UNIVERSITY OF TAMPERE

# Development of a Protein Conservation Analysis Pipeline and Application to Carbonic Anhydrase IV

Master's Thesis

**Harlan Barker**

**6/12/2013**

## Acknowledgements

My studies in Finland have been some of the most enjoyable and rewarding times of my life. I have the utmost respect for a culture that encourages advancement by providing higher education to its citizens. The fact that this benefit has been extended to motivated foreign students speaks even more highly of the Finnish people. The country of Finland has my greatest of thanks for allowing me the opportunity to study and earn my Master's degree here.

Individually, I would like to thank my advisor Martti Tolvanen, whose mentoring began during a summer project which eventually evolved into my thesis work, a greater understanding of bioinformatics, and the root of future endeavors. I would like to thank Seppo Parkkila for his introduction to, and guidance through, the world of carbonic anhydrases, and for a warm welcoming to his research group.

I would like to thank my mother for showing me the value of family, and for her unfailing support, without which I would be a very different person today. My dear Bettina provided daily encouragement which allowed me to press forward during the most difficult parts of this year. Thanks to my good friend Payam, the hardest working student I have ever met, for setting the bar so high. Finally, thanks to those friends and family who have made a difference these past two years.

# Master's Thesis


| | |
|---|---|
| Place: | University of Tampere<br>School of Medicine<br>Institute of Biomedical Technology |
| Author | BARKER, HARLAN REID |
| Title | Development of a protein Conservation Analysis Pipeline and application to Carbonic Anhydrase IV |
| Pages | |
| Supervisors | Professor Seppo Parkkila<br>University Lecturer Martti Tolvanen |
| Reviewers | Professor Seppo Parkkila<br>Professor Matti Nykter |
| Date | June 2013 |


## Abstract


### Background and Aims

Conservation is a hallmark of inherent valuable function. Computational analysis of conservation of each amino acid in a protein provides targets for future computational or experimental research.

Carbonic anhydrases (CA) reversibly catalyze the carbon dioxide to bicarbonate reaction. Despite the apparent simplicity of the reaction, the α-CA protein family exists as more than 15 different isoforms in mammals and its members are present in a wide array of tissues and perform a variety of functions.

The overall goal of this research is identification of all residues of functional significance in CA-IV through utilization of gene prediction, comparative genomics, and conservation analysis. The expanded goal is to make this process applicable to any protein group.

### Methods

Automated methods were created, using Python scripting, to extract orthologs for human carbonic anhydrases. Predictions were made for incomplete orthologs, and conservation analysis was performed on a codon alignment of the final set. Additional python scripts created three dimensional (3D) models of conservation values, within specific taxa or as comparisons between them.

## Results

A pipeline was created for automated gene annotation, 3D image generation, and comparative analysis between different taxa. A total of 499 residue positions were altered in 55 carbonic anhydrase proteins from 6 isozyme types. Key amino acids have been identified in a region potentially related to CA-IV ion channel binding.

## Conclusions

$K_a/K_s$ conservation analysis can be applied in a quick and automated manner to produce models and images from large numbers of orthologs, allowing for determination of critical amino acid residues in proteins.

# Abbreviations

| | |
|---|---|
| 3D | Three Dimensional |
| 4-MI | 4-methylimidazole |
| BLAST | Basic Local Alignment Search Tool |
| BSDP | Bounded Sparse Dynamic Programming |
| CA | Carbonic Anhydrase |
| CARP | Carbonic Anhydrase Related Protein |
| DNA | Deoxyribonucleic Acid |
| DP | Dynamic Programming |
| DSSP | Dictionary of Protein Secondary Structure |
| EBI | European Bioinformatics Institute |
| ER | Endoplasmic Reticulum |
| EST | Expressed Sequence Tag |
| GCRMA | Gene Chip Robust Multiarray Averaging |
| GPI | Glycophosphatidylinositol |
| HMM | Hidden Markov Model |
| HSP | High-scoring Segment Pair |
| MSA | Multiple Sequence Alignment |
| PH | Proton Hole |
| RNA | Ribonucleic Acid |
| RSA | Relative Solvent Accessibility |
| SAR | Sub-Alignment Region |
| SS | Secondary Structure |

# Contents

# 1. Introduction

At first glance, a carbonic anhydrase protein is of middling size, resembles half of an opened walnut, and implies no great stature in the pantheon of proteins currently categorized. Yet, this protein, in a variety of incarnations, performs a basic function most essential to life in all of its myriad levels of complexity, that of acid/base maintainence. It has even been proposed that CA activity in ancient cyano-bacteria enabled $CO_2$ sequestration that allowed for the rise to dominance of higher forms of life on earth (Kupriyanova & Pronina, 2011).

Amino acids which are conserved in a protein provide structural or functional properties. Examination of aligned protein sequences allows for identification of conserved amino acids. $K_a/K_s$ analysis of codon-aligned DNA sequences allows a more precise determination of conservation. Comparison of conservation between taxa provides a view of which amino acids or regions are unique to protein function within each.

In total, 16 known isoforms of α-carbonic anhydrase help adjust acid-base balances within mammals. The carbonic anhydrase IV (CA-IV) protein is present in most vertebrates and is linked by Glycophosphatidylinositol (GPI) to the outside surface of the cell. It is most strongly expressed in the lung (Wu, et al., 2009) in air breathing vertebrates, and promotes release of $CO_2$ from bicarbonate laden blood. CA-IV is also found well expressed in many other tissues, including thyroid, heart, colon, retina, kidney, and cerebellum (Wu, et al., 2009).

No conservation study of the CA-IV protein has been previously made using this number of sequences, the MEC codon model, or methods of taxa comparison.

# 2. Literature Review

## 2.1 Carbonic Anhydrase Background

The metal ion activated CA enzyme (metalloenzyme) catalyzes the interconversion of $HCO_3^-$ + $H^+ \leftrightarrow CO_2 + H_2O$ at a very high rate of up to 1 x $10^6$ reactions/s (Lindskog, 1997). Since the first investigation of its enzymatic activity, and proposal for naming as 'Carbonic Anhydrase' in 1932, the understanding of this protein family has increased dramatically (Brinkman, Margaria, Meldrum, & al, 1932). The Ensembl database alone categorizes over 480 CA proteins among 61 species (primarily mammalian). There are currently 16 carbonic anhydrase isoforms identified as present in mammalian species; only 15 of these are active in primates. Three of these proteins are carbonic anhydrase related proteins (CARPs) and are catalytically inactive. The remaining active isozymes are categorized into five subgroups based on localization (Leggat, Dixon, Saleh, & al, 2005):

- Cytoplasmic (I, II, III, VII, and XIII)
- Mitochondrial (VA and VB)
- Secreted (VI)
- Transmembrane (IX, XII, XIV)
- GPI-linked (IV, XV)

The CA isozymes currently present in the mammalian genome are believed to have been created through both whole genome and individual gene duplication. Despite the large time scale of differentiation, the isozymes, as a group, still possess regions of high similarity. Conversely, CA proteins vary in their locations and structure in functionally significant ways. Some of the many noted physiological processes to which they contribute are: acid-base balance, respiration, calcification, and bone resorption. They are also involved in the formation of cerebrospinal fluid, saliva, and gastric acid

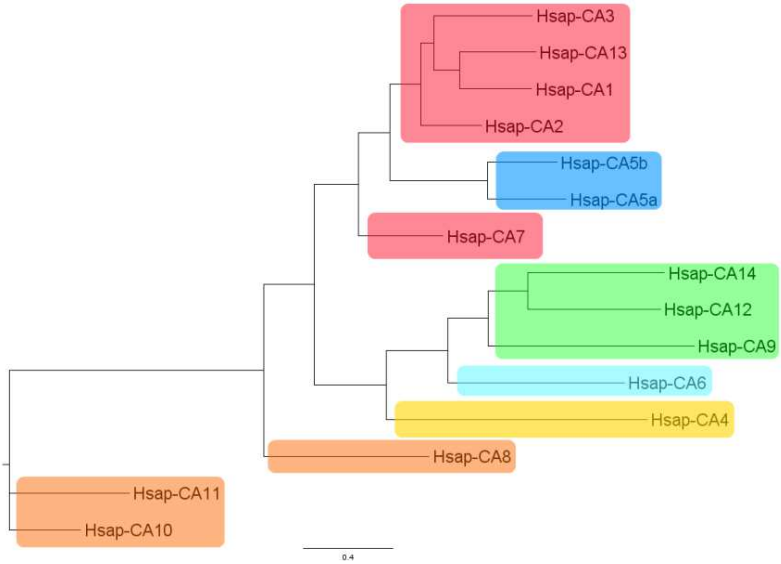


Figure 1 - Phylogenetic tree depicting relationships of all CA isoforms current known to be present in humans. Protein sequences were aligned using Clustal Omega (Sievers, et al., 2011), a codon alignment generated using Pal2Nal (Suyama, Torrents, & Bork, 2006). The tree was constructed using Maximum Likelihood analysis program PhyML (Guindon & Gascuel, 2003). The codon substitution model GTR was used and the parameters set using by analysis of the codon aligned nucleotide sequences; 100 bootstrap replications were performed. The tree was

(Entrez).

The diversity of isoforms, widespread distribution in nearly all tissue types, and presence in all known vertebrate genomes indicates a fundamental necessity for the enzymatic action that CAs provide. Proteins that are catalytically similar to α-CAs are also present in bacteria, and other lower life-forms, in the β, γ, δ, and ζ CA families. Regardless of whether the presence of this enzymatic activity in two branches of life are due to the CA family possessing a root occurring before their divergence, or due to instances of convergent evolution, it is apparent that the reaction is vital (Tashian, 1989) (Gu, 1997). Indeed, it is thought likely that CAs, from one family or another, are present in all living cells (Gilmour, 2010).

