Proteins (CARPs) which are CARP VIII, X and XI. These isoforms are inactive in terms of CA catalytic activity due to the substitution of some key residues involved directly in the active site of the CA enzymes.

The cytosolic CAs form the largest group consisting of five isozymes distributed in various compartments at the intracellular environment. The CA1, CA2, CA3 and CA13 genes are located in the same chromosome 8 in humans while CA7 gene is in a different chromosome. And moreover, the former four genes share highest sequence identity with each other than with any other isoforms as well as a phylogenetic analysis shows a cluster of these four proteins together while CA VII lying more distantly with them than the mitochondrial isoforms (Barker 2013). CA II is among the most widely studied isozymes and there are much more crystal structures of CA II than any other CAs in the PDB repositories. The deficiency in CA II has often been highly linked with a syndrome called as Osteopetrosis with renal acidosis and cerebral calcification (Borthwick et al. 2003; Sly, Sato, and Zhu 1991). The disease is an autosomal recessive disorder, caused due to several different loss-of-function mutations in the CA2 gene. In a single study by direct sequencing method, Shah et al have identified eleven novel mutations in patients with the CA II deficiency syndrome and the mutations were found to be scattered over the exons of CA2 gene (Shah et

al. 2004), whereas there have been previously twelve different mutations identified as well in several studies (Venta et al. 1991; Roth et al. 1992; Hu et al. 1992; Hu, Waheed, and Sly 1995; Soda et al. 1995; Soda et al. 1996; Hu et al. 1997).

Likewise, the expression of CA3 gene is highly tissue-specific, found to be differentially expressed in Type-I muscle fibers in human skeletal muscle tissue (Shima et al. 1983) and hence often called as muscle-specific CA. Patients with Myasthenia gravis, a neuromuscular disease, were found to have specifically an insufficient level of CA III in skeletal muscles (Du et al. 2009). While patients with progressive muscular dystrophy conditions have significantly elevated level of CA3 than the normal ones, specifically in Duchene muscular dystrophy (Mokuno et al. 1985; Carter et al. 1983). And similarly, autoantibodies to CA3 were detected to be markedly higher in Rheumatoid arthritis patients (Liu et al. 2012). Studies such as these are indications that CA III might be a useful marker for muscle-related diseases. In a recent de novo whole-genome sequencing study of Amur tiger (*Panthera tigris altaica*) along with comparative analyses of genomic sequences of other Panthera-lineage felines (big-cats), various genetic signatures reflecting the specific molecular adaptions to big-cats' hypercarnivorous diet and muscle strength were reported (Cho et al. 2013). The study identified various tiger genes evolving under positive selection which provided the evidences of rapid evolution of genes (MYH7, TPM4, TNNC2, MYO1A, ACTN4) that were involved in development of muscle contraction and actin cytoskeleton. Here, in CA III sequence comparison of tiger, cat, dog, giant panda, polar bear, human, mouse and opossum, six unique substitutions were found in tiger sequence among which two seem to be meaningful ones. The variations V217R (hydrophobic aa to hydrophilic aa) and D220L (hydrophilic aa to hydrophobic aa) in tiger sequence with reference to all other sequences (including Cat) might also have some potential significant roles in functional changes of CA III activity in Pantherinae sub-lineage of Felidae (excluding Cat which is from Felinae sub-lineage) which could have added to the distinct muscle strength evolution in big-cats.

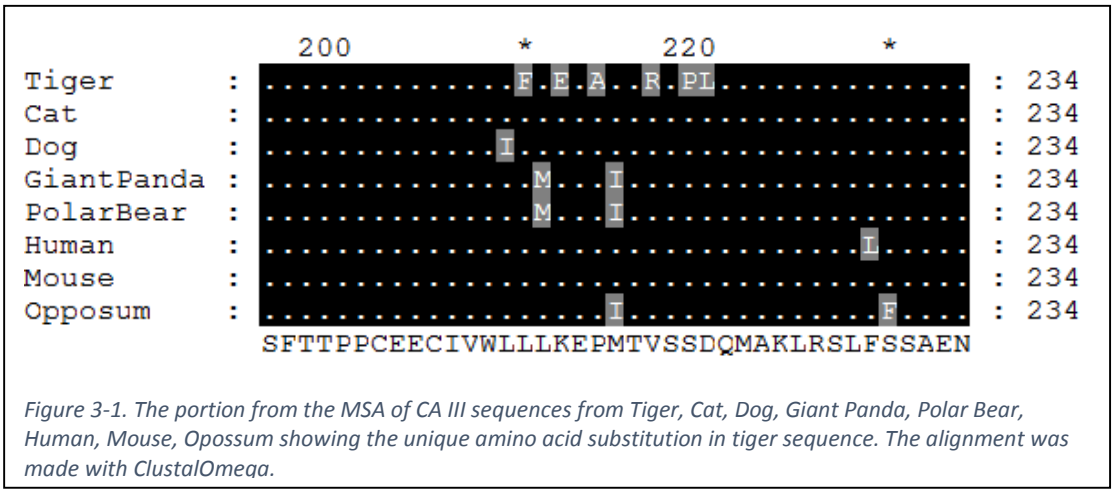


Figure 3-1. The portion from the MSA of CA III sequences from Tiger, Cat, Dog, Giant Panda, Polar Bear, Human, Mouse, Opossum showing the unique amino acid substitution in tiger sequence. The alignment was made with ClustalOmega.

The two mitochondrial CA homologues Va and Vb show highest sequence similarity among each other, however, the genes encoding the proteins are located in two different chromosomes. The CA 5a gene maps to chromosome 16 while CA 5b gene maps to chromosome X in humans. Both of the homologues, CA Va and Vb, possess a leader sequence which localizes them to mitochondria of the cell (Fujikawa-Adachi et al. 1999a). Despite their sequence similarity and same localization, CA Vb has broader tissue

distribution than CA Va which is confined mostly to the liver, skeletal muscle and kidney. And moreover, phylogenetic analysis estimates the two homologues of the CA V in mammals had diverged from a single ancestral gene around 90 million years ago (Shah et al. 2000), and since then, the mammalian CA Vb has been evolving much more slowly than CA Va. The differences in tissue-specific distribution, chromosomal location and variable evolutionary constraints among the two homologues also suggest that they have evolved to acquire different physiological roles.

One group of extracellular CAs include the GPI-anchored CAs which are bound to plasma membrane peripherally. Glycosyl-phosphatidyl-inositol (GPI) is a glycolipid that gets attached to the C-terminus of a protein during post-translational modification, thus the protein originally consists of a C-terminus signal peptide targeting it to the Endoplasmic reticulum (ER) which is then cleaved off, and the carboxyl group of the new terminal amino acid residue of the protein is anchored with amino group of ethanolamine residue of GPI precursor, which then gets transported to the cellular membrane via Golgi apparatus as a lipid rafts, and reside at the exterior leaflet of the membrane (Ikezawa 2002). The CA IV and CA XV are bound to the cellular membrane via GPI-anchor and typically appear on the apical membrane (Zhu and Sly 1990; Hilvo et al. 2005). CA XV is the youngest member in mammalian α-CA family which was characterized and investigated during database searches by (Hilvo et al. 2005), and most probably the final addition to the family, as no any other CA-like homologues were found in the database search. An interesting thing about this isoform is that it was detected in most of the mammalian genomes except for humans and chimpanzees, where it exist as a mere pseudogene that does not have any function and is rather never expressed. The phylogenetic analysis estimated that the CA XV is closely related to CA IV (Hilvo et al. 2005). In the same study, inspection of a low resolution Rhesus macaque (*Macaca mulatta*) genome also provided sufficient hints that it has also become pseudogene in the macaque, suggesting that the orthologues of CA15 gene in primates might have lost the function during early evolution itself (Hilvo et al. 2005). Additionally, further investigation on evolutionary analyses of these isozymes by a co-author of the previous study found out another novel GPI-linked isoform, CA XVII, in vertebrates while it has been lost in mammals (Tolvanen et al. 2013). Another property of these three GPI-linked isozymes is that they consist of multiple N-linked glycosylation sites.

CARPs are the group of inactive isozymes that does not have essential catalytic activity of CA enzymes, but however they have a potential alternative physiological function in the body. Each of the CARP isozymes (CARP VIII, X and XI) possess either one or more substitution of the three Histidine (His94, His96, His119) residues in its active site which co-ordinate the Zinc atom. Most of the CARP isozymes have been shown to have wide expression profiles in and around different tissues of the brain in humans and mice (Fujikawa-Adachi et al. 1999b; Taniuchi et al. 2002). The distinct expression profiles of CARPs in human and mouse brain have suggested its important functions in the development of the brain and nervous system (Taniuchi et al. 2002). These were made evident by some studies, where an Iraqi family with mild mental retardation, quadrupedal gait and ataxia were found to possess a defect in their CA8 gene (Turkmen et al. 2009). Another earlier experimental study on waddles mice showed that the CARP-VIII deficiency was associated with a distinctive lifelong gait disorder (Jiao et al. 2005). The sequences of each CARP isozymes were found to be highly conserved among each of the respective orthologues, the identities percentage was higher in all CARPs than in any of the other active CAs (Aspatwar, Tolvanen, and Parkkila 2010). The fact that the CARP sequences are very well conserved throughout many vertebrate taxa also suggests that their biological role have a definite significance during the evolution, despite losing the CA activity.

3.3 Transmembrane CAs

Transmembrane proteins are integral membrane proteins that span the entirety of the biological membrane as oppose to the GPI-linked proteins which reside peripherally at the extracellular half of the lipid bilayer membrane as mentioned earlier. Transmembrane proteins can have extracellular and intracellular domains along linked by the membrane-spanning domain which allows the firm attachment of the protein to the cell membrane aided by a special class of membrane lipids called annular lipid shell. The structures of the transmembrane domains are basically of two types: alpha-helical and beta-barrels. About 1/3rd of all the proteins in humans have been estimated to be alpha-helical membrane proteins (Almen et al. 2009), and nevertheless, the transmembrane CAs also possess an alpha-helical and C-

Feature key	Position(s)	Length	Description	Graphical view
Signal peptide	1 – 37	37	1 Publication (N-terminal Signal peptide)	
Topological domain	38 – 414	377	Extracellular (Proteoglycan + Catalytic CA IX)	
Transmembrane	415 – 435	21	Helical Sequence Analysis (alpha helical)	
Topological domain	436 – 459	24	Cytoplasmic	

Table 3-1. Table depicting sequence topology of human CA IX protein derived from Uniprot (http://www.uniprot.org/uniprot/Q16790), modified by Prajwol Manandhar.

terminal transmembrane domain with extracellular CA catalytic domain and intracellular cytoplasmic domain. Transmembrane CAs are the second largest groups of active α-CAs after cytoplasmic CAs comprising of three isozymes CA IX, XII and XIV. The sequence topology of these proteins consists of ~15-37 amino acid N-terminal signal-peptide, then main CA catalytic domain of ~275-377 amino acid which resides outside of the cell, ~22 amino acid transmembrane domain nearby C-terminus, and finally a small cytoplasmic domain of ~24-32 amino acid residues (Table 3-1).

3.3.1 Carbonic anhydrase IX

The first transmembrane CA to be identified was CA IX, which was rather recognized initially as a novel tumor-associated antigen named as MN (Pastorekova et al. 1992), subsequently later whose cDNA cloning revealed a large CA-like domain in the sequence (Pastorek et al. 1994), and finally was characterized by sequence analysis in 1996 as the ninth addition to the alpha CA family, named as CA IX (Opavsky et al. 1996). The transmembrane CA IX is a glycoprotein (Pastorekova et al. 1992), as it comprises of a distinct proteoglycan domain in the N-terminus which is closely related to the keratan sulfate binding domain of a large aggregating proteoglycan aggrecan (Doege et al. 1991), then the main CA domain, followed by a transmembrane helix and short intra-cytoplasmic tail (Opavsky et al. 1996). It also possesses a signal peptide in its N-terminus, while it is the only CA isozyme to possess such proteoglycan domain.

The N-terminal region of the protein is found to possess similarity with helix-loop-helix (HLH) family of DNA binding proteins, and moreover, DNA-cellulose chromatography experiment determined the protein to have affinity for binding DNA (Pastorek et al. 1994). In the earlier study by the same group, it is mentioned that the MN protein (CA IX) has two peptides of 54 kDa and 58 kDa molecular mass and

localized on the cell membrane in addition to the nucleus too (Pastorekova et al. 1992). Further, in the radioimmunoassay of MN-specific antibodies, the protein was visualized particularly in nucleoli of the nucleus (Zavada et al. 1993). Similarly, immunoreactivity of MN-protein in cervix carcinomas with glandular differentiation was found to be localized to some nuclei of neoplastic cells, the study was however focused on pathogenic and prognostic significance of MN-protein as cancer-biomarker (Costa, Ndoye, and Trelford 1995). The role of CA9 gene and its protein product as an important cancer biomarker has been always of a great interest to researchers since the beginning of its discovery, but the faint hints of its possible roles in nucleus seem to have always been overlooked. Similarly, another immunohistochemical study of a cancer-type under hypoxic condition has found expression of CA IX in perinuclear location in 46 patients and determined to associate with poor prognosis, while 3 patients among them also had nuclear CA IX expression (Swinson et al. 2003). Likewise, relatively with these findings, a nuclear protein with CA activity was determined in several rat tissues. The polypeptide of apparent 66 kDa mass was recognized by CA II antibodies itself and later determined by sequence analysis to be nonO/p54 which is an RNA and DNA binding transcription factor. The polypeptide was found to bind with CA inhibitor and have detectable CA activity (25 units/mg), higher than previously determined for CA III and CA Va. The transcriptional factor was denoted as non-classical CA, considering its CA activity might function in the maintenance of pH homeostasis in the nucleus (Karhumaa et al. 2000). Contemplating these interesting findings of DNA binding property, nuclear localization occurrences mostly under the influence of tumorigenesis, prognostic variable of perinuclear appearances of the CA IX and observation of a nuclear factor with CA activity, its plausible to speculate that CA IX could have a function in the nucleus and even might act as a transcription factor inducing cancer progression or cell proliferation.

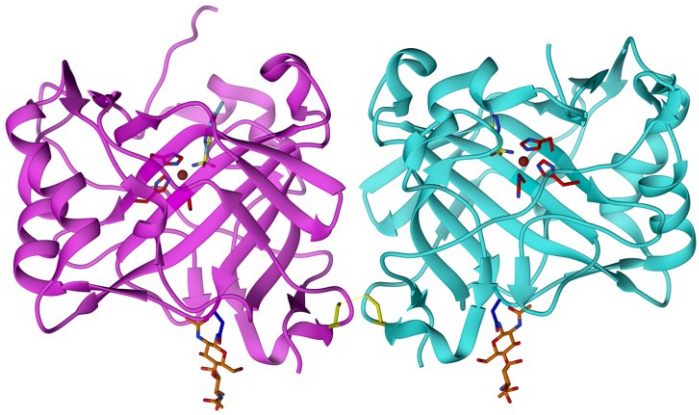


Figure 3-2: The dimer of CA IX structure (pdb: 3IAI), the two chains are shown in magenta and cyan, showing the active site Histidines (red) and Zinc (brown), disulfide linked Cysteines (yellow). The glucosamine (orange) sugar is shown to be attached at the bottom of two subunits linked with Arginines (blue).

The x-ray crystallographic structure of the catalytic domain of human CA IX has been resolved with a resolution of 2.20 Å and R-value 0.157 in complex with a classical sulfonamide CA inhibitor acetazolamide. The crystal structure unveils typical alpha-CA folds, which, however, differs significantly from other isozymes when the quaternary structure of the enzyme is considered (Alterio et al. 2009). The oligomerization and stability of the enzyme had been previously investigated too, where recombinant proteins were found in dimeric forms stabilized by intermolecular disulfide bond(s). The recombinant proteins were produced in baculovirus system in two forms of either catalytic domain only (CA form) or proteoglycan and catalytic domains (PG + CA form) (Hilvo et al. 2008). The PG domains and active site pockets of the dimeric enzyme are located on its one face, while the C-termini where transmembrane regions anchor the protein to cell membrane are located on the opposite face. The PDB structure 3IAI consists of mutation in Cys-41/Ser which is involved in an interchain disulfide bond. Hence, the Ser-41 residues were replaced with suitable rotamers of Cys residues in UCSF Chimera for the

depiction of disulfide linkage (Figure 3-2). The mass spectrometry experiments of the extracellular portions (PG + CA domains) of the CA IX recombinants in murine cell line expression system demonstrated unique N-linked (Asn-309) and additional O-linked (Thr-78) glycosylation sites, while the nature of oligosaccharides were also characterized (Hilvo et al. 2008; Alterio et al. 2009). The resolved structure provides an important suggestion for the CA IX specific inhibitor drug design, provided that the inhibition of the isozyme could aid in antitumor activity.

3.3.2 Carbonic anhydrase XII

Another transmembrane isozyme, CA XII, was characterized just a few years after the first transmembrane CA IX in two independent studies (Tureci et al. 1998; Ivanov et al. 1998). Similar to CA9, the expression of CA12 has also been found to be associated with tumor mainly induced by hypoxia but to a lesser extent (Watson et al. 2003). The human CA XII protein is a 354 amino acid polypeptide coded by the CA12 gene located at chromosome 15 and the protein sequence consists of 29 amino acid signal peptide, 261 amino acid CA catalytic domain, a short extracellular juxtamembrane segment, followed by 26 amino acid transmembrane helix and a 29 amino acid cytoplasmic tail. The molecular weight of the protein expressed in COS-7 cells was reported as 43-44 kDa and is reduced to 39 kDa upon PNGase treatment which was consistent with removal of two oligosaccharide chains indicating the protein has two N-linked glycosylation sites (Tureci et al. 1998). An overall structure of CA XII is broadly similar to that of CA IX. The crystal structure of CA XII is found to exist in the dimeric form which was also elucidated from its electrophoresis profile where the mature form of the enzyme in solution had molecular mass of 60 kDa suggesting its dimeric organization (Whittington et al. 2001). There is a single disulfide linkage between Cys-23 and Cys-203, the similar pairs are also conserved in CA IX and CA IV and it helps in stabilizing Pro-201-Thr-202 cis-peptide linkage and anchoring the loop containing Thr-199 (Stams et al. 1996). Unlike CA IX, the CA XII sequence does not possess extra Cys-41 responsible for dimer stabilization in CA IX. Rather, the sequence analysis of CA XII earlier have revealed that its transmembrane segment consists of the signature motifs GxxxG and GxxxS which serve as the framework for dimerization of transmembrane helices in transmembrane proteins (Senes, Gerstein, and Engelman 2000; Russ and Engelman 2000). The crystallization study have speculated that the presence of the signature motifs in the transmembrane segment of CA XII mediates dimerization which persists within the membrane in the full-length protein (Whittington et al. 2001). This might lead into stabilizing the dimer formation in the quaternary structure of CA XII.

3.3.3 Carbonic anhydrase XIV

The final transmembrane isozyme to be discovered was CA XIV that was characterized in 1999 (Mori et al. 1999). The study has found its broad expression in various tissues such as kidney, heart, brain, skeletal muscle and liver. The CA XIV is expressed in apical and basolateral membrane of hepatocytes in mouse liver (Parkkila et al. 2002), while its strong expression was also seen in neuronal membranes and axons in the human and mouse brain (Parkkila et al. 2001). The human CA14 gene located in chromosome 1 encodes for a polypeptide of 337-amino acids whose molecular mass was found to be 37.6 kDa. The topology of the protein sequence is similar to other two transmembrane CAs consisting of a 15-amino acid signal peptide, 275-amino acid extracellular catalytic domain, 21-amino acid transmembrane helix and a short 26-amino acid cytoplasmic tail. The crystal structure of the extracellular domain of human CA XIV was reported much later and recently than the other membrane associated alpha CAs. The structure is resolved with a resolution of 2.00 Å and the arrangement was found to be in a monomeric form unlike the two previous transmembrane CAs (Alterio et al. 2014). This was supported by bioinformatics as well

as gel filtration analysis. Similar to CA XII, the CA XIV structure also possess Cys-23 and Cys-203 disulfide pair as well as no Cys-41 that serves the disulfide link between two chains in CA IX.

3.4 Secreted CA

Based on the sub-cellular localization, there has been only one isoform among all the alpha-CA isozymes in vertebrates that exist as the secreted form, the CA VI.

3.4.1 Carbonic anhydrase VI

The CA VI was first characterized from the ovine parotid gland while investigating bicarbonate hydration in the parotid gland of the sheep (Fernley, Wright, and Coghlan 1979). However, it had been already isolated as zinc protein from parotid saliva by gel filtration and ion-exchange chromatography and due to its association to taste perception, it was named Gustin (Henkin et al. 1975). Although these two studies went on in parallel for almost two decades until it was finally discovered in 1998 as identical protein (Thatcher et al. 1998). The CA VI is known to be expressed exclusively in the serous acinar and ductal cells of the parotid, submandibular glands following its secretion into the saliva (Parkkila et al. 1994). The salivary enzyme was first purified from human saliva and characterized by (Murakami and Sly 1987), each molecule of the enzyme had two N-linked oligosaccharide chains which were found to be of complex type. A specific immunoflurometric and radioimmunoassays for human salivary CA VI was developed (Parkkila et al. 1993), which allowed accurate quantification of CA VI in saliva and serum. The application of the competitive time-resolved assay later revealed that the secretion of CA VI into saliva followed a circadian pattern i.e. its concentration being very low while sleeping and increasing rapidly to the daytime after awakening (Parkkila, Parkkila, and Rajaniemi 1995). Likely, the secretion of saliva which is controlled by the autonomous nervous system also follows the rhythms in circadian periodicity (Helm et al. 1982; Dawes 1972). Previous speculations on CA VI that it helps in regulating pH of saliva was disregarded, instead the salivary enzyme has been demonstrated to be localized in the dental pellicle, a protein film on the surface of enamel, on which the biofilm of bacterial plaque develops. The pellicle has the function of protecting teeth from continuous ions deposition from saliva and the acids produced by oral microbes. Hence, the CA VI located at the most favorable sites on the dental surface plays the role in catalyzing the salivary bicarbonate and microbe-delivered hydrogen ions to carbon dioxide and water (Leinonen et al. 1999). This speculation was supported by a study which found out that the lower salivary CA VI concentrations are associated with the prevalence of increased caries in teeth (Kivela et al. 1999). It was also suggested that CA VI provides protection in the esophageal and gastric epithelium from acid accumulation as symptoms of acid-peptic disease were observed in patients with lower concentration of CA VI in their saliva than the healthy subjects (Parkkila et al. 1997). Nevertheless, CA VI has been found as one of the elementary factors in mammary gland secretions, milk, of human and rat suggesting it is an essential factor in the normal growth and development of the infant alimentary tract (Karhumaa et al. 2001). Furthermore, the fact that Gustin/CA VI may contribute in the growth and development of taste buds or, in other words, taste sensation is also very intriguing (Henkin, Martin, and Agarwal 1999). Despite the studies for over three decades, investigations towards its exact function and physiological role still remain uncertain and invites huge obsession towards more research.

The cDNA of the gene encoding CA VI in humans was cloned and characterized by (Aldred et al. 1991) and it was mapped to chromosome 1. The isozyme's subunit molecular weight is 42 kDa while the molecule was found to have two complex type of N-linked oligosaccharide chains (Murakami and Sly 1987). Later, it was found to possess three potential N-linked glycosylation sites and two cysteine residues, Cys-25 and

Cys-207 (Aldred et al. 1991), which are also conserved in other isozymes already described earlier. There are small extensions of hydrophilic residues in the C-terminus of the CA VI (Jiang and Gupta 1999).

Bioinformatics analyses of CA VI orthologues have discovered a novel domain in the CA VI of certain vertebrate species. An unpublished observation of CA VI sequences of some species such as frog, fish and chicken found out a different type of domain attached to the C-terminus of the CA VI protein. A further investigation on this was done later with the availability of numerous genome sequences. In the thesis research by Patrikainen, sequence analysis of CA VI orthologues from multiple species was performed where it was found out that the novel domain is present on all the species except for mammals (Patrikainen 2012). This novel domain was found to be related to Pentraxin proteins, and so forth it was concluded that the secretory CA VI in non-mammalian vertebrates is a multi-domain protein. The sequence analysis also confirmed the presence of signal peptide in the secretory isozyme, and very highly conserved N-linked glycosylation sites in the analyzed orthologue sequences. Additional experiments were done to produce a construct of zebrafish CA VI protein in bacterial and insect cells and to observe the morphology of knockdown zebrafish model (Patrikainen 2012). The sequencing of the template DNA verified that it codes for the correct CA VI protein, while the knockdown zebrafish embryos and fry showed malformations of the swim bladder and the stomach area. With the interesting findings from the research, aspiring studies have been undertaking in our research group since then.

A phylogenetic study has shown that CA VI is closely related to transmembrane CAs (Hewett-Emmett and Tashian 1996) and a speculation have been made from an unpublished observation in our research group that the CA VI lost its transmembrane and cytoplasmic domain early in vertebrate evolution and attached Pentraxin, while subsequently losing Pentraxin later during mammalian divergence [Tolvanen, unpublished observation].

3.5 Pentraxin

In the CA VI of non-mammalian vertebrates, there is a different type of domain that is related to the Pentraxin (PTX) proteins. This domain is a novel type discovered in any alpha CAs. Pentraxins are a superfamily of evolutionarily conserved proteins that are characterized by a distinct structural motif which is known as the pentraxin domain, usually lying at the C-terminal region. These proteins are multimeric pattern recognition receptors (PRR) that are mainly made up of about five identical subunits. Based on the primary structure of the monomer, these proteins are mainly divided into two groups called short pentraxins and long pentraxins. Short pentraxins comprise of C-reactive protein (CRP) and serum amyloid P component (SAP). CRP is the first PRR to be identified, and similarly SAP are classic short pentraxins produced in the liver in response to Interleukin (IL)-6. Long pentraxins comprise of rather numerous identified ones, such as PTX3, PTX4, neuronal protein (NP) 1, NP2, NPR. Basically, the primary structure of the short pentraxins are composed of a classic pentraxin domain of about 200 amino acid residues with a short N-terminal signal peptide, while the long pentraxins have starting-unrelated sequence of about 170 amino acid residues in N-terminal region followed by the regular pentraxin domain. The Pentraxin domains are highly conserved across different lineages including mammals. They have also been characterized in invertebrates such as arthropods (*Limulus polyphemus* and *Tachypleus tridentatus*, the horseshoe crabs and *Drosophila melanogaster*, the fruitfly), and in lower vertebrates (*Xenopus laevis* African clawed frog, *Danio rerio* Zebrafish, *Takifugu rubripes* Pufferfish) (Garlanda et al. 2005).

Sequence analysis have identified these proteins to have the Pentraxin domain of ~200 amino acid residues in C-terminus with an 8 amino acid long conserved signature sequence motif (HxCxS/TWxS,

where x is any amino acid). In a phylogenetic analysis, clusters of mainly five different groups have been identified. Short pentraxins are mostly clustered as a single group of the molecules. It has been observed that short pentraxins have diverged from others early in the evolution and that the CRP and SAP may have formed from duplication event just following the divergence, as both can be found in vertebrates as well as in arthropods (Garlanda et al. 2005). Congruently, it has been deduced previously that Human SAP as a close relative of CRP for the amino acid sequence homology (51%) as well as for the similar appearance of annular disc-like structure with pentameric symmetry in electron microscopy (Szalai et al. 1999; Pepys and Hirschfield 2003; Breviario et al. 1992). The different types of long pentraxins are clustered in other four separate groups. One group consisting of neuronal pentraxins NP1, NP2 and NPR found in mammals and in lower vertebrates. Another group includes PTX3, identified in mammals, birds (*Gallus gallus*) and ancient ray-finned fishes, the puffer fish and distantly related to the PTX3, Swiss cheese protein of fruitfly represents another single group. And, the last group consisting of recently characterized PTX4 that have been found in mammals as well as in zebrafish. The study speculated that groups originated independently through multiple fusion events between the ancestral pentraxin domain gene and other unrelated sequences.

Although, their proper physiological role have not been identified, however, the availability of the information regarding short pentraxin CRP, SAP and long pentraxin PTX3 to have different ligand specificity, these proteins are suggested to provide the innate immune system with a repertoire of diverse receptors of distinct specificity (Garlanda et al. 2005). Different forms of CRP and SAP identified in the anthropod *Limulus polyphemus* were found to be the abundant constituents in haemolymph that are involved in recognizing and destroying pathogens (Shrive et al. 1999). Similarly, CRP administration in mice have shown to provide protection against pathogens like *Streptococcus pneumonia*, *Haemophilus influenza*, *Salmonella enterica* (Szalai, Briles, and Volanakis 1995; Weiser et al. 1998; Lysenko et al. 2000; Szalai et al. 2000). Likewise, SAP and PTX3 have been found to bind apoptotic cells releasing nuclear components, regulating their clearance and gate the activation of autoimmunity (Rovere et al. 2000). The neuronal pentraxins are named as such because they are involved in neuronal functions like regulation of neurodegeneration. NP1 was identified originally in snake venom neurotoxin as a protein binding taipoxin (Schlimgen et al. 1995). While, the prototype long pentraxin, PTX3, are known to be produced by dendritic cells and macrophages in response to Toll-like receptor engagement and inflammatory cytokines. Additionally, PTX3 is also considered essential in female fertility as they act as a nodal point for the assembly of the cumulus oophorus hyaluronan-rich extracellular matrix. The actual functions of this protein superfamily still remains elusive, however the studies point out Pentraxins as multifunctional PRRs at the crossroads between innate, adaptive immunity, inflammation, matrix deposition and female fertility (Garlanda et al. 2005).

The CA VI-related Pentraxins have not been studied about in detail in any published sources, so the functional role they might have along with the CA activity of the CA domain remains unknown. Some preliminary analysis have detected the CA VI-related Pentraxins to be related to the short pentraxins despite their multi-domain structure resembling the long pentraxins [Tolvanen, unpublished observation].

3.6 Homology modeling

In structural biology, one of the most frequently tackled problems is a functional characterization of protein which is usually confronted by an experimental three-dimensional (3-D) structure of the studied

protein. Although, the protein structures are best determined experimentally, it is not always possible and convenient in terms of cost, time, and purpose of the study. And moreover, there are computational methods to predict the structure from available experimental structures. Comparative or homology modeling provides the method to predict a useful 3-D model for a protein based on one or more related proteins of known structure. The sequence of the protein of unknown structure (target) is used to find the homologous proteins of known structure (template), and based on the 3-D structure of the template and the alignment between template and target sequences, the target protein structure is modeled. In homologous or closely related proteins, it has been found that the folds and overall structural orientations are more conserved than the sequences, however the distantly related sequences (less than 20% sequence identity) have very different structures (Chothia and Lesk 1986). In other words, the tertiary structures of homologous proteins are evolutionarily more conserved than their primary structure. It has also been shown that threading potentials and proper packing in the homologous protein secondary and tertiary structures are evolutionarily more strongly conserved than the sequence homology measured alone (Kaczanowski and Zielenkiewicz 2010). Hence, if sufficient similarity at sequence level is detected between two proteins, their structural similarity can usually be assumed. Approximately one-third of all protein sequences are estimated to be related to at least one protein of known structure (Rost and Sander 1996).

The comparative modeling methods are usually divided into several steps in different literature, but mainly the process consists of four different steps: template selection, target-template alignment, model building and model assessment. The templates are mostly identified based on the sequence alignment, so the first two steps are often performed together. However, alignments produced in these extensive database search methods are usually made through heuristic approaches that prioritize speed over quality. Here is a brief about steps in comparative modeling as described in (Eswar et al. 2006).