Figures depicting expression values for each human CA isoform, for a variety of tissues, have been gathered from the BioGPS website and are located in Appendix A (Wu, et al., 2009). As the measurements of expression are based on fluorescence intensity values from Affymetrix microarray chips, and different probe sets were used when testing different isoforms, the values are best interpreted when only comparing tissues from a single probeset, and therefore isoform. This fact remains despite the fact that BioGPS microarray datasets have been normalized (using GCRMA). However, in cases where there are significant expression value differences between same tissue in different probesets, assumptions become safer. An abbreviated expression chart containing only those tissues and values which were greater than 3x the mean value of expression for the whole isoform (Table 1). While most of the CAs have a low level of expression in all tissues tested in the microarray, in 7 of the isoforms at least one tissue shows expression levels 10x that of the mean of all tissues.

| CA | **Mean Exp.** | >3x mean Tissue(**Value**) | >10x mean Tissue(Value) | Greatest Value | Kcat/s | Chrom. | Subcellular Localization or Category |
|---|---|---|---|---|---|---|---|
| CAI | **4.15** | Skin(13.70) | | Skin(13.70) | $20 \times 10^4$ | 8 | Cytoplasmic |
| CAII | 261.4 | CD71+ Early Erythroid(9468.15)<br>Colon(3395.85)<br>CD105+ Endothelial(2884.00) | CD71+ Early Erythroid(9468.15)<br>Colon(3395.85)<br>CD105+ Endothelial(2884.00) | CD71+ Early Erythroid(9468.15) | $140 \times 10^4$ | 8 | |
| CAIII | 22.8 | Skeletal Muscle(112.85)<br>Thyroid(493.25) | Thyroid(493.25) | Thyroid(493.25) | $1.3 \times 10^4$ | 8 | |
| CAVII | 5.7 | Atrioventricular Node(19.45) | | Atrioventricular Node(19.45) | $95 \times 10^4$ | 16 | |
| CAXIII | 6.84 | None | None | | $15 \times 10^4$ | 8 | |
| CAVI | 135.6 | Salivary Gland(11373.40) | Salivary Gland(11373.40) | Salivary Gland(11373.40) | $34 \times 10^4$ | 1 | Secreted |
| CAIX | 10.68 | | | | $38 \times 10^4$ | 9 | Transmembrane |
| CAXII | 36.9 | Kidney(1069.25)<br>Bronchial Epithelial Cells(182.50)<br>Smooth Muscle(387.55)<br>Colon(566.30)<br>Caudatenucleus(161.05) | Kidney(1069.25)<br>Smooth Muscle(387.55)<br>Colon(566.30) | Kidney(1069.25) | $42 \times 10^4$ | 15 | |
| CAXIV | 11.1 | Retina(33.65) | | Retina(33.65) | $31 \times 10^4$ | 1 | |
| CAIV | 28.3 | Thyroid(295.35)<br>Lung(1054.50)<br>Heart(105.05) | Thyroid(295.35)<br>Lung(1054.50) | Lung(1054.50) | $110 \times 10^4$ | 17 | GPI-Linked |
| CAVIII | 5.3 | Cerebellum(47.20) | | Cerebellum(47.20) | NA | 8 | CARPs |
| CAX | 16.5 | Pineal (night)(222.94)<br>Cerebellum(99.55)<br>Cerebellum(Peduncles)(58.20) | Pineal (night)(222.94) | Pineal (night)(222.94) | NA | 17 | |
| CAXI | 95.7 | Whole Brain(1030.95)<br>Amygdala(776.45)<br>Prefrontal Cortex(1183.15)<br>Caudatenucleus(347.40)<br>Parietal Lobe(305.85)<br>Medualla Oblongata(358.75)<br>Cingulate Cortex(473.15)<br>Occipital Lobe(337.85)<br>Temporal Lobe(434.10)<br>Cerebellum(443.05)<br>Cerebellum (Peduncles)(529.30) | Whole Brain(1030.95)<br>Prefrontal Cortex(1183.15) | Prefrontal Cortex(1183.15) | NA | 19 | |
| CAVA | 4.73 | None | None | | $29 \times 10^4$ | 16 | Mitochondria |
| CAVB | 5.11 | None | None | | $95 \times 10^4$ | X | |

**Table 1 - Microarray data analysis tissue expression values of significance. Experimentally derived catalytic activity rates are shown in standardized units of 10,000 reactions per second (Hilvo, et al., 2008). Chromosome and sub-cellular localizations are obtained from Ensembl (Flicek, Amode, Barrell, & al., 2012). Rates of catalysis (Hilvo, et al., 2008).**

## 2.2 Cytoplasmic CAs

As the name suggests, this largest group of CAs is found in various locations of the interior of cells. The CA1, CA2, CA3, and CA13 genes are located on chromosome number 8 in humans and within a ~261kb region between 86,132,816-86,393,722 (Stelzer, Dalah, Stein, & al, 2011). In addition to being located on the same chromosome these four genes share highest identity with each other, as compared to the other isoforms, indicating that they likely came about from duplication events on chromosome 8. CAII has the highest average identity with the other cytoplasmic CAs at 58.25%, and with all other CAs at 39.6% (Hassan, Shajee, Waheed, Ahmad, & Sly, 2013). Additionally, phylogenetic analysis (Figure 1) shows that these four proteins cluster together, while the final cytoplasmic protein CA-VII appears to be less closely related than the mitochondrial CA proteins.

### 2.2.1 CA-I

According to microarray analysis available at the BioGPS website, the carbonic anhydrase I protein demonstrates highest expression levels in the skin (Wu, et al., 2009) (Su, Wiltshire, Batalov, Lapp, & al., 2004). However, antibody staining presented at proteinatlas.org shows highest expression in hematopoietic cells of the bone marrow and a subcellular localization of the Golgi apparatus (Uhlen, et al., 2010). This difference in expression highlights the difference between these two methods with microarray generally considered to be more accurate. CA-I is not secreted and therefore cannot be localized in the Golgi. At 200,000 de/hydrations per second, it is the third slowest CA protein (Hilvo, et al., 2008).

### 2.2.2 CA-II

The carbonic anhydrase II protein is the most widely studied isoform, has the highest rate of catalysis at roughly 1.4 million de/hydrations per second (Hilvo, et al., 2008), and there are over 400 3D protein models in the Protein Data Bank that depict it. Microarray analysis shows high levels of expression in early erythroid, colon, and endothelial cells (Wu, et al., 2009). Antibody staining support cytoplasmic subcellular localization and indicates highest expression in gastrointestinal tract, kidney, and glial cells (Uhlen, et al., 2010).

### 2.2.3 CA-III

The carbonic anhydrase III protein has evolved to be the slowest of the CA isozymes with a rate of catalysis equal to 13,000 de/hydrations per second (Hilvo, et al., 2008). According to microarray studies of expression the primary point of action for CA-III is in the thyroid gland, and secondarily in skeletal muscle (Wu, et al., 2009). Antibody staining shows expression in

skeletal muscle and adipocytes (Uhlen, et al., 2010). Likely tied to the slow rate of catalysis, when it comes to inhibition by sulfonamides the CA-III protein is the least reactive, with an acetazolamide $K_i$ value (concentration needed to reduce catalytic activity by half) of 300,000 nM (Lindskog, 1997) (Lehtonen, et al., 2004).

### 2.2.4 CA-VII

Of all the cytoplasmic α-CA genes, CA7 is the only one that does not reside on chromosome 8. Like most of the CA proteins, CA7 has low expression levels in most tissues, however there are virtually no tissues in which it is very strongly expressed. The highest of those noted in microarray analysis was the thyroid gland (Wu, et al., 2009). Antibody staining identified the presence of the protein in most of the tissues searched, with the strongest expression in colon, cervix, uterus, vulva, and esophagus (Uhlen, et al., 2010).

### 2.2.5 CA-XIII

For the CA-XIII protein, published microarray data at the BioGPS web database appears to have been normalized improperly as expression levels across many tissues are identical. Antibody staining showed the presence of CA13 in a variety of cells, however, most strongly in: lower stomach, small intestine, appendix, colon, rectum, and gallbladder (Uhlen, et al., 2010). At 150,000 de/hydrations per second the CA-XIII protein is the second slowest of all (Hilvo, et al., 2008). Interestingly, CA-XIII was experimentally determined to be the most susceptible to inhibition by acetazolamide, with a $K_i$ value of 17nM (Lehtonen, et al., 2004).

## 2.3 Mitochondrial CAs

Despite both of the mitochondrial CA isozymes possessing subcellular localization within mitochondria, they are located on two different chromosomes. The CA5A gene is located on chromosome 16 and CA5B is located on the X chromosome (Flicek, Amode, Barrell, & al., 2012). Phylogenetic analysis of both genes in (Shah, et al., 2000) predicts that they resulted from duplication of a CA5 proto-gene roughly 200-300 million years ago.

### 2.3.1 CA-VA

Microarray data for mitochondrial CA-VA reveals no areas of high expression, but does present low levels of expression in all tissues and the most significant expression in the liver (Wu, et al., 2009); RNA-seq analysis confirms highest expression in liver, while antibody staining is not available (Uhlen, et al., 2010).

### 2.3.2 CA-VB

Similar to CA-VA, microarray data for the CA-VB protein shows low levels of expression across many tissues, with the highest level appearing in fat tissue (Wu, et al., 2009). Antibody staining for this isozyme shows expression across a variety of tissues with high levels occurring in hematopoietic cells of the lymph node, tonsil, and spleen (Uhlen, et al., 2010).

## 2.4 Secreted CA

### 2.4.1 CA-VI

There is currently only a single CA that has been identified to exist in a secreted form. Expression levels of CA-VI in saliva dramatically exceed that of other tissues, by a factor of >2,000x, as determined by microarray analysis (Wu, et al., 2009); this is supported by antibody staining (Uhlen, et al., 2010). A 2006 study of human patients, whose oral CA-VI was inhibited by acetazolamide, showed that pH was significantly lower in plaque of subjects receiving the inhibitor (Kimoto, Kishino, Yura, & Ogawa, 2006). The conclusion of the experimenters was that CA-VI function provides protection against caries through neutralization of oral pH. However, a contradictory 2011 study found that Car6 knockout mice exhibited reduced incidence of caries caused by oral Streptococcus mutans and cariogenic diet (Culp, et al., 2011).

## 2.5 Transmembrane CAs

The three transmembrane localized CAs (CA-IX, CA-XII, CA-XIV) are the second largest grouping of active CA isoforms. These three protein sequences possess an introductory ~289-410 amino acids which resides outside the cell, a ~22 amino acid transmembrane domain near the C-terminal end of the protein sequence, finally followed by a ~24-26 amino acid region that resides inside the cell (Figure 2). Unlike the cytoplasmic group, none of the transmembrane genes are located on the same chromosome.

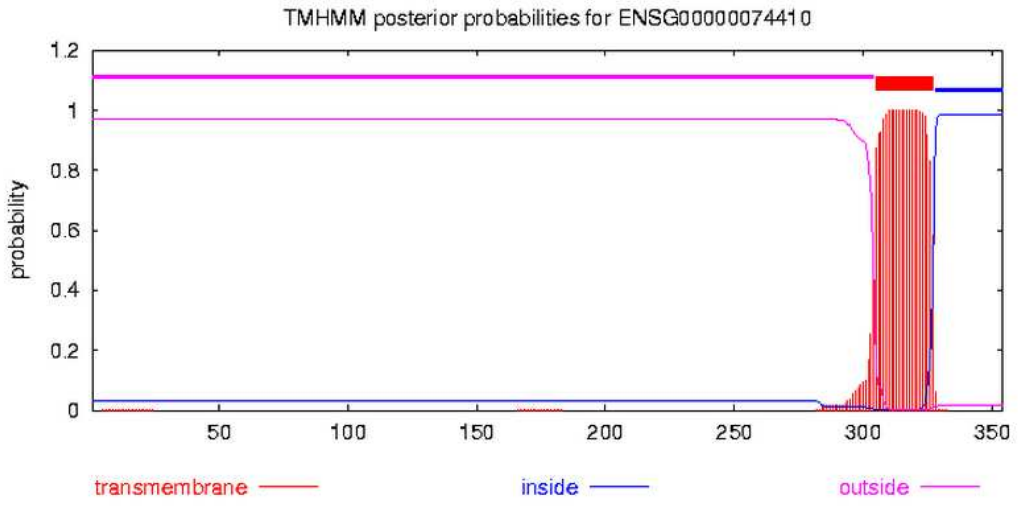


Figure 2 - Transmembrane prediction for CA-XII by TMHMM webserver (Sonnhammer, Heijne, & Krogh, 1998)

### 2.5.1 CA-IX

Like other CAs, microarray data shows CA-IX is expressed at low levels across many tissues, with highest levels in the testis and skin; general expression levels appear to be roughly twice as high for low expression tissues of other CAs (Wu, et al., 2009). Antibody staining shows expression only in digestive system cells of stomach, duodenum, small intestine, and gall bladder in addition to liver (Uhlen, et al., 2010). Of all the CA proteins the CA-IX isozyme is of the greatest interest as relates to cancer. Hypoxia associated with neoplasms, in particular, has been noted as a trigger for *CA9* gene expression. Positive signal for hypoxia-induced CA IX has been identified as an important biomarker for poor prognosis in several cancers (Masayuki Nakao, 2009). In particular, recent studies have shown this holds true for the cancers derived from the following tissues: lung, bladder, rectum, uterine cervix, mesothelium, brain, and breast (Masayuki Nakao, 2009) (Sherwood, Colquhoun, & D., 2007) (Rasheed, Harris, & Tekkis, 2008) (Liao, Darcy, Randall, & al, 2010) (Kivela, Knuuttila, Sihvo, & al, 2012) (Dungwa, Hunt, Ramani, & al, 2012) (Hsieh, Chen, Chiou, & al, 2010).

### 2.5.2 CA-XII

In addition to low level expression across many tissues, microarray data for CA-XII shows high levels of expression in cells of the kidney, colon, smooth muscle, bronchial epithelial cells, and caudate nucleus (Wu, et al., 2009). Antibody staining shows highest presence of CA-XII in appendix, colon, rectum, pancreas, and kidney (Uhlen, et al., 2010). Like CA-IX this isozyme has been associated, though to a lesser extent, with a number of cancers (Table 3).

### 2.5.3 CA-XIV

The CA-XIV protein is shown, by microarray data, to have low levels of expression in many tissues with higher levels occurring in the retina. However, the microarray data also shows that, as a whole, various tissues in the brain exhibited enhanced expression of the CA14 gene over other tissue groups (Wu, et al., 2009). Antibody staining for the protein confirms its presence in many tissues with highest levels occurring in skin (Uhlen, et al., 2010). A 2002 study by Kaunisto et al. suggested that a primary function of CA-XIV is that it plays a complementary role, with CA-IV, in renal acidification (Kaunisto, et al., 2002).

## 2.6 GPI-Linked CAs

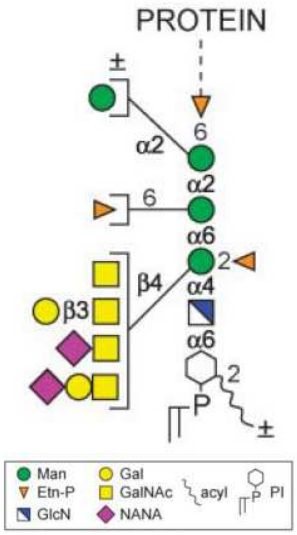


Figure 3 - Schematic diagram of glycophosphatidylinositol (Orlean & Menon, 2007)

Glycophosphatidylinositol (GPI) is a glycolipid which, in the case of GPI-linked proteins, forms a link between the C-terminus of a protein and the membrane of a cell (Figure 3). Commonly referred to as an 'anchor', GPI is irreversibly attached to a protein inside the endoplasmic reticulum (ER) (Orlean & Menon, 2007). Unlike transmembrane proteins, GPI penetrates only through the first half of the lipid bilayer, and preferentially attaches in detergent resistant areas called lipid rafts, so called for their infusion with cholesterol (Orlean & Menon, 2007) (Pike, 2009). CA-XV is present in some mammals, such as mouse, rat, cat, dog, elephant, ferret, and squirrel (Flicek, Amode, Barrell, & al., 2012). However, in humans it is present only as a pseudogene and its functionality is suspected to have been lost in all primates (Tolvanen, et al., 2012).

In a previous study, to which this author contributed, phylogenetic analysis allowed for designation of a new group of GPI-linked CA isozymes present in fishes called CA-XVII (Tolvanen, et al., 2012). In this study, analysis of GPI-linked CAs IV, XV, and XVII

revealed a significant number of probable N-glycosylation sites (defined by the Asn-X-Ser/Thr motif (Shakin-Eshleman, Spitalnik, & Kasturi, 1996)) present on all of the isozymes in at least some species; however, human CA-IV expresses none of these sites (Tolvanen, et al., 2012). N-glycosylation involves the addition of oligosaccharides and like GPI-linkage these addition is made in the lumen of the ER (Shakin-Eshleman, Spitalnik, & Kasturi, 1996). While the function of N-glycosylation is unknown in relation to the GPI-linked CAs, it is potentially significant to note that the sites are more numerous on CA-XVII, which has only been observed in fishes and one lizard species.

### 2.6.1 CA IV

Microarray data shows at least low level expression in all tissues sampled, very high levels of CA-IV expression in the lung and significant expression in thyroid, heart, colon, and retina (Wu, et al., 2009). Antibody staining confirms presence in the lung (specifically capillaries), and rectum (Uhlen, et al., 2010).

The varied distribution of CA-IV indicates that it performs a number of functions. The high expression of CA-IV in the capillaries of the lung indicates it is crucial in the dehydration of blood-born bicarbonate during the process of respiration (Zhu & Sly, 1990). Expression of CA-IV in the kidney allows for reabsorption of bicarbonate (Sterling, Alvarez, & Casey, 2002). A 2009 study identified CA-IV as the primary sensor for $CO_2$ in the mouth, allowing an organism to identify sour tastes (Chandrashekar, et al., 2009).

Before attachment of GPI to CA-IV, inside the ER, a 28 amino acid segment is cleaved from its C-terminus (Lindskog, 1997). This cleavage reveals Ser-284 as the C-terminus attachment point for GPI. In an experiment by Okuyama et al., a site specific mutation S284F produced an un-cleaved and inactive CA unbound to the cell surface while the G285F mutant produced normal CA-IV expression (Okuyama, Waheed, Kusumoto, Zhu, & Sly, 1995). Despite the apparent extracellular position of a matured CAIV, a 2013 study by Schneider et al. determined that pre-mature CAIV protein provided intracellular activity in frog eggs before its anchorage to the cell surface (Schneider, et al., 2013).