Template search and selection

Using the target protein sequence as the query, the experimental 3-D models of homologous proteins are searched in the database of known protein structures such as PDB (Deshpande et al. 2005), SCOP (Andreeva et al. 2004), DALI (Dietmann et al. 2001), and CATH (Pearl et al. 2005). Usually, sequence comparison methods like BLAST and FASTA are used for detecting similarity which usually quantifies results in terms of sequence identity or statistical measures such as E-value or z-score. Occasionally, numerous templates availability makes it possible for utilizing more sensitive searching methods like profile matching and Hidden Markov Models (HMM) (Gribskov, McLachlan, and Eisenberg 1987; Krogh et al. 1994). Whereas, to detect more distantly related homologs, other sensitive methods based on MSA such as PSI-BLAST are utilized. Another method evaluates the compatibility of the target sequence with each of the structures in the database, called protein-threading, achieved by fold recognition or 3D-1D alignment (Marti-Renom et al. 2000; Peng and Xu 2011). It applies sequence-structure fitness function such as low-resolution, knowledge-based force-fields to evaluate potential target-template matches which generally does not rely on sequence similarity. As a result, it often allows identification of structural similarity among proteins with no significant sequence similarity i.e. distantly related proteins (Dunbrack et al. 1997). Though, in general, the heuristic method, BLAST search, is a reliable approach that identifies hits with sufficiently low E-value reflecting its sufficiently close evolutionary relatedness for making a reliable homology model. A template with very poor E-value is generally not recommended even when that is the only available one, since it can lead to a generation of a misguided model.

Once a several potential template structures are identified by one or more of the template searching methods, the next task of selecting appropriate template structure for the modeling process becomes necessary. Usually, it's the sequence similarity criteria that is taken into consideration while selecting a template, as it's assumed that higher the sequence similarity between the target and the template sequences, better will be the template to be the desired one for the target. However, there are few other factors too that need to be taken into account before selecting a template. The first one, from the list of different templates, an analysis could be done simply to relate the proteins and select a template that is closest to the target sequence (Felsenstein 1985). While secondly, the physiological condition where the target supposedly exists should also be considered to look at in template's native physiological background such as solvent, pH, ligands, quaternary interactions and the like. Lastly, the most important factor in template selection underlies in the experimental quality of the template structure. The accuracy of a crystallographic structure depends on the variables such as the resolution and R-factor while for a nuclear magnetic resonance (NMR) structure, the number of restraints per residue is the indicative factor for the structure's accuracy.

Target-template alignment

After the target has been selected, all comparative modeling programs rely on sequence alignment to ascertain structural equivalences between template and target residues for constructing a homology model. Although such alignments are already constructed by template search methods, these procedures are not based on producing optimal alignment. The search methods utilize mostly heuristic approaches which often sacrifice quality of the alignment over speed. Hence, a specialized alignment methods need to be applied to construct a proper alignment after template selection. Most often, the best possible alignment depends on the sequence identity of template and target. If the target-template sequence identity is above 40%, an accurate alignment would be produced from any standard alignment methods. But when the target-template sequence identity is lower than 40%, the alignment generally has gaps and hence, careful manual interventions would often be necessary so as to minimize the occurrences of misaligned residues. Some alignment methods even take structural information from the template into account, especially this helps in avoiding gaps in secondary-structure elements, in buried regions, or between two residues that are far apart in space.

Model building

Once the starting target-template alignment is ready, the 3-D model construction can proceed through either of the three main methods used for this process. The initially and still most widely used method is called modeling by rigid-body assembly (Blundell et al. 1987; Browne et al. 1969; Greer 1981), in which the model is generated through few core regions, loops and sidechains obtained by dissecting the structures. Secondly, a method called modeling by segment matching utilizes the approximate position of matched atoms from the templates to determine coordinates of other atoms (Jones and Thirup 1986; Unger et al. 1989; Claessens et al. 1989; Levitt 1992). The third and the latest method called as modeling by satisfaction of spatial restraints uses a technology similar to the experimental NMR method. By estimating the spatial restraints from the alignment of target sequence with template structure, the method implies to satisfy the restraints variables using either distance geometry or optimization techniques (Havel and Snow 1991; Srinivasan, March, and Sudarsanam 1993; Sali and Blundell 1993; Brocklehurst and Perham 1993; Aszodi and Taylor 1996). Despite the different types of methods for generating models, their accuracies are relatively similar when considered optimally. While, the initial steps of template selection and alignment generation usually have a stronger impact on the model

accuracy especially when the models are constructed based on sequence identity of 40% and lesser. However, the model building method allows a degree of flexibility and automation that is important in generating better models.

Model assessment

After the model has been built, its assessment is necessary to determine the quality of the model. Usually, the models are first structurally compared with the experimental template structure. The most common method uses the root-mean-square deviation (RMSD) metric for measuring the mean distance between the corresponding Cα-atoms in the backbone chain of the two structures by superimposing them. It gives an idea about how structurally related the modeled target is to the template structure, assuming the structures of homologous proteins are evolutionarily better conserved. Although, this method does have limitations in measuring the accuracy under certain conditions such as when the core region of the model is correctly modeled, while some flexible loop regions in the surfaces remain inaccurate. The possibility of such unreliable regions can be assessed by composite score approaches which partly use atomic statistical potentials and energy profiles of the model. Often those information which can be extracted from the predicted model alone are considered more appropriate accuracy measure since it seems essential in estimating the accuracy of the model in the absence of the known reference structure. In situations when the sequence identity is poorer (< 30%), it becomes necessary to evaluate whether the template chosen was the correct one. For this, the environment of each residue in a model in reference to the expected conditions in native experimental structures are evaluated by assessing the compatibility between the sequence and predicted model. Several methods which applies 3-D profiles and statistical potentials (Sippl 1990; Luthy, Bowie, and Eisenberg 1992; Melo, Sanchez, and Sali 2002) such as VERIFY3D (Luthy, Bowie, and Eisenberg 1992), PROSAII (Sippl 1993), HARMONY (Topham et al. 1994), ANOLEA (Melo and Feytmans 1998) and DFIRE (Zhou and Zhou 2002) are used for this assessment. Additionally, evaluations of self-consistency of the model are also necessary to ensure that it satisfies the spatial restraints used while calculating it. Several variables relating to the stereochemical properties of the model such as bond-lengths, bond-angles, torsion angles and non-bonded contacts (Vander Waal) can be evaluated with tools such as PROCHECK (Laskowski et al. 1993) and WHATCHECK (Hooft et al. 1996).

3.7 Data driven protein-protein docking

To study biomolecular interaction through structural insights has become one of the main interests in understanding functionality of biological complexes that exist in huge bundles in the living system. Considering these huge mass of complexes and their often weak and flexible nature, the experimental methods such as X-ray crystallography and NMR spectroscopy are not always conventionally sufficient to get the structural insight into the complexes. The development and implementations of computational methods in structural biology field have improved and made possible to complement such difficulties posed by the classical structural methods. Moreover, the availability of a wealth of biochemical and/or biophysical data about the complexes have helped in implementing these information to computationally drive the docking of complexes in what is known as data-driven molecular docking. In simple terms, molecular docking can be called as the process of modeling the 3-D structure of a biological complex from its known subunits or other constituents. The data required to drive the molecular docking can be usually determined by studying the biomolecular interactions in the complexes through various biochemical and biophysical experimental procedures. Using the information generated through such experiments, models

of biomolecular complexes can be constructed from its constituents as starting raw materials. The constituents of the complexes should have a known structure or it could be computationally built models such as those produced from comparative modeling approach. In contrast, the ab-initio docking approaches which generally do not use any kind of experimental data are known to have difficulty in constructing a consistently reliable models of the complexes. Likewise, the crystallization of larger complexes would be the main daunting task in x-ray crystallography experiments while due to severe line broadening caused by large complexes in NMR experiment, it also has certain limits for solving structures of large complexes (van Dijk, Boelens, and Bonvin 2005). Hence, data-driven docking has proved to be a very convenient method of computational modeling of complex molecules.

In general, molecular docking is the computational modeling of the quaternary structure of biological complexes such as those formed by interacting macromolecules. The multi-domain proteins are the most common candidates for such computational methods, although protein-small ligand complexes, as well as protein-nucleic acids complexes, also fall under the targeted attempts of molecular docking approaches. Docking methods predict the organization of the complexes based on the structures of its constituents in close possible orientation for which the predicted structures are ranked with scoring functions to determine the best possible naturally occurring orientations. To do this computationally, two things are necessary, the first is to generate all possible orientations which is called sampling and the second is to decide which of the generated structure would resemble its natural occurrence that is done by scoring all structures and ranking them accordingly. A preliminary output of a docking basically consists of large number of solutions, from which high ranking models are chosen for further processing that are considered to correspond to the real structure, while low-ranking ones are discarded (van Dijk, Boelens, and Bonvin 2005). During the docking itself, there are usually two ways on which the conformational changes of constituents are allowed for interaction. When the orientations of interacting molecules are allowed to vary relatively but keeping their internal geometry intact, the type of modeling is referred to as rigid body docking. Here, the bond angles, bond lengths and torsion angles of the interactors are not allowed to modify at any stage during the docking but this method would be sometimes inadequate when considering occurrences of substantial conformational changes within the molecules during the natural formation of complexes. Hence, the docking procedures have been developed to allow conformational changes in the internal geometry of the models of interacting molecules in what is called as flexible body docking. Here, the scoring of all possible conformational changes will, however, be prohibitively expensive in computation time, which is why the flexible docking procedures requires to intelligently select only small subset of possible conformational changes for scoring. Additionally, combining various kinds of biochemical and biophysical data with docking would relatively reduce the stress on computation time as well as ensure a reliable conformation of the generated models by directing of multidomain docking more evidently.

The sources of data which provide information about residues in the interface regions of the complex macromolecules can be obtained from various biochemical and/or biophysical experiments. Different types of experiment provide a different level of detail and reliability in the data i.e. specificity in interface residue level. One of the most frequently used and reliable methods considered is mutagenesis experiment. The general idea of this method is that the mutation of interface residues will affect the interaction of the constituent molecules while mutation in non-interface residues will have no effect. The target residues for mutagenesis are usually selected from surface residues on the basis of conservation while in-depth systematic scan such as alanine scanning mutagenesis studies are also available (Clackson

and Wells 1995; DeLano 2002). Another source is from Mass spectrometry (MS) that are among frequently used tool in structural biology for studying biomolecular complexes (Hanson and Robinson 2004; Hernández and Robinson 2001). The popular approach used here is Hydrogen/Deuterium exchange where the rate of exchange is used to infer knowledge about the accessibility of the residue in question. The idea is that the free and unbound forms have different rates which indicates that a given residue is protected from the complex formation and probably must be involved in the interaction (Lanman and Prevelige 2004; Garcia, Pantazatos, and Villarreal 2004). Another approach used in MS is by covalently linking two molecules by the use of cross-linking reagent following subjecting the resulting complex to peptide mass fingerprinting or other protein identification methods to detect residues that are close in space (Back et al. 2003). Despite the methods being promising, due to the difficulty in the cross-linking reaction and its nontrivial interpretation as well as detection of the cross-linked residues, the data from such methods have not often been combined with docking approaches (van Dijk, Boelens, and Bonvin 2005). Different from the conventional use of NMR methods in structure determination, it can also be used to map interface regions of biomolecular complexes with a method called as chemical shift perturbation (CSP) experiments (Zuiderweg 2002). There are other NMR techniques that are able to provide similar information about interface regions such as Hydrogen/Deuterium exchange, cross-saturation and residual dipolar couplings, pseudocontact shift (Takahashi et al. 2000; Bax 2003).

There are several other experimental techniques for identifying information about the interface regions. The data produced through whichever mechanisms should be reliable for use in the computational docking approaches. As discussed earlier, there are generally two stages in this computational method, the generation of all possible conformations of the complexes or sampling and the scoring of the generated models. Under various criteria, different type of sampling methods can be used. A rigid body docking are performed when using a grid representation of molecules by calculating correlations of surface complementarity with fast Fourier transform methods (Gabb, Jackson, and Sternberg 1997; Mandell et al. 2001). Likewise, when explicit representation by atomic models of proteins are used, sampling methods such as Monte Carlo (Knegtel, Boelens, and Kaptein 1994), molecular dynamics methods (Dominguez, Boelens, and Bonvin 2003) or genetic algorithms (Morris et al. 1998) in combination with simulated annealing schemes can be used. While scoring is typically done using force fields (Mackerell 2004) of atom-atom or residue-residue pair's energy assignment, following the calculation of overall energy for a given configuration. Force fields can be derived based on physical factors or can be based on statistical information from database i.e. by identifying how often a given pair occurs in experimental structure database. The data-driven docking approach has advantage over ab-initio docking which basically have to deal with unnecessary sampling and/or scoring of tens to hundreds of false positives. The data-driven docking can relax strains on both these stages by producing more relevant conformations during sampling while improving the ranking of true positives in the scoring stage. The main difference between various methods that utilizes this approach exist in this scenario that whether the experimental data are incorporated in sampling stage or only during scoring stage in order to filter the generated samples. Although this docking approach has considerable advantages, every method has disadvantages too, and computational docking in itself is an unsolved problem. The field of docking in structural bioinformatics is still in active development, and no any docking method produces the perfect solution, rather they are just the close-possible ones.

3.8 Tools and theory

3.8.1 Ensembl

Ensembl is a genome browser (www.ensembl.org) that features the sequenced genomic information of mostly vertebrates with a particular focus on key model organisms such as humans, mouse and zebrafish (Flicek et al. 2013). It is a project jointly undertaken by European Bioinformatics Institute and Wellcome Trust Sanger Institute in 1999 with the Human Genome Project around the corner to be completed. The genome annotations are mostly done through an automated gene annotation system that predicts the gene locations which is made available for access through a MySQL database for further analysis as well as viewing. The database houses genomic, transcriptomic and proteomic sequence data that can be accessed through different tools available in the browser and also programmatic interfaces. There are new releases of the database every 3 months which involve changes mostly the inclusion of new species sequences. The database is equipped with a Perl API (Application Programming Interface) through which the data could be accessed programmatically, that is further divided into sections. Such as for comparative genomic study, the compara API is used for predicting different homolog relationships. As for the sequence retrieval of orthologs, a recent implementation of a Representational State Transfer (REST) API was utilized as an updated version to the 'Orthologer' script (described in Section 4.1) (Yates et al. 2015; Barker 2013). The Ensembl REST enables the easy retrieval of different data from the database with the use of various other programming languages other than Perl.

3.8.2 Python

Python is a high-level programming language which is an open-source free software (www.python.org). Its codes are highly readable and the syntax requires comparatively fewer lines of code than with other languages such as C, C++ or Java. Unlike the curly braces used by most other languages, it uses whitespace indentation for delimiting blocks. A start of certain statement body is followed by an increase in indentation while a decrease in the indentation signifies its end. All programmatic scripting done in this thesis for different tasks were written in Python language. Python v2.7 was used despite the existence of newer version Python v3, basically for the full compatibility of a bioinformatics-based supplementary subroutine called biopython.

3.8.3 Biopython

Biopython is a collaborative project of an international team of developers, which is an open-source collection of various Python tools designed to perform tasks related to computational molecular biology or bioinformatics. It is a collection of Python modules which consists of classes that can be used programmatically for various bioinformatics analysis tasks. Basically, it is a platform designed for bioinformaticians to easily access the tools and perform various analysis through python scripts (Cock et al. 2009). Sequence file formats like *fasta* and alignment file formats can be easily parsed through scripts for reading and writing and to perform various analyses. There are modules that are designed to deal with 3D protein model structures, and other bioinformatics tools such as BLAST, ClustalW and EMBOSS can also be accessed. Online databases of biological information such as NCBI can also be accessed through a programmatic interface with the help of biopython modules. Different sequence analyses done in this thesis have frequently utilized biopython based scripts.

3.8.4 Clustal Omega

It is a stand-alone program designed to perform multiple sequence alignment (MSA) of protein and nucleotide sequences in much faster and reliable way. Homologous sequences often need to be aligned

to infer any relationships and information required for sequence analysis and also inferring phylogeny. Most methods compute the alignments of larger datasets while compensating the accuracy, whereas others produce better quality alignments but at the expense of ease of computation as the number of sequences increase. The algorithm of Clustal Omega have been developed to produce quality alignments while also allowing infinite sample sizes with great computation power (Sievers et al. 2011). In the Clustal Omega, a modified version of mBed (Blackshields et al. 2010) with a complexity of O(N log N) is implemented that produces accurate guide trees. All sequences are embedded in a space of n dimensions (n ~ log N) and then replaced by an n element vector where each element is the distance to one of n reference sequences. Then these vectors can be clustered extremely quickly by standard methods such as UPGMA or K-means. Finally, the alignments are computed by HHalign package (Soding 2005), it aligns two profile hidden Markov models (Eddy 1998) which produce very accurate alignments. Additional feature includes allowing users to specify an HMM profile from an aligned MSA of homologous sequences of the input set. All the sequence alignment tasks performed during various analyses during this thesis research was performed with the Clustal Omega, available in webserver (http://www.ebi.ac.uk/Tools/msa/clustalo/) at European Bioinformatics Institute (EBI).

3.8.5 Prediction webservers

3.8.5.1 NetNES 1.1 Server

This webserver tool is a program developed to perform throughput analysis of protein sequences for predicting Nuclear Export Signals (NES). The NetNES program employs two different machine learning algorithms, a significant improvement over the generally used consensus patterns, for predicting NESs (la Cour et al. 2004). In eukaryotic cells, nuclear compartments are separate from the cytoplasm, which is why it becomes essential for the active nucleo-cytoplasmic transport of macromolecules across these compartments. This program predicts the leucine-rich NESs in eukaryotic proteins and for it, the program combines hidden Markov models (HMM) and artificial neural network (ANN) methods. These models are trained with datasets contained in NESbase 1.0 (la Cour et al. 2003) which consists of 64 proteins having 67 high-confidence NESs. Each residue in a submitted sequence is assigned an HMM score and an ANN score calculated through the trained models, and another algorithm post-processes the scores to assign a final score to every sequence position. The output of the program displays all three HMM, ANN and post-processed 'NES' scores. This allows for manual inspection of individual scores for possible misprediction due to either of the models failing to recognize or falsely recognizing the NES motif. The NetNES predictor is made available at (http://www.cbs.dtu.dk/services/NetNES) webserver.

3.8.5.2 NucPred

This is also a webserver tool designed for analyzing patterns in eukaryotic protein sequences and predicting nuclear localization signals (NLS). The NLS are a stretch of positively charged basic amino acid sequences in those proteins that are destined to the nucleus. This novel tool NucPred for predicting NLS is based on a different machine learning algorithm called genetic programming (GP) that induces regular expression matching and multiple program classifiers (Brameier, Krings, and MacCallum 2007). The GP is such type of machine learning method that spontaneously develops computer programs in an artificial evolutionary process (Koza 1992). The predictors incorporate multiple regular expressions that are evolved together with actual classification rules for matching against the input sequence. In this way, the predictor uses the evolutionary information from the sequence profiles and these evolved sequence motifs are not restricted to a predefined set of NLS patterns, so this program has potential to discover even novel and unknown NLSs. For user-submitted multiple sequences, the webserver has two different

ways where a likelihood score is calculated in either input of a 15 related sequence sets at one time producing alignment, or up to 1000 sequences in batch mode. In the multiple sequences, the scores are calculated for each residue and are colored in a spectrum of red to blue based on the influence of the residue on nuclear classification. The closer the subsequence color lies at the red end of the spectrum, the more positive is its effect for nuclear location, while vice-versa when it lies towards the blue end of the spectrum. The web interface to the NucPred tool is made available at (https://www.sbc.su.se/~maccallr/nucpred/) webserver.

3.8.5.3 TMHMM Server v.2.0

This is a program developed for predicting transmembrane helices in integral membrane proteins. The program TMHMM is also housed in a webserver and its algorithm is based on hidden Markov model (HMM). This software is devised to detect full topology of the protein sequences with the total number of transmembrane helices and their in/out orientation with respect to the cell membrane. This program incorporates various criteria of transmembrane helices potentiality into a specialized model that consists of submodels designed to model specific regions of a membrane protein through several HMM states. Different criterion for prediction of transmembrane helices like hydrophobicity signal, charge bias signal, helix lengths and grammatical constraints are employed in the HMM which provides high degree of accuracy, reliability and better prediction (Krogh et al. 2001). An evaluation of several methods found TMHMM to be the best-performing transmembrane prediction program (Moller, Croning, and Apweiler 2001).

3.8.6 RCSB Protein Data Bank (PDB)

The PDB is a repository of experimental 3-D structures of macromolecules such as proteins, nucleic acids and macromolecular complexes. The database is the single archive worldwide for structural data of the biological molecules, mostly protein structures (about 90% of the total data), determined through experimental methods such as X-ray crystallography, NMR, cryoelectron microscopy and theoretical modeling (Berman et al. 2000). It was first established at Brookhaven National Laboratories (BNL) in 1971, which is currently overseen by a central organization called Worldwide Protein Data Bank (wwPDB). The parent organization consists of host members viz. PDB in Europe (PDBe, http://www.ebi.ac.uk/pdbe/), PDB Japan (PDBj, http://pdbj.org/) and Research Collaboratory for Structural Bioinformatics (RCSB, http://www.rcsb.org). RCSB has the sole responsibility of the management of PDB since 1998, and it is the most popular and major host of wwPDB that is mostly accessed (Berman et al. 2000). As the member, RCSB curates and annotates data that are deposited by diverse groups of researchers such as biologists, biochemists, computer scientists from around the world. The data are made publicly accessible on the Internet via three different websites of its member organizations (PDBe, PDBj, and RCSB). The structural data of the macromolecules are universally written in pdb file format. The file compiles the parameters obtained from the experimental methods in a textual script format that consists of description and annotation of the macromolecule (protein, nucleic acid) structures such as its atomic coordinates, observed sidechain rotamer(s), secondary structure assignments and atomic connectivity. The structures are often deposited along with other molecules used in the experiment settings such as water, ions, ligands, inhibitor molecule which are also described in the pdb format file. This file format can be accessed for structural simulation or analysis through different molecular visualization software such as UCSF Chimera.

3.8.7 UCSF Chimera with MODELLER interface

UCSF Chimera is a molecular visualization software package developed by the Resources for Biocomputing, Visualization and Informatics (RBVI) at the University of California, San Francisco. Chimera can be used for various purposes of visualizations and analyses of protein 3-D model structures, nucleotide structures. Various types of structural editing can be performed using Chimera for producing very high-resolution images of publication quality that are essential in this thesis research. The structural analyses such as hydrogen bond, contact and clash detection; distance, angles, surface area, volume measurements; structure building and bond rotation; morphing between different conformations of proteins etc. are some of the basic capabilities of the software package. Along with these, Chimera is also implemented with further extensible visualization tools. In general, the package is segmented into a core which provides basic services and visualization facilities as mentioned above, and extensions for even higher level functionality. Such architecture satisfies the demands of third-party developers who wish to incorporate new features in their products (Pettersen et al. 2004). One example of such extension is MultiAlign Viewer (Meng et al. 2006) for analyzing multiple sequence alignment and associated structures that is applicable in various purposes such as structure comparison based on sequence conservation and also for comparative modeling.

Another extension tool in the Chimera that is an important part of this research is MODELLER interface. The MODELLER is a platform for doing comparative modeling of protein three-dimensional structures (Eswar et al. 2006; Marti-Renom et al. 2000). The homology models are produced based on the user-submitted initial alignment of the target and template sequences along with a script defining the required parameters for the modeling. The incorporation of this suite in Chimera provides more comprehensive graphical user interface (GUI) for setting up of input data and parameters in simplified way required for the modeling process, evaluating results and also performing various types of refinement during model generation. The BLAST web-service for searching potential templates from the PDB database is itself integrated into the tool as well as the alignment of the target and selected template will be made through ClustalW algorithm in the MultiAlign Viewer. And finally based on the target and template sequence alignment with the known structure of the template, the MODELLER suite can be called either locally or via web-service. The MODELLER runs in the background and generates models within few minutes which are directly displayed in the Chimera interface along with their associated scores. The whole interface is very user-friendly that greatly facilitate modeling tasks as well as maximizes the coverage, resolution, accuracy and efficiency for the structural characterization of macromolecular assemblies.

3.8.8 HADDOCK webserver

HADDOCK is an information-driven flexible docking platform which incorporates prior experimental information or predicted protein interface data in ambiguous interaction restraints (AIRs) for performing the docking between proteins, nucleic acids and any other biomolecules. The HADDOCK docking suite differentiates from other *ab-initio* docking programs in the fact that it utilizes information from identified or predicted protein interfaces to drive the docking process (de Vries, van Dijk, and Bonvin 2010). The data-driven approach implemented in the docking suite can support a wide range of experimental data, such as from mutagenesis, mass spectrometry or various NMR techniques (chemical shift perturbation (CSP), residual dipolar couplings (RDCs) or hydrogen/deuterium exchange). The determined information are entered in the form of active and passive residues that are interpreted as AIRs by the program and used for directing the docking by automatically generating the topology of the molecules to be docked. This docking is driven in a similar manner as AIRs based on classical NMR distance restraints data which

drive the structural calculation of an NMR structure (Linge et al. 2003). The docking protocol follows mainly three stages, firstly a rigid-body energy minimization, secondly semi-flexible refinement in torsion angle space and finally, a refinement in explicit solvent. The structures are scored and ranked after each of the stages and accordingly kept for the next stage. The HADDOCK program allows for full structural flexibility of both side chains and backbone of the biomolecules. Few other docking programs built prior to the HADDOCK can only allow the structural flexibility but only at the side chain level, as well as only some could deal with nucleic acids. The HADDOCK server is powered up by a dedicated cluster because of which a docking run typically takes a few hundreds of CPU hours. While additionally, the server can also make use of GRID computing resources deployed across Europe within FP7 e-NMR e-Infrastructure European project (http://www.enmr.eu) when the main HADDOCK server is too busy or down for maintenance/upgrading. The HADDOCK webserver is available at (http://haddock.chem.uu.nl/haddock) and the GRID enabled version at (http://haddock.chem.uu.nl/enmr). Both requires registration through a web form for access.

The HADDOCK webserver can also incorporate bioinformatics predictions interface data when the experimental information is sparse or absent. A tool called as CPORT (Consensus Prediction of Interface Residues in Transient complexes) is implemented in the web-server for predicting interface residues of the protein structures (de Vries and Bonvin 2011). The CPORT is developed to integrate the interface predictions suitable for use in the HADDOCK program by combining six interface prediction webservers into a consensus method. The CPORT predictor generates a list of active and passive residues that can directly be entered into the HADDOCK webserver as input parameters. However, the interface residues are usually over predicted by CPORT, which could be further improved by interface post-prediction, by using any other information regarding the probable docking solutions. The CPORT is very helpful when no experimental information are available and thus also presents an alternative to *ab initio* docking methods. The CPORT takes in PDB file of the structure and along with the lists of the interface residues, it also generates a PDB file containing the information regarding the prediction. The active and passive predictions are incorporated in the B-factor (temperature) column of the PDB file, which can be easily visualized in various molecular graphics software for e.g. UCSF Chimera through rendering by B-factor attribute.

4 Research methodologies

4.1 Sequence retrieval

All the protein sequences required for various bioinformatics analyses were retrieved from Ensembl database using a pipeline, written in python language, called *'Orthologer'* developed in our research group by Harlan Barker (Barker 2013). The python script employs over twenty python modules including biopython, fetches data from the Ensembl database through REST API, while calling multiple external bioinformatics programs and retrieves and processes data before finally producing usable sequence datasets. It has to be operated through Unix/Linux systems for full accessibility because some external programs are unavailable in Windows system. After the sequence retrieval from the database, each orthologues are processed for 'good' or 'bad' quality analysis on the basis that it is a 'bad' sequence if it does not have starting-Methionine, an 'X' in the sequence, or if its length deviates more than 5% from the corresponding human orthologous sequence. And optionally for the sequences assigned as 'bad', predictions can be made with Exonerate (Slater and Birney 2005) program, called within the pipeline, for extracting a better quality protein sequence from the genome than that annotated in the public database.

The Ensembl Ids of the required proteins are given as input to the *Orthologer* to retrieve their orthologous sequences from the database. Here, the Ids of human genes for each CA isozymes: CA IX, CA XII, CA XIV and CA VI are used to retrieve the orthologous protein sequence datasets of the respective isoforms. And similarly for the Pentraxins, the human gene Ids of Pentraxin proteins CRP and SAP were used to retrieve their corresponding orthologous sequences. The Ensembl Gene Ids for each protein that were used as input are shown in the Table 4-1.

S.N.	Protein	Ensembl Gene Ids
1.	CA IX	*ENSG00000107159*
2.	CA XII	*ENSG00000074410*
3.	CA XIV	*ENSG00000118298*
4.	CA VI	*ENSG00000131686*
5.	CRP	*ENSG00000132693*
6.	SAP	*ENSG00000132703*

Table 4-1. The Protein names and their Ensembl Gene Ids

After the retrieval, the 'bad quality' assigned sequences were manually inspected in Multiple Sequence Alignment made with ClustalOmega. Since, the quality analysis algorithm of *Orthologer* will also skip out any sequences which just do not have initial 'M' despite rest of the features being very fit for further analyses, the bad quality sequences of transmembrane CAs were manually inspected and selected for analysis based on conditions that the sequence did not have any 'X' character, and the transmembrane and/or cytoplasmic portions were intact.

4.2 Sequence analyses

In preliminary analyses, predictions of sub-cellular localization and transmembrane helices for orthologues of transmembrane CAs: CA IX, CA XII and CA XIV were performed. The following programs were used in the World Wide Web (www) for the predictions. TargetP 1.1 Server for secretory signal peptides, mitochondrial targeting peptides or other location (cytoplasmic) in non-plant organism groups of eukaryotes using the cutoffs of 95% specificity was implemented for sub-cellular location prediction.

The potential cleavage site of the predicted presequence was also predicted in the TargetP analysis for which SignalP is used. Likewise, TMHMM Server v.2.0 is used to predict transmembrane helices and along with it, the analysis generates some statistics and locations of the predicted transmembrane helices and the topologies for the intervening loop regions whether it lies outside or inside of the membrane.

The subsequent major sequence analyses in the research include the prediction of nucleo-cytoplasmic transport signals in the orthologous protein sequences of transmembrane CAs: CA IX, CA XII and CA XIV. For predicting Nuclear export signals (NES) in the protein sequences, the analysis was performed in the program NetNES 1.1 Server available at (http://www.cbs.dtu.dk/services/NetNES) that uses ANN and HMM algorithms. As stated by (Buanne et al. 2013), the regions encompassing only TM and IC portions of the human CA IX sequence were used mainly for the simplicity during their NES prediction analysis. But the technical explanation to the statement is that the algorithms of the program would have to analyze considerably larger set of parameters when the longer amino acid segments are fed and that would require correspondingly large training datasets which are not available in the current development of the program (la Cour et al. 2004). So when the full sequences were used, it produced a lot of irrelevant noises with detections of false-positive faint signals along with misprediction of the actual NES signal. Hence, all the sequences were trimmed to contain only last 60 amino acids segment i.e. sixty residues starting from the C-terminal end of the sequences that would include the TM and IC portions, and it was achieved programmatically through a Python script utilizing biopython module for parsing sequence objects. These truncated sequence datasets were submitted to the NetNES webserver for prediction of the NES signals.

As for the prediction of Nuclear localization signal (NLS) in the protein sequences, the program called NucPred was used in its server available at (https://www.sbc.su.se/~maccallr/nucpred/) which uses genetic programming algorithm. For the multiple sequence submission, the program has limits up to 15 sequences at a time. So as required, the orthologous sequence datasets were divided into batches of subsets containing 15 or fewer sequences, which was performed programmatically through a Python language script utilizing biopython module for parsing sequence objects. The subsets of sequence were then submitted to the NucPred Server under *Protein family* service that takes up to 15 related sequences and generates an MSA made from ClustalW which is colored according to the scores predicted by NucPred.

Other minor analyses performed are explained as follows. The orthologous sequences of CA XII and CA XIV were analyzed for dimerization motifs in the transmembrane regions of the proteins. The CA XII and CA XIV sequences were aligned separately in ClustalOmega and the MSA were visually inspected for the presence and conservation of the dimerization sequence motifs in the transmembrane helices portions of the sequences. And lastly, the orthologous sequences of non-mammalian CA VI which consists of Pentraxin domain were aligned with the orthologous sequences of CRP and SAP in ClustalOmega to analyze sequence variation and conservation especially for certain conserved regions of the Pentraxin proteins such as Cys residues that has functionality in formation of disulfide bridges within the sub-unit.

4.3 Homology modeling of Zebrafish CA VI and Pentraxin domain

The homology modeling of two different domains: CA VI catalytic domain (hereafter called CA VI domain) and Pentraxin domain of the Zebrafish were each performed separately using MODELLER interface tool in UCSF Chimera. However, the procedures for the modeling involves pretty much the same principle except

for some additional analyses performed as needed in case of the CA VI catalytic domain which is described in detail below. The methodologies for each domain are described in separate sections.

4.3.1 Homology modeling of CA VI domain

The PDB database was searched for homologous proteins of zebrafish CA VI domain with known structures using the zebrafish CA VI sequence as a query. This was performed using the Chimera. The query sequence was loaded in the Chimera, which opens in MultiAlign Viewer tool. The query sequence can be BLASTed against PDB database from menu *Info -> Blast Protein* using all the default Blast parameters such as E-value threshold of 1e-3 and Blosum62 matrix. The 3-D structure of human CA VI with PDB Id: 3FE4 was found to be the best hit with lowest E-value (lower the better) of 5e-79.

Template evaluation and refinement

The 3FE4 is the only structure of any CA VI isozyme and it has been resolved with resolution of a 1.9 Å (Pilka et al. 2012). A structural evaluation of the crystal structure was done. The crystallography study included a fragment from 21-290 amino acids (GeneBank entry 21706434) of the catalytic domain and TEV-cleavable His_6-tag was incorporated at the C-terminal end for purification (Pilka et al. 2012). Later, the cleavage has left extra seven residues in the end (AENLYFQ) (Figure 4-1), which are not any part of CA VI sequence. Also, in the final construct, the structural data is not available at all for the last 18 residues. A preliminary analysis on this missing C-terminal sequence fragment had predicted it to resemble an amphipathic (AP) helix (Figure 5-4). Before the structure could be used as the template, a refinement process of the structure to include the helical region was performed. The presumed amphipathic helix was modeled from I-TASSER (Iterative Threading ASSEmbly Refinement) Server (Zhang 2008) and it was docked to the C-terminal face of the PDB structure 3FE4 (described detail in Section 4.4) in HADDOCK

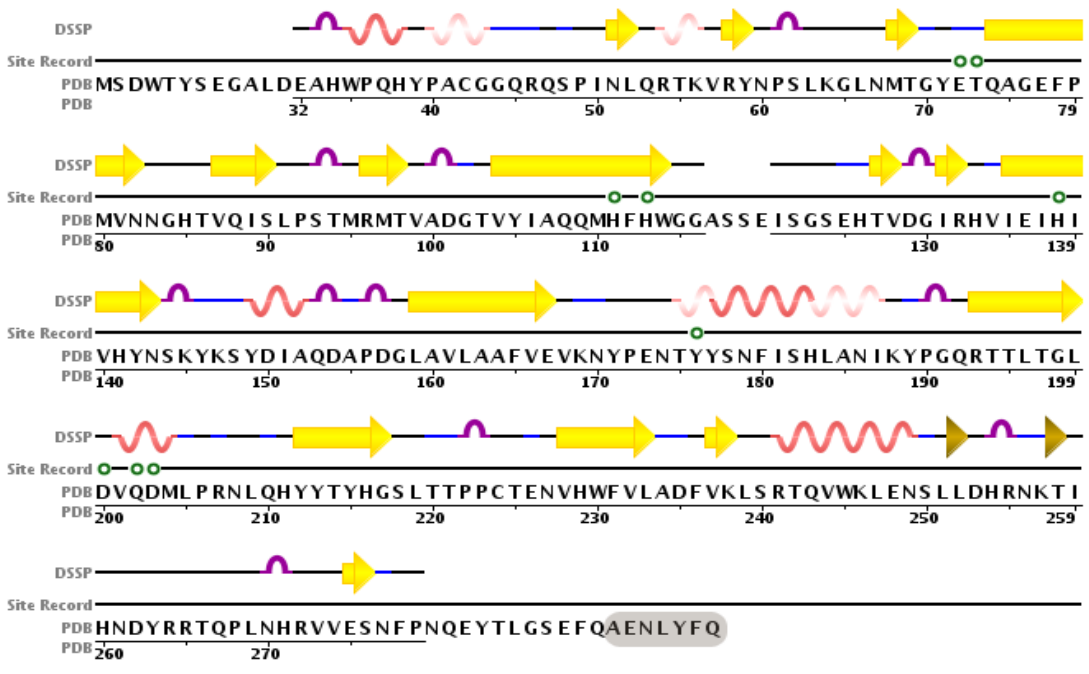


Figure 4-1. Secondary structure of the human CA VI (PDB Id: 3FE4, Pilka et al, 2012). Residues without structural data are non-underlined, the last seven residues in grey box is non-CA VI sequence left by His-tag added during purification.

Server and the final model saved in PDB format file. The refined structure of CA VI (3FE4) with the amphipathic helical region was used as the template for modeling the zebrafish CA VI domain.

Template-target alignment

The second task in the modeling is to make a pair-wise alignment of the target and template sequence. The alignment was produced with ClustalOmega for a better-quality alignment, as a good alignment serves as main drive into generating a good model, and it was brought into an optimal alignment (Figure 4-2). The N-terminal signal peptide fragments of both target and template sequence were trimmed from the alignment as it was not part of the template structure. Similarly, the last 5 residues (RRALN) from

Figure 4-2. An optimal alignment between target and template sequences is constructed with ClustalOmega. For the residues in red box, no structure data is available.

human CA VI sequence was removed as there was no structural data (Figure 4-2) and the Pentraxin domain region was also trimmed from zebrafish CA VI sequence. Finally, the sequence alignment is used to determine the orientation of corresponding residues in the target by exploiting the information from the template structure.

Model generation

The alignment is loaded in Chimera which opens in MultiAlign Viewer (MAV). The interface to the MODELLER can be started either locally or via web-service from MAV menu option: *Structure -> Modeller (homology)*. The MODELLER relies on the principle of constructing models by the satisfaction of spatial restraints. The information such as spacing between atoms, bond lengths, bond angles, dihedral angles, etc. are extracted from the template structure in the form of spatial restraints and then the target structure is constructed by satisfying all these restraints in best possible way. The MODELLER takes about 3-4 minutes on the Modeller web-service to construct the homology model and generates five (default) comparative models with their associated quality scores, GA341 (Melo, Sanchez, and Sali 2002) and zDOPE (Shen and Sali 2006). GA341 score is derived from statistical potentials, and a value greater than 0.7 indicates a reliable model with more than 95% chances of having the correct fold. Similarly, zDOPE or normalized Discrete Optimized Protein Energy is a statistical score derived from the atomic distance, and negative values indicate better models. In addition to this, the quality of models can also be assessed by structural comparison to the template structure, by using the MatchMaker tool in Chimera (*Tools -> Structure comparison -> Matchmaker*), Cα RMSDs scores were calculated for each of the comparative models. And lastly, unfavorable clashes or contacts such as close contacts between atoms or direct interactions of polar and nonpolar residues was calculated for each of the comparative models by using Find Clashes/Contacts tool in Chimera (*Tools -> Structure Analysis -> Find Clashes/Contacts*). Agreement

on the best model was made on the basis of the majority in best scores from these four different quality analyses.

4.3.2 Homology modeling of Pentraxin domain

In a similar way, the sequence part of the Pentraxin domain of zebrafish CA VI was used as a query for searching potential template structures in PDB database. As the chimera enables a comfortable access to the BLAST portal service, the target Pentraxin sequence was loaded in Chimera that opens up in MAV tool. From the MAV tool, the query sequence was subjected to BLAST (*Info -> Blast Protein*) with default setting of E-value threshold 1e-3 and Blosum62 matrix, that returned numerous homologous structures of short Pentraxin proteins from *Homo sapiens* and *Limulus polyphemus* available in the PDB database. Based on the highest resolution of 1.4 Å, the crystal structure of hSAP protein of PDB Id: 4AVS was chosen as the best template for homology modeling.

Target-template alignment

For producing a good quality pair-wise alignment, the target and template sequences were aligned using ClustalOmega. The CA VI domain from the zebrafish sequence was trimmed and the N-terminal signal peptide from the hSAP sequence and its corresponding aligned amino acids in Pentraxin (zebrafish-CA VI) sequence were also trimmed for further structural analysis. The optimal alignment with sequence identity of 30.2% was finally produced, that could be used in the further procedure of driving model generation.

Figure 4-3. An optimal alignment between template (hSAP) and target (zPtx-CAVI) is made with ClustalOmega. The residues in white box of zebrafish sequence are trimmed out, since there is no corresponding residues in the human sequence.

Model generation

The target-template sequence alignment was loaded in Chimera that opens up in MAV tool. The interface to the Modeller in Chimera was started by choosing *Structure -> Modeller (homology)* from the MAV menu option. Then by choosing the target and the template, the Modeller was run via web-service. Again, the Modeller builds the comparative models based on satisfaction of the spatial restraints as explained before. With 3-4 minutes of computation time on its server clusters, the Modeller returned five different homology models along with their statistical measures of model accuracy, GA341 (Melo, Sanchez, and Sali 2002) and zDOPE (Shen and Sali 2006). An additional structural comparison scores are calculated which are displayed in the Reply log dialog box of the Chimera, the Cα RMSDs scores for each of the models are calculated with reference to the template structure by superimposing the target models with the template structure. As a final analysis, a clash/contacts were calculated from Find clash/contacts option that can be chosen from *Tools -> Structural analysis -> Find clashes/contacts* in the Chimera menu option. The

agreement on choosing the best final model was made based on the models having the majority of best scores from these four different structural analysis.

4.3.3 Model assessment

For statistical potentials or physics-based calculation assessment of the modeled structures, PROCHECK was used. The PDB files of the models were uploaded to the PROCHECK webserver, and the statistical report of the analysis was generated by the software package, which produces various physiochemical statistical data of the models. For structural comparison, the modeled target structures are structurally superimposed with each of the corresponding template structures to measure the root mean square deviation (RMSD) between corresponding atoms in the two structures. Energy minimization was performed in Chimera with Structure minimization tool. The final resulting models were used for further docking experiments.

4.4 The docking of CA VI and Pentraxin domains

The model of full-length zebrafish CA VI + Pentraxin complex was built by docking them using the HADDOCK web-server (http://haddock.science.uu.nl) with its Easy interface since it is the most convenient one as of the scope of this research. The program utilizes user-defined input parameters of active and passive residues to drive the docking process. Active residues are those residues in the interface of each subunit of the complex that are directly involved in the interaction, whereas passive residues correspond to their solvent-accessible surrounding surface residues. The input data was also predicted

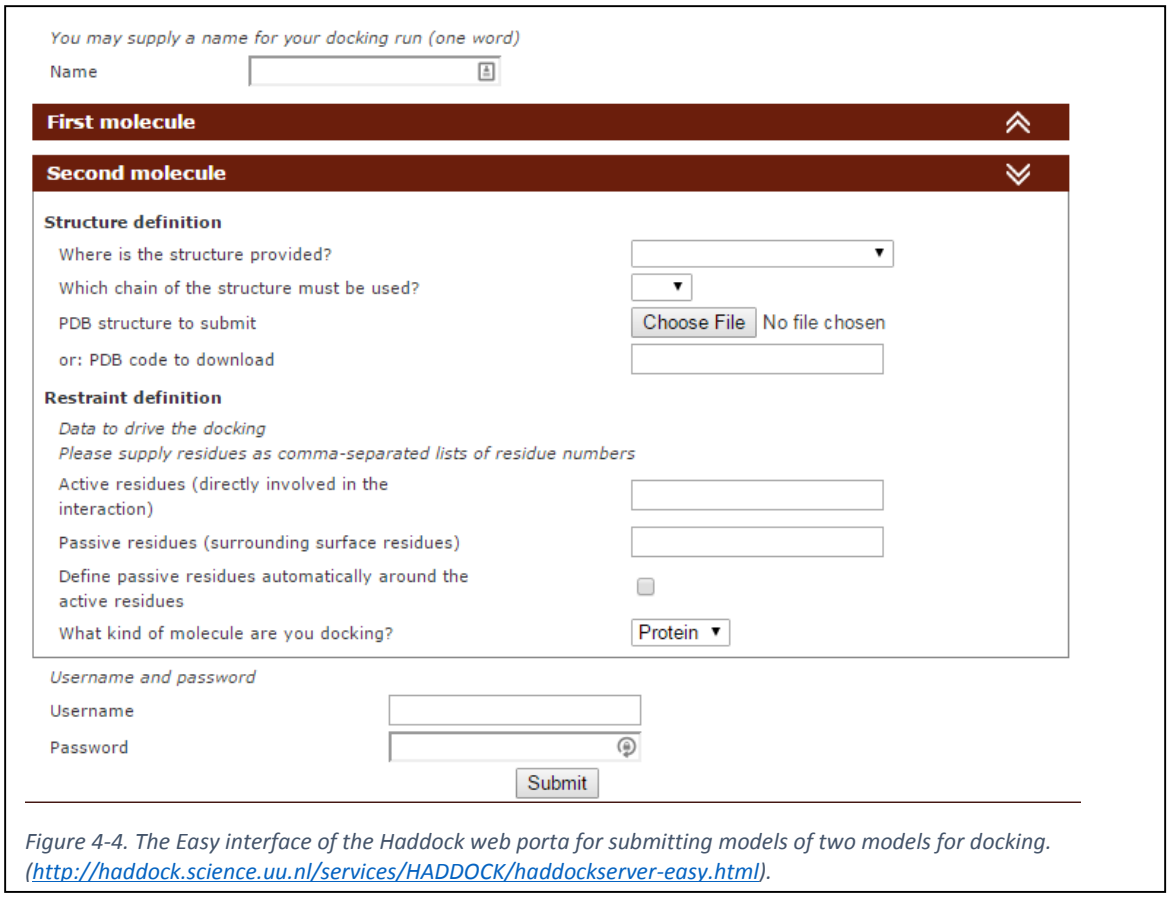


Figure 4-4. The Easy interface of the Haddock web porta for submitting models of two models for docking. (http://haddock.science.uu.nl/services/HADDOCK/haddockserver-easy.html).

through computational approach using CPORT program available in the Haddock web-server. CPORT implements a protocol to predict interface residues that can be used as input data into the Haddock's data-driven docking suite. This tool is extremely helpful in our docking step as no any experimental information about the interface residues are available and this thesis research was fully based on bioinformatics approaches. The CPORT combines six different interface prediction web servers into a consensus method and generates a list of predicted interface residues. The PDB files of each protein/amphipathic helix produced during previous modeling step were submitted to the CPORT server (http://haddock.science.uu.nl/services/CPORT) with threshold option 'Very sensitive' that is recommended for HADDOCK. The server returns output as lists of active and passive residues which can be directly included in the HADDOCK docking procedure. The CPORT also incorporate these information into B-factor column of the PDB file, which can be visualized in Chimera by rendering the B-factor occupancies in the surface coloring that in actual is active and passive residues parameters. HADDOCK interprets these information of active and passive residues in the form of Ambiguous Interaction Restraints (AIRs) to drive the docking process.

The CPORT predicted information were noted down for each of the molecules, and later used during docking. There were two different runs of docking required in this thesis. Firstly, a docking was done between the 3FE4 structure and the predicted Amphipathic helix structure of the human CA VI in order to generate its complete structural model which can be used further as template in the homology modeling of zebrafish CA VI domain. It was obvious in a near-approximate way that the AP-helix would fit on the C-terminal face of the 3FE4 structure. So the only those predicted residues lying on the C-terminal face of the 3FE4 structure were given as input, and similarly the residues in a hydrophobic part of the AP-helix was given as input in the HADDOCK web portal.

And secondly, a docking was performed between the models of zebrafish CA VI domain and Pentraxin domain that were produced from the comparative modeling in the previous step (Section 4.3). Here, it was obvious that the C-terminal face of the CA VI domain would interact with the N-terminal face of the Pentraxin domain. Hence, only those predicted residues lying at the particular aforementioned surfaces of each domain were provided as input to the HADDOCK submission portal.

Using this information in the form of AIRs, the HADDOCK performs the docking in the clusters of its servers and usually it takes about approximately 5 hours of computation time to complete the docking, depending on the queue pending jobs in their server. Finally, an ensemble or clusters of models are produced with their quality scores. From the cluster having the best score, a top model of the docked protein complex is chosen for further analysis or interpretation.

5 Results

5.1 Retrieval of sequences from Ensembl

The orthologous sequences of CAs IX, XII, XIV, VI and short Pentraxins CRP and SAP were pulled from the Ensembl database with *Orthologer* Python-script.

Species	CA IX	CA XII	CA XIV	CA VI	CRP	SAP
Homo sapiens	X	X	X	X	X	X
Gorilla gorilla		X	X	X	X	X
Pan troglodytes	X	X	X		X	X
Pongo abelii	X	X		X		X
Nomascus leucogenys	X	X	X	X	X	X
Macaca mulatta	X	X	X	X	X	X
Microcebus murinus		X			X	X
Callithrix jacchus	X	**X**	X		X	X
Tarsius syrichta					X	X
Otolemur garnettii	X	**X**	X		X	X
Felis catus	X	**X**	X	X		X
Procavia capensis	X		X		X	X
Equus caballus	X	X				X
Loxodonta africana	X	**X**	X			X
Echinops telfairi	X					X
Bos taurus	X	**X**	X		X	X
Vicugna pacos			**X**			X
Ailuropoda melanoleuca	X	**X**	X			X
Canis familiaris	X	**X**	**X**			
Sus scrofa	X				X	X
Ochotona princeps	X		X		X	X
Oryctolagus cuniculus	X	X	X		X	X
Mus musculus	X	X	X		X	X
Rattus norvegicus	X	X	X			X
Ictidomys tridecemlineatus	X	**X**	X		X	X
Monodelphis domestica	X	X	X		X	X
Tursiops truncatus		X				
Mustela putorius furo		X	X			X
Pteropus vampyrus	**X**	X	X			X
Myotis lucifugus	**X**	**X**	**X**	X		
Sorex araneus	**X**					
Erinaceus europaeus				X		
Sarcophilus harrisii	**X**	X				X
Cavia porcellus	**X**	**X**	X		X	X
Dipodomys ordii	**X**		X			X
Sarcophilus harrisii			X			
Dasypus novemcinctus			**X**			
Macropus eugenii			**X**			
Danio rerio	**X**			P		
Meleagris gallopavo	**X**	X		P		
Gallus gallus	**X**	X		P		
Xenopus tropicalis		**X**	X	P		
Latimeria chalumnae	**X**	**X**	X	P		
Tetraodon nigroviridis		**X**				
Xiphophorus maculatus			X	P		
Oryzias latipes			**X**	P		
Oreochromis niloticus			**X**			
Pelodiscus sinensis		**X**		P		
Anas platyrhynchos		**X**				
Gadus morhua				P		
Ficedula albicollis	**X**	**X**		P		
Takifugu rubripes		**X**		P		
Taeniopygia guttata				P		
TOTAL SPECIES	**33**	**35**	**33**	**20**	**18**	**30**

Table 5-1. The table showing orthologous sequences of different species retrieved from Ensembl for each of the proteins. Bold X represents the ones that were originally bad-quality sequence, P represents those CA VI consisting of Pentraxin domain.

The good and usable bad quality sequences were collected for other bioinformatics analyses. In total, there were 33, 35, 33, 20, 18 and 30 sequences for CA IX, CA XII, CA XIV, CA VI, CRP and SAP respectively. The Table 5-1 lists the species names from which the orthologous sequences of each of the proteins were retrieved for further analyses in this thesis.

5.2 Sub-cellular localization and Transmembrane helices prediction

TargetP predictions

The transmembrane CA sequences were submitted to *TargetP* server for prediction of sub-cellular localization signal. Among the 33 CA IX sequences, three sequences were not predicted to have any signaling pre-sequence while rest 30 sequences were predicted to have N-terminal secretory signal peptide. The length of the predicted signal peptide was 37 in the majority of sequences whereas it ranged from 31 to 50. Similarly, for 35 analyzed CA XII sequences, only 22 sequences were predicted to have N-terminal secretory signal peptide, while the rest were not predicted to have any signal sequence. And lastly, for 33 CA XIV sequences, 28 were predicted to have N-terminal secretory signal peptide, 4 were not predicted of any signals, and interestingly one sequence was predicted to be targeted to Mitochondria. The results output are given in Appendix I.

TMHMM predictions

All the orthologues of CA IX, CA XII and CA XIV were predicted to possess transmembrane helix which is an obvious observation for transmembrane proteins. The majority of the transmembrane helices were predicted to be composed of 22 amino acids, with an exception of 18 amino acids in *Canis familiaris* CA IX sequence. The CA IX of *Ictidomys tridecemlineatus* and CA XIV of *Latimeria chalumnae* sequences had predictions of two transmembrane helices, but the extra ones were found to be N-terminal signal peptides mispredicted as transmembrane helices when visually inspected in MSA. The results output are given in Appendix II.

5.3 NES and NLS motifs in transmembrane CAs

NetNES Prediction

The transmembrane CAs orthologues sequences were subjected for the Nuclear export signal in the NetNES Server. The sequences were truncated to contain last 60 amino acids from the C-terminal end for submitting to the NetNES server. There were positive predictions in most of the transmembrane CAs sequences. Among CA IX, 32 out of 33 sequences showed NES predictions where most of them had very strong signals. Similarly for CA XII, 28 out of 35 sequences showed the NES predictions. Likewise, 8 sequences of CA XIV did not show positive predictions, while remaining 25 sequences showed strong to moderate signals. The server calculates a consensus NES score from HMM and ANN scores, all the three scores are shown in the output result. Some mispredictions could be the result of either of two algorithms failing to recognize the NES motif site completely or partially. From the MSA, there was very high degree of conservation observed in the region of predicted NES motif sites (Figure 6-1, Figure 6-2, Figure 6-3). The results output are given in Appendices III and IX.

NucPred

The transmembrane CAs were also subjected to NLS site prediction in a different server called NucPred, where almost every sequences were predicted to contain NLS subsequence motif. The server creates a

colorful MSA of submitted set of sequences based on a scale of whether the subsequence regions are predicted to have nuclear signals or not. The scale goes from blue which is negative (non-nuclear) signal towards red which is positive (nuclear) signal. In CA IX sequences, signals were moderately detected towards the C-terminal region. A single cluster of basic amino acids Arg and Lys are observed to be highly conserved among all species. These clusters appear to occur into the end region of predicted location of the transmembrane domain and starting of the cytoplasmic domain. In contrast, the CA XII sequences have very brightly colored region of the positive nuclear signals being detected. In most of the sequences, the detection are observed to the extent that the clusters of the highly conserved basic amino acids are even underlined in the MSA. Lys is mostly conserved in all the species while there are moderate levels of conservation of Arg too. Similarly also in CA XIV sequences, there are even very highly detected signals for nuclear localization with almost all the detected sequence motifs as underlined in the MSA output. Both Arg and Lys seem to be equivalently conserved among most species and the detected sequence motif is also at the similar location as of CA IX and CA XII i.e. starting position of cytoplasmic domain (Figure 6-1, Figure 6-2, Figure 6-3). The result output are given in Appendices IV and IX.

5.4 Dimerization signal in transmembrane helix

It is previously reported that in the transmembrane region of human CA XII, there are GxxxG and GxxxS sequence motifs that signal for transmembrane helix dimerization (Whittington et al. 2001; Senes, Gerstein, and Engelman 2000; Russ and Engelman 2000). The sequence motifs for dimerization were analyzed in all orthologous sequences of CA XII in the MSA. It was observed there is a high degree of conservation of the sequence motifs in most of the orthologous sequences, especially in the good quality sequences. These signal motifs of those CA XII sequences having 100% conservation are highlighted in the MSA (Figure 5-1).

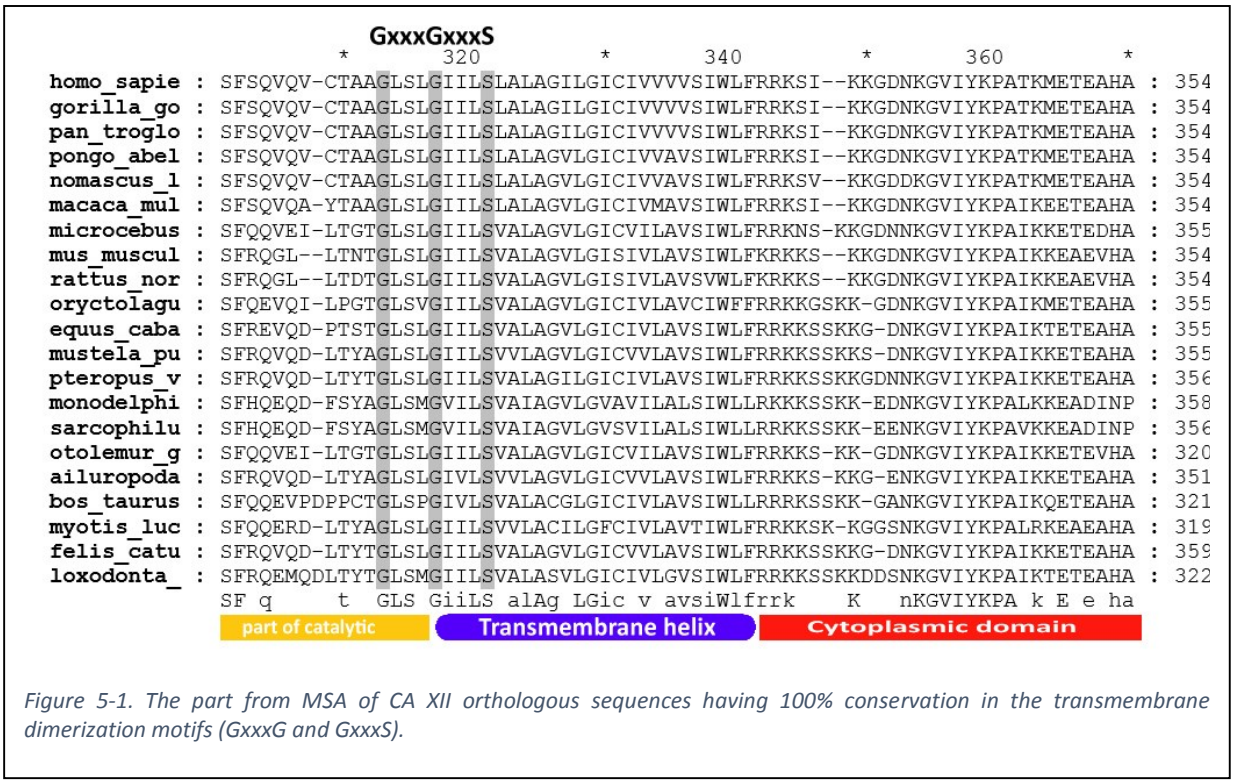


Figure 5-1. The part from MSA of CA XII orthologous sequences having 100% conservation in the transmembrane dimerization motifs (GxxxG and GxxxS).

Additionally, the CA IX and CA XIV sequences were also visually analyzed for the presence of the dimerization signal motifs. It was not observed in CA IX sequences, but the CA XIV of most of the species were found to have a very good conservation of the motifs, but in lesser extent to that found in CA XII sequences. The MSA from those CA XIV sequences is shown in Figure 5-2 having fully conserved dimerization signal motifs.

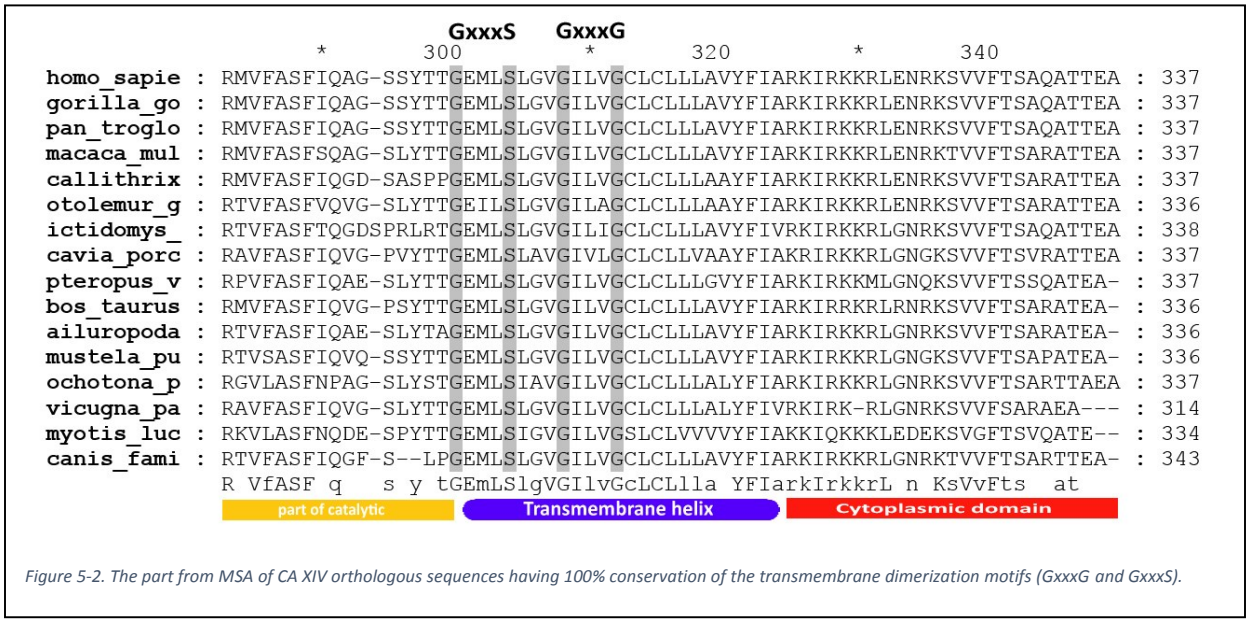


Figure 5-2. The part from MSA of CA XIV orthologous sequences having 100% conservation of the transmembrane dimerization motifs (GxxxG and GxxxS).

5.5 Sequence analysis of Pentraxin domain

The sequences of Pentraxins proteins: CRP, SAP and CA VI Pentraxin domain were aligned with ClustalOmega. Only five representative sequences from each of the proteins were included during alignment. The sequence features such as Pentraxin signature motif HxCxS/TWxS and conserved Cys residues involved in disulfide bridge formation were visually inspected in the alignment. There was high degree of conservation of the signature motif in almost all sequences. The two Cys residues known to form disulfide linkage in CRP and SAP structure are also found to be conserved in the CA VI Pentraxin domain of the five species. While additionally, there were two other Cys residues fully conserved among the CA VI Pentraxin sequences and was not present in other short Pentraxin proteins. The two novel Cys residues conserved in the CA VI Pentraxins could have a role in disulfide bond formations. The alignment with annotation is shown in Figure 5-3.

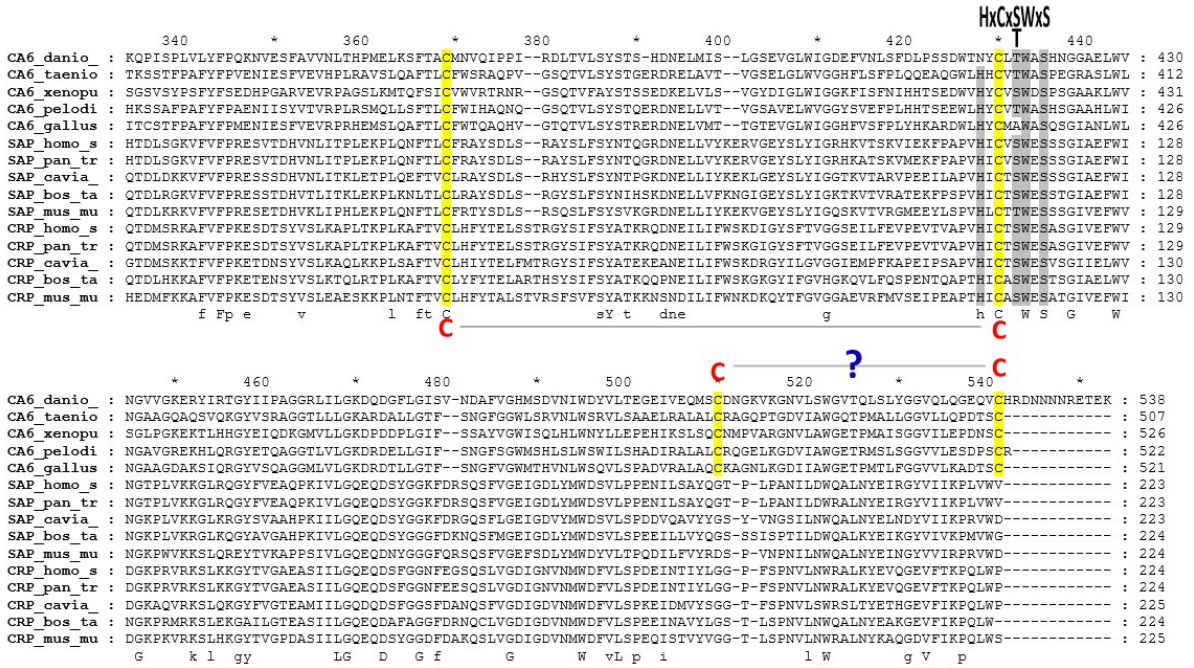


Figure 5-3. The Multiple sequence alignment of some of representatives of each CAVI Pentraxin domain, CRP and SAP; made with ClustalOmega. The horizontal lines drawn to connect 'C's depicts disulfide bridges formed between two Cys residues. The Cys columns are highlighted with yellow, the HxCxS/TWxS motif is highlighted in grey.