## 2.7 Carbonic Anhydrase Related Proteins

There are three carbonic anhydrase isoforms (CA-VIII, CA-X, and CA-XI) which have no catalytic activity due to inactivation of the active site caused by substitution of some key residues. Foremost among these changes is the substitution of at least one of three Histidines (His94, His96, His119) which coordinate the zinc atom present in all active CAs. Site specific mutation of these three positions, to His, restores catalytic activity to each CARP (Nishimori, et al., 2012). The function of the three CARPs has not been determined although it is a safe assumption that

their purpose still revolves around $CO_2$. Like their active counterparts the CARP proteins are expressed at low levels across all tissues sampled by microarray with each present in higher quantities in at least one specific tissue. The CARP-VIII protein is expressed significantly in only one tissue, as identified in microarray analysis, the cerebellum (Wu, et al., 2009); antibody staining confirms cerebellum as the primary tissue and localizes to the Purkinje cells (Uhlen, et al., 2010). According to microarray analysis, CARP-X is found expressed at significant levels in pineal gland, cerebellum, and cerebellar peduncles (Wu, et al., 2009). The final CARP, CA-XI, is highly expressed in many neural tissues and retina; within the brain the more significantly expressed tissues are prefrontal cortex and amygdala (Wu, et al., 2009).

| CAs | CA-I | CA-II | CA-III | CA-IV | CA-Va | CA-Vb | CA-VI | CA-VII | CA-VIII | CA-IX | CA-X | CA-XI | CA-XII | CA-XIII | CA-XIV |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CA-I | 100 | | | | | | | | | | | | | | |
| CA-II | 60 | 100 | | | | | | | | | | | | | |
| CA-III | 53 | 58 | 100 | | | | | | | | | | | | |
| CA-IV | 30 | 33 | 31 | 100 | | | | | | | | | | | |
| CA-Va | 47 | 50 | 45 | 23 | 100 | | | | | | | | | | |
| CA-Vb | 46 | 52 | 43 | 23 | 58 | 100 | | | | | | | | | |
| CA-VI | 31 | 33 | 32 | 26 | 27 | 24 | 100 | | | | | | | | |
| CA-VII | 50 | 56 | 49 | 31 | 48 | 39 | 34 | 100 | | | | | | | |
| CA-VIII | 39 | 40 | 38 | 27 | 33 | 34 | 28 | 41 | 100 | | | | | | |
| CA-IX | 31 | 33 | 20 | 25 | 27 | 26 | 35 | 35 | 31 | 100 | | | | | |
| CA-X | 26 | 28 | 27 | 23 | 24 | 24 | 19 | 27 | 27 | 20 | 100 | | | | |
| CA-XI | 25 | 30 | 28 | 22 | 24 | 26 | 18 | 26 | 28 | 21 | 50 | 100 | | | |
| CA-XII | 35 | 34 | 31 | 26 | 27 | 25 | 32 | 37 | 28 | 36 | 21 | 19 | 100 | | |
| CA-XIII | 59 | 59 | 57 | 28 | 46 | 47 | 33 | 52 | 39 | 34 | 27 | 27 | 34 | 100 | |
| CA-XIV | 34 | 35 | 34 | 25 | 29 | 25 | 32 | 35 | 27 | 37 | 19 | 23 | 40 | 37 | 100 |

Table 2 - Percent identity of all human CAs and CARPs [14].

## 2.8 Disease Related to Carbonic Anhydrase

A variety of diseases are related to expression levels of the various carbonic anhydrase isoforms. As evidenced by the preceding section nearly all of the CA proteins are expressed to some level in the tissues examined by microarray data analysis. It becomes easy to speculate that it is due to this redundancy there are not significantly more conditions associated with this protein family. However, the association with a number of cancer types, in numerous CA types, makes the CA family a serious target for cancer research. A table outlining some of the conditions and cancers associated with carbonic anhydrase follows.

| CA gene | Associated Conditions | Associated Cancers |
|---|---|---|
| *CA1* | hyperthyroidism, erythroleukemia, thyroid diseases (Stelzer, Dalah, Stein, & al, 2011) | thyroid (Odcikin, Ozdemir, Ciftci, & al, 2002); non-small cell lung cancer (Chiang, Chu, Yang, & al, 2002); adenocarcinomas of prostate, stomach, breast, and ovary (Ozensoy, Kockar, Arslan, & al, 2006);oral squamous cell carcinoma (Liu, et al., 2012) |
| *CA2* | acidosis renal tubular, osteopetrosis, calcification, glaucoma, biliary cirrhosis, Sjogrens syndrome, chronic pancreatitis (Stelzer, Dalah, Stein, & al, 2011) | squamous cell carcinoma (Chiang, Chu, Yang, & al, 2002); adenocarcinomas of prostate, lung, stomach, breast, and ovary (Ozensoy, Kockar, Arslan, & al, 2006) |
| *CA3* | reflux, rhabdomyolysis, neuromuscular diseases (Stelzer, Dalah, Stein, & al, 2011) | |
| *CA4* | retinitis pigmentosa, glaucoma (Stelzer, Dalah, Stein, & al, 2011) | |
| *CA8* | | non-small cell lung cancer, colorectal carcinoma (Stelzer, Dalah, Stein, & al, 2011) |
| *CA9* | | renal clear cell carcinoma (Stelzer, Dalah, Stein, & al, 2011); carcinomas of lung, bladder, rectum, colon, uterine cervix, and breast, mesothelioma, and brain cancer (Masayuki Nakao, 2009) (Sherwood, Colquhoun, & D., 2007) (Rasheed, Harris, & Tekkis, 2008) (Liao, Darcy, Randall, & al, 2010) (Hsieh, Chen, Chiou, & al, 2010) (Kivela, Knuuttila, Sihvo, & al, 2012) (Dungwa, Hunt, Ramani, & al, 2012) |
| *CA12* | | renal clear cell carcinoma, colorectal and breast carcinomas, brain cancer (Stelzer, Dalah, Stein, & al, 2011); uterus (Hynninen, et al., 2011); oral (Chien, et al., 2012) |
| *CA13* | | Colorectal cancer (Kummola, et al., 2005) |

**Table 3 - Conditions and cancers associated with each carbonic anhydrase.**

# 3. Active Site

The reversible hydration of $CO_2$, occurring in the active site of carbonic anhydrase, is a two-step process. In the first step, during hydration, zinc-bound water is converted to $OH^-$ when His64 shuttles an $H^+$ ion out of the active site into bulk solution. In the second step the newly activated $OH^-$ attacks $CO_2$ creating $HCO_3^-$ (bicarbonate). Water then replaces bicarbonate which is released into bulk solution (Lindskog, 1997) (Fisher, et al., 2007) (Zheng, Avvaru, Tu, & Silverman, 2008).

1) **Buffer + His64 + Zn-$H_2O$ ↔ Buffer + $H^+$-His64 + Zn-$OH^-$ ↔ $H^+$-Buffer + His64 + Zn-$OH^-$**
2) **Zn-$OH^-$ + $CO_2$ ↔ Zn-$HCO_3$ + $H_2O$ ↔ Zn-$H_2O$ + $HCO_3$**

Protons leaving and entering the active site are passed through a series of three water molecules, commonly called W1, W2, and W3a/W3b, whose position is maintained by hydrogen bonding with amino acids lining the active site (Fisher, et al., 2007) (Maupin & Voth, 2010) (Fisher, et al., 2005). Furthest from the site of catalysis, and after W2, the water chain bifurcates into W3A and W3B.

The movement of hydrogen into and out of the active site of carbonic anhydrase, for the purpose of catalysis, is well established. However, the method of this movement is not perfectly understood and more than one hypothesis of this action exists which are discussed in the following sections.

## 3.1 Proton Shuttle

### 3.1.1 Grotthuss Mechanism

In many experimental papers on carbonic anhydrase, and more commonly, the Grotthuss mechanism of proton transfer is assumed to facilitate proton shuttling. In a paper published in 1806, German born Theodor Grotthuss hypothesized, in French, on the transfer of charge through a solution of water by conducting experiments with battery and solution (Cukierman, 2006). Grotthuss surmised that positive and negative properties must both exist within the water molecule, and correctly associated the charges with H and O. Despite his misunderstanding of the water molecule, in his belief that it consisted of a single O and H,

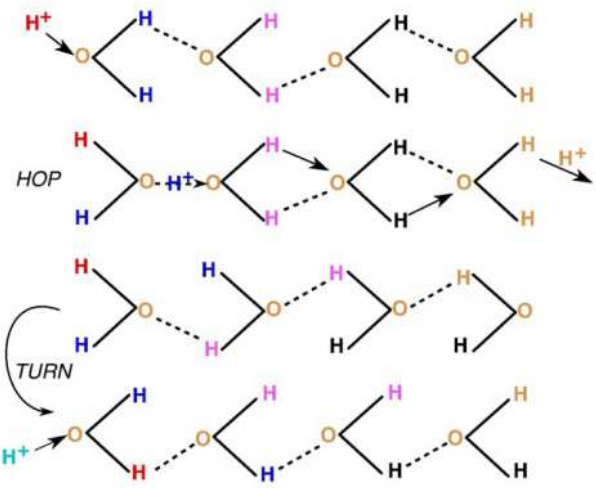


Figure 4 - Proton Shuttling in Grotthus Mechanism – Cuckierman "et tu grotthuss" (Cukierman, 2006).

he envisaged a chain of temporary molecular unions as dissociated charged atoms traveled through water towards their respective complementary electrodes (Cukierman, 2006). The understanding of this process has evolved since its first elucidation, and an underlying principle has taken shape, that proton shuttling occurs through temporary formation of a Hydronium ion passing like a wave through adjacent water molecules in a chain later described a 'wire' (Nagle & Morowitz, 1978). Modern interpretation of the Grotthuss mechanism additionally describes hydrogen bonding between H and O in sequential waters (Cukierman, 2006). This wire consists of a hydronium and three waters, the composite being called an Eigen cation $H_3O^+$-$(H_2O)_3$ (Figure 4 row 1) (Cukierman, 2006) (Wraight, 2006) (Riccardi, et al., 2006). As a free $H^+$ bonds to the first accepting water ($H_2O^1$) one of its hydrogen's, whichever has previously hydrogen-bonded to O in the next water ($H_2O^2$), will detach to form the next Hydronium, this process is visualized in (Figure 4 row 2). In the intermediate step, before a $H^+$ is passed to the next water, a complex known as a Zundel cation $H_2O$-$H^+$-$OH_2$ forms (Figure 4 row 2) (Wraight, 2006) (Riccardi, et al., 2006). This process is repeated until the proton has left the final water. At the completion of this process the reoriented molecules will need to rotate to the starting position in order to accept a new $H^+$ (Cukierman, 2006) (Kale, Herfeld, Dai, & Blank, 2012).

### 3.1.2 Proton Hole Mechanism

A counter view to the Grotthuss mechanism is the "Proton Hole" mechanism (PH). In this scenario the proton shuttling action begins at the final water of the group which releases an $H^+$. The $OH^-$ proton hole must then be filled by the preceding waters of the wire (Riccardi, et al., 2006). In this fashion a proton will move in the same direction as the Grotthuss mechanism however the alteration of water moves in opposite direction (receptor to donor). Additionally, this approach accounts for the creation of a wave of negatively, instead of positively, charged molecules. Riccardi et al. propose that the PH mechanism is potentially as commonplace as the Grotthuss mechanism. They assert that, in cases when the final water in a proton shuttling wire inhabits a basic environment, energetically, a primary release of $H^+$ will be favored (Riccardi, et al., 2006). The opposite is assumed in cases where the final water inhabits an acidic environment. Predictions about the nature of proton shuttling, and in particular that of CAII, were generated using a complex model based on both quantum mechanical and molecular mechanical (QM/MM) simulations (Riccardi, et al., 2006). The results showed that the serial conversion of water to $OH^-$ was favored for the transfer of protons, and that polar residues in the active site favorably stabilized $OH^-$ (Riccardi, et al., 2006). While acknowledging the significant potential for error in complex simulations the researchers support their supposition with a relevant fact. Owing to the function of CA to interconvert the negatively charged bicarbonate, it is likely the active site has evolved to this purpose and therefore favors negatively charged molecules.

### 3.1.3 Combination Grotthuss & Proton Hole

A combinatory explanation of proton transfer was proposed by Kale et al., after simulation using their less computationally expensive LEWIS model. In this scenario the final proton receptor, receives $H^+$ from the preceding water characteristic of the PH mechanism. At the same time the Hydronium ion at the proton donating end transfers $H^+$ to the water in position two. The proton and proton hole travel towards one another until meeting in the middle of the water wire (Kale, Herfeld, Dai, & Blank, 2012).

## 3.2 CA Active Site Amino Acids

Aside from the internal structure supporting amino acids, those that occupy the active site are quite understandably some of the most important within the whole protein. It is for this reason that a number of these amino acids are strongly conserved both among species' orthologs and across isoforms. In the following sections are discussed those active site amino acids which have been noted to be of functional significance when it comes to catalytic activity of the often studied CA-II protein. The majority of studies reference the CA-II number system, as a result all following references to CA sequence position shall be made based on that scale unless otherwise noted.

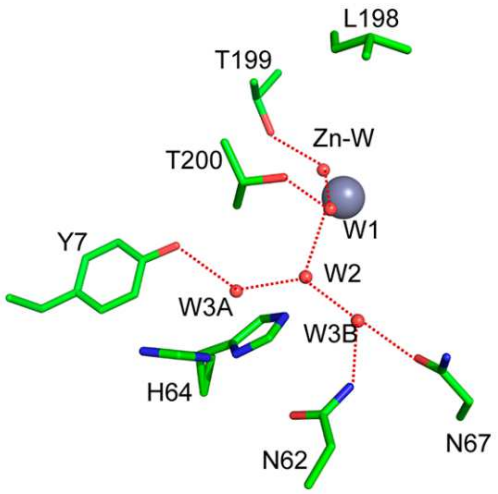


**Figure 5 - Important amino acids of the wild type HCA-II active site. The large central grey sphere represents a zinc atom while the smaller red spheres represent water and are thus labeled (W1, W2, W3A, and W3B). The dashed red lines are hydrogen bonds** (Mikulski, et al., 2012).

### 3.2.1 Tyrosine7

Previous experiments have shown that substitution for Tyr7 has a dramatic effect on proton shuttling into and out of the active site while having no effect on de/hydration. Mikulski et al. replaced Tyr7 with seven amino acids (I, A, W, D, N, R, and S) resulting in proton transfer rates ranging between 0.8-2.5 $\mu s^{-1}$ compared to 0.8 $\mu s^{-1}$ for wild type (Mikulski, et al., 2011). An even more significant gain was achieved with a Y7F replacement, as performed by Fisher et al. (Fisher, et al., 2007), resulting in a seven-fold increase in rate at 7 $\mu s^{-1}$.

Experiments performed by Fisher et al. tested the effect on proton shuttling ability of mutations at the amino acids which maintain the W3a and W3b water molecules, specifically

Y7F, N62L and N67L (Fisher, et al., 2007). The previously mentioned proton transfer rate of 7 $\mu s^{-1}$, achieved by Y7F replacement, was attributed to three factors: Observed inward orientation of His64 side-chain, collapse of the W3a water position, and increased acidity of His64 (Fisher, et al., 2007).

In support of the first supposition, An et al. showed in 2002 that the outward orientation of His64 side-chain non-critical for functioning of the proton shuttle (An, et al., 2002). Rescue of catalytically inactive H64A HCA II occurs through addition of proton donor 4-methylimidazole (4-MI). The Trp5 side-chain occupies the 'out' position of His64 in HCA II wild-type and is a known binding site for 4-MI. Replacement of Trp5 with residues blocking 4-MI binding did not disrupt rescue of the H64A mutant, suggesting that the 'out' position does not promote $H^+$ shuttling in HCA II (An, et al., 2002). Despite the enhanced proton shuttling ability of the Y7F mutant the Tyr7 residue is widely conserved among chordates. The reduced thermo-stability of a Y7F protein, by ~4°C and higher in other replacements (Mikulski, et al., 2011), is a potential reason for retention of Tyrosine at this position.

Regarding the second supposition, the bifurcation of the proton shuttling pathway that occurs between W2 and W3A/B may result in reduced strength of hydrogen bonding between them. This, comparatively, reduced affinity could limit the speed of proton transport at that juncture. Supporting this idea is a recent experiment in which a double mutant human CA-II (Y7F and N67Q) also observed a collapse of the W3A water position and catalytic activity as high as 9 $\mu s^{-1}$ (Mikulski, et al., 2012); in this experiment the authors note that a reduced distance between elements of the water chain resulted in stronger hydrogen bonds.

### 3.2.2 Asn62

The hydrophilic asparagine residue at position 62 plays an important role in both supporting the water network and configuration of the His64 side-chain. Specifically, the Asn62 position supports water position W3B as can be seen in Figure 5. An experiment by Zheng et al. showed that site mutations of this Asn amino acid in HCA-II (N62A, N62V, N62L, and N62D) produced a CA protein with either predominantly inward or outward oriented His64 side-chain (Zheng, Avvaru, Tu, & Silverman, 2008). Fractional catalytic activity was observed from the one mutant (N62D) that produced a majority outward orientation.

### 3.2.3 Asn67

The asparagine located at position 67 is also involved in maintenance of the water chain and side chain position of His64. A N67L mutant CA-II, created by Maupin et al., was shown to also cause His64 to favor an outward orientation while at the same time disrupting the water

network (Maupin, et al., 2009). Along with Asn62, the Asn67 position supports the water chain at position W3B (Figure 6).

### 3.2.4 His64

Arguably the most important residue involved in the transfer of protons to and from the active site is histidine at position 64. In wild type human CA-II the His64 amino acid side chain can occupy a space between water chain positions W3A and W3B (Figure 5). From this location it is able to receive from, or donate to, either of these entry points to the water chain. Many experiments, both in-vivo and in-silico, have demonstrated the effect that loss or reorientation of His64 has on the catalytic activity of CA (primarily CAII) (Zheng, Avvaru, Tu, & Silverman, 2008) (An, et al., 2002) (Riccardi, et al., 2006) (Mikulski, et al., 2011) (Fisher, et al., 2005) (Becker, Klier, Schuler, McKenna, & Deitmer, 2011). It has not been conclusively demonstrated whether or not the imidazole side-chain of His64 needs to be oriented towards the active site, the 'in' position, or have flexibility to switch between both 'in' and 'out'. However, experiments have shown that when the side-chain is maintained in an outward position the catalytic activity of CA drops dramatically (Zheng, Avvaru, Tu, & Silverman, 2008) (Maupin, et al., 2009) (Zheng, Avvaru, Tu, & Silverman, 2008).

### 3.2.5 Thr199

A threonine at position 199 performs a very significant function and yet is not studied as often as other active site residues. This residue is highly conserved in human CAs present in all isoforms except CA-XI. Located directly next to the zinc in the deepest reaches of the active site, this amino acid forms a hydrogen bond with the $H_2O$ or $OH^-$ which is bound to zinc. Previous experiments have shown that substitution at this location for a hydrophobic alanine results in a rate of catalysis reduced by a factor of 100 (Liang, Xue, Behravan, Jonsson, & Lindskog, 1993). Additionally, the amine group of the Thr199 position forms a hydrogen bond to a water molecule most greatly recessed in the active site, and aptly called the 'deep water', which the zinc bound water also binds with (Lindskog, 1997). Owing to its deep location and multiple hydrogen bonds, the Thr199 position has been identified as a location where CA inhibitors bind (Lindskog, 1997).