5.6 The CA VI has amphipathic helix at C-terminus

To model a zebrafish CA VI domain, the only closest homologous structure available in PDB was the structure of human CA VI (3FE4). The crystal structure of the human CA VI has 29 amino acids missing from its C-terminal end. Of the missing sequence fragment, 18 residues were predicted to have an amphipathic helical property as shown in (Figure 5-4, A) by HelicalWheel Projection (http://rzlab.ucr.edu/scripts/wheel/wheel.cgi). The putative amphipathic helix region was modeled by I-TASSER program as an alpha helical structure and the secondary structure prediction of the residues in C-terminal end was also shown to be helical with very high confidence scores of 8 or 9. The overall C-score of the model was calculated as -1.49 in a scale of [-5, 2], where higher value signifies a model with high confidence. The visualization of the helix structure in Chimera clearly exhibited the predicted amphipathic conformation as shown in (Figure 5-4, B). In correspondence to this, a structural inspection of 3FE4 pdb structure in Chimera highlighted hydrophobic patches in the C-terminal face of the protein 3-D structure (Figure 5-5, A), where the hypothetical helical structure would fit. The CPORT analysis also predicted interface residues which coincide well with the hydrophobic patches (Figure 5-5, B). The restraint parameters i.e. active and passive residues for docking were selected from overlapping residues of the hydrophobic patches and the CPORT prediction in the C-terminal face of the CA VI (3FE4) structure (Figure 5-5, C). The selected parameters for active residues are A117, A234, V134, and I121; and for passive residues are V57, I131, V168, and F236 in 3FE4 pdb file.

The docking of the AP-helix and 3FE4 structures were done in HADDOCK server. The docking was completed in about 3 hours of computation time in the server, and returned seven different clusters of models. The top model from cluster 1 having best Haddock score of -109.3 +/- 4.5 was chosen for further analysis (the one with lowest Haddock score is the best). The hydrophobic contacts calculated between the predicted interface residues in 3FE4 and AP-helix was found to be 45 hydrophobic or pseudo bonds (Appendix VII). The visualization of the model in Chimera shows that the helix fits seemingly perfectly in the hydrophobic patches supporting our hypothesis. The full and refined 3-D model of human CA VI structure was usable as a template for modeling the zebrafish CA VI domain.

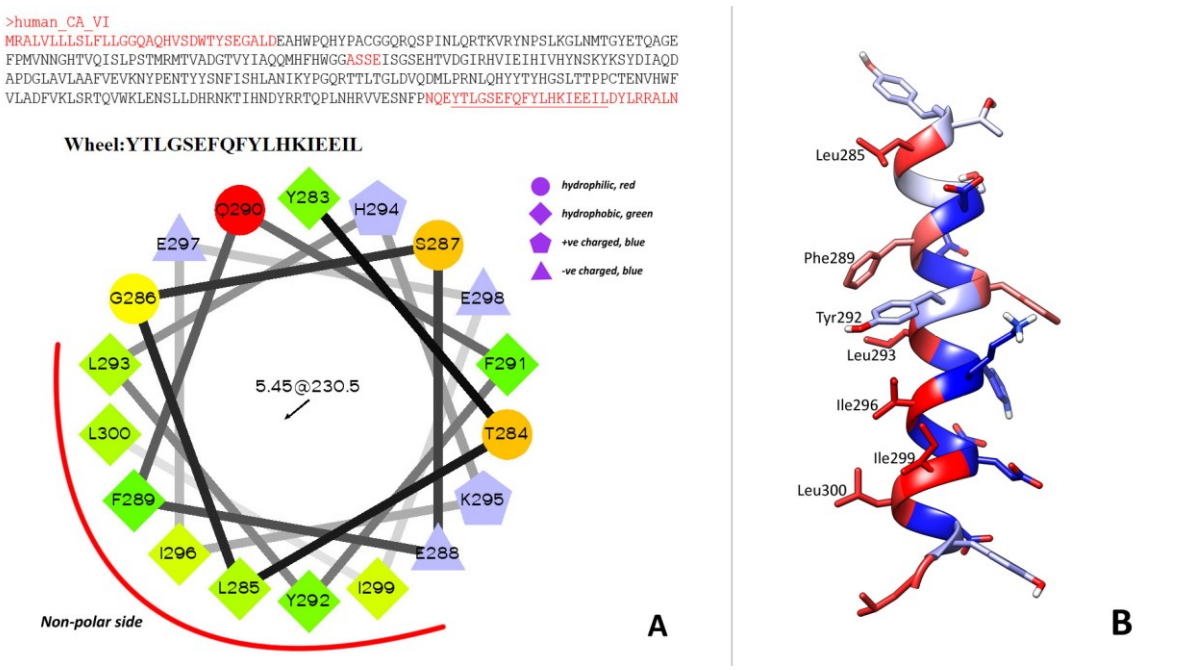


Figure 5-4. ***[A]*** *The helical wheel diagram of sequence fragment (underlined) from C-terminus of human CA VI, red colored amino acids in the sequence are residues with no structural data available in 3FE4 pdb file;* ***[B]*** *The structure of Amphipathic helix modeled by I-tasser and visualized in Chimera, the amino acids lying in the non-polar face of the helix are labelled, the residues are colored based on hydrophobicity from highest-red, mid-white to lowest-blue.*

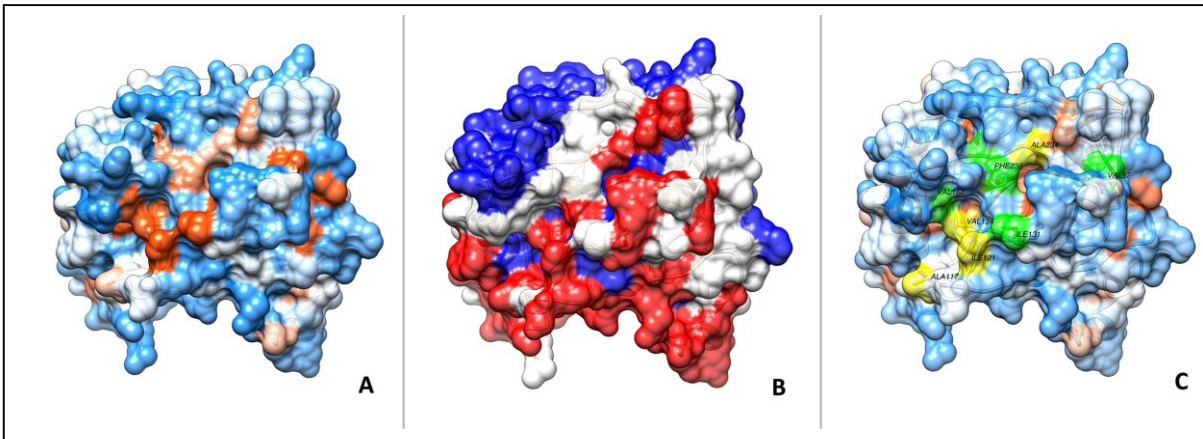


Figure 5-5. The visualization of human CA VI structure (PDB id 3FE4), the surface facing towards the reader is the C-terminus face, based on following criterias: ***[A]*** *The surface is colored according to hydropathy (red to blue represents hydrophobic to hydrophilic);* ***[B]*** *Here the surface is colored based on interacting residues predicted by CPORT (red represents active residues, white passive residues, and blue as non-interacting);* ***[C]*** *The surface is primarily colored based on hydropathy in similar way as in A, but the interacting residues from B that overlaps with the hydrophobic patches in A are highlighted here differently (yellow represents active residues and green passive residues), and the predicted interacting residues are labelled too.*

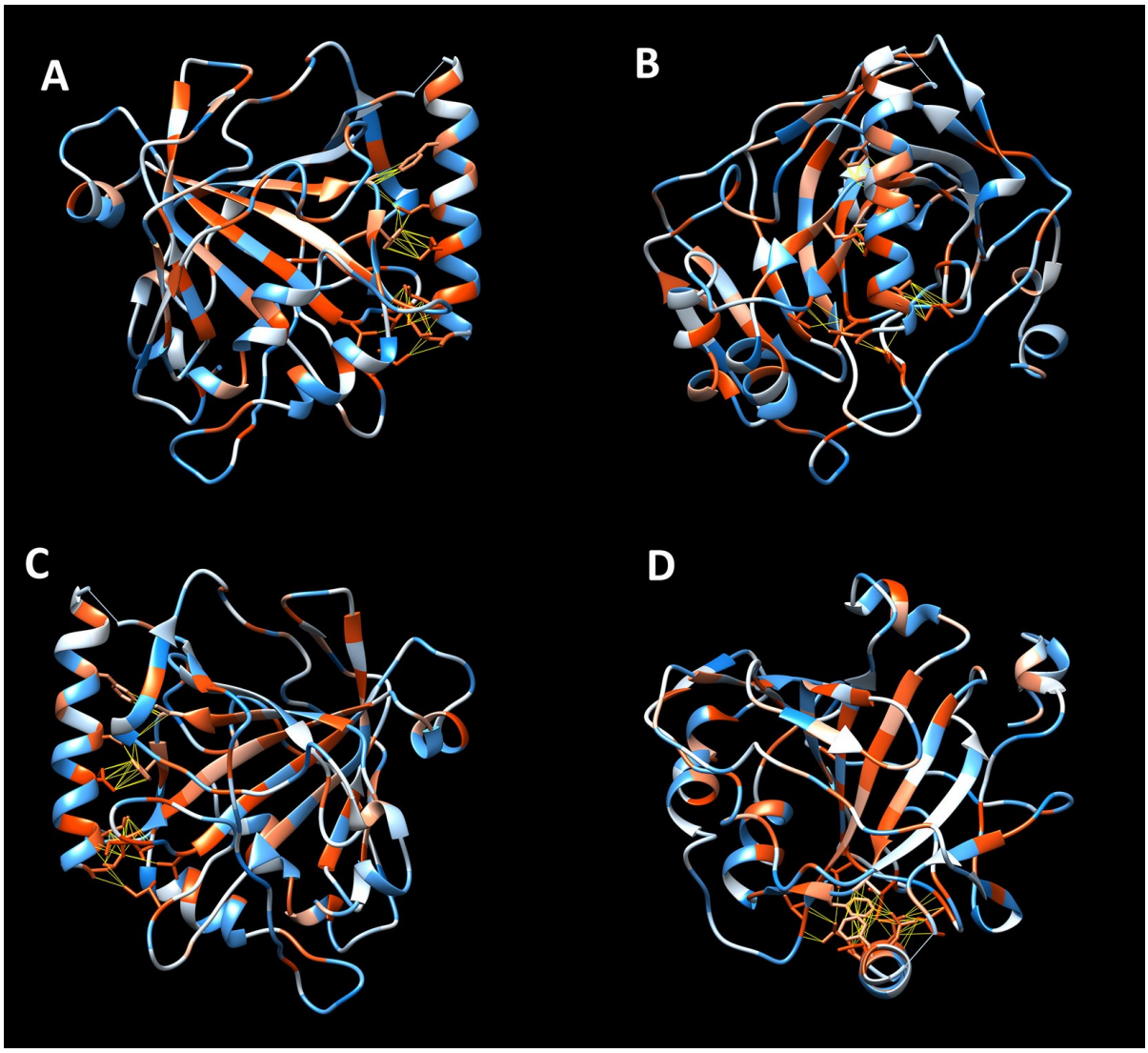


Figure 5-6. The hydrophobic contacts was calculated with default contact cutoffs of -0.4A with an allowance of 0.0A. The visualization showing the pseudobonds are shown in different angles: ***[A]*** *The side view with reference to the AP-helix;* ***[B]*** *The direct front view showing the AP-helix in front;* ***[C]*** *The 180' rotation vertically of A;* ***[D]*** *The 90' rotation horizontally of B or top view with reference to the main axis of AP-helix. The residues involved in hydrophobic contacts are displayed and the pseudobonds between them are shown as yellow connecting lines.*

5.7 The modeled 3-D structure of zebrafish CA VI with Pentraxin domain

The refined human CA VI structure consisting of AP-helix was used as the template for modeling the zebrafish CA VI domain using GUI to Modeller in Chimera. The Modeller generated five comparative models with individual statistical scores GA341 and zDOPE. Additionally, the RMSDs was measured for each of the models through structural comparison with template structure and intramolecular Clashes were calculated in Chimera. The analysis scores are shown in Table 5-2. Based on the majority of winning scores, the best comparative model #1.4 was chosen for further analyses. The model was subsequently subjected to energy minimization in Chimera with Structure minimization tool. The energy minimization was performed with 100 steps of steepest descent minimization followed by 10 steps of conjugate gradient minimization. The potential energy at the beginning of minimization was -1554.76 kJ/mol, while, at the end of the process, it was calculated as -14035.22 kJ/mol. There were no any intramolecular Clashes detected after the minimization process. Likewise, the physiochemical properties of the model were analyzed in Procheck webserver. The overall quality analysis of the model seemed to be good. There were no residues observed in the disallowed regions, 3 out of 277 residues were found in generously allowed regions, 27 in additional allowed regions while 218 in most favored regions of Ramachandran plot. The detailed result of Ramachandran plot analysis from the program is given in Appendix V.

Similarly, the Pentraxin domain of zebrafish CA VI was modeled using human SAP (4AVS) as the template in Modeller GUI in Chimera. The Modeller generated five homology models with the scores for each. And, the RMSDs with template structure and intramolecular Clashes were calculated by Matchmaker and Find Clashes/Contacts tools respectively in Chimera. The scores of the analysis are given in Table 5-3 on the basis of which the best model was chosen with the same criteria as for the CA VI domain explained above. The model #1.3 was chosen as the best model for further analyses. The selected model was subsequently subjected to energy minimization with 100 steps of steepest descent minimization followed by 10 steps of conjugate gradient minimization. The initial potential energy of the model was measured to be 7090.57 kJ/mol, while, after the minimization, it was found to be -2168.71 kJ/mol. The intramolecular Clashes reduced to 3 from 41 after the local energy minimization. The minimized structure was evaluated for physiochemical properties in Procheck webserver. The quality analysis showed 87.7% of residues were in the most favored regions, 9.4% in additional allowed regions, 2.3% in generously allowed regions and 1 residue found in disallowed region of the Ramachandran plot. The detailed graphical Ramachandran plot is given in Appendix VI.

Model	GA341	zDOPE	RMSD	Clashes
#1.1	1.00	-0.57	0.236	71
#1.2	1.00	-0.60	0.277	90
#1.3	1.00	-0.51	0.235	67
#1.4	1.00	-0.70	0.254	45
#1.5	1.00	-0.63	0.303	58

Table 5-2. Structural analysis scores of five comparative models of CA VI domain

Model	GA341	zDOPE	RMSD	Clashes
#1.1	1.00	-0.89	0.289	65
#1.2	1.00	-0.80	0.315	51
#1.3	1.00	-0.93	0.288	41
#1.4	1.00	-0.82	0.348	60
#1.5	1.00	-0.88	0.295	53

Table 5-3. Structural analysis scores of five comparative models of Pentraxin domain.

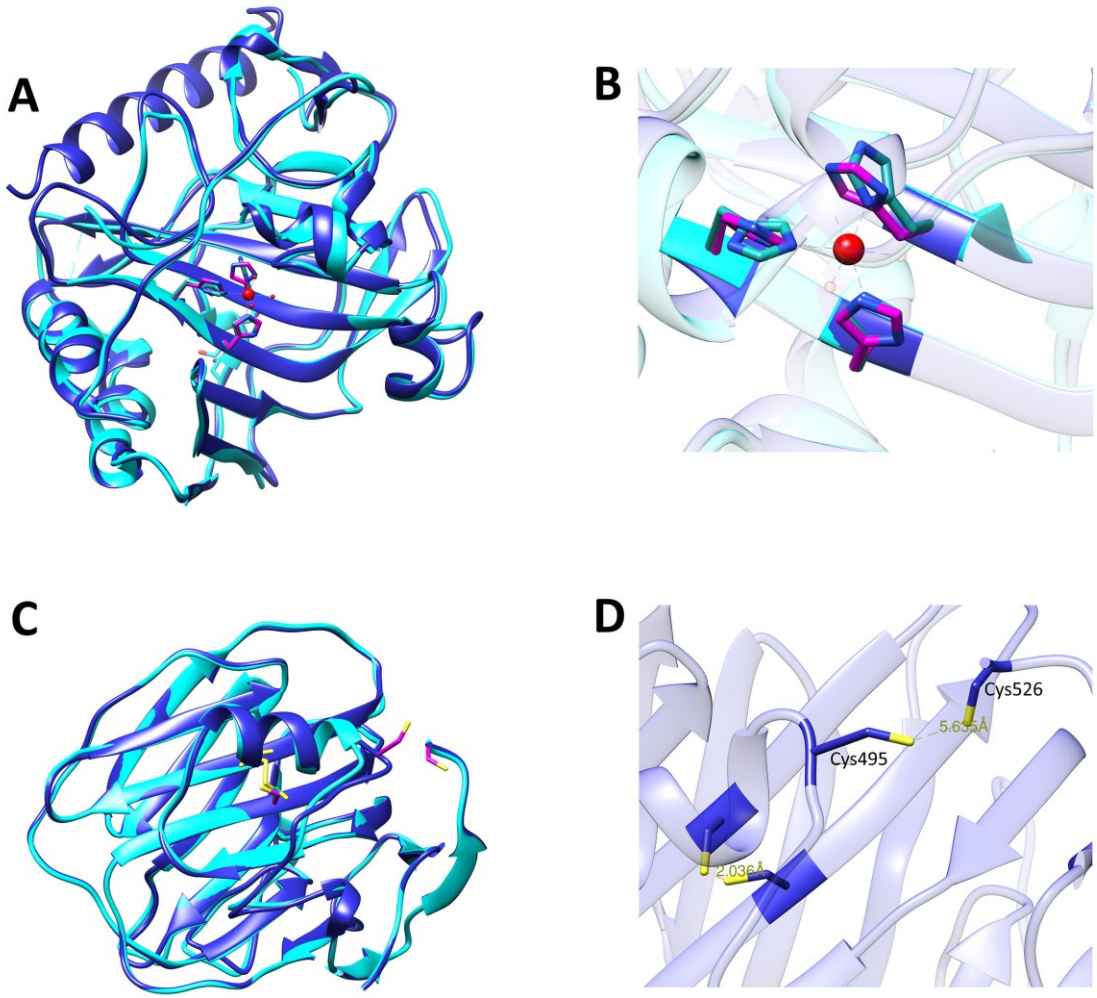


Figure 5-7. ***[A]*** *The superimposition of 3FE4 (cyan) and zebrafish CA VI model (blue) structures, showing direct view of the active site;* ***[B]*** *The magnification of the active site region highlighting the 3 Histidines of hCAVI (green) and zebrafishCAVI (magenta) actively involved in coordinating with the metal ion (red);* ***[C]*** *The superimposition of 4AVS (cyan) and zebrafish PTX domain model (blue) structures, the Cysteine are displayed;* ***[D]*** *The magnification of the region showing the Cysteine residues in zebrafish PTX domain (model #1.5), focusing on the extra pair of Cys residues (Cys495, Cys526, zebrafish CAVI numbering) conserved only in CA VI related PTX, which may form disulfide bridge.*

The models of CA VI and PTX domains after their quality assessment, was subjected to docking in HADDOCK. For input data required in docking, the information were predicted through CPORT interface prediction tool. From the list of interface residues predicted by CPORT, the residues lying in the C-terminal face of the CA VI domain and N-terminal face of the PTX domain were selected as it is highly likely that these faces would interact and also help in reducing number of sampling during the docking procedure. The parameters for active and passive residues thus selected for CA VI and PTX domains are as follows.

The active residues predicted by CPORT that were selected from the C-terminus face of CA VI domain are D120, L121, E122, G288, M289, C290, E294, L307, S308, L309, D310 and the passive residues are D131, G132, G172, G287, E293, A297, S304, L305. Similarly for Pentraxin domain, the active residues selected from the N-terminus face of the domain are P330, L331, L512, S513, L514, Y515, G516, Q519, G522 and passive residues are K325, Q326, P327, I328, N347, T510, Q511, G517, V518, L520, Q521, E523. These active and passive residues are visualized in the structure of CA VI and Pentraxin domains in Figure 5-9.

With the docking parameters in place, the two domains of zebrafish CA VI were assigned for docking in HADDOCK webserver. Normally, the docking process takes about 5 hours if there is not a queue in the server. The docking experiment projected ensembles of 127 models in 10 clusters that represented 63.5% of the water-refined models reliable according to HADDOCK. The HADDOCK score of the best cluster was -115.9 +/- 5.2 with a cluster size of 9 and Z-score of -2.5. The details about the scores are given in Figure 5-8. The Z-score of a cluster indicates how many standard deviation it is located from the average in terms of score, the more negative the better. A top model from this cluster with the lowest (best) score was chosen for further visualization and interpretation. The pseudo contacts such as Vander Waal's force, hydrophobic bonds were calculated between the two different domains and found to be 155 pseudo contacts, the parameters of this analysis is shown in Appendix VIII. The dropbox link to the pdb file of the docked complex molecule is provided in Appendix IX.

HADDOCK score	-115.9 +/- 5.2
Cluster size	9
RMSD from the overall lowest-energy structure	0.6 +/- 0.4
Van der Waals energy	-72.2 +/- 3.5
Electrostatic energy	-181.8 +/- 21.4
Desolvation energy	-15.9 +/- 4.1
Restraints violation energy	86.0 +/- 21.42
Buried Surface Area	1987.6 +/- 51.9
Z-Score	-2.5

Figure 5-8. HADDOCK score details of the best cluster from the docking experiment of CA VI and PTX domain of zebrafish.

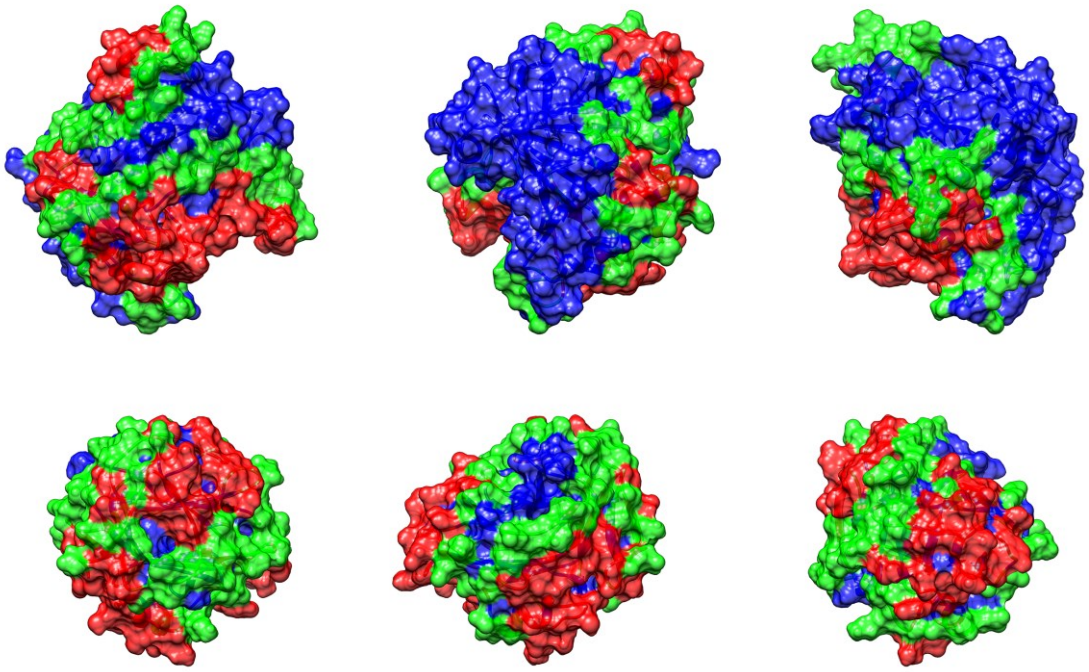


Figure 5-9. The CA VI (first row) and Pentraxin (second row) domains are visualized in three angles and the surface coloration is done based on CPORT prediction. The first column of CA VI domain is its C-terminal face and Pentraxin domain is its N-terminal face. The second and third column depicts the protein rotated 120' along vertical axis. The red color is for active residues, green for passive residues and blue for no predictions from CPORT.

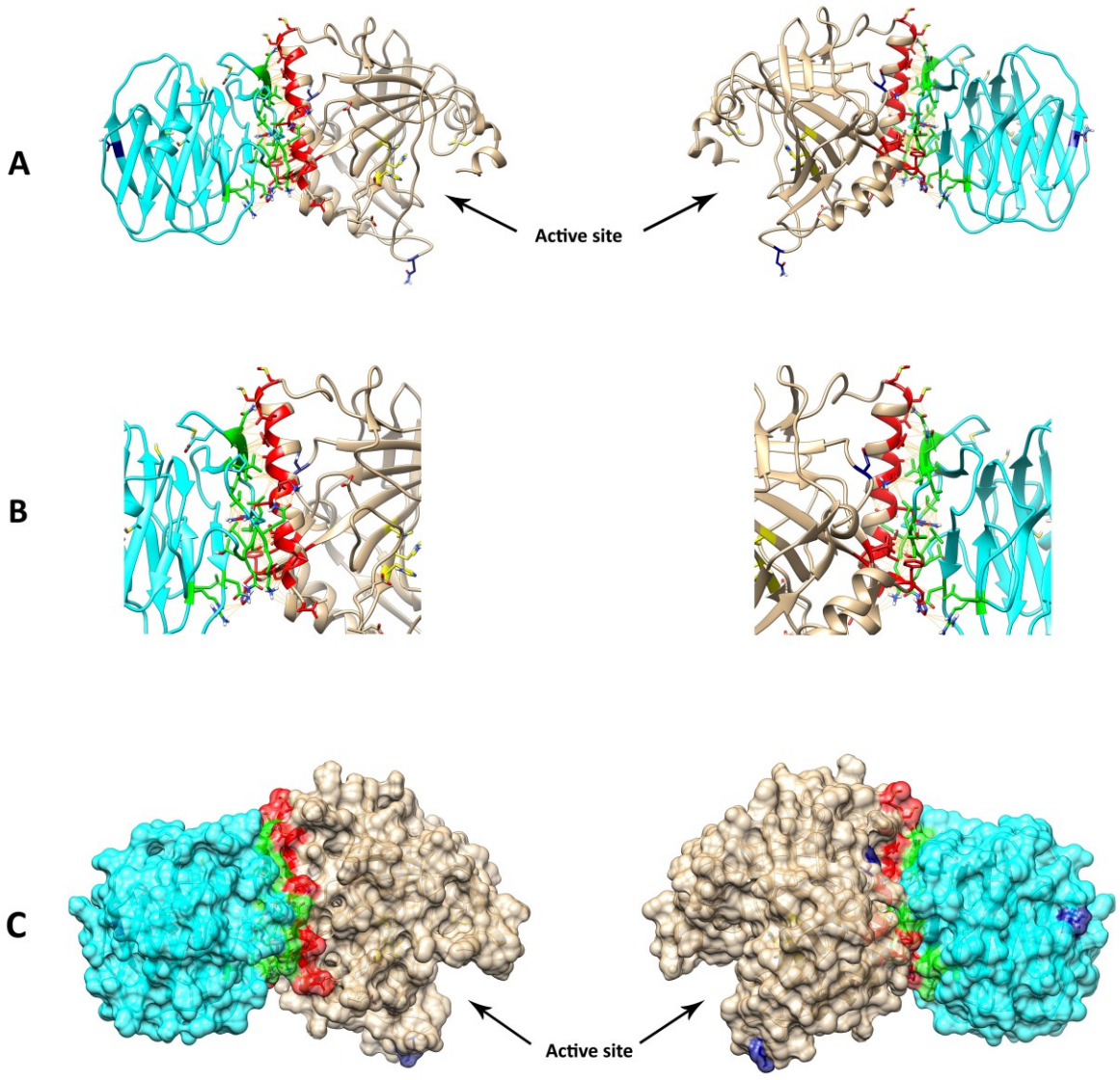


Figure 5-10. The visualization of the structure of docked complex of zebrafish CA VI with Pentraxin. The 2nd column is the horizontally rotated orientation of 1st column. The active site Histidines are shown in yellow. [A] The ribbon representation of the whole complex, CA VI domain in tan and PTX domain in cyan. [B] The magnification at the interface region of the docked complex. Interface residues from CA VI domain are in red, while from PTX domain are in green and the faint lines in yellow represents the pseudobonds. [C] Surface representation of the complex, the surface lying on the interface regions are colored red and green for CA VI and PTX domains respectively.

6 Discussion

6.1 The transmembrane CAs have possible secondary roles in nucleus

All the transmembrane CAs are destined to extracellular location while them possessing the transmembrane helices in their sequence would mean their final targeted destination would be in the cell membrane. As described in literature as well as from the topology prediction of TMHMM, the catalytic CA domain of the transmembrane CAs lie outside the cell, followed by its membrane-spanning transmembrane helix anchoring the enzyme on the cell membrane, and finally a small extension of C-terminal fragment or cytoplasmic domain lying inside the cell. All those sequences in which secretory peptide were not observed were the bad-quality assigned sequences, and most of those were either lacking the N-terminal presequence portion completely or the presequences were of bad quality that looked completely different from the other sequences. So it is plausible to state that these were merely just mispredictions caused due to bad-quality of the sequences. Hence, it is very evident to state that the transmembrane CAs are localized primarily to the plasma membrane of the cell. However, another type of signals that were detected mostly in the transmembrane and cytoplasmic regions of the sequences give an insight to the possible secondary localization of the transmembrane CAs. Although there were high degree of variation of the nucleo-cytoplasmic signal predicted among the analyzed sequences, such as some sequences having very strong signals while others having faint to no signals predicted at all, the patterns are clearly visible while inspecting the MSA visually. The signal motifs predicted very strongly in many of the sequences are highly if not fully conserved among all the analyzed sequences. The weak or no prediction in some sequences might have been caused due to various constraints the algorithms of the programs had to deal with in different sequences.

Though the possible function of the transmembrane CAs in nucleus and/or in what kind of situations do they get transported into the nucleus are yet to be discovered that would require a design of different laboratory-based studies, but from the results of nucleo-cytoplasmic predictions obtained in this thesis, we can at least interpret how the transport of proteins that are almost locked in the phospholipid bilayer of the plasma membrane would occur. From the observation, it is clearly observed that the NLS motifs are located in the cytoplasmic portion, just where the transmembrane region ends and the cytoplasmic region starts. It is plausible to expect that the nuclear transporting proteins such as Importins would have quite trivial access for recognizing these signals that are lying in the cytoplasmic area and carrying out the necessary mechanisms to pull out the proteins out of the cell membrane and transport into the nucleus where their undiscovered role would be awaiting. And the NES motifs lying in the transmembrane portions would also have no any hindrance for the Exportins from recognizing and binding with, since the transmembrane helices of the enzymes are not bound with or sealed inside any cell membrane as they are inside nucleus now.

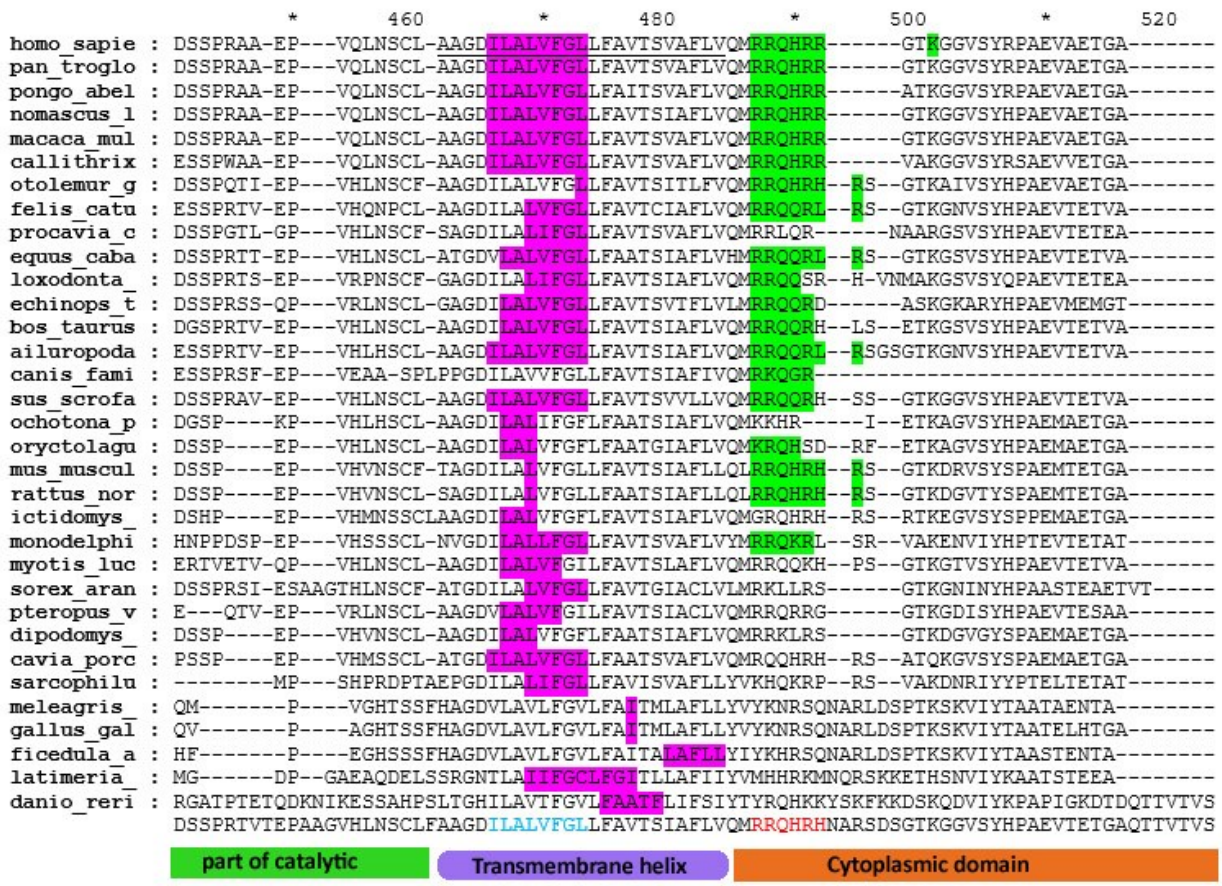


Figure 6-1. The C-terminal portion from the MSA of orthologues of CA IX and the predicted NLS (green) and NES (pink) sites highlighted in the respective sequences. The transmembrane and cytoplasmic domain boundary are indicated by the figures annotated below the alignment.

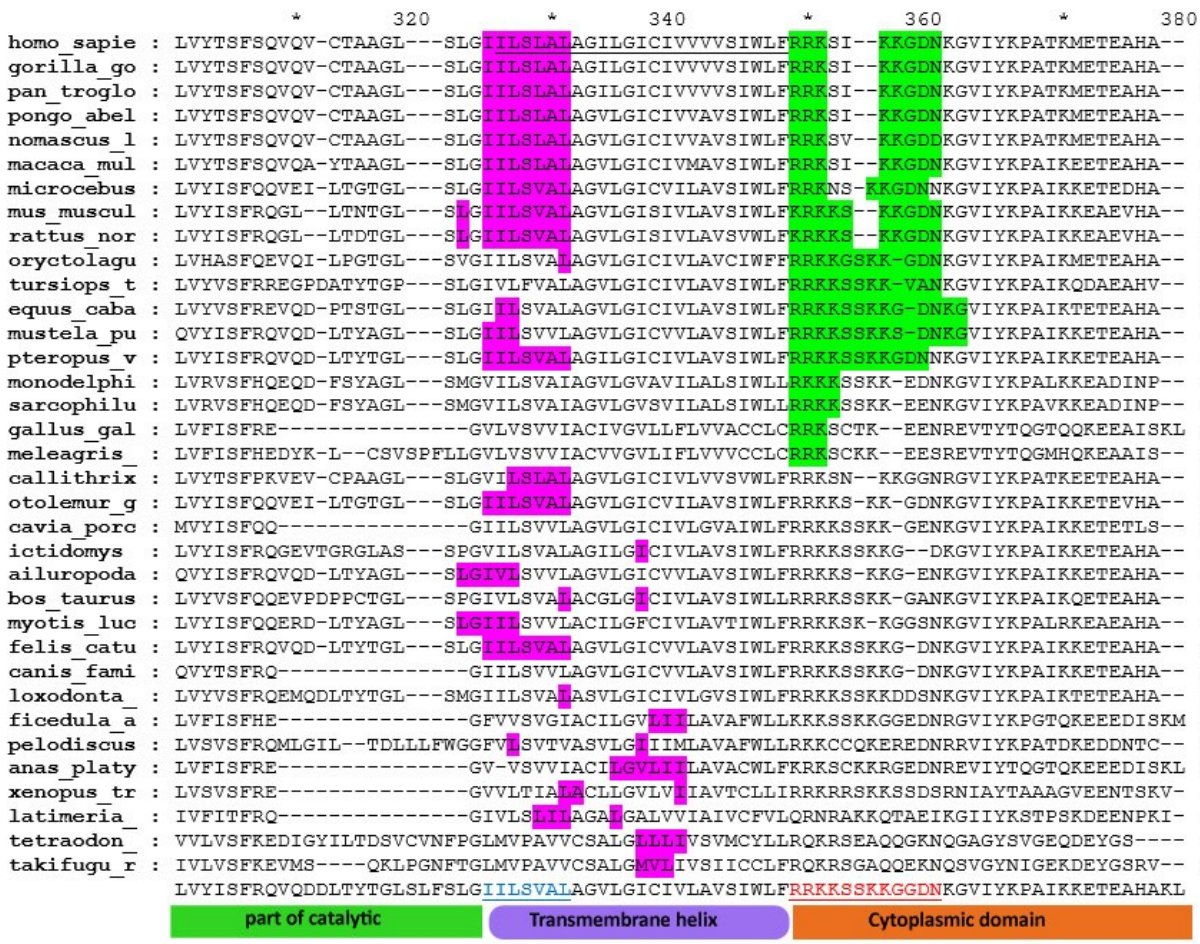


Figure 6-2. The portion from C-terminal region of the MSA of CA XII and the predicted NLS (green) and NES (pink) highlighted in respective sequences. The transmembrane and cytoplasmic domain boundary are depicted by figures below the alignment.

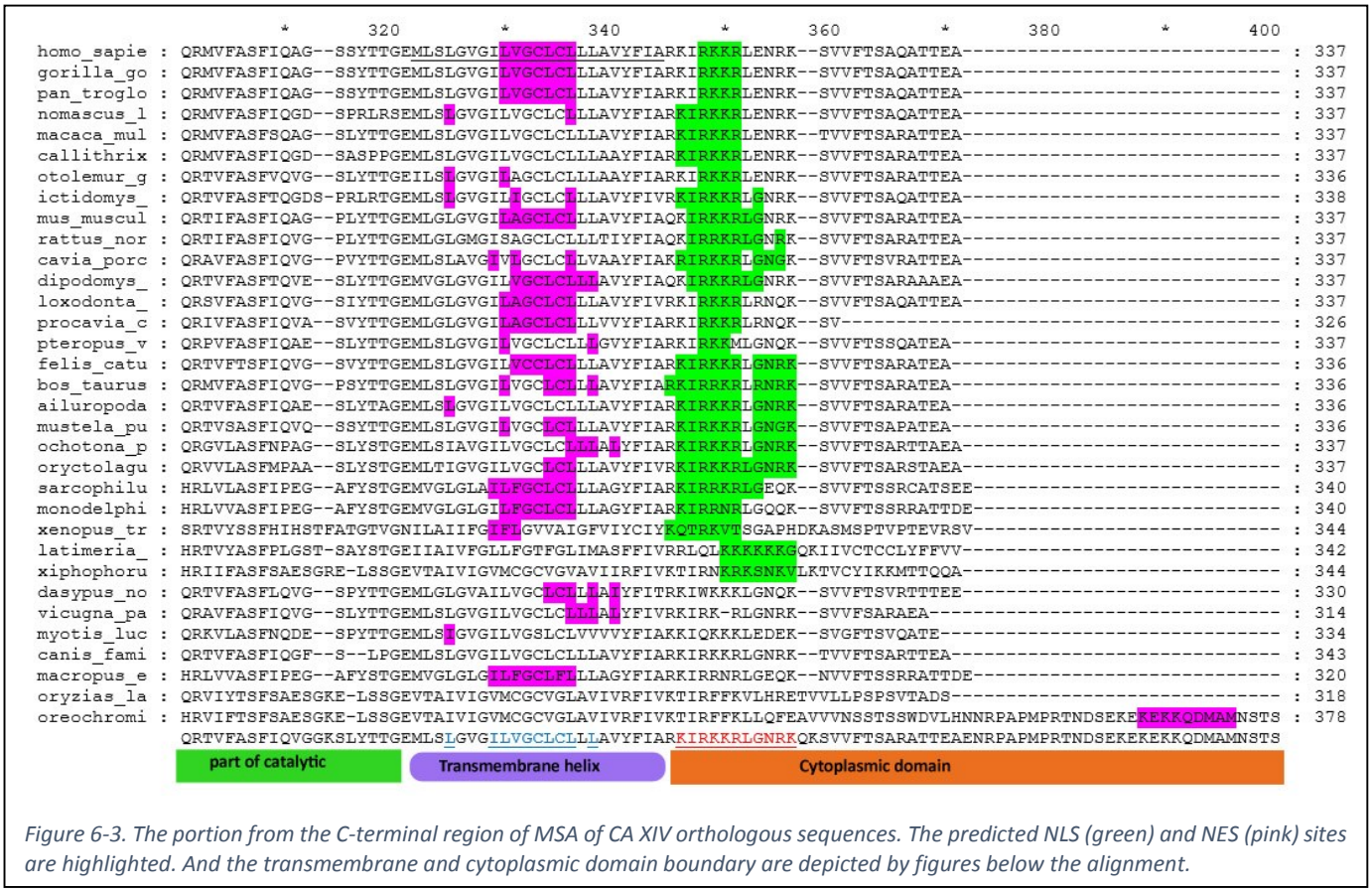


Figure 6-3. The portion from the C-terminal region of MSA of CA XIV orthologous sequences. The predicted NLS (green) and NES (pink) sites are highlighted. And the transmembrane and cytoplasmic domain boundary are depicted by figures below the alignment.

6.2 The CA XII and CA XIV can form dimers

The sequence motifs GxxxG and GxxxS have been identified as the signatures which mediate the dimerization in transmembrane alpha helices forming coiled-coil helices. The same motifs have been demonstrated to mediate helix-helix association in the NMR structures of Glycophorin A dimer (MacKenzie, Prestegard, and Engelman 1997; Smith et al. 2001). Likewise, the isologous dimer architecture of human CA XII dimer structure in x-ray crystallography study was also speculated to persist within the transmembrane segment of the enzyme where the sequence motifs are present (Whittington et al. 2001). The conservation of the sequence motifs GxxxG and GxxxS observed in the transmembrane domains of most of the orthologous sequences of CA XII and CA XIV would suggest that these isozymes can also persevere as a dimer. The stabilization of dimeric form of the transmembrane CAs, also including CA IX where Cys disulfide bridge mediates the dimerism, would support our hypothesis that these groups of CA isozymes have some role in the nucleus which probably would have something to do with DNA or else just a normal CA activity inside nucleus. From the proposed model of full-length CA IX and CA XII structures as illustrated by (Alterio et al. 2009) and (Whittington et al. 2001) and CA XIV in Figure 6-4, it is visually reflective that the architecture of the transmembrane helix nearly resembles the geometry of coiled-coil helices. The coiled-coil helices are known to have an important function in DNA-binding proteins such as transcription factors where the two helical secondary structure would interlock onto the major groove of DNA molecules. If our hypothesis of nuclear localization of transmembrane CAs, as predicted by the bioinformatics programs, are in fact true and the dimeric organization of these enzymes are stable, then the dimeric architecture of transmembrane helix would possibly function as coiled-coil helix for binding with DNA in the nucleus, which would mean that the transmembrane CAs might have a role as transcription factor mediating transcription of some unsuspecting proteins, maybe those proteins that have important role in tumorigenesis. However, these are just our hypothesis.

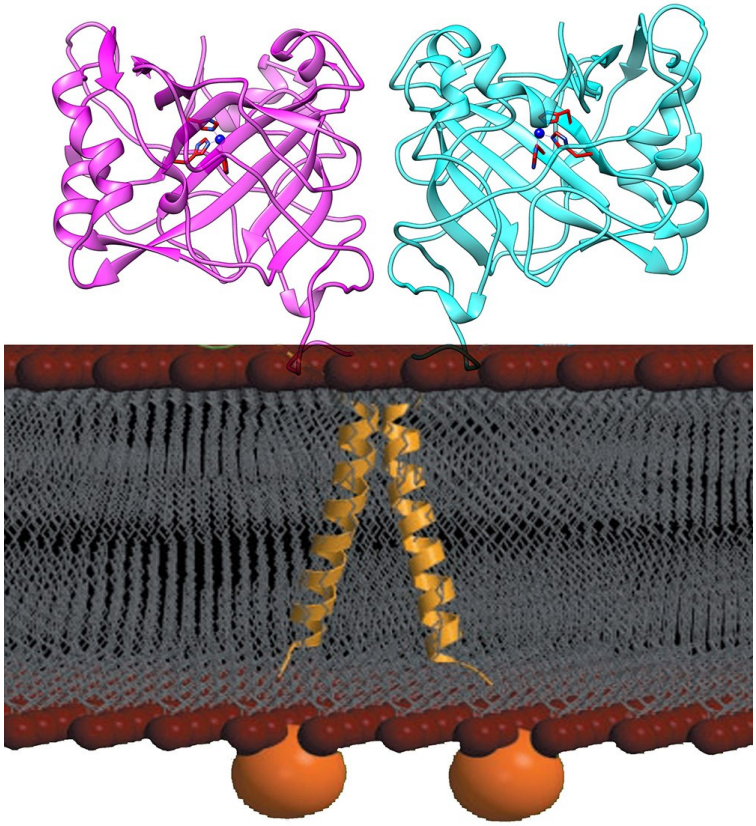


Figure 6-4. Schematic representation of a hypothetical model of full length CA XIV dimer, CA XIV structure from pdb: 4LU3 (Alterio et al. 2014). The two extracellular catalytic domains are shown in magenta and cyan along with active site Histidines (red) and Zinc (blue). The extra-drawings of transmembrane helix (yellow) and cytoplasmic domain (orange spheres) and the plasma membrane was derived from (Whittington et al. 2001). The dimerization motifs GxxxG and GxxxS in the transmembrane helices would mediate the dimer organization in the two moieties.

6.3 The zebrafish CA VI has double domain and may exist as oligomer

The novel type of domain discovered in the CA VI of non-mammalian vertebrates has provided new insight into the CA evolution of the higher animals. This novel domain that are related to Pentraxins is also new to the Pentraxin family of proteins which have never been characterized before. Though not yet clearly known, the Pentraxin proteins potentially have some roles in the immune system. These facts make it interesting that the CA VI - related Pentraxins might also possess a potential functionality in the immune system in addition to their regular CA activity. And moreover, a structural insight would provide researchers to speculate or investigate any hypothesis regarding the undiscovered physiological role of these isozymes. The computer-generated model of full CA VI of zebrafish along with the associated Pentraxin domain is the first ever 3-D model of any CA VI isozyme which have the Pentraxin domain in its sequence. The organism zebrafish was chosen here in this study as it is a model organism that has been popularly in the interest of molecular biologists and hence it could serve as an important model for any structural interpretation for the researchers. As observed from the docking experiment performed computationally in this thesis, the model of CA VI with Pentraxin domain looks to be stable since the catalytic pocket of CA domain is as equally accessible as it would be for the lone CA domain, as well as the N-linked glycosylation sites of both Pentraxin and CA VI domains also lie on the extreme surfaces of the bi-domain structure exposing it to the sugar molecules. This could provide evidence confirming that the enzyme is functional as a double domain protein in zebrafish. Adding to this, the experimental structures of the Pentraxin proteins in the PDB database are mostly found to exist as oligomeric organizations of three or five or ten monomers (Chen et al. 2015; Kolstoe et al. 2014; Ramadan et al. 2002). Hence, it is plausible to say that the Pentraxin domain-bearing CA VI proteins in non-mammalian vertebrates might also organize in the similar fashion in nature. Then, the Pentraxin domains would organize as a certain oligomer while the CA VI domains would lie hanging from the Pentraxin domains around the circumference of oligomer structure.

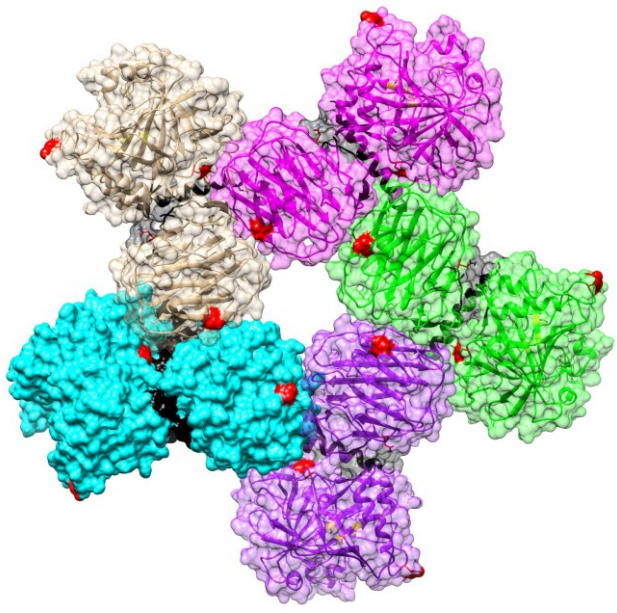


Figure 6-5. An ideal pentameric model of the zebrafish CA VI+PTX complex, as constructed by aligning 5 units with each of the subunit of human SAP structure (4AVS). Each particular units are in distinct colors (cyan, tan, magenta, green, purple), the interface regions between the CA VI and PTX domains are in black, the N-glycosylation sites are shown in red.

6.4 Possible sources of error

The main task in this study involved the structural modeling of a protein through the computational approach. The homology modeling as well as docking themselves are not the ultimate solutions to determine a structural model of any macromolecules, rather it is a method to predict a near-close naturally resembling structure. Although, the computational approaches have gone through major development and today's softwares are far more powerful and reliable in performing such tasks, but they still require numerous improvements from methodological point of view. Relying completely on computational model for most interpretation is not recommended, however, the models produced from such methods can still make a good use in providing idea to experimental biologists regarding how their experiment setup could be designed so as to minimize chances of any unsuccessful attempt. The interface predicting tool CPORT also basically over-predicts the interface residues of the protein structure, which could be another issue that can cause unnecessary sampling while docking resulting in poorly scored structures. Hence, a manual intervention was done mainly to omit those parts from the over predicted interface residues which would normally do not seem to interact i.e. the either end and start terminal surfaces of each interacting molecules would face each other directly while interaction, thus only those residues from C-terminal surface of CA VI domain and N-terminal surface of Pentraxin domain were selected. This enforcement is completely hypothetical and does not guarantee the actual orientation, however, it is so far a near possible orientation. Nevertheless, the docked model of the full length of zebrafish CA VI with Pentraxin can help researchers in our research group and also other groups working on this protein into studying various structural properties in the absence of its experimental structure. Moreover, this computational model is a first ever model of any CA isozyme having a recently discovered novel Pentraxin domain, which in itself is a new member in the Pentraxin family of proteins.

The possibility of errors in the sequence analysis part was clearly observed. Even though it was observed from the sequence alignment that the motifs predicted very strongly in most sequences was conserved throughout all the orthologues, there were a considerable amount of mispredictions clearly seen in the highly conserved motifs. There could be several reasons behind this that has to do with the technicality of the algorithms of the program. It also depends on how input sequences were fed in the program such as in NetNES, only last 60 amino acids residues from each of the proteins were fed so as to reduce the restrain for the program to have to deal with lesser amount of data. However, the mispredictions do not mean that the signals are definitely not present in those sequences. Again, the bioinformatics tools are mostly used for predicting, not as an ultimate confirmation solution. Henceforth, experimental investigations to the nucleo-cytoplasmic localization of transmembrane CAs should be greatly anticipated in the future research on these isozymes.

7 Conclusion

The primary goal of this study was to identify nucleo-cytoplasmic localization signals in the transmembrane CAs: CA IX, CA XII and CA XIV. The bioinformatics tools NetNES and NucPred were used to predict the NES and NLS sequence motifs respectively in a datasets of orthologous sequences from vertebrate species of each isozyme form. The results obtained have given new insight into the possible secondary roles of the transmembrane family of CAs. The elusiveness of NES and NLS signals that were lying hidden to researchers for decades have been identified from not just a single, but multiple vertebrate species which shows that these signals have been conserved early on evolution. Moreover, these signals have been identified in the transmembrane and cytoplasmic regions of these isozymes that are a distinct features of these isozyme family, whose functional characterization was largely lacking beside its function in anchoring the proteins in the cell membrane. The findings have opened a new door to investigations for understanding these possible nuclear roles of the transmembrane CAs that could be achieved through experiments.

The other main goal of this study was to model a full-length structure of zebrafish CA VI consisting of the Pentraxin domain along with the regular CA domain. The both softwares MODELLER and HADDOCK are among the best in comparative modeling and docking macromolecules respectively. Each of the domains were modeled computationally through homology modeling using templates structures of their respective homologous proteins and subsequently docked to generate a complete model of the macromolecular complex of CA VI and Pentraxin enzyme of zebrafish *Danio rerio*, an experimental model organism. The model of the complex would be useful for researchers to investigate structural properties of the CA isozyme that has the novel Pentraxin domain and may help to interpret the functional role that Pentraxin domain could interplay in normal functionality of secretory CA VI isozyme of non-mammalian vertebrates.

8 Bibliography

Alber, B. E. and J. G. Ferry. 1994. "A Carbonic Anhydrase from the Archaeon Methanosarcina Thermophila." *Proceedings of the National Academy of Sciences of the United States of America* 91 (15): 6909-6913.

Aldred, P., P. Fu, G. Barrett, J. D. Penschow, R. D. Wright, J. P. Coghlan, and R. T. Fernley. 1991. "Human Secreted Carbonic Anhydrase: cDNA Cloning, Nucleotide Sequence, and Hybridization Histochemistry." *Biochemistry* 30 (2): 569-575.

Almen, M. S., K. J. Nordstrom, R. Fredriksson, and H. B. Schioth. 2009. "Mapping the Human Membrane Proteome: A Majority of the Human Membrane Proteins can be Classified According to Function and Evolutionary Origin." *BMC Biology* 7: 50-7007-7-50.

Alterio, Vincenzo, Mika Hilvo, Anna Di Fiore, Claudiu T. Supuran, Peiwen Pan, Seppo Parkkila, Andrea Scaloni, et al. 2009. "Crystal Structure of the Catalytic Domain of the Tumor-Associated Human Carbonic Anhydrase IX." *Proceedings of the National Academy of Sciences}* 106 (38): 16233-16238.

Alterio, Vincenzo, Peiwen Pan, Seppo Parkkila, Martina Buonanno, Claudiu T. Supuran, Simona M. Monti, and Giuseppina De Simone. 2014. "The Structural Comparison between Membrane-Associated Human Carbonic Anhydrases Provides Insights into Drug Design of Selective Inhibitors." *Biopolymers* 101 (7): 769-778.

Andreeva, A., D. Howorth, S. E. Brenner, T. J. Hubbard, C. Chothia, and A. G. Murzin. 2004. "SCOP Database in 2004: Refinements Integrate Structure and Sequence Family Data." *Nucleic Acids Research* 32 (Database issue): D226-9.

Aspatwar, A., M. E. Tolvanen, and S. Parkkila. 2010. "Phylogeny and Expression of Carbonic Anhydrase-Related Proteins." *BMC Molecular Biology* 11: 25-2199-11-25.

Aszodi, A. and W. R. Taylor. 1996. "Homology Modelling by Distance Geometry." *Folding & Design* 1 (5): 325-334.

Back, J. W., L. de Jong, A. O. Muijsers, and C. G. de Koster. 2003. "Chemical Cross-Linking and Mass Spectrometry for Protein Structural Modeling." *Journal of Molecular Biology* 331 (2): 303-313.

Barker, Harlan Reid. 2013. "Development of a Protein Conservation Analysis Pipeline and Application to Carbonic Anhydrase IV." Master of Science, Institute of Biomedical Technology, University of Tampere.

Bax, A. 2003. "Weak Alignment Offers New NMR Opportunities to Study Protein Structure and Dynamics." *Protein Science : A Publication of the Protein Society* 12 (1): 1-16.

Berman, H. M., J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. 2000. "The Protein Data Bank." *Nucleic Acids Research* 28 (1): 235-242.

Blackshields, G., F. Sievers, W. Shi, A. Wilm, and D. G. Higgins. 2010. "Sequence Embedding for Fast Construction of Guide Trees for Multiple Sequence Alignment." *Algorithms for Molecular Biology : AMB* 5: 21-7188-5-21.

Blundell, T. L., B. L. Sibanda, M. J. E. Sternberg, and J. M. Thornton. 1987. "Knowledge-Based Prediction of Protein Structures and the Design of Novel Molecules." *Nature* 326 (6111): 347-352.

Borthwick, K. J., N. Kandemir, R. Topaloglu, U. Kornak, A. Bakkaloglu, N. Yordam, S. Ozen, et al. 2003. "A Phenocopy of CAII Deficiency: A Novel Genetic Explanation for Inherited Infantile Osteopetrosis with Distal Renal Tubular Acidosis." *Journal of Medical Genetics* 40 (2): 115-121.

Brameier, M., A. Krings, and R. M. MacCallum. 2007. "NucPred--Predicting Nuclear Localization of Proteins." *Bioinformatics (Oxford, England)* 23 (9): 1159-1160.

Breviario, F., E. M. d'Aniello, J. Golay, G. Peri, B. Bottazzi, A. Bairoch, S. Saccone, R. Marzella, V. Predazzi, and M. Rocchi. 1992. "Interleukin-1-Inducible Genes in Endothelial Cells. Cloning of a New Gene Related to C-Reactive Protein and Serum Amyloid P Component." *The Journal of Biological Chemistry* 267 (31): 22190-22197.

Brocklehurst, S. M. and R. N. Perham. 1993. "Prediction of the Three-Dimensional Structures of the Biotinylated Domain from Yeast Pyruvate Carboxylase and of the Lipoylated H-Protein from the Pea Leaf Glycine Cleavage System: A New Automated Method for the Prediction of Protein Tertiary Structure." *Protein Science : A Publication of the Protein Society* 2 (4): 626-639.

Browne, W. J., A. C. North, D. C. Phillips, K. Brew, T. C. Vanaman, and R. L. Hill. 1969. "A Possible Three-Dimensional Structure of Bovine Alpha-Lactalbumin Based on that of Hen's Egg-White Lysozyme." *Journal of Molecular Biology* 42 (1): 65-86.

Buanne, P., G. Renzone, F. Monteleone, M. Vitale, S. M. Monti, A. Sandomenico, C. Garbi, et al. 2013. "Characterization of Carbonic Anhydrase IX Interactome Reveals Proteins Assisting its Nuclear Localization in Hypoxic Cells." *Journal of Proteome Research* 12 (1): 282-292.

Carter, N. D., R. Heath, S. Jeffery, M. J. Jackson, D. J. Newham, and R. H. Edwards. 1983. "Carbonic Anhydrase III in Duchenne Muscular Dystrophy." *Clinica Chimica Acta; International Journal of Clinical Chemistry* 133 (2): 201-208.

Chen, R., J. Qi, H. Yuan, Y. Wu, W. Hu, and C. Xia. 2015. "Crystal Structures for Short-Chain Pentraxin from Zebrafish Demonstrate a Cyclic Trimer with New Recognition and Effector Faces." *Journal of Structural Biology* 189 (3): 259-268.

Cho, Yun Sung, Li Hu, Haolong Hou, Hang Lee, Jiaohui Xu, Soowhan Kwon, Sukhun Oh, et al. 2013. "The Tiger Genome and Comparative Analysis with Lion and Snow Leopard Genomes." *Nat Commun* 4.

Chook, Y. M. and G. Blobel. 2001. "Karyopherins and Nuclear Import." *Current Opinion in Structural Biology* 11 (6): 703-715.

Chothia, C. and A. M. Lesk. 1986. "The Relation between the Divergence of Sequence and Structure in Proteins." *The EMBO Journal* 5 (4): 823-826.

Clackson, T. and J. A. Wells. 1995. "A Hot Spot of Binding Energy in a Hormone-Receptor Interface." *Science (New York, N.Y.)* 267 (5196): 383-386.

Claessens, M., E. Van Cutsem, I. Lasters, and S. Wodak. 1989. "Modelling the Polypeptide Backbone with 'Spare Parts' from Known Protein Structures." *Protein Engineering* 2 (5): 335-345.

Cock, Peter J. A., Tiago Antao, Jeffrey T. Chang, Brad A. Chapman, Cymon J. Cox, Andrew Dalke, Iddo Friedberg, et al. 2009. "Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics." *Bioinformatics* 25 (11): 1422-1423.

Costa, Michael J., Assane Ndoye, and John D. Trelford. 1995. "MN Protein Immunolocalization in Uterine Cervix Carcinoma with Glandular Differentiation: A Clinicopathologic Study of a New Cancer-Specific Biomarker." *International Journal of Surgical Pathology}* 3 (2): 73-82.

Dawes, C. 1972. "Circadian Rhythms in Human Salivary Flow Rate and Composition." *The Journal of Physiology* 220 (3): 529-545.

de Vries, S. J. and A. M. Bonvin. 2011. "CPORT: A Consensus Interface Predictor and its Performance in Prediction-Driven Docking with HADDOCK." *PloS One* 6 (3): e17695.

de Vries, Sjoerd J., Marc van Dijk, and Alexandre M. J. J. Bonvin. 2010. "The HADDOCK Web Server for Data-Driven Biomolecular Docking." *Nat.Protocols* 5 (5): 883-897.

DeLano, W. L. 2002. "Unraveling Hot Spots in Binding Interfaces: Progress and Challenges." *Current Opinion in Structural Biology* 12 (1): 14-20.

Deshpande, N., K. J. Addess, W. F. Bluhm, J. C. Merino-Ott, W. Townsend-Merino, Q. Zhang, C. Knezevich, et al. 2005. "The RCSB Protein Data Bank: A Redesigned Query System and Relational Database Based on the mmCIF Schema." *Nucleic Acids Research* 33 (Database issue): D233-7.

Dietmann, S., J. Park, C. Notredame, A. Heger, M. Lappe, and L. Holm. 2001. "A Fully Automatic Evolutionary Classification of Protein Folds: Dali Domain Dictionary Version 3." *Nucleic Acids Research* 29 (1): 55-57.

Dingwall, C., J. Robbins, S. M. Dilworth, B. Roberts, and W. D. Richardson. 1988. "The Nucleoplasmin Nuclear Location Sequence is Larger and More Complex than that of SV-40 Large T Antigen." *The Journal of Cell Biology* 107 (3): 841-849.

Doege, K. J., M. Sasaki, T. Kimura, and Y. Yamada. 1991. "Complete Coding Sequence and Deduced Primary Structure of the Human Cartilage Large Aggregating Proteoglycan, Aggrecan. Human-Specific Repeats, and Additional Alternatively Spliced Forms." *The Journal of Biological Chemistry* 266 (2): 894-902.