### 3.2.6 Thr200

Threonine at position 200 is responsible for stabilizing the water chain at position W1 (Figure 5). Unlike Thr199 this amino acid is significantly more variable in human CAs and is substituted with histidine (CA-I), valine (CA-XIII), and isoleucine (CA-VIII, CA-X). A 1990 study by Behravan

et al. made the assertion that Arg, Val, Ile, or His in this position resulted in activity approximating that of CA-I (Behravan, Jonsson, & Lindskog, 1991). Interestingly, those isoforms lacking Thr at this position, including two catalytically inactive CARPs (CA-VIII, CA-X), are those with the 2nd and 3rd slowest catalytic rates (CA-XIII, CA-I).

## 3.3 Tools and Theory

The entirety of the work performed in this thesis was completed on a laptop based Ubuntu Linux installation. Of the software explained in the following sections some versions are available across multiple platforms, in at least some level of functionality, while others are currently available only for Linux.

### 3.3.1 Ensembl

The Ensembl online database (http://www.ensembl.org) provides both browser-based and MySQL server access to genome data for 72 species as of its most current release number 72 (made available in April 2013). The species available are primarily chordate and mammalian, however a variety of fish and bird species are present as well as a few invertebrates. Additional releases are generally based on the addition of new species to the database and occur roughly every 3 months while previous releases remain accessible indefinitely. Annotated genes are easily searched within all, or particular, species. Homolog relationships are predicted by the Compara pipeline which can be used to easily compare both intra and interspecies sequence variations.

Ensembl release number 68 (made available in July 2012) was the primary source of data for this thesis as it provided the best compatibility with the PyCogent library for data retrieval (discussed in a later section). The Ensembl database was used to retrieve: protein sequences of orthologs, raw genomic data surrounding genes, and expressed sequence tags (EST) associated with target genes. ESTs have been derived from mRNA sequences and are retrieved from the UniGene database.

### 3.3.2 Python

All scripting created in order to complete this thesis was written using the freely available, open source, and dynamic Python programming language (Python.org). For complete compatibility with existing supplementary modules, such as Biopython, Python version 2.7 was used despite the existence of the most current version 3.3.

### 3.3.3 Biopython

The widely used, freely available, and open source Biopython libraries have enabled handling of, and performing functions on, DNA and protein sequences within the Python environment since 1999 (Cock, et al., 2009). The Biopython package has a wide array of functionality, and in addition to creating simple sequence objects will also: load Multiple Sequence Alignments (MSA), query various sequence databases (Entrez, ExPASy, InterPro, KEGG, and SCOP), and perform phylogenetic related functions. For the purposes of this thesis Biopython was used to create sequence records and write/handle FASTA sequence files.

### 3.3.4 PyCogent

Perhaps the single most import set of libraries involved in this thesis are those described by **Py**thon the **C**omparative **GEN**omic **T**oolkit, or PyCogent. Similar to Biopython, this library toolkit provides a wide variety of bioinformatics tools to be called by the Python programming language. Some of the tools present in this library include: access to a number of online databases (NCBI, KEGG, PDB, Rfam, and most notably Ensembl), phylogenetic analysis, and sequence alignment (Knight, et al., 2007).

While supporting a variety of analyses, PyCogent was initially designed to focus on comparative analysis of sequences by providing a wealth of tools for retrieving sequences, phylogenetic analysis with a variety of models, and visualization. For the purposes of this thesis PyCogent was primarily utilized to query the Ensembl database for orthologs, genomic regions, and ESTs. Additionally, the PyCogent phylogenetic tools were utilized to determine relatedness of orthologs using the F81 maximum likelihood model.

### 3.3.5 GeneWise

The primary use of the GeneWise algorithm is to predict a gene from genomic data using a related sequence as a template. The algorithm is available as a download in the Wise2 software package at the European Bioinformatics Institute (EBI) (www.ebi.ac.uk/Tools/psa/genewise/help/). It is also available for direct use on a number of public university webservers and at EBI. Most importantly, GeneWise is extensively used by the Ensembl project to model genes based on homologous sequences in other species. To predict genes by mapping cDNA and EST data to a species genome, the related, and newer, Genomewise algorithm is used (Birney, Clamp, & Durbin, 2004). The GeneWise webserver was used in the beginning steps of this thesis work for gene prediction before settling on the more robust Exonerate software; it was also the primary software for gene prediction in the manuscript serving as a seed for initiation of this thesis (Tolvanen, et al., 2012).

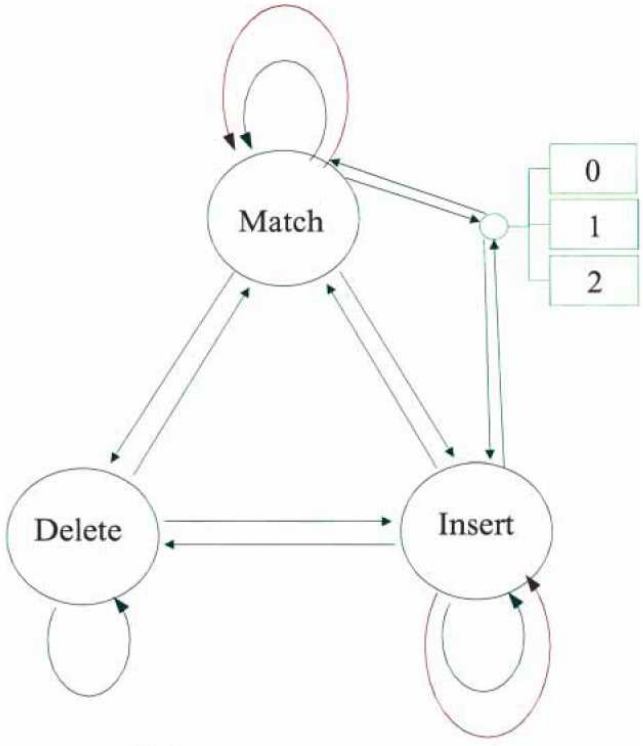


Figure 6 - Diagram of GeneWise6:23 HMM model. States (6) are represented by both circles and rectangles (intron states). Transitions (23) are represented by arrowed lines (Birney, Clamp, & Durbin, 2004).

GeneWise addresses the gene prediction problem, and that of alignment, through the utilization of Hidden Markov Models (HMM) which are statistical methods often initially used in speech recognition (Rabiner, 1989). HMMs find their root in Markov processes, which originated in the 1950's (Meyer, 2009). HMMs can be used to predict the likelihood of protein or DNA sequence occurring after training. The training of an HMM involves analysis of a set of sequences, thus allowing for derivation of frequencies of particular amino acids, or dinucleotide bases, at each position (Figure 7) (Krogh, 1998). Similarly, when comparing two sequences during an alignment, a probability score is generated. The GeneWise algorithm makes comparisons twice and thus uses two separate HMMs. The first HMM compares the genomic data, which is a sequence of DNA, to the six possible protein sequences that can be derived from it (reading frames 1,2,3 in both forward and reverse strands). The second HMM compares these six sequences to the ortholog, or related protein sequence, which has been provided as a template for what the gene may look like (Birney, Clamp, & Durbin, 2004).

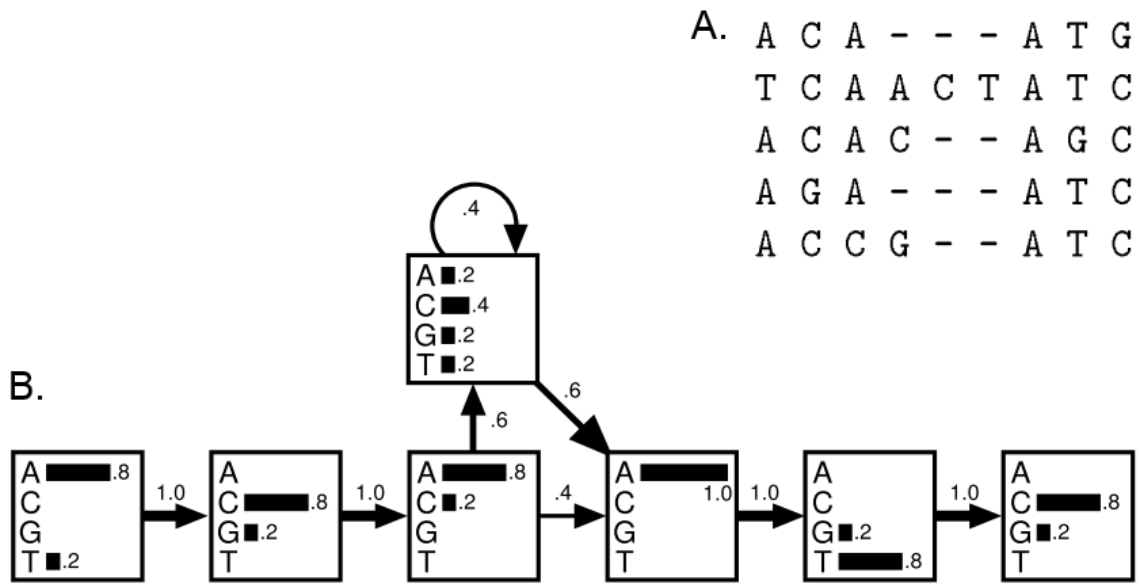


Figure 7 - A) A short segment of MSA used for training an HMM. B) An HMM trained by the sequences in A. At each position in the chain, called a state, the probability of having each base (A,C,G,T) is noted. Arrowed lines indicate transitions. Using this HMM the probability that a new sequence is related to the profile can be calculated (Krogh, 1998).

While multiple GeneWise models exist for user choice (GeneWise21:93, GeneWise6:23, GeneWise4:21), the default model, and most heavily tested, is GeneWise6:23. In this model there are 6 states and 23 transitions (Figure 6).