Dominguez, C., R. Boelens, and A. M. Bonvin. 2003. "HADDOCK: A Protein-Protein Docking Approach Based on Biochemical Or Biophysical Information." *Journal of the American Chemical Society* 125 (7): 1731-1737.

Du, A. L., H. M. Ren, C. Z. Lu, J. L. Tu, C. F. Xu, and Y. A. Sun. 2009. "Carbonic Anhydrase III is Insufficient in Muscles of Myasthenia Gravis Patients." *Autoimmunity* 42 (3): 209-215.

Dunbrack, R. L.,Jr, D. L. Gerloff, M. Bower, X. Chen, O. Lichtarge, and F. E. Cohen. 1997. "Meeting Review: The Second Meeting on the Critical Assessment of Techniques for Protein Structure Prediction (CASP2), Asilomar, California, December 13-16, 1996." *Folding & Design* 2 (2): R27-42.

Eddy, S. R. 1998. "Profile Hidden Markov Models." *Bioinformatics (Oxford, England)* 14 (9): 755-763.

Eswar, N., B. Webb, M. A. Marti-Renom, M. S. Madhusudhan, D. Eramian, M. Y. Shen, U. Pieper, and A. Sali. 2006. "Comparative Protein Structure Modeling using Modeller." *Current Protocols in Bioinformatics / Editoral Board, Andreas D.Baxevanis ...[Et Al.]* Chapter 5: Unit 5.6.

Felsenstein, Joseph. 1985. "Confidence Limits on Phylogenies: An Approach using the Bootstrap." *Evolution* 39 (4): 783-791.

Fernley, R. T., R. D. Wright, and J. P. Coghlan. 1979. "A Novel Carbonic Anhydrase from the Ovine Parotid Gland." *FEBS Letters* 105 (2): 299-302.

Flicek, Paul, Ikhlak Ahmed, M. Ridwan Amode, Daniel Barrell, Kathryn Beal, Simon Brent, Denise Carvalho-Silva, et al. 2013. "Ensembl 2013." *Nucleic Acids Research* 41 (D1): D48-D55.

Fujikawa-Adachi, K., I. Nishimori, T. Taguchi, and S. Onishi. 1999a. "Human Mitochondrial Carbonic Anhydrase VB. cDNA Cloning, mRNA Expression, Subcellular Localization, and Mapping to Chromosome X." *The Journal of Biological Chemistry* 274 (30): 21228-21233.

Fujikawa-Adachi, K., I. Nishimori, T. Taguchi, K. Yuri, and S. Onishi. 1999b. "cDNA Sequence, mRNA Expression, and Chromosomal Localization of Human Carbonic Anhydrase-Related Protein, CA-RP XI." *Biochimica Et Biophysica Acta* 1431 (2): 518-524.

Gabb, H. A., R. M. Jackson, and M. J. Sternberg. 1997. "Modelling Protein Docking using Shape Complementarity, Electrostatics and Biochemical Information." *Journal of Molecular Biology* 272 (1): 106-120.

Garcia, R. A., D. Pantazatos, and F. J. Villarreal. 2004. "Hydrogen/Deuterium Exchange Mass Spectrometry for Investigating Protein-Ligand Interactions." *Assay and Drug Development Technologies* 2 (1): 81-91.

Garlanda, C., B. Bottazzi, A. Bastone, and A. Mantovani. 2005. "Pentraxins at the Crossroads between Innate Immunity, Inflammation, Matrix Deposition, and Female Fertility." *Annual Review of Immunology* 23: 337-366.

Goldfarb, D. S., A. H. Corbett, D. A. Mason, M. T. Harreman, and S. A. Adam. 2004. "Importin Alpha: A Multipurpose Nuclear-Transport Receptor." *Trends in Cell Biology* 14 (9): 505-514.

Greer, J. 1981. "Comparative Model-Building of the Mammalian Serine Proteases." *Journal of Molecular Biology* 153 (4): 1027-1042.

Gribskov, M., A. D. McLachlan, and D. Eisenberg. 1987. "Profile Analysis: Detection of Distantly Related Proteins." *Proceedings of the National Academy of Sciences of the United States of America* 84 (13): 4355-4358.

Hanson, Charlotte L. and Carol V. Robinson. 2004. "Protein-Nucleic Acid Interactions and the Expanding Role of Mass Spectrometry." *Journal of Biological Chemistry* 279 (24): 24907-24910.

Havel, T. F. and M. E. Snow. 1991. "A New Method for Building Protein Conformations from Sequence Alignments with Homologues of Known Structure." *Journal of Molecular Biology* 217 (1): 1-7.

Helm, J. F., W. J. Dodds, W. J. Hogan, K. H. Soergel, M. S. Egide, and C. M. Wood. 1982. "Acid Neutralizing Capacity of Human Saliva." *Gastroenterology* 83 (1 Pt 1): 69-74.

Henkin, R. I., R. E. Lippoldt, J. Bilstad, and H. Edelhoch. 1975. "A Zinc Protein Isolated from Human Parotid Saliva." *Proceedings of the National Academy of Sciences of the United States of America* 72 (2): 488-492.

Henkin, R. I., B. M. Martin, and R. P. Agarwal. 1999. "Efficacy of Exogenous Oral Zinc in Treatment of Patients with Carbonic Anhydrase VI Deficiency." *The American Journal of the Medical Sciences* 318 (6): 392-405.

Hernández, Helena and Carol V. Robinson. 2001. "Dynamic Protein Complexes: Insights from Mass Spectrometry." *Journal of Biological Chemistry* 276 (50): 46685-46688.

Hewett-Emmett, D. and R. E. Tashian. 1996. "Functional Diversity, Conservation, and Convergence in the Evolution of the Alpha-, Beta-, and Gamma-Carbonic Anhydrase Gene Families." *Molecular Phylogenetics and Evolution* 5 (1): 50-77.

Hilvo, M., L. Baranauskiene, A. M. Salzano, A. Scaloni, D. Matulis, A. Innocenti, A. Scozzafava, et al. 2008. "Biochemical Characterization of CA IX, One of the most Active Carbonic Anhydrase Isozymes." *The Journal of Biological Chemistry* 283 (41): 27799-27809.

Hilvo, M., M. Tolvanen, A. Clark, B. Shen, GÂ N Shah, A. Waheed, P. Halmi, et al. 2005. "Characterization of CA XV, a New GPI-Anchored Form of Carbonic Anhydrase." *Biochemical Journal* 392 (Pt 1): 83-92.

Hooft, R. W., G. Vriend, C. Sander, and E. E. Abola. 1996. "Errors in Protein Structures." *Nature* 381 (6580): 272.

Hu, P. Y., E. J. Lim, J. Ciccolella, P. Strisciuglio, and W. S. Sly. 1997. "Seven Novel Mutations in Carbonic Anhydrase II Deficiency Syndrome Identified by SSCP and Direct Sequencing Analysis." *Human Mutation* 9 (5): 383-387.

Hu, P. Y., D. E. Roth, L. A. Skaggs, P. J. Venta, R. E. Tashian, P. Guibaud, and W. S. Sly. 1992. "A Splice Junction Mutation in Intron 2 of the Carbonic Anhydrase II Gene of Osteopetrosis Patients from Arabic Countries." *Human Mutation* 1 (4): 288-292.

Hu, P. Y., A. Waheed, and W. S. Sly. 1995. "Partial Rescue of Human Carbonic Anhydrase II Frameshift Mutation by Ribosomal Frameshift." *Proceedings of the National Academy of Sciences of the United States of America* 92 (6): 2136-2140.

Ikezawa, H. 2002. "Glycosylphosphatidylinositol (GPI)-Anchored Proteins." *Biological & Pharmaceutical Bulletin* 25 (4): 409-417.

Ivanov, S., S. Y. Liao, A. Ivanova, A. Danilkovitch-Miagkova, N. Tarasova, G. Weirich, M. J. Merrill, et al. 2001. "Expression of Hypoxia-Inducible Cell-Surface Transmembrane Carbonic Anhydrases in Human Cancer." *The American Journal of Pathology* 158 (3): 905-919.

Ivanov, Sergey V., Igor Kuzmin, Ming-Hui Wei, Svetlana Pack, Laura Geil, Bruce E. Johnson, Eric J. Stanbridge, and Michael I. Lerman. 1998. "Down-Regulation of Transmembrane Carbonic Anhydrases in Renal Cell Carcinoma Cell Lines by Wild-Type Von Hippel-Lindau Transgenes." *Proceedings of the National Academy of Sciences* 95 (21): 12596-12601.

Jenkins, Y., M. McEntee, K. Weis, and W. C. Greene. 1998. "Characterization of HIV-1 Vpr Nuclear Import: Analysis of Signals and Pathways." *The Journal of Cell Biology* 143 (4): 875-885.

Jiang, W. and D. Gupta. 1999. "Structure of the Carbonic Anhydrase VI (CA6) Gene: Evidence for Two Distinct Groups within the Alpha-CA Gene Family." *The Biochemical Journal* 344 Pt 2: 385-390.

Jiao, Y., J. Yan, Y. Zhao, L. R. Donahue, W. G. Beamer, X. Li, B. A. Roe, M. S. Ledoux, and W. Gu. 2005. "Carbonic Anhydrase-Related Protein VIII Deficiency is Associated with a Distinctive Lifelong Gait Disorder in Waddles Mice." *Genetics* 171 (3): 1239-1246.

Jones, T. A. and S. Thirup. 1986. "Using Known Substructures in Protein Model Building and Crystallography." *The EMBO Journal* 5 (4): 819-822.

Kaczanowski, Szymon and Piotr Zielenkiewicz. 2010. "Why Similar Protein Sequences Encode Similar Three-Dimensional Structures?" *Theoretical Chemistry Accounts* 125 (3-6): 643-650.

Karhumaa, P., S. Parkkila, A. Waheed, A. K. Parkkila, K. Kaunisto, P. W. Tucker, C. J. Huang, W. S. Sly, and H. Rajaniemi. 2000. "Nuclear NonO/p54(Nrb) Protein is a Nonclassical Carbonic Anhydrase." *The Journal of Biological Chemistry* 275 (21): 16044-16049.

Karhumaa, Pepe, Jukka Leinonen, Seppo Parkkila, Kari Kaunisto, Juha Tapanainen, and Hannu Rajaniemi. 2001. "The Identification of Secreted Carbonic Anhydrase VI as a Constitutive Glycoprotein of Human and Rat Milk." *Proceedings of the National Academy of Sciences of the United States of America* 98 (20): 11604-11608.

Kivela, J., S. Parkkila, A. K. Parkkila, and H. Rajaniemi. 1999. "A Low Concentration of Carbonic Anhydrase Isoenzyme VI in Whole Saliva is Associated with Caries Prevalence." *Caries Research* 33 (3): 178-184.

Knegtel, Ronald M. A., Rolf Boelens, and Robert Kaptein. 1994. "Monte Carlo Docking of Protein-DNA Complexes: Incorporation of DNA Flexibility and Experimental Data." *Protein Engineering* 7 (6): 761-768.

Koepp, D. M. and P. A. Silver. 1998. "Nucleocytoplasmic Transport and Cell Proliferation." *Biochimica Et Biophysica Acta* 1377 (2): M39-47.

Kolstoe, S. E., M. C. Jenvey, A. Purvis, M. E. Light, D. Thompson, P. Hughes, M. B. Pepys, and S. P. Wood. 2014. "Interaction of Serum Amyloid P Component with Hexanoyl Bis(D-Proline) (CPHPC)." *Acta Crystallographica.Section D, Biological Crystallography* 70 (Pt 8): 2232-2240.

Koza, JR. 1992. "Genetic Programming–On the Programming of Computer Programs by Natural Selection." *MIT Press* 1 (2): 12.

Krogh, A., M. Brown, I. S. Mian, K. Sjolander, and D. Haussler. 1994. "Hidden Markov Models in Computational Biology. Applications to Protein Modeling." *Journal of Molecular Biology* 235 (5): 1501-1531.

Krogh, A., B. Larsson, G. von Heijne, and E. L. Sonnhammer. 2001. "Predicting Transmembrane Protein Topology with a Hidden Markov Model: Application to Complete Genomes." *Journal of Molecular Biology* 305 (3): 567-580.

Kumar, R. S. and J. G. Ferry. 2014. "Prokaryotic Carbonic Anhydrases of Earth's Environment." *Sub-Cellular Biochemistry* 75: 77-87.

Kutay, U. and S. Guttinger. 2005. "Leucine-Rich Nuclear-Export Signals: Born to be Weak." *Trends in Cell Biology* 15 (3): 121-124.

la Cour, T., R. Gupta, K. Rapacki, K. Skriver, F. M. Poulsen, and S. Brunak. 2003. "NESbase Version 1.0: A Database of Nuclear Export Signals." *Nucleic Acids Research* 31 (1): 393-396.

la Cour, T., L. Kiemer, A. Molgaard, R. Gupta, K. Skriver, and S. Brunak. 2004. "Analysis and Prediction of Leucine-Rich Nuclear Export Signals." *Protein Engineering, Design & Selection : PEDS* 17 (6): 527-536.

Lange, A., R. E. Mills, C. J. Lange, M. Stewart, S. E. Devine, and A. H. Corbett. 2007. "Classical Nuclear Localization Signals: Definition, Function, and Interaction with Importin Alpha." *The Journal of Biological Chemistry* 282 (8): 5101-5105.

Lanman, J. and P. E. Prevelige Jr. 2004. "High-Sensitivity Mass Spectrometry for Imaging Subunit Interactions: Hydrogen/Deuterium Exchange." *Current Opinion in Structural Biology* 14 (2): 181-188.

Laskowski, R. A., M. W. MacArthur, D. S. Moss, and J. M. Thornton. 1993. "\it PROCHECK: A Program to Check the Stereochemical Quality of Protein Structures." *Journal of Applied Crystallography* 26 (2): 283-291.

Lee, B. J., A. E. Cansizoglu, K. E. Suel, T. H. Louis, Z. Zhang, and Y. M. Chook. 2006. "Rules for Nuclear Localization Sequence Recognition by Karyopherin Beta 2." *Cell* 126 (3): 543-558.

Leinonen, J., J. Kivela, S. Parkkila, A. K. Parkkila, and H. Rajaniemi. 1999. "Salivary Carbonic Anhydrase Isoenzyme VI is Located in the Human Enamel Pellicle." *Caries Research* 33 (3): 185-190.

Levitt, M. 1992. "Accurate Modeling of Protein Conformation by Automatic Segment Matching." *Journal of Molecular Biology* 226 (2): 507-533.

Lindskog, S. and J. E. Coleman. 1973. "The Catalytic Mechanism of Carbonic Anhydrase." *Proceedings of the National Academy of Sciences of the United States of America* 70 (9): 2505-2508.

Linge, J. P., M. Habeck, W. Rieping, and M. Nilges. 2003. "ARIA: Automated NOE Assignment and NMR Structure Calculation." *Bioinformatics (Oxford, England)* 19 (2): 315-316.

Liu, C., Y. Wei, J. Wang, L. Pi, J. Huang, and P. Wang. 2012. "Carbonic Anhydrases III and IV Autoantibodies in Rheumatoid Arthritis, Systemic Lupus Erythematosus, Diabetes, Hypertensive Renal Disease, and Heart Failure." *Clinical & Developmental Immunology* 2012: 354594.

Luthy, R., J. U. Bowie, and D. Eisenberg. 1992. "Assessment of Protein Models with Three-Dimensional Profiles." *Nature* 356 (6364): 83-85.

Lysenko, E., J. C. Richards, A. D. Cox, A. Stewart, A. Martin, M. Kapoor, and J. N. Weiser. 2000. "The Position of Phosphorylcholine on the Lipopolysaccharide of Haemophilus Influenzae Affects Binding and Sensitivity to C-Reactive Protein-Mediated Killing." *Molecular Microbiology* 35 (1): 234-245.

MacKenzie, K. R., J. H. Prestegard, and D. M. Engelman. 1997. "A Transmembrane Helix Dimer: Structure and Implications." *Science (New York, N.Y.)* 276 (5309): 131-133.

Mackerell, A. D.,Jr. 2004. "Empirical Force Fields for Biological Macromolecules: Overview and Issues." *Journal of Computational Chemistry* 25 (13): 1584-1604.

Mandell, J. G., V. A. Roberts, M. E. Pique, V. Kotlovyi, J. C. Mitchell, E. Nelson, I. Tsigelny, and L. F. Ten Eyck. 2001. "Protein Docking using Continuum Electrostatics and Geometric Fit." *Protein Engineering* 14 (2): 105-113.

Marti-Renom, M. A., A. C. Stuart, A. Fiser, R. Sanchez, F. Melo, and A. Sali. 2000. "Comparative Protein Structure Modeling of Genes and Genomes." *Annual Review of Biophysics and Biomolecular Structure* 29: 291-325.

McGinn, P. J. and F. M. Morel. 2008. "Expression and Regulation of Carbonic Anhydrases in the Marine Diatom Thalassiosira Pseudonana and in Natural Phytoplankton Assemblages from Great Bay, New Jersey." *Physiologia Plantarum* 133 (1): 78-91.

Melo, F. and E. Feytmans. 1998. "Assessing Protein Structures with a Non-Local Atomic Interaction Energy." *Journal of Molecular Biology* 277 (5): 1141-1152.

Melo, F., R. Sanchez, and A. Sali. 2002. "Statistical Potentials for Fold Assessment." *Protein Science : A Publication of the Protein Society* 11 (2): 430-448.

Meng, E. C., E. F. Pettersen, G. S. Couch, C. C. Huang, and T. E. Ferrin. 2006. "Tools for Integrated Sequence-Structure Analysis with UCSF Chimera." *BMC Bioinformatics* 7: 339.

Mokuno, K., S. Riku, Y. Matsuoka, I. Sobue, and K. Kato. 1985. "Serum Carbonic Anhydrase III in Progressive Muscular Dystrophy." *Journal of the Neurological Sciences* 67 (2): 223-228.

Moller, S., M. D. Croning, and R. Apweiler. 2001. "Evaluation of Methods for the Prediction of Membrane Spanning Regions." *Bioinformatics (Oxford, England)* 17 (7): 646-653.

Mori, K., Y. Ogawa, K. Ebihara, N. Tamura, K. Tashiro, T. Kuwahara, M. Mukoyama, et al. 1999. "Isolation and Characterization of CA XIV, a Novel Membrane-Bound Carbonic Anhydrase from Mouse Kidney." *The Journal of Biological Chemistry* 274 (22): 15701-15705.

Moroianu, J. 1998. "Distinct Nuclear Import and Export Pathways Mediated by Members of the Karyopherin Beta Family." *Journal of Cellular Biochemistry* 70 (2): 231-239.

Morris, Garrett M., David S. Goodsell, Robert S. Halliday, Ruth Huey, William E. Hart, Richard K. Belew, and Arthur J. Olson. 1998. "Automated Docking using a Lamarckian Genetic Algorithm and an Empirical Binding Free Energy Function." *Journal of Computational Chemistry* 19 (14): 1639-1662.

Murakami, H. and W. S. Sly. 1987. "Purification and Characterization of Human Salivary Carbonic Anhydrase." *The Journal of Biological Chemistry* 262 (3): 1382-1388.

Okada, N., Y. Ishigami, T. Suzuki, A. Kaneko, K. Yasui, R. Fukutomi, and M. Isemura. 2008. "Importins and Exportins in Cellular Differentiation." *Journal of Cellular and Molecular Medicine* 12 (5B): 1863-1871.

Opavsky, R., S. Pastorekova, V. Zelnik, A. Gibadulinova, E. J. Stanbridge, J. Zavada, R. Kettmann, and J. Pastorek. 1996. "Human MN/CA9 Gene, a Novel Member of the Carbonic Anhydrase Family: Structure and Exon to Protein Domain Relationships." *Genomics* 33 (3): 480-487.

Parisi, G., M. Perales, M. S. Fornasari, A. Colaneri, N. Gonzalez-Schain, D. Gomez-Casati, S. Zimmermann, et al. 2004. "Gamma Carbonic Anhydrases in Plant Mitochondria." *Plant Molecular Biology* 55 (2): 193-207.

Parkkila, S., A. J. Kivela, K. Kaunisto, A. K. Parkkila, J. Hakkola, H. Rajaniemi, A. Waheed, and W. S. Sly. 2002. "The Plasma Membrane Carbonic Anhydrase in Murine Hepatocytes Identified as Isozyme XIV." *BMC Gastroenterology* 2: 13.

Parkkila, S., A. K. Parkkila, T. Juvonen, and H. Rajaniemi. 1994. "Distribution of the Carbonic Anhydrase Isoenzymes I, II, and VI in the Human Alimentary Tract." *Gut* 35 (5): 646-650.

Parkkila, S., A. K. Parkkila, J. Lehtola, A. Reinila, H. J. Sodervik, M. Rannisto, and H. Rajaniemi. 1997. "Salivary Carbonic Anhydrase Protects Gastroesophageal Mucosa from Acid Injury." *Digestive Diseases and Sciences* 42 (5): 1013-1019.

Parkkila, S., A. K. Parkkila, and H. Rajaniemi. 1995. "Circadian Periodicity in Salivary Carbonic Anhydrase VI Concentration." *Acta Physiologica Scandinavica* 154 (2): 205-211.

Parkkila, S., A. K. Parkkila, H. Rajaniemi, G. N. Shah, J. H. Grubb, A. Waheed, and W. S. Sly. 2001. "Expression of Membrane-Associated Carbonic Anhydrase XIV on Neurons and Axons in Mouse and Human Brain." *Proceedings of the National Academy of Sciences of the United States of America* 98 (4): 1918-1923.

Parkkila, S., A. K. Parkkila, T. Vierjoki, T. Stahlberg, and H. Rajaniemi. 1993. "Competitive Time-Resolved Immunofluorometric Assay for Quantifying Carbonic Anhydrase VI in Saliva." *Clinical Chemistry* 39 (10): 2154-2157.

Pastorek, J., S. Pastorekova, I. Callebaut, J. P. Mornon, V. Zelnik, R. Opavsky, M. Zat'ovicova, S. Liao, D. Portetelle, and E. J. Stanbridge. 1994. "Cloning and Characterization of MN, a Human Tumor-Associated Protein with a Domain Homologous to Carbonic Anhydrase and a Putative Helix-Loop-Helix DNA Binding Segment." *Oncogene* 9 (10): 2877-2888.

Pastorekova, S., Z. Zavadova, M. Kostal, O. Babusikova, and J. Zavada. 1992. "A Novel Quasi-Viral Agent, MaTu, is a Two-Component System." *Virology* 187 (2): 620-626.

Patrikainen, Maarit Susanna. 2012. "Pentraxin-Carbonic Anhydrase VI: A Novel Multidomain Protein." Master of Science, Institute of Biomedical Technology, University of Tampere.

Pearl, F., A. Todd, I. Sillitoe, M. Dibley, O. Redfern, T. Lewis, C. Bennett, et al. 2005. "The CATH Domain Structure Database and Related Resources Gene3D and DHS Provide Comprehensive Domain Family Information for Genome Analysis." *Nucleic Acids Research* 33 (Database issue): D247-51.

Peng, J. and J. Xu. 2011. "RaptorX: Exploiting Structure Information for Protein Alignment by Statistical Inference." *Proteins* 79 (Suppl 10): 161-171.

Pepys, M. B. and G. M. Hirschfield. 2003. "C-Reactive Protein: A Critical Update." *The Journal of Clinical Investigation* 111 (12): 1805-1812.

Pettersen, E. F., T. D. Goddard, C. C. Huang, G. S. Couch, D. M. Greenblatt, E. C. Meng, and T. E. Ferrin. 2004. "UCSF Chimera--a Visualization System for Exploratory Research and Analysis." *Journal of Computational Chemistry* 25 (13): 1605-1612.

Pilka, E. S., G. Kochan, U. Oppermann, and W. W. Yue. 2012. "Crystal Structure of the Secretory Isozyme of Mammalian Carbonic Anhydrases CA VI: Implications for Biological Assembly and Inhibitor Development." *Biochemical and Biophysical Research Communications* 419 (3): 485-489.

Poon, I. K. and D. A. Jans. 2005. "Regulation of Nuclear Transport: Central Role in Development and Transformation?" *Traffic (Copenhagen, Denmark)* 6 (3): 173-186.

Ramadan, M. A., A. K. Shrive, D. Holden, D. A. Myles, J. E. Volanakis, L. J. DeLucas, and T. J. Greenhough. 2002. "The Three-Dimensional Structure of Calcium-Depleted Human C-Reactive Protein from Perfectly Twinned Crystals." *Acta Crystallographica.Section D, Biological Crystallography* 58 (Pt 6 Pt 2): 992-1001.

Robertson, N., C. Potter, and A. L. Harris. 2004. "Role of Carbonic Anhydrase IX in Human Tumor Cell Growth, Survival, and Invasion." *Cancer Research* 64 (17): 6160-6165.

Rost, B. and C. Sander. 1996. "Bridging the Protein Sequence-Structure Gap by Structure Predictions." *Annual Review of Biophysics and Biomolecular Structure* 25: 113-136.

Roth, D. E., P. J. Venta, R. E. Tashian, and W. S. Sly. 1992. "Molecular Basis of Human Carbonic Anhydrase II Deficiency." *Proceedings of the National Academy of Sciences of the United States of America* 89 (5): 1804-1808.

Rovere, P., G. Peri, F. Fazzini, B. Bottazzi, A. Doni, A. Bondanza, V. S. Zimmermann, et al. 2000. "The Long Pentraxin PTX3 Binds to Apoptotic Cells and Regulates their Clearance by Antigen-Presenting Dendritic Cells." *Blood* 96 (13): 4300-4306.

Russ, W. P. and D. M. Engelman. 2000. "The GxxxG Motif: A Framework for Transmembrane Helix-Helix Association." *Journal of Molecular Biology* 296 (3): 911-919.

Saarnio, J., S. Parkkila, A. K. Parkkila, K. Haukipuro, S. Pastorekova, J. Pastorek, M. I. Kairaluoma, and T. J. Karttunen. 1998. "Immunohistochemical Study of Colorectal Tumors for Expression of a Novel Transmembrane Carbonic Anhydrase, MN/CA IX, with Potential Value as a Marker of Cell Proliferation." *The American Journal of Pathology* 153 (1): 279-285.

Sali, A. and T. L. Blundell. 1993. "Comparative Protein Modelling by Satisfaction of Spatial Restraints." *Journal of Molecular Biology* 234 (3): 779-815.

Schlimgen, A. K., J. A. Helms, H. Vogel, and M. S. Perin. 1995. "Neuronal Pentraxin, a Secreted Protein with Homology to Acute Phase Proteins of the Immune System." *Neuron* 14 (3): 519-526.

Senes, A., M. Gerstein, and D. M. Engelman. 2000. "Statistical Analysis of Amino Acid Patterns in Transmembrane Helices: The GxxxG Motif Occurs Frequently and in Association with Beta-Branched Residues at Neighboring Positions." *Journal of Molecular Biology* 296 (3): 921-936.

Shah, G. N., G. Bonapace, P. Y. Hu, P. Strisciuglio, and W. S. Sly. 2004. "Carbonic Anhydrase II Deficiency Syndrome (Osteopetrosis with Renal Tubular Acidosis and Brain Calcification): Novel Mutations in CA2 Identified by Direct Sequencing Expand the Opportunity for Genotype-Phenotype Correlation." *Human Mutation* 24 (3): 272.

Shah, G. N., D. Hewett-Emmett, J. H. Grubb, M. C. Migas, R. E. Fleming, A. Waheed, and W. S. Sly. 2000. "Mitochondrial Carbonic Anhydrase CA VB: Differences in Tissue Distribution and Pattern of Evolution from those of CA VA Suggest Distinct Physiological Roles." *Proceedings of the National Academy of Sciences of the United States of America* 97 (4): 1677-1682.

Shen, M. and A. Sali. 2006. "Statistical Potential for Assessment and Prediction of Protein Structures." *Protein Science : A Publication of the Protein Society* 15 (11): 2507-2524.

Shima, K., K. Tashiro, N. Hibi, Y. Tsukada, and H. Hirai. 1983. "Carbonic Anhydrase-III Immunohistochemical Localization in Human Skeletal Muscle." *Acta Neuropathologica* 59 (3): 237-239.

Shrive, A. K., A. M. Metcalfe, J. R. Cartwright, and T. J. Greenhough. 1999. "C-Reactive Protein and SAP-Like Pentraxin are both Present in Limulus Polyphemus Haemolymph: Crystal Structure of Limulus SAP." *Journal of Molecular Biology* 290 (5): 997-1008.

Sievers, F., A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, et al. 2011. "Fast, Scalable Generation of High-Quality Protein Multiple Sequence Alignments using Clustal Omega." *Molecular Systems Biology* 7: 539.

Sippl, M. J. 1990. "Calculation of Conformational Ensembles from Potentials of Mean Force. an Approach to the Knowledge-Based Prediction of Local Structures in Globular Proteins." *Journal of Molecular Biology* 213 (4): 859-883.

Sippl, M. J. 1993. "Recognition of Errors in Three-Dimensional Structures of Proteins." *Proteins* 17 (4): 355-362.

Slater, G. S. and E. Birney. 2005. "Automated Generation of Heuristics for Biological Sequence Comparison." *BMC Bioinformatics* 6: 31.

Sly, W. S., S. Sato, and X. L. Zhu. 1991. "Evaluation of Carbonic Anhydrase Isozymes in Disorders Involving Osteopetrosis and/Or Renal Tubular Acidosis." *Clinical Biochemistry* 24 (4): 311-318.

Smith, K. S., C. Jakubzick, T. S. Whittam, and J. G. Ferry. 1999. "Carbonic Anhydrase is an Ancient Enzyme Widespread in Prokaryotes." *Proceedings of the National Academy of Sciences of the United States of America* 96 (26): 15184-15189.

Smith, S. O., D. Song, S. Shekar, M. Groesbeek, M. Ziliox, and S. Aimoto. 2001. "Structure of the Transmembrane Dimer Interface of Glycophorin A in Membrane Bilayers." *Biochemistry* 40 (22): 6553-6558.

Soda, H., S. Yukizane, I. Yoshida, S. Aramaki, and H. Kato. 1995. "Carbonic Anhydrase II Deficiency in a Japanese Patient Produced by a Nonsense Mutation (TAT-->TAG) at Tyr-40 in Exon 2, (Y40X)." *Human Mutation* 5 (4): 348-350.

Soda, H., S. Yukizane, I. Yoshida, Y. Koga, S. Aramaki, and H. Kato. 1996. "A Point Mutation in Exon 3 (His 107-->Tyr) in Two Unrelated Japanese Patients with Carbonic Anhydrase II Deficiency with Central Nervous System Involvement." *Human Genetics* 97 (4): 435-437.

Soding, J. 2005. "Protein Homology Detection by HMM-HMM Comparison." *Bioinformatics (Oxford, England)* 21 (7): 951-960.

Srinivasan, S., C. J. March, and S. Sudarsanam. 1993. "An Automated Method for Modeling Proteins on Known Templates using Distance Geometry." *Protein Science : A Publication of the Protein Society* 2 (2): 277-289.

Stams, T., S. K. Nair, T. Okuyama, A. Waheed, W. S. Sly, and D. W. Christianson. 1996. "Crystal Structure of the Secretory Form of Membrane-Associated Human Carbonic Anhydrase IV at 2.8-A Resolution." *Proceedings of the National Academy of Sciences of the United States of America* 93 (24): 13589-13594.

Supuran, C. T. 2008. "Carbonic Anhydrases--an Overview." *Current Pharmaceutical Design* 14 (7): 603-614.

Svastova, E., A. Hulikova, M. Rafajova, M. Zat'ovicova, A. Gibadulinova, A. Casini, A. Cecchi, et al. 2004. "Hypoxia Activates the Capacity of Tumor-Associated Carbonic Anhydrase IX to Acidify Extracellular pH." *FEBS Letters* 577 (3): 439-445.

Swietach, P., R. D. Vaughan-Jones, and A. L. Harris. 2007. "Regulation of Tumor pH and the Role of Carbonic Anhydrase 9." *Cancer Metastasis Reviews* 26 (2): 299-310.

Swinson, D. E., J. L. Jones, D. Richardson, C. Wykoff, H. Turley, J. Pastorek, N. Taub, A. L. Harris, and K. J. O'Byrne. 2003. "Carbonic Anhydrase IX Expression, a Novel Surrogate Marker of Tumor Hypoxia, is Associated with a Poor Prognosis in Non-Small-Cell Lung Cancer." *Journal of Clinical Oncology : Official Journal of the American Society of Clinical Oncology* 21 (3): 473-482.

Szalai, A. J., A. Agrawal, T. J. Greenhough, and J. E. Volanakis. 1999. "C-Reactive Protein: Structural Biology and Host Defense Function." *Clinical Chemistry and Laboratory Medicine : CCLM / FESCC* 37 (3): 265-270.

Szalai, A. J., D. E. Briles, and J. E. Volanakis. 1995. "Human C-Reactive Protein is Protective Against Fatal Streptococcus Pneumoniae Infection in Transgenic Mice." *Journal of Immunology (Baltimore, Md.: 1950)* 155 (5): 2557-2563.

Szalai, A. J., J. L. VanCott, J. R. McGhee, J. E. Volanakis, and W. H. Benjamin Jr. 2000. "Human C-Reactive Protein is Protective Against Fatal Salmonella Enterica Serovar Typhimurium Infection in Transgenic Mice." *Infection and Immunity* 68 (10): 5652-5656.

Takahashi, H., T. Nakanishi, K. Kami, Y. Arata, and I. Shimada. 2000. "A Novel NMR Method for Determining the Interfaces of Large Protein-Protein Complexes." *Nature Structural Biology* 7 (3): 220-223.

Taniuchi, K., I. Nishimori, T. Takeuchi, K. Fujikawa-Adachi, Y. Ohtsuki, and S. Onishi. 2002. "Developmental Expression of Carbonic Anhydrase-Related Proteins VIII, X, and XI in the Human Brain." *Neuroscience* 112 (1): 93-99.

Thatcher, B. J., A. E. Doherty, E. Orvisky, B. M. Martin, and R. I. Henkin. 1998. "Gustin from Human Parotid Saliva is Carbonic Anhydrase VI." *Biochemical and Biophysical Research Communications* 250 (3): 635-641.

Thiry, A., J. M. Dogne, B. Masereel, and C. T. Supuran. 2006. "Targeting Tumor-Associated Carbonic Anhydrase IX in Cancer Therapy." *Trends in Pharmacological Sciences* 27 (11): 566-573.

Tolvanen, M. E., C. Ortutay, H. R. Barker, A. Aspatwar, M. Patrikainen, and S. Parkkila. 2013. "Analysis of Evolution of Carbonic Anhydrases IV and XV Reveals a Rich History of Gene Duplications and a New Group of Isozymes." *Bioorganic & Medicinal Chemistry* 21 (6): 1503-1510.

Topham, C. M., N. Srinivasan, C. J. Thorpe, J. P. Overington, and N. A. Kalsheker. 1994. "Comparative Modelling of Major House Dust Mite Allergen Der P I: Structure Validation using an Extended Environmental Amino Acid Propensity Table." *Protein Engineering* 7 (7): 869-894.

Truant, R. and B. R. Cullen. 1999. "The Arginine-Rich Domains Present in Human Immunodeficiency Virus Type 1 Tat and Rev Function as Direct Importin Beta-Dependent Nuclear Localization Signals." *Molecular and Cellular Biology* 19 (2): 1210-1217.

Tureci, O., U. Sahin, E. Vollmar, S. Siemer, E. Gottert, G. Seitz, A. K. Parkkila, et al. 1998. "Human Carbonic Anhydrase XII: cDNA Cloning, Expression, and Chromosomal Localization of a Carbonic Anhydrase Gene that is Overexpressed in some Renal Cell Cancers." *Proceedings of the National Academy of Sciences of the United States of America* 95 (13): 7608-7613.

Turkmen, S., G. Guo, M. Garshasbi, K. Hoffmann, A. J. Alshalah, C. Mischung, A. Kuss, N. Humphrey, S. Mundlos, and P. N. Robinson. 2009. "CA8 Mutations Cause a Novel Syndrome Characterized by Ataxia and Mild Mental Retardation with Predisposition to Quadrupedal Gait." *PLoS Genetics* 5 (5): e1000487.

Unger, R., D. Harel, S. Wherland, and J. L. Sussman. 1989. "A 3D Building Blocks Approach to Analyzing and Predicting Structure of Proteins." *Proteins* 5 (4): 355-373.

van Dijk, A. D., R. Boelens, and A. M. Bonvin. 2005. "Data-Driven Docking for the Study of Biomolecular Complexes." *The FEBS Journal* 272 (2): 293-312.

Venta, P. J., R. J. Welty, T. M. Johnson, W. S. Sly, and R. E. Tashian. 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.

Watson, P. H., S. K. Chia, C. C. Wykoff, C. Han, R. D. Leek, W. S. Sly, K. C. Gatter, P. Ratcliffe, and A. L. Harris. 2003. "Carbonic Anhydrase XII is a Marker of Good Prognosis in Invasive Breast Carcinoma." *British Journal of Cancer* 88 (7): 1065-1070.

Weiser, J. N., N. Pan, K. L. McGowan, D. Musher, A. Martin, and J. Richards. 1998. "Phosphorylcholine on the Lipopolysaccharide of Haemophilus Influenzae Contributes to Persistence in the Respiratory Tract and Sensitivity to Serum Killing Mediated by C-Reactive Protein." *The Journal of Experimental Medicine* 187 (4): 631-640.

Whittington, D. A., A. Waheed, B. Ulmasov, G. N. Shah, J. H. Grubb, W. S. Sly, and D. W. Christianson. 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.

Xu, Y., L. Feng, P. D. Jeffrey, Y. Shi, and F. M. Morel. 2008. "Structure and Metal Exchange in the Cadmium Carbonic Anhydrase of Marine Diatoms." *Nature* 452 (7183): 56-61.

Yates, Andrew, Kathryn Beal, Stephen Keenan, William McLaren, Miguel Pignatelli, Graham R. S. Ritchie, Magali Ruffier, Kieron Taylor, Alessandro Vullo, and Paul Flicek. 2015. "The Ensembl REST API: Ensembl Data for any Language." *Bioinformatics* 31 (1): 143-145.

Yokoya, F., N. Imamoto, T. Tachibana, and Y. Yoneda. 1999. "Beta-Catenin can be Transported into the Nucleus in a Ran-Unassisted Manner." *Molecular Biology of the Cell* 10 (4): 1119-1131.

Zavada, J., Z. Zavadova, S. Pastorekova, F. Ciampor, J. Pastorek, and V. Zelnik. 1993. "Expression of MaTu-MN Protein in Human Tumor Cultures and in Clinical Specimens." *International Journal of Cancer.Journal International Du Cancer* 54 (2): 268-274.

Zhang, Y. 2008. "I-TASSER Server for Protein 3D Structure Prediction." *BMC Bioinformatics* 9: 40-2105-9-40.

Zhou, H. and Y. Zhou. 2002. "Distance-Scaled, Finite Ideal-Gas Reference State Improves Structure-Derived Potentials of Mean Force for Structure Selection and Stability Prediction." *Protein Science : A Publication of the Protein Society* 11 (11): 2714-2726.

Zhu, X. L. and W. S. Sly. 1990. "Carbonic Anhydrase IV from Human Lung. Purification, Characterization, and Comparison with Membrane Carbonic Anhydrase from Human Kidney." *The Journal of Biological Chemistry* 265 (15): 8795-8801.

Zuiderweg, E. R. 2002. "Mapping Protein-Protein Interactions in Solution by NMR Spectroscopy." *Biochemistry* 41 (1): 1-7.

9 Appendices

Appendix I – TargetP results

A. TargetP results of CA IX orthologues

TargetP v1.1 prediction results
Number of query sequences: 33
Cleavage site predictions included.
Using NON-PLANT networks.

Name	Len	mTP	SP	other	Loc	RC	TPlen
Homo_sapiens	459	0.007	0.981	0.118	S	1	37
Pan_troglodytes	453	0.007	0.98	0.114	S	1	37
Pongo_abelii	465	0.01	0.922	0.275	S	2	37
Nomascus_leucogenys	447	0.009	0.966	0.166	S	2	37
Macaca_mulatta	447	0.007	0.98	0.121	S	1	37
Callithrix_jacchus	451	0.007	0.981	0.122	S	1	37
Otolemur_garnettii	455	0.008	0.985	0.103	S	1	37
Felis_catus	454	0.011	0.921	0.192	S	2	37
Procavia_capensis	451	0.006	0.992	0.063	S	1	37
Equus_caballus	443	0.006	0.957	0.242	S	2	37
Loxodonta_africana	455	0.014	0.932	0.159	S	2	37
Echinops_telfairi	454	0.038	0.543	0.432	S	5	36
Bos_taurus	449	0.015	0.922	0.154	S	2	37
Ailuropoda_melanoleuca	449	0.008	0.92	0.255	S	2	39
Canis_familiaris	440	0.008	0.97	0.16	S	1	37
Sus_scrofa	442	0.007	0.973	0.141	S	1	37
Ochotona_princeps	437	0.014	0.912	0.21	S	2	37
Oryctolagus_cuniculus	440	0.018	0.881	0.208	S	2	36
Mus_musculus	437	0.012	0.904	0.256	S	2	31
Rattus_norvegicus	437	0.006	0.982	0.117	S	1	31
Ictidomys_tridecemlineatus	437	0.004	0.986	0.115	S	1	33
Monodelphis_domestica	456	0.035	0.924	0.066	S	1	33
Myotis_lucifugus	434	0.01	0.871	0.29	S	3	37
Sorex_araneus	265	0.455	0.099	0.451	*	5	-
Pteropus_vampyrus	434	0.006	0.987	0.095	S	1	35
Dipodomys_ordii	433	0.005	0.987	0.106	S	1	33
Cavia_porcellus	448	0.04	0.613	0.539	S	5	50
Sarcophilus_harrisii	455	0.66	0.368	0.073	*	4	-
Meleagris_gallopavo	314	0.158	0.074	0.837	_	2	-
Gallus_gallus	367	0.219	0.709	0.023	S	3	21
Ficedula_albicollis	367	0.122	0.851	0.016	S	2	21
Latimeria_chalumnae	368	0.025	0.943	0.076	S	1	22
Danio_rerio	384	0.028	0.937	0.056	S	1	22
cutoff		0.78	0	0.73			

B. TargetP results of CA XII orthologues

TargetP v1.1 prediction results
Number of query sequences: 35
Cleavage site predictions included.
Using NON-PLANT networks.

Name	Len	mTP	SP	other	Loc	RC	TPlen
Homo_sapiens	354	0.078	0.805	0.041	S	2	24
Gorilla_gorilla	354	0.078	0.805	0.041	S	2	24
Pan_troglodytes	354	0.078	0.805	0.041	S	2	24
Pongo_abelii	354	0.077	0.812	0.041	S	2	24
Nomascus_leucogenys	354	0.076	0.766	0.054	S	2	24
Macaca_mulatta	354	0.039	0.877	0.068	S	1	24
Microcebus_murinus	355	0.042	0.865	0.068	S	2	24
Mus_musculus	354	0.328	0.739	0.016	S	3	24
Rattus_norvegicus	354	0.025	0.93	0.082	S	1	24
Oryctolagus_cuniculus	355	0.206	0.815	0.029	S	2	24
Tursiops_truncatus	356	0.039	0.917	0.063	S	1	24
Equus_caballus	355	0.05	0.867	0.064	S	1	24
Mustela_putorius_furo	355	0.028	0.955	0.039	S	1	24
Pteropus_vampyrus	356	0.027	0.897	0.081	S	1	24
Monodelphis_domestica	358	0.058	0.906	0.037	S	1	27
Sarcophilus_harrisii	356	0.056	0.91	0.056	S	1	25
Gallus_gallus	347	0.108	0.836	0.035	S	2	26
Meleagris_gallopavo	357	0.094	0.698	0.09	S	2	31
Callithrix_jacchus	319	0.082	0.175	0.853	_	2	-
Otolemur_garnettii	320	0.074	0.154	0.874	_	2	-
Cavia_porcellus	310	0.086	0.14	0.849	_	2	-
Ictidomys_tridecemlineatus	320	0.097	0.157	0.782	_	2	-
Ailuropoda_melanoleuca	351	0.047	0.761	0.21	S	3	21
Bos_taurus	321	0.047	0.225	0.89	_	2	-
Myotis_lucifugus	319	0.078	0.212	0.814	_	2	-
Felis_catus	359	0.108	0.059	0.862	_	2	-
Canis_familiaris	309	0.097	0.133	0.843	_	2	-
Loxodonta_africana	322	0.1	0.124	0.839	_	2	-
Ficedula_albicollis	320	0.053	0.09	0.931	_	1	-
Pelodiscus_sinensis	326	0.057	0.216	0.867	_	2	-
Anas_platyrhynchos	343	0.021	0.881	0.108	S	2	17
Xenopus_tropicalis	337	0.015	0.95	0.089	S	1	18
Latimeria_chalumnae	335	0.091	0.907	0.033	S	1	20
Tetraodon_nigroviridis	357	0.313	0.21	0.208	*	5	-
Takifugu_rubripes	331	0.044	0.068	0.95	_	1	-
cutoff	0.78	0	0.73				

C. TargetP results of CA XIV

TargetP v1.1 prediction results
Number of query sequences: 33
Cleavage site predictions included.
Using NON-PLANT networks.

Name	Len	mTP	SP	other	Loc	RC	TPlen
Homo_sapiens	337	0.019	0.921	0.133	S	2	18
Gorilla_gorilla	337	0.019	0.921	0.134	S	2	18
Pan_troglodytes	337	0.02	0.92	0.132	S	2	18
Nomascus_leucogenys	337	0.019	0.921	0.133	S	2	18
Macaca_mulatta	337	0.019	0.921	0.133	S	2	18
Callithrix_jacchus	337	0.019	0.921	0.131	S	2	18
Otolemur_garnettii	336	0.019	0.928	0.123	S	1	18
Ictidomys_tridecemlineatus	338	0.02	0.921	0.133	S	2	18
Mus_musculus	337	0.028	0.898	0.114	S	2	15
Rattus_norvegicus	337	0.028	0.904	0.113	S	2	15
Cavia_porcellus	337	0.028	0.883	0.185	S	2	15
Dipodomys_ordii	337	0.027	0.892	0.161	S	2	16
Loxodonta_africana	337	0.021	0.921	0.134	S	2	16
Procavia_capensis	326	0.041	0.784	0.294	S	3	15
Pteropus_vampyrus	337	0.018	0.947	0.117	S	1	18
Felis_catus	336	0.017	0.947	0.111	S	1	18
Bos_taurus	336	0.017	0.949	0.109	S	1	18
Ailuropoda_melanoleuca	336	0.016	0.955	0.095	S	1	18
Mustela_putorius_furo	336	0.017	0.953	0.099	S	1	18
Ochotona_princeps	337	0.03	0.849	0.165	S	2	17
Oryctolagus_cuniculus	337	0.02	0.941	0.119	S	1	18
Sarcophilus_harrisii	340	0.037	0.88	0.182	S	2	16
Monodelphis_domestica	340	0.028	0.887	0.185	S	2	16
Xenopus_tropicalis	344	0.067	0.777	0.187	S	3	17
Latimeria_chalumnae	342	0.016	0.97	0.057	S	1	29
Xiphophorus_maculatus	344	0.112	0.501	0.349	S	5	18
Dasypus_novemcinctus	330	0.108	0.181	0.773	_	3	-
Vicugna_pacos	314	0.053	0.041	0.956	_	1	-
Myotis_lucifugus	334	0.019	0.949	0.088	S	1	18
Canis_familiaris	343	0.89	0.022	0.154	M	2	22
Macropus_eugenii	320	0.057	0.058	0.948	_	1	-
Oryzias_latipes	318	0.037	0.147	0.946	_	2	-
Oreochromis_niloticus	394	0.011	0.984	0.051	S	1	20
cutoff	0.78	0	0.73				

Appendix II – TMHMM results

A. TMHMM results of CA IX

						start_last_60	size_IC_domain	TM_size
Homo_sapiens	len=459	ExpAA=26.92	First60=4.97	PredHel=1	Topology=o411-433i	12	26	22
Pan_troglodytes	len=453	ExpAA=26.97	First60=5.00	PredHel=1	Topology=o405-427i	12	26	22
Pongo_abelii	len=465	ExpAA=22.74	First60=0.87	PredHel=1	Topology=o417-439i	12	26	22
Nomascus_leucogenys	len=447	ExpAA=26.90	First60=4.89	PredHel=1	Topology=o399-421i	12	26	22
Macaca_mulatta	len=447	ExpAA=27.54	First60=5.29	PredHel=1	Topology=o399-421i	12	26	22
Callithrix_jacchus	len=451	ExpAA=23.84	First60=2.02	PredHel=1	Topology=o403-425i	12	26	22
Otolemur_garnettii	len=455	ExpAA=29.54	First60=7.43	PredHel=1	Topology=o405-427i	10	28	22
Felis_catus	len=454	ExpAA=21.90	First60=0.15	PredHel=1	Topology=o404-426i	10	28	22
Procavia_capensis	len=451	ExpAA=23.38	First60=1.77	PredHel=1	Topology=o403-425i	12	26	22
Equus_caballus	len=443	ExpAA=21.30	First60=0.04	PredHel=1	Topology=o393-415i	10	28	22
Loxodonta_africana	len=455	ExpAA=21.56	First60=0.06	PredHel=1	Topology=o404-426i	9	29	22
Echinops_telfairi	len=454	ExpAA=21.62	First60=0.00	PredHel=1	Topology=o407-429i	13	25	22
Bos_taurus	len=449	ExpAA=26.50	First60=4.72	PredHel=1	Topology=o399-421i	10	28	22
Ailuropoda_melanoleuca	len=449	ExpAA=22.29	First60=0.47	PredHel=1	Topology=o397-419i	8	30	22
Canis_familiaris	len=440	ExpAA=21.55	First60=1.28	PredHel=1	Topology=o415-433i	35	7	18
Sus_scrofa	len=442	ExpAA=28.32	First60=6.50	PredHel=1	Topology=o392-414i	10	28	22
Ochotona_princeps	len=437	ExpAA=22.46	First60=0.29	PredHel=1	Topology=o390-412i	13	25	22
Oryctolagus_cuniculus	len=440	ExpAA=21.81	First60=0.02	PredHel=1	Topology=o390-412i	10	28	22
Mus_musculus	len=437	ExpAA=21.60	First60=0.15	PredHel=1	Topology=o387-409i	10	28	22
Rattus_norvegicus	len=437	ExpAA=29.42	First60=8.43	PredHel=1	Topology=o387-409i	10	28	22
Ictidomys_tridecemlineatus	len=437	ExpAA=40.25	First60=18.24	PredHel=2	Topology=i7-29o387-409i	10	28	22
Monodelphis_domestica	len=456	ExpAA=23.21	First60=1.77	PredHel=1	Topology=o407-429i	11	27	22
Myotis_lucifugus	len=434	ExpAA=22.60	First60=0.64	PredHel=1	Topology=o384-406i	10	28	22
Sorex_araneus	len=265	ExpAA=21.43	First60=0.00	PredHel=1	Topology=o219-238i	14	27	19
Pteropus_vampyrus	len=434	ExpAA=35.78	First60=14.05	PredHel=1	Topology=o386-408i	12	26	22
Dipodomys_ordii	len=433	ExpAA=29.13	First60=7.44	PredHel=1	Topology=o385-407i	12	26	22
Cavia_porcellus	len=448	ExpAA=30.50	First60=9.11	PredHel=1	Topology=o398-420i	10	28	22
Sarcophilus_harrisii	len=455	ExpAA=32.98	First60=11.05	PredHel=1	Topology=o407-429i	12	26	22
Meleagris_gallopavo	len=314	ExpAA=22.41	First60=0.09	PredHel=1	Topology=o264-286i	10	28	22
Gallus_gallus	len=367	ExpAA=21.98	First60=0.12	PredHel=1	Topology=o316-338i	9	29	22
Ficedula_albicollis	len=367	ExpAA=21.86	First60=0.07	PredHel=1	Topology=o317-339i	10	28	22
Latimeria_chalumnae	len=368	ExpAA=22.76	First60=0.57	PredHel=1	Topology=o319-341i	11	27	22
Danio_rerio	len=384	ExpAA=22.46	First60=0.02	PredHel=1	Topology=o326-348i	2	36	22

B. TMHMM results of CA XII orthologues

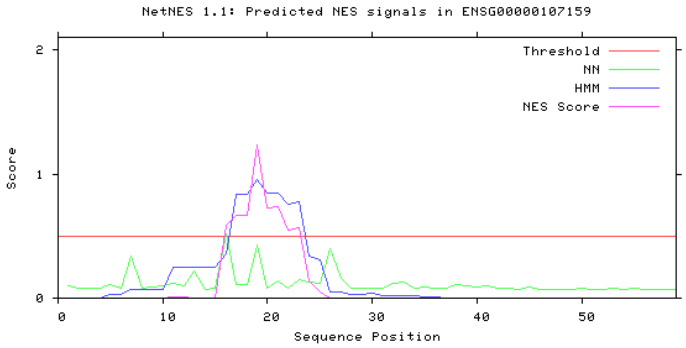
						start_last_60	size_IC	TM_size
Homo_sapiens	len=354	ExpAA=23.68	First60=0.00	PredHel=1	Topology=o305-327i	11	27	22
Gorilla_gorilla	len=354	ExpAA=23.68	First60=0.00	PredHel=1	Topology=o305-327i	11	27	22
Pan_troglodytes	len=354	ExpAA=23.68	First60=0.00	PredHel=1	Topology=o305-327i	11	27	22
Pongo_abelii	len=354	ExpAA=23.69	First60=0.00	PredHel=1	Topology=o305-327i	11	27	22
Nomascus_leucogenys	len=354	ExpAA=23.64	First60=0.01	PredHel=1	Topology=o305-327i	11	27	22
Macaca_mulatta	len=354	ExpAA=24.25	First60=0.00	PredHel=1	Topology=o305-327i	11	27	22
Microcebus_murinus	len=355	ExpAA=23.64	First60=0.00	PredHel=1	Topology=o305-327i	10	28	22
Mus_musculus	len=354	ExpAA=23.67	First60=0.00	PredHel=1	Topology=o305-327i	11	27	22
Rattus_norvegicus	len=354	ExpAA=23.42	First60=0.00	PredHel=1	Topology=o305-327i	11	27	22
Oryctolagus_cuniculus	len=355	ExpAA=23.52	First60=0.05	PredHel=1	Topology=o304-326i	9	29	22
Tursiops_truncatus	len=356	ExpAA=23.19	First60=0.00	PredHel=1	Topology=o306-328i	10	28	22
Equus_caballus	len=355	ExpAA=23.03	First60=0.00	PredHel=1	Topology=o305-327i	10	28	22
Mustela_putorius_furo	len=355	ExpAA=25.79	First60=1.51	PredHel=1	Topology=o305-327i	10	28	22
Pteropus_vampyrus	len=356	ExpAA=23.73	First60=0.01	PredHel=1	Topology=o305-327i	9	29	22
Monodelphis_domestica	len=358	ExpAA=25.90	First60=2.72	PredHel=1	Topology=o308-330i	10	28	22
Sarcophilus_harrisii	len=356	ExpAA=23.64	First60=0.47	PredHel=1	Topology=o306-328i	10	28	22
Gallus_gallus	len=347	ExpAA=24.92	First60=2.22	PredHel=1	Topology=o295-317i	8	30	22
Meleagris_gallopavo	len=357	ExpAA=24.97	First60=1.28	PredHel=1	Topology=o305-327i	8	30	22
Callithrix_jacchus	len=319	ExpAA=23.20	First60=0.00	PredHel=1	Topology=o270-292i	11	27	22
Otolemur_garnettii	len=320	ExpAA=23.63	First60=0.00	PredHel=1	Topology=o271-293i	11	27	22
Cavia_porcellus	len=310	ExpAA=23.06	First60=0.00	PredHel=1	Topology=o259-281i	9	29	22
Ictidomys_tridecemlineatus	len=320	ExpAA=22.96	First60=0.00	PredHel=1	Topology=o271-293i	11	27	22
Ailuropoda_melanoleuca	len=351	ExpAA=24.30	First60=0.00	PredHel=1	Topology=o302-324i	11	27	22
Bos_taurus	len=321	ExpAA=22.96	First60=0.00	PredHel=1	Topology=o271-293i	10	28	22
Myotis_lucifugus	len=319	ExpAA=24.04	First60=0.00	PredHel=1	Topology=o269-291i	10	28	22
Felis_catus	len=359	ExpAA=23.75	First60=0.00	PredHel=1	Topology=o309-331i	10	28	22
Canis_familiaris	len=309	ExpAA=22.94	First60=0.01	PredHel=1	Topology=o258-280i	9	29	22
Loxodonta_africana	len=322	ExpAA=23.54	First60=0.00	PredHel=1	Topology=o271-293i	9	29	22
Ficedula_albicollis	len=320	ExpAA=22.99	First60=0.00	PredHel=1	Topology=o266-288i	6	32	22
Pelodiscus_sinensis	len=326	ExpAA=27.08	First60=0.00	PredHel=1	Topology=o274-296i	8	30	22
Anas_platyrhynchos	len=343	ExpAA=22.93	First60=0.01	PredHel=1	Topology=o290-312i	7	31	22
Xenopus_tropicalis	len=337	ExpAA=23.46	First60=0.49	PredHel=1	Topology=o284-306i	7	31	22
Latimeria_chalumnae	len=335	ExpAA=26.43	First60=3.41	PredHel=1	Topology=o282-304i	7	31	22
Tetraodon_nigroviridis	len=357	ExpAA=22.94	First60=0.23	PredHel=1	Topology=o308-330i	11	27	22
Takifugu_rubripes	len=331	ExpAA=22.92	First60=0.00	PredHel=1	Topology=o280-302i	9	29	22

C. TMHMM results of CA XIV orthologues

						start_last_60	size_IC	TM_size
Homo_sapiens	len=337	ExpAA=22.26	First60=0.04	PredHel=1	Topology=o290-312i	13	25	22
Gorilla_gorilla	len=337	ExpAA=22.22	First60=0.02	PredHel=1	Topology=o290-312i	13	25	22
Pan_troglodytes	len=337	ExpAA=22.31	First60=0.11	PredHel=1	Topology=o290-312i	13	25	22
Nomascus_leucogenys	len=337	ExpAA=22.06	First60=0.04	PredHel=1	Topology=o290-312i	13	25	22
Macaca_mulatta	len=337	ExpAA=22.27	First60=0.04	PredHel=1	Topology=o290-312i	13	25	22
Callithrix_jacchus	len=337	ExpAA=22.10	First60=0.04	PredHel=1	Topology=o290-312i	13	25	22
Otolemur_garnettii	len=336	ExpAA=22.67	First60=0.06	PredHel=1	Topology=o289-311i	13	25	22
Ictidomys_tridecemlineatus	len=338	ExpAA=22.37	First60=0.16	PredHel=1	Topology=o291-313i	13	25	22
Mus_musculus	len=337	ExpAA=22.44	First60=0.02	PredHel=1	Topology=o290-312i	13	25	22
Rattus_norvegicus	len=337	ExpAA=22.35	First60=0.05	PredHel=1	Topology=o290-312i	13	25	22
Cavia_porcellus	len=337	ExpAA=22.20	First60=0.00	PredHel=1	Topology=o290-312i	13	25	22
Dipodomys_ordii	len=337	ExpAA=22.41	First60=0.03	PredHel=1	Topology=o290-312i	13	25	22
Loxodonta_africana	len=337	ExpAA=22.63	First60=0.05	PredHel=1	Topology=o290-312i	13	25	22
Procavia_capensis	len=326	ExpAA=22.51	First60=0.00	PredHel=1	Topology=o290-312i	24	14	22
Pteropus_vampyrus	len=337	ExpAA=22.03	First60=0.01	PredHel=1	Topology=o291-313i	14	24	22
Felis_catus	len=336	ExpAA=22.29	First60=0.05	PredHel=1	Topology=o290-312i	14	24	22
Bos_taurus	len=336	ExpAA=22.39	First60=0.12	PredHel=1	Topology=o290-312i	14	24	22
Ailuropoda_melanoleuca	len=336	ExpAA=22.14	First60=0.04	PredHel=1	Topology=o290-312i	14	24	22
Mustela_putorius_furo	len=336	ExpAA=22.09	First60=0.05	PredHel=1	Topology=o290-312i	14	24	22
Ochotona_princeps	len=337	ExpAA=22.29	First60=0.01	PredHel=1	Topology=o290-312i	13	25	22
Oryctolagus_cuniculus	len=337	ExpAA=22.80	First60=0.16	PredHel=1	Topology=o290-312i	13	25	22
Sarcophilus_harrisii	len=340	ExpAA=22.25	First60=0.00	PredHel=1	Topology=o292-314i	12	26	22
Monodelphis_domestica	len=340	ExpAA=22.41	First60=0.00	PredHel=1	Topology=o292-314i	12	26	22
Xenopus_tropicalis	len=344	ExpAA=22.99	First60=0.00	PredHel=1	Topology=o294-316i	10	28	22
Latimeria_chalumnae	len=342	ExpAA=45.92	First60=22.76	PredHel=2	Topology=i2-24o293-315i	11	27	22
Xiphophorus_maculatus	len=344	ExpAA=22.73	First60=0.00	PredHel=1	Topology=o295-317i	11	27	22
Dasypus_novemcinctus	len=330	ExpAA=22.08	First60=0.00	PredHel=1	Topology=o283-305i	13	25	22
Vicugna_pacos	len=314	ExpAA=22.55	First60=0.00	PredHel=1	Topology=o271-293i	17	21	22
Myotis_lucifugus	len=334	ExpAA=21.76	First60=0.04	PredHel=1	Topology=o289-311i	15	23	22
Canis_familiaris	len=343	ExpAA=22.52	First60=0.00	PredHel=1	Topology=o297-319i	14	24	22
Macropus_eugenii	len=320	ExpAA=22.31	First60=0.00	PredHel=1	Topology=o272-294i	12	26	22
Oryzias_latipes	len=318	ExpAA=22.59	First60=0.00	PredHel=1	Topology=o270-292i	12	26	22
Oreochromis_niloticus	len=394	ExpAA=26.94	First60=4.22	PredHel=1	Topology=o300-322i	-34	72	22