### 3.3.6 Exonerate

A novel algorithm named Bounded Sparse Dynamic Programming (BSDP) was developed by Guy Slater and Ewan Birney of the European Bioinformatics Institute (EBI). Optimal alignment building between two sequences, such as that achieved by Dynamic Programming (DP) in the Needleman-Wunsch algorithm is computationally expensive. Given two sequences of length *m* and *n*, time complexity for alignment is represented by *m* x *n* or if the two sequences are the same length $m^2$. This is known as quadratic time ($O(m^2)$) (Slater & Birney, 2005). The objective of the algorithm is to reduce time complexity by only calling for the more computationally intensive DP alignment method over short distances instead of the full length of two sequences. This algorithm builds High-scoring Segment Pairs (HSP) in the same fashion as the traditional Basic Local Alignment Search Tool (BLAST). Once an HSP has been generated the character possessing the median score within an HSP is located. This location is identified such that half the total score of the HSP falls on either side (Slater & Birney, 2005). In joining HSPs to form a continuous alignment a Sub-Alignment Region (SAR) is created as a rectangle whose opposing corners are defined by the previously identified median score character (Figure 8).

When there are multiple possible combinations of HSPs, an upper score limit, or bound, is calculated for each combination. If a score for an HSP combination is calculated that is higher than the upper possibility for the other combinations, a full calculation is not made for the remaining possibilities and they are discarded. This approach dramatically reduces the time required for an alignment. The BSDP approach identifies five possible HSP scenarios and defines a specific approach for each. Most relevant to this thesis work is the connection of HSPs on either side of an intron. In this case the SAR scores are carried from one HSP to the next. Additionally, splice site prediction is built into the algorithm.

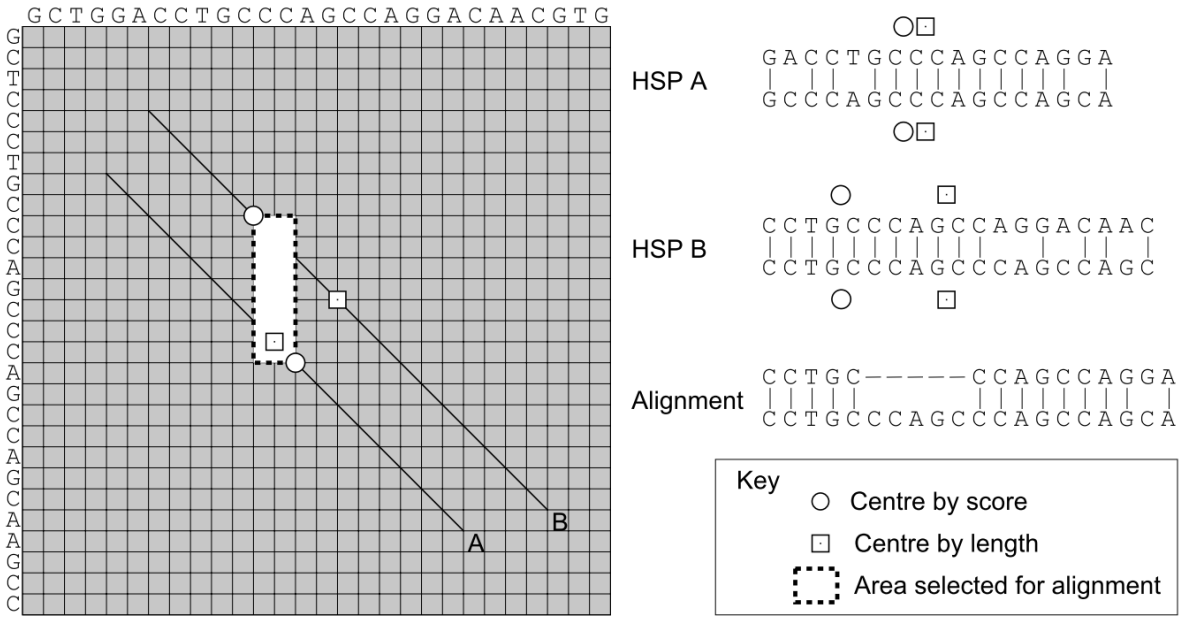


**Figure 8 – BSDP joining of two HSPs. A Sub-Alignment Region is generated as bounded by the center defined by total score of each HSP. A DP approach will only be applied within this region.** (Slater & Birney, 2005)

The Exonerate software package created by Guy Slater and Ewan Birney of EBI implements a number of models based on alignment algorithms ranging from simple ungapped alignments of protein or DNA sequences to alignment of very large sequences on the scale of whole chromosomes, named as follows:

- ungapped
- affine
- est2genome
- ner
- protein2dna
- protein2genome
- coding2coding
- cdna2genome
- genome2genome

The specific model utilized in this thesis allows for comparison of a query protein sequence to a target genomic sequence (protein2genome). When Ensembl contained only a partial protein record for a CA protein for a certain species the genomic region containing the partial protein for that species was compared to the full protein sequence of a closely related species.

### 3.3.7 Clustal Omega

In their 2011 paper Sievers et al. extoll the benefits of the algorithms used in their new Clustal Omega Multiple Sequence Alignment (MSA) package (Sievers, et al., 2011). The algorithm performs comparably with high quality alignment algorithms when aligning low numbers of sequences. However, Clustal Omega outperforms all other algorithms when large numbers of sequences, or particularly long sequences, are aligned. Furthermore, the package is able to operate on nearly indefinite alignment sizes by today's standards. Computational complexity for an optimal alignment of more than two sequences is defined by $O(L^N)$ where $L$ is the length of sequences and $N$ is the number of sequences (Sievers, et al., 2011). A time demand of this complexity is prohibitive for excessively long or numerous sequences. Indeed, when Needleman and Wunsch first demonstrated their alignment algorithm their paper was limited to 10 runs of 5 sequences of length ~150 amino acids due to processing time constraints (Needleman & Wunsch, 1970). Since the development of early, exhaustive alignment algorithms a variety of new approaches have been built that greatly increase the speed of alignment while attempting to maintain as much of their accuracy as possible. However, the majority of these approaches either cannot handle aligning more than a few thousand sequences and/or are not accurate when handling larger alignments. Sievers et al. state that their program has successfully handled up to 190,000 sequences in an alignment performed by a single processor in "a few hours" (Sievers, et al., 2011). While Clustal Omega does not rank the highest in quality of alignment it does offer the best ratio of quality and speed and it is for this reason that it was used for general MSAs in this thesis.

### 3.3.8 DSSP

In 1983, when the number of known protein structures was just over 100, Wolfgang Kabsch and Christian Sander described a Dictionary of Protein Secondary Structure (DSSP) (Kabsch & Sander, 1983). The objective of the DSSP was identification of secondary structure from atomic coordinates created during x-ray crystallography. At the time the determination of secondary structure (SS) was based on either a subjective manual approach or an algorithm inferior in accuracy; Kabsch and Sander decided to base their algorithm on H-bonding patterns (Kabsch & Sander, 1983). In addition to SS, Kabsch and Sander also describe how "solvent exposure" for

any residue is calculated based on the number of water molecules (W) that can interact with its surface. Using the following equation:

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

This determination becomes very useful when deciding which residues of a protein are accessible to the surface and therefore likely play an interacting role with other proteins, compounds, or water. However, an additional step is required after solvent exposure has been determined. In this step the solvent exposure of the residue is divided by the total surface area of the specific amino acid, thus deriving the Relative Solvent Accessibility (RSA). Assignment of a protein integrated amino acid to either 'surface' or 'buried' is based on RSA, however no strict guidelines have been set. Generally an RSA of 0.25 to 0.30 is used to demarcate the boundary between these two categories, where a value lower indicates the amino acid is buried inside of the protein and a higher value indicates the amino acid is on the surface and accessible for interaction. However, a recent study by Amir Momen-Roknabadi et al. proposes that when predicting secondary structure utilizing RSA values, a single threshold for buried/exposed assignment does not perform as well as using thresholds unique to each amino acid (Momen-Roknabadi, Sadeghi, Pezeshk, & Marashi, 2008). In a 2012 study of 587 yeast genes, Scherrer et al. presented their findings that the RSA and conservation of amino acids were linearly related (Scherrer, Meyer, & Wilke, 2012). The relationship highlighted increased conservation of residues with low RSA (buried location) and increased variability of residues having a higher RSA (surface location). This observation is logical under the premise that residues near the center of a protein often perform a structural function, and more greatly impact proper folding, while those at the surface do so to a lesser extent.

Since the publication of the DSSP in 1983 a program has been produced for automated processing of PDB files. The program is available online (http://swift.cmbi.ru.nl/gv/dssp/) and is written for use on both Windows and Linux. Additionally, there are a number of databases which contain pre-computed data for all current PDB files as described by (Joosten, et al., 2010).

### 3.3.9 PAL2NAL

The PAL2NAL algorithm is available for use online in its webserver version at the Bork group website (http://www.bork.embl.de/pal2nal/), and as downloadable program, coded in Perl, which can be utilized on Linux. The program's primary function is to take as input both aligned protein sequences (Figure 9) and their corresponding unaligned DNA sequences (Figure 10), and output codon-aligned DNA sequences (Figure 11); this pair of sequence alignments (codon-aligned DNA and protein) is essential when computing $K_a/K_s$ values for conservation analysis

and for constructing more refined phylogenies (Suyama, Torrents, & Bork, 2006). A few conservation analysis programs that utilize codon alignments are:

- gKaKs (Zhang, Wang, Long, & Fan, 2013)
- Selecton (Doron-Faigenboim, Stern, Mayrose, Bacharach, & Pupko, 2005)
- WSPMaker (Lee, Kim, Kang, Chung, & Shin, 2008)
- SNAP (HIV Databases, 2013) (Korber, 2000)
- HyPhy/DataMonkey (Pond, Frost, & Muse, 2005) (Delport, Poon, Frost, & Pond, 2010)

Phylogeny calculating programs which utilize codon alignment are:

- PAML (Yang, Phylogenetic Analysis by Maximum Likelihood (PAML), 1997)
- MrBayes (Ronquist, et al., 2011)
- PhyML (Guindon & Gascuel, 2003)

While a number of programs exist that will make a codon alignment for perfectly translatable DNA sequences, PAL2NAL allows for the wholly common discrepancies that can exist between DNA and amino acid sequences that purportedly represent the same protein. The algorithm will also account for such features as untranslated regions (UTRs) and polyA tails, which are removed in an initial step (Suyama, Torrents, & Bork, 2006).