Appendix III – NetNES output sample of human CA IX sequence

>ENSG00000107159 - NetNES 1.1 prediction



#Seq-Pos-Residue	ANN	HMM	NES	Predicted
ENSG00000107159-1-A	0.100	0.000	0.000	-
ENSG00000107159-2-A	0.084	0.000	0.000	-
ENSG00000107159-3-E	0.080	0.000	0.000	-
ENSG00000107159-4-P	0.083	0.000	0.000	-
ENSG00000107159-5-V	0.111	0.028	0.000	-
ENSG00000107159-6-Q	0.075	0.028	0.000	-
ENSG00000107159-7-L	0.335	0.071	0.000	-
ENSG00000107159-8-N	0.082	0.071	0.000	-
ENSG00000107159-9-S	0.091	0.071	0.000	-
ENSG00000107159-10-C	0.095	0.071	0.000	-
ENSG00000107159-11-L	0.121	0.248	0.008	-
ENSG00000107159-12-A	0.096	0.244	0.008	-
ENSG00000107159-13-A	0.217	0.244	0.000	-
ENSG00000107159-14-G	0.070	0.244	0.000	-
ENSG00000107159-15-D	0.076	0.244	0.000	-
ENSG00000107159-16-I	0.517	0.356	0.586	Yes
ENSG00000107159-17-L	0.105	0.837	0.663	Yes
ENSG00000107159-18-A	0.109	0.837	0.666	Yes
ENSG00000107159-19-L	0.426	0.959	1.230	Yes
ENSG00000107159-20-V	0.084	0.850	0.724	Yes
ENSG00000107159-21-F	0.135	0.851	0.736	Yes
ENSG00000107159-22-G	0.084	0.756	0.545	Yes
ENSG00000107159-23-L	0.152	0.772	0.571	Yes
ENSG00000107159-24-L	0.132	0.336	0.138	-
ENSG00000107159-25-F	0.122	0.306	0.048	-
ENSG00000107159-26-A	0.399	0.051	0.000	-
ENSG00000107159-27-V	0.163	0.053	0.000	-
ENSG00000107159-28-T	0.078	0.028	0.000	-
ENSG00000107159-29-S	0.082	0.028	0.000	-
ENSG00000107159-30-V	0.082	0.036	0.000	-
ENSG00000107159-31-A	0.076	0.023	0.000	-
ENSG00000107159-32-F	0.116	0.023	0.000	-
ENSG00000107159-33-L	0.126	0.020	0.000	-
ENSG00000107159-34-V	0.075	0.016	0.000	-
ENSG00000107159-35-Q	0.085	0.015	0.000	-
ENSG00000107159-36-M	0.084	0.015	0.000	-
ENSG00000107159-37-R	0.079	0.000	0.000	-
ENSG00000107159-38-R	0.110	0.000	0.000	-
ENSG00000107159-39-Q	0.096	0.000	0.000	-
ENSG00000107159-40-H	0.085	0.000	0.000	-
ENSG00000107159-41-R	0.097	0.000	0.000	-
ENSG00000107159-42-R	0.082	0.000	0.000	-
ENSG00000107159-43-G	0.079	0.000	0.000	-
ENSG00000107159-44-T	0.074	0.000	0.000	-
ENSG00000107159-45-K	0.090	0.000	0.000	-
ENSG00000107159-46-G	0.072	0.000	0.000	-
ENSG00000107159-47-G	0.068	0.000	0.000	-
ENSG00000107159-48-V	0.070	0.000	0.000	-
ENSG00000107159-49-S	0.065	0.000	0.000	-
ENSG00000107159-50-Y	0.077	0.000	0.000	-
ENSG00000107159-51-R	0.072	0.000	0.000	-
ENSG00000107159-52-P	0.072	0.000	0.000	-
ENSG00000107159-53-A	0.075	0.000	0.000	-
ENSG00000107159-54-E	0.068	0.000	0.000	-
ENSG00000107159-55-V	0.075	0.000	0.000	-
ENSG00000107159-56-A	0.071	0.000	0.000	-
ENSG00000107159-57-E	0.073	0.000	0.000	-
ENSG00000107159-58-T	0.069	0.000	0.000	-
ENSG00000107159-59-G	0.073	0.000	0.000	-
ENSG00000107159-60-A	0.074	0.000	0.000	-

//

[Note: See **Appendix IX** for full result files]

Appendix IV – NucPred output sample of CA XII orthologues

NucPred - multiple sequences
Got 15 sequences with 5044 residues
Calculating NucPred scores 0.59 0.59 0.57 0.63 0.62 0.60 0.53 0.57 0.55 0.53 0.47 0.53 0.51 0.49 0.42
Running ClustalW (please be patient)

NucPred coloured multiple alignment (warning - the alignment may be inaccurate in places because we have aligned full-length sequences which may have different domain organisation)

```
ENSG00000118298     0.59 QLEKLQGTLFSTEEEPSKLLVQNYRALQPLNQRMVFASFIQAG-SSYTTGEMLSLGVGIL
ENSGGOG00000017116 0.59 QLEKLQGTLFSTEEEPSKLLVQNYRAPQPLNQRMVFASFIQAG-SSYTTGEMLSLGVGIL
ENSPTRG00000001249 0.57 QLEKLQGTLFSTEEEPSKLLVQNYRALQPLNQRMVFASFIQAG-SSYTTGEMLSLGVGIL
ENSNLEG00000009787 0.63 QLEKLQGTLFSTEEGPSKLLVQNYRAPQPLNQRMVFASFIQGD-SPRLRSEMLSLGVGIL
ENSMMUG00000021385 0.62 QLEKLQGTLFSTEEEPSKLLVQNYRAPQPLNQRMVFASFSQAG-SLYTTGEMLSLGVGIL
ENSCJAG00000009756 0.60 QLEKLQETLFSTEEEPSEPLVQNYRAPQPLNQRMVFASFIQGD-SASPPGEMLSLGVGIL
ENSOGAG00000004842 0.53 QLEMLQETLFSTEE-PSNLLAQNYRAPQPLNQRTVFASFVQVG-SLYTTGEILSLGVGIL
ENSSTOG00000027815 0.57 QLEKLQETLFSTEEEPSEPLIQNYRAPQPLNQRTVFASFTQGDSPRLRTGEMLSLGVGIL
ENSMUSG00000038526 0.55 QLEKLQETLSSTEEDPSEPLVQNYRVPQPLNQRTIFASFIQAG-PLYTTGEMLGLGVGIL
ENSRNOG00000023162 0.53 QLEKLQETLSSTEEDPSEPLVQNYRVPQPLNQRTIFASFIQVG-PLYTTGEMLGLGMGIS
ENSCPOG00000006079 0.47 QLERLQQTLFSTEEEPSEALVQNYRAPQPLNQRAVFASFIQVG-PVYTTGEMLSLAVGIV
ENSDORG00000011692 0.53 QLERLQETLFSTEEESSEPLVQNYRAPQPLNQRTVFASFTQVE-SLYTTGEMVGLGVGIL
ENSLAFG00000004703 0.51 QLEKLQETLFSTEEEPSELLVQNYRAPQPLNQRSVFASFIQVG-SIYTTGEMLGLGVGIL
ENSPCAG00000007282 0.49 QLEKLQETLSSTEEEPSKPLVQNYRAPQPLNQRIVFASFIQVA-SVYTTGEMLGLGVGIL
ENSPVAG00000013464 0.42 QLEKLQETLFSSETDPSELLVQNYRAPQPLNQRPVFASFIQAE-SLYTTGEMLSLGVGIL
                    cons QLEkLQeTLfStEeepSelLvQNYRapQPLNQRmvFASFiQ g slyttgEmlsLgvGIl

ENSG00000118298     0.59 VGCLCLLLAVYFIARKIRKKRLENRKSVVFTSAQATTEA
ENSGGOG00000017116 0.59 VGCLCLLLAVYFIARKIRKKRLENRKSVVFTSAQATTEA
ENSPTRG00000001249 0.57 VGCLCLLLAVYFIARKIRKKRLENRKSVVFTSAQATTEA
ENSNLEG00000009787 0.63 VGCLCLLLAVYFIARKIRKKRLENRKSVVFTSAQATTEA
ENSMMUG00000021385 0.62 VGCLCLLLAVYFIARKIRKKRLENRKTVVFTSARATTEA
ENSCJAG00000009756 0.60 VGCLCLLLAAYFIARKIRKKRLENRKSVVFTSARATTEA
ENSOGAG00000004842 0.53 AGCLCLLLAAYFIARKIRKKRLENRKSVVFTSARATTEA
ENSSTOG00000027815 0.57 IGCLCLLLAVYFIVRKIRKKRLGNRKSVVFTSAQATTEA
ENSMUSG00000038526 0.55 AGCLCLLLAVYFIAQKIRKKRLGNRKSVVFTSARATTEA
ENSRNOG00000023162 0.53 AGCLCLLLTIYFIAQKIRRKRLGNRKSVVFTSARATTEA
ENSCPOG00000006079 0.47 LGCLCLLVAAYFIAKRIRKKRLGNGKSVVFTSVRATTEA
ENSDORG00000011692 0.53 VGCLCLLLAVYFIAQKIRKKRLGNRKSVVFTSARAAAEA
ENSLAFG00000004703 0.51 AGCLCLLLAVYFIVRKIRKKRLRNQKSVVFTSAQATTEA
ENSPCAG00000007282 0.49 AGCLCLLLVVYFIARKIRKKRLRNQKSV----------
ENSPVAG00000013464 0.42 VGCLCLLLGVYFIARKIRKKMLGNQKSVVFTSSQATEA-
                    cons vGCLCLL|avYFIarkIRkKrLeNrKsVvftsa attea
```

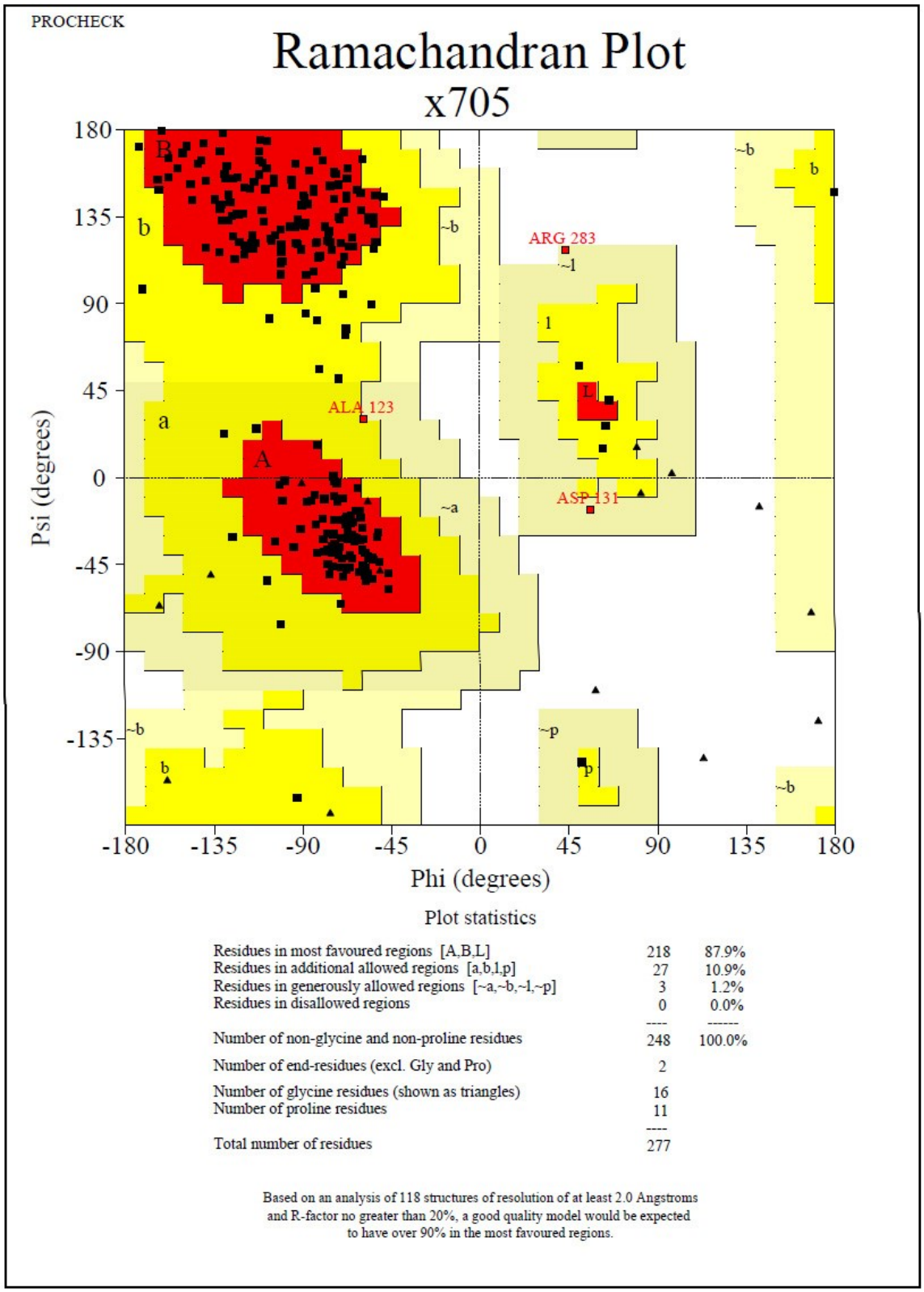
The consensus ('cons') is calculated as follows: the most frequent character in each column is shown in lowercase (unless it is a gap character); uppercase letters represent a column containing just one amino acid and no gaps.

Positively and negatively influencing subsequences are coloured according to the following scale:

(non-nuclear) negative positive (nuclear)

[Note: See **Appendix IX** for full result files]

Appendix V – Ramachandran plot of the comparative model of zebrafish CA VI domain

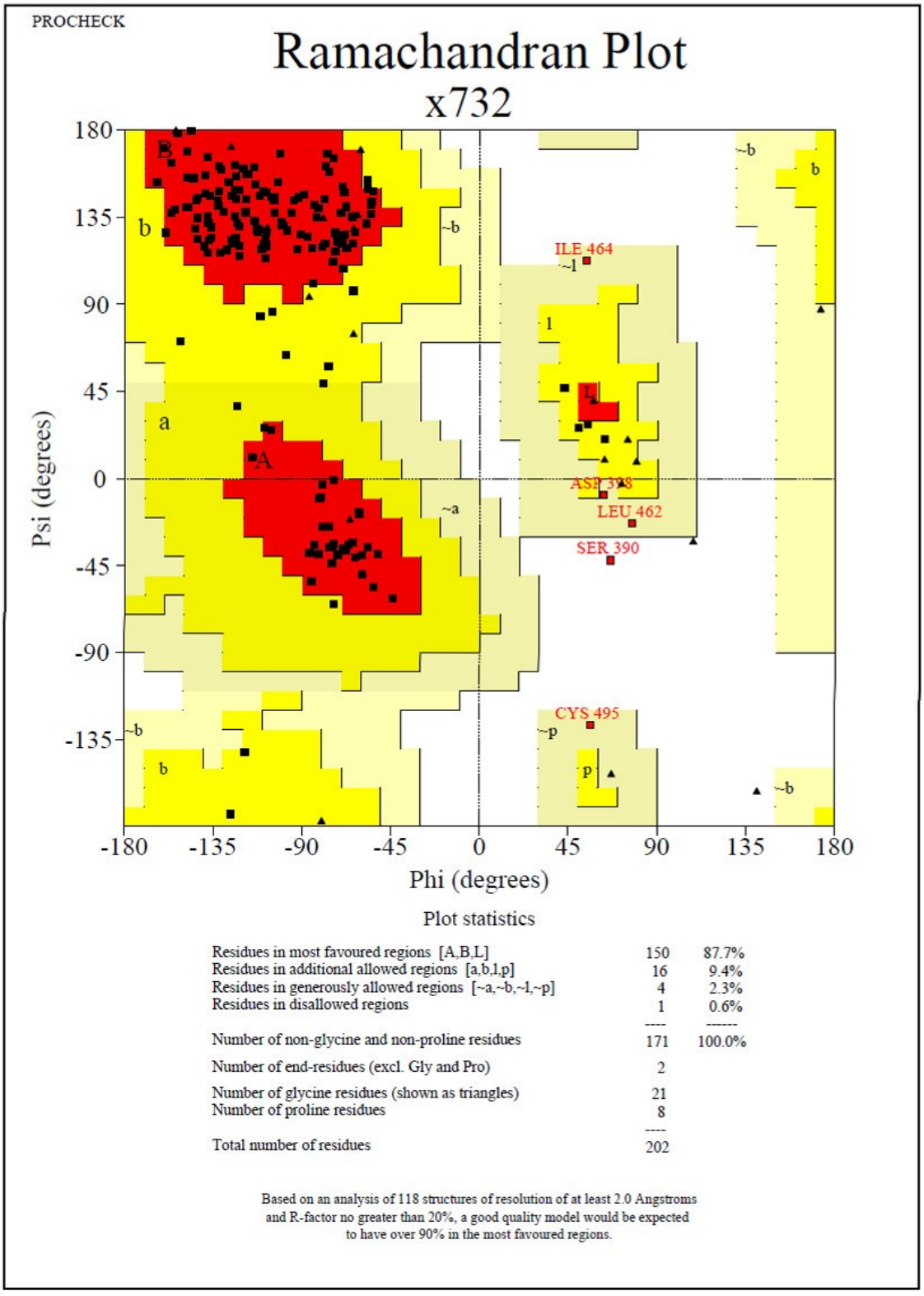


Plot statistics

Residues in most favoured regions [A,B,L]	218	87.9%
Residues in additional allowed regions [a,b,l,p]	27	10.9%
Residues in generously allowed regions [~a,~b,~l,~p]	3	1.2%
Residues in disallowed regions	0	0.0%
Number of non-glycine and non-proline residues	248	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	16	
Number of proline residues	11	
Total number of residues	277	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

Appendix VI – Ramachandran plot of the comparative model of zebrafish Pentraxin domain



Plot statistics

Residues in most favoured regions [A,B,L]	150	87.7%
Residues in additional allowed regions [a,b,l,p]	16	9.4%
Residues in generously allowed regions [~a,~b,~l,~p]	4	2.3%
Residues in disallowed regions	1	0.6%
Number of non-glycine and non-proline residues	171	100.0%
Number of end-residues (excl. Gly and Pro)	2	
Number of glycine residues (shown as triangles)	21	
Number of proline residues	8	
Total number of residues	202	

Based on an analysis of 118 structures of resolution of at least 2.0 Angstroms and R-factor no greater than 20%, a good quality model would be expected to have over 90% in the most favoured regions.

Appendix VII – Pseudocontact parameters in docked CA VI (3FE4) and AP-helix

The table showing the 45 hydrophobic (pseudo) contact parameters calculated between the selected residues of 3FE4 and AP-helix structures after they were docked:

atom1	atom2	overlap	distance
ALA 234.A CB	PHE 289.A CE2	0.355	3.285
ILE 131.A CB	LEU 300.A CD2	0.297	3.463
PHE 236.A CD1	ILE 296.A CG1	0.284	3.356
VAL 168.A O	LEU 303.A CD1	0.236	3.124
ALA 234.A O	PHE 289.A CE2	0.206	3.034
ALA 234.A C	PHE 289.A CE2	0.19	3.27
PHE 236.A CE1	LEU 293.A CD1	0.176	3.464
PHE 236.A CE1	ILE 296.A CG1	0.149	3.491
ILE 131.A CB	LEU 300.A CB	0.106	3.654
PHE 236.A CE1	ILE 296.A CB	0.076	3.564
ILE 131.A CG2	LEU 300.A O	0.067	3.293
VAL 168.A C	LEU 303.A CD1	0.066	3.514
PHE 236.A CZ	LEU 293.A CD1	0.062	3.578
ILE 131.A CG2	LEU 300.A CB	0.024	3.736
ILE 131.A O	LEU 300.A CD2	0.023	3.337
PHE 236.A CE1	ILE 296.A CG2	0.018	3.622
VAL 168.A HN	LEU 303.A CD2	-0.002	2.882
ALA 234.A O	LEU 293.A CD1	-0.015	3.375
PHE 236.A CZ	ILE 296.A CG2	-0.03	3.67
PHE 236.A CD1	ILE 296.A CG2	-0.048	3.688
ILE 131.A CB	LEU 300.A CG	-0.063	3.823
VAL 168.A CG2	LEU 303.A CD2	-0.088	3.848
ALA 234.A CB	PHE 289.A CD2	-0.103	3.743
VAL 134.A CG2	LEU 303.A CD2	-0.109	3.869
ILE 121.A CD1	LEU 300.A O	-0.128	3.488
ILE 121.A CD1	LEU 303.A OXT	-0.134	3.494
PHE 236.A CE2	ILE 296.A CG2	-0.136	3.776
ILE 131.A C	LEU 300.A CD2	-0.149	3.729
PHE 236.A CD1	ILE 296.A CB	-0.188	3.828
PHE 236.A CD2	ILE 296.A CG2	-0.198	3.838
ILE 131.A CG2	LEU 300.A C	-0.2	3.78
ALA 234.A O	PHE 289.A CZ	-0.213	3.453
PHE 236.A CG	ILE 296.A CG2	-0.223	3.803
ILE 131.A CG2	LEU 300.A CG	-0.225	3.985
ALA 234.A CA	PHE 289.A CE2	-0.237	3.877
ALA 234.A C	PHE 289.A CZ	-0.277	3.737
VAL 168.A N	LEU 303.A CD2	-0.312	3.817
ILE 121.A CD1	LEU 303.A CB	-0.313	4.073
PHE 236.A CE1	LEU 293.A CG	-0.319	3.959
ALA 234.A CA	LEU 293.A CD1	-0.322	4.082
ILE 131.A CA	LEU 300.A CD2	-0.324	4.084
PHE 236.A CZ	ILE 296.A CB	-0.326	3.966
PHE 236.A CZ	LEU 293.A CG	-0.333	3.973
ILE 131.A CG1	LEU 300.A CB	-0.34	4.1
ILE 131.A CG2	LEU 300.A CD2	-0.395	4.155

Appendix VIII – Pseudo contact parameters between docked CA VI and Pentraxin

The table showing 155 pseudo contacts parameters calculated between CA VI and Pentraxin domain of the docked model of zebrafish CA VI + Pentraxin protein:

atom1	atom2	overlap	distance
LEU307.AO	LYS 325.B HZ1	0.795	1.685
LYS300.AHZ1	GLN 326.B OE1	0.7	1.78
GLU170.AOE1	SER 513.B HG	0.691	1.789
GLU170.AOE1	LEU 514.B HN	0.651	1.829
GLU293.AOE2	GLY 522.B HN	0.626	1.854
CYS290.ACA	GLN 521.B HE22	0.508	2.372
SER304.ACB	GLN 326.B CB	0.427	3.333
GLU294.ACA	LEU 520.B O	0.418	2.942
ASP310.AO	LYS 325.B HZ3	0.398	2.082
ALA297.ACB	LEU 331.B CD2	0.394	3.366
LEU307.AO	LYS 325.B NZ	0.391	2.714
GLY172.ACA	SER 513.B OG	0.354	3.026
GLU170.AOE1	LEU 514.B N	0.342	2.763
LYS300.ANZ	GLN 326.B OE1	0.341	2.764
GLU293.AOE2	GLY 522.B N	0.34	2.765
ALA297.ACB	LEU 331.B CD1	0.324	3.436
ASP310.AO	LYS 325.B NZ	0.295	2.81
GLU293.ACB	GLN 521.B NE2	0.289	3.216
CYS290.ACA	GLN 521.B NE2	0.284	3.221
ALA297.ACA	LEU 331.B CD1	0.271	3.489
GLU293.ACB	GLN 521.B HE21	0.264	2.616
GLU170.AOE1	SER 513.B OG	0.248	2.732
ALA297.ACB	LEU 520.B CB	0.236	3.524
LYS298.ACG	LEU 514.B CD2	0.221	3.539
GLU294.ACG	GLN 519.B CB	0.202	3.558
LEU305.ACD1	LEU 512.B O	0.193	3.167
CYS290.ACA	GLN 521.B OE1	0.172	3.188
LEU305.ACD2	SER 513.B CA	0.138	3.622
ARG301.ANE	LEU 512.B CB	0.129	3.376
SER308.ACA	GLN 511.B OE1	0.124	3.236
SER304.ACB	GLN 326.B HN	0.106	2.774
GLU293.AOE1	GLN 521.B HE21	0.105	2.375
LEU305.ACD1	LEU 512.B C	0.1	3.48
HIS173.AO	LEU 462.B CD2	0.099	3.261
ARG301.ANH2	THR 510.B CG2	0.095	3.41
ARG301.ACZ	ILE 328.B CD1	0.078	3.502
ARG301.ANH2	ILE 328.B CG2	0.056	3.449
GLY172.ACA	SER 513.B CB	0.047	3.713
ARG301.AHH21	THR 510.B CG2	0.045	2.835
LYS298.ACA	LEU 520.B CD1	0.042	3.718
GLU294.ACB	LEU 520.B O	0.04	3.32
LEU307.AO	LYS 325.B CD	0.038	3.322

GLU294.ACD	GLN 519.B CB	0.022	3.558
CYS290.ACB	GLN 521.B OE1	0.009	3.351
GLU294.ACG	LEU 520.B HN	0.004	2.876
SER308.AOG	GLN 511.B CG	0.003	3.377
LEU307.ACD1	LYS 325.B CB	0.003	3.757
SER304.ACB	GLN 326.B N	-0.011	3.516
GLU293.ACD	GLN 521.B HE21	-0.018	2.718
ARG301.AHH21	ILE 328.B CG2	-0.019	2.899
CYS290.ACA	GLN 521.B CD	-0.02	3.6
THR240.ACB	TYR 515.B CD1	-0.049	3.689
THR240.ACG2	TYR 515.B CB	-0.053	3.813
SER304.ACA	GLN 326.B CB	-0.054	3.814
GLU293.ACD	GLN 521.B CG	-0.065	3.645
LYS300.ACD	GLN 326.B HE22	-0.066	2.946
SER308.AOG	GLN 511.B CB	-0.068	3.448
ALA297.ACB	LEU 331.B CG	-0.07	3.83
LYS300.ACB	LEU 331.B CD1	-0.071	3.831
ARG301.AHE	LEU 512.B CB	-0.072	2.952
ARG301.ANH1	ILE 328.B CG2	-0.08	3.585
THR240.ACG2	TYR 515.B CD1	-0.082	3.722
GLU293.AOE2	GLY 522.B CA	-0.088	3.448
LYS298.ACE	VAL 518.B O	-0.09	3.45
LYS300.ACG	GLN 326.B NE2	-0.095	3.6
HIS173.ANE2	ASN 347.B CG	-0.098	3.423
LYS300.AHZ1	GLN 326.B CD	-0.1	2.8
LYS300.ACD	LEU 331.B CD1	-0.102	3.862
PHE174.ACZ	TYR 515.B CZ	-0.107	3.567
LYS298.ACE	LEU 514.B CD2	-0.112	3.872
ARG301.AHE	LEU 512.B CD1	-0.112	2.992
SER308.ACB	GLN 511.B CG	-0.113	3.873
GLU293.AOE2	GLN 521.B CG	-0.116	3.476
PHE174.ACE2	TYR 515.B CE1	-0.118	3.638
ARG301.AHE	LEU 512.B CG	-0.129	3.009
GLU170.ACD	SER 513.B HG	-0.131	2.831
ARG301.ACG	ILE 328.B CD1	-0.132	3.892
ARG301.ANH1	ILE 328.B CD1	-0.136	3.641
LEU307.AO	LYS 325.B CE	-0.143	3.503
SER308.ACB	GLN 511.B CD	-0.146	3.726
ARG301.ACG	LEU 520.B CD2	-0.148	3.908
LEU307.AC	LYS 325.B HZ1	-0.157	2.857
MET289.AO	GLN 521.B HE22	-0.164	2.644
HIS173.AHE2	ASN 347.B CG	-0.176	2.876
HIS173.AHE2	ASN 347.B ND2	-0.178	2.803
ARG301.AHH11	ILE 328.B CG2	-0.182	3.062
SER308.ACB	GLN 511.B OE1	-0.182	3.542
HIS173.ANE2	ASN 347.B OD1	-0.185	3.29
HIS173.ACE1	ARG 451.B NE	-0.185	3.57
LEU305.ACD1	SER 513.B CA	-0.186	3.946
SER304.ACB	GLN 326.B CA	-0.187	3.947

LEU305.ACD2	LEU 512.B O	-0.187	3.547
GLU170.AOE1	SER 513.B CA	-0.19	3.55
GLU293.ACD	GLY 522.B HN	-0.191	2.891
CYS290.AO	GLN 521.B NE2	-0.192	3.297
LYS300.ACD	GLN 326.B NE2	-0.192	3.697
GLU170.ACD	LEU 514.B HN	-0.196	2.896
SER304.ACB	LYS 325.B N	-0.198	3.718
ARG301.ANH1	ILE 328.B CB	-0.199	3.704
LEU307.ACB	LYS 325.B CD	-0.202	3.962
ARG301.ACD	LEU 512.B CB	-0.205	3.965
ARG301.ACZ	ILE 328.B CG2	-0.218	3.798
GLU293.AOE1	GLN 521.B NE2	-0.238	3.343
THR240.AOG1	TYR 515.B CD1	-0.249	3.509
LYS300.ACG	GLN 326.B CD	-0.257	3.837
ASP310.AO	LYS 325.B HZ1	-0.259	2.739
CYS290.AO	GLN 521.B CD	-0.261	3.441
ARG301.ANE	ILE 328.B CD1	-0.27	3.775
LYS298.ACG	LEU 520.B CD1	-0.27	4.03
LYS300.ACG	LEU 331.B CD1	-0.273	4.033
ALA297.AO	LEU 331.B CD1	-0.273	3.633
PHE174.ACZ	TYR 515.B OH	-0.276	3.536
ARG301.ANE	LEU 512.B CG	-0.281	3.786
ARG301.ANE	LEU 512.B CD1	-0.287	3.792
GLU294.ACG	LEU 520.B O	-0.292	3.652
MET289.AO	GLN 521.B NE2	-0.292	3.397
GLU294.ACG	GLN 519.B CA	-0.294	4.054
LYS300.ACG	GLN 326.B HE22	-0.299	3.179
LYS300.ACG	GLN 326.B OE1	-0.301	3.661
SER308.ACA	GLN 511.B CD	-0.312	3.892
GLU293.ACG	GLN 521.B HE21	-0.313	3.193
LEU305.ACD1	SER 513.B N	-0.314	3.819
SER304.AO	LYS 325.B HN	-0.315	2.795
GLU170.AOE1	SER 513.B CB	-0.315	3.675
ARG301.ANH2	ILE 328.B CD1	-0.315	3.82
GLU294.ACG	LEU 520.B N	-0.316	3.821
GLU293.ACD	GLN 521.B NE2	-0.321	3.646
CYS290.AO	GLN 521.B OE1	-0.323	3.283
PHE174.ACE1	TYR 515.B CE2	-0.326	3.846
PHE174.ACZ	TYR 515.B CE1	-0.329	3.849
GLU170.AOE1	LEU 514.B CA	-0.34	3.7
HIS173.ANE2	ASN 347.B ND2	-0.341	3.591
GLU294.ACB	GLN 519.B NE2	-0.348	3.853
LYS300.ACD	GLN 326.B OE1	-0.349	3.709
HIS173.AHE2	ASN 347.B OD1	-0.35	2.83
HIS173.ACE1	ARG 451.B CZ	-0.35	3.81
ARG301.ACD	LEU 512.B CD1	-0.358	4.118
LEU307.ACD1	LYS 325.B CA	-0.361	4.121
CYS290.AC	GLN 521.B HE22	-0.362	3.062
CYS290.AN	GLN 521.B HE22	-0.368	2.993

PHE174.ACE1	TYR 515.B CZ	-0.368	3.828
SER308.ACB	GLN 511.B CB	-0.373	4.133
HIS173.ACE1	ARG 451.B NH2	-0.373	3.758
CYS290.AC	GLN 521.B NE2	-0.374	3.699
ALA297.ACB	LEU 520.B C	-0.375	3.955
PHE174.ACD2	TYR 515.B CE1	-0.375	3.895
LEU307.ACB	LYS 325.B CB	-0.375	4.135
LEU305.ACD1	LEU 512.B N	-0.381	3.886
HIS173.ACD2	ASN 347.B CG	-0.383	3.843
MET289.AC	GLN 521.B HE22	-0.386	3.086
ALA297.ACB	GLN 521.B N	-0.387	3.892
LYS298.ACB	LEU 520.B CD1	-0.393	4.153
GLU170.AOE1	SER 513.B C	-0.393	3.573
ARG301.ACZ	LEU 512.B CB	-0.394	3.974
GLY172.ACA	SER 513.B HG	-0.398	3.278

Appendix IX – Dropbox links of supplementary files

NetNES output of CA IX orthologues:
https://www.dropbox.com/s/fxek2orwyfdgumf/NetNES_output_CAIX.pdf?dl=0

NetNES output of CA XII orthologues:
https://www.dropbox.com/s/0yzloz02qg1ryza/NetNES_output_CAXII.pdf?dl=0

NetNES output of CA XIV orthologues:
https://www.dropbox.com/s/d42340wa0s5fbhy/NetNES_output_CAXIV.pdf?dl=0

NucPred output of CA IX orthologues:
https://www.dropbox.com/s/2dyscw9n94nxksf/NucPred_output_CAIX.pdf?dl=0

NucPred output of CA XII orthologues:
https://www.dropbox.com/s/rwi979w5pvmtez7/NucPred_output_CAXII.pdf?dl=0

NucPred output of CA XIV orthologues:
https://www.dropbox.com/s/hm5mnnzfksvxhwm/NucPred_output_CAXIV.pdf?dl=0

Pdb file of docked model of zebrafish CA VI and Pentraxin domains:
https://www.dropbox.com/s/trykeo5ch0u0zep/Danio_CAviPtx.pdb?dl=0