Programmatically and algorithmically the PAL2NAL is both simple and clever in its approach. The amino acid sequences are converted to regular expression based on the redundancy of the codon table; options for both 'universal' and 'vertebrate mitochondria' are available. For example, a short amino acid sequence **MNG** would be represented as **(A(T|U)G) (AA(T|U|C)) (GG.)**. Frame shifts due to insertion/deletion are easily overcome by changing the regular expression pattern to allow for the changed number of bases between two known residues, however this is only addressed within coding regions (Suyama, Torrents, & Bork, 2006).

After codon-alignment of the DNA sequences the PAL2NAL program is itself capable of performing a $K_a/K_s$ analysis through calling of codeml program included in the PAML package (Yang, Phylogenetic Analysis by Maximum Likelihood (PAML), 1997). However, the program limits the analysis to only two sequences due to the computational complexity of the analysis.

A

```
CLUSTAL W multiple sequence alignment

BC070280        MVGSLNCIVAVSQNMGIGKNGDLPWPPLRNEFRYFQRMTTTSSVEGKQNLVIMGKKTWFSIPEKNRPLKGRINLVLSR
pseudogene      ----LNCIVNVSQKMGIIRNGDLP*PQLKNKF2-FQRMTTPSSAEGKENLVFLIRKNWFSITEKNQPLKYIINLVVSR
                    ##################            ##########################################

BC070280        ELKEPPQGAHFLSRSLDDALKLTEQPELANKVDMLWIVGGSSVYKEAMNHPGHLKLFVTRIMQDFESDTFF-PEIDLE
pseudogene      ESKEPPQRPPFLD*SLGDALKRIEQLKLANKQDVFFTVGGSSVYKESMN*-DHFKLFVTWIMQDFQSDTFFS4EGDLE
                              ######################

BC070280        KYKLLPEYP-GVLSDVQEEKGIKYKFEVYEKND
pseudogene      KYKLLPEYPQGVVSDVEEEKGIKYKFEVYEKND
```

**Figure 9 – An alignment of protein sequences, at least two, are required as input to begin the codon alignment process within PAL2NAL** (Suyama, Torrents, & Bork, 2006).

B

```
>BC0700280  dihydrofolate reductase (human)
TGTAACGAGC GGGCTCGGAG GTCCTCCCGC TGCTGTCATG GTTGGTTCGC TAAACTGCAT CGTCGCTGTG TCCCAGAACA TGGGCATCGG
CAAGAACGGG GACCTGCCCT GGCCACCGCT CAGGAATGAA TTCAGATATT TCCAGAGAAT GACCACAACC TCTTCAGTAG AAGGTAAACA
GAATCTGGTG ATTATGGGTA AGAAGACCTG GTTCTCCATT CCTGAGAAGA ATCGACCTTT AAAGGGTAGA ATTAATTTAG TTCTCAGCAG
AGAACTCAAG GAACCTCCAC AAGGAGCTCA TTTTCTTTCC AGAAGTCTAG ATGATGCCTT AAAACTTACT GAACAACCAG AATTAGCAAA
TAAAGTAGAC ATGcTCTGGA TAGTTGGTGG CAGTTCTGTT TATAAGGAAG CCATGAATCA CCCAGGCCAT CTTAAACTAT TTGTGACAAG
GATCATGCAA GACTTTGAAA GTGACACGTT TTTTCCAGAA ATTGATTTGG AGAAATATAA ACTTCTGCCA GAATACCCAG GTGTTCTCTC
TGATGTCCAG GAGGAGAAAG GCATTAAGTA CAAATTTGAA GTATATGAGA AGAATGATTA ATATGAAGGT GTTTTCTAGT TTAAGTTGTT
CCCCCTCCCT CTGAAAAAAG TATGTATTTT TACATTAGAA AAGGTTTTTT GTTGACTTTA GATCTATAAT TATTTCTAAG CAACTTGTTT
TTATTCCCCA CTACTCTTGT CTCTATCAGA TACCATTTAT GAGACATTCT TGCTATAACT AAGTGCTTCT CCAAGACCCC AACTGAGTCC
CCAGCACCTG CTACAGTGAG CTGCCATTCC ACACCCATCA CATGTGGCAC TCTTGCCAGT CCTTGACATT GTCGGGCTTT TCACATGTTG
GTAATATTTA TTAAAGATGA AGATCCACAT ACCCTTCAAA AAAAAAAAAA AAAAAAAAAA AAAAAAA
>pseudogene  dihydrofolate reductase pseudogene (human)
CTAAACTGCA TTGTCAATGA TTCCCAGAAG ATGGGCATCA TCAGGAATGG GGACCTGCCC TGACCTCAGC TCAAAAATAA ATTCGATTCC
AAAGAATGAC CACACCCTCT TCAGCAGAGG GTAAAGAAAA TTTAGTATTT TTAATTAGGA AGAACTGGTT CTCGATTACT GAGAAGAATC
AACCTTTAAA GTATATAATT AATTTAGTTG TCAGTAGAGA ATCCAAGGAA CCACCGCAAA GACCTCCTTT TCTTGACTAA AGTCTGGGTG
ATGCCTTAAA ACGTATTGAG CAACTAAAAT TAGCAAATAA ACAAGACGTG TTTTTTACAG TGGGAGGCAG TTCTGTTTAT AAGGAATCCA
TGAATTGAGA CCATTTTAAA CTATTTGTGA CATGGATCAT GCAGGACTTT CAAAGTGACA CGTTTTTTTC CCCTAGAAGG TGATTTAGAG
AAATATAAAC TTCTCCCAGA ATACCCACAA GGTGTTGTCT CTGATGTGGA GGAGGAGAAA GGCATTAAGT ACAAATTTGA AGTATATGAA
AAGAATGAT
```

**Figure 10 – The corresponding unaligned DNA sequences are additionally required as input where the UTR and Poly-A tails are identified (blue) and removed** (Suyama, Torrents, & Bork, 2006).

C

```
#---------------------------------------------------------------------#
#  WARNING: pseudogene pepAlnPos 11: V does not correspond to GAT
#---------------------------------------------------------------------#

CLUSTAL W multiple sequence alignment

BC070280        ATGGTTGGTTCGCTAAACTGCATCGTCGCTGTGTCCCAGAACATGGGCATCGGCAAGAACGGGGACCTGCCCTGGCCA
pseudogene      ------------CTAAACTGCATTGTCAATGATTCCCAGAAGATGGGCATCATCAGGAATGGGGACCTGCCCTGACCT
                            ##################################################

BC070280        CCGCTCAGGAATGAATTCAGATATTTCCAGAGAATGACCACAACCTCTTCAGTAGAAGGTAAACAGAATCTGGTGATT
pseudogene      CAGCTCAAAAATAAATTCGA----TTCCAAAGAATGACCACACCCTCTTCAGCAGAGGGTAAAGAAAATTTAGTATTT
                                      ##################################################

BC070280        ATGGGTAAGAAGACCTGGTTCTCCATTCCTGAGAAGAATCGACCTTTAAAGGGTAGAATTAATTTAGTTCTCAGCAGA
pseudogene      TTAATTAGGAAGAACTGGTTCTCGATTACTGAGAAGAATCAACCTTTAAAGTATATAATTAATTTAGTTGTCAGTAGA
                ########################################################################

BC070280        GAACTCAAGGAACCTCCACAAGGAGCTCATTTTCTTTCCAGAAGTCTAGATGATGCCTTAAAACTTACTGAACAACCA
pseudogene      GAATCCAAGGAACCACCGCAAAGACCTCCTTTTCTTGACTAAAGTCTGGGTGATGCCTTAAAACGTATTGAGCAACTA
                                                                  ############################

BC070280        GAATTAGCAAATAAAGTAGACATGCTCTGGATAGTTGGTGGCAGTTCTGTTTATAAGGAAGCCATGAATCACCCAGGC
pseudogene      AAATTAGCAAATAAACAAGACGTGTTTTTTACAGTGGGAGGCAGTTCTGTTTATAAGGAATCCATGAATTGA---GAC
                ###################################

BC070280        CATCTTAAACTATTTGTGACAAGGATCATGCAAGACTTTGAAAGTGACACGTTTTTT---CCA---GAAATTGATTTG
pseudogene      CATTTTAAACTATTTGTGACATGGATCATGCAGGACTTTCAAAGTGACACGTTTTTTTCCCCTA--GAAGGTGATTTA
```

**Figure 11 - The final output of the PAL2NAL program are the codon aligned DNA sequences** (Suyama, Torrents, & Bork, 2006).

### 3.3.10 Selecton

The Selecton algorithm is available both on a webserver located at (http://selecton.tau.ac.il/) or as a download for both Windows and Linux. Taking as input codon aligned DNA sequences, the Selecton webserver will perform a $K_a/K_s$ analysis, generate a conservation score from 1-7 (where higher indicates greater conservation), and map these scores onto the residues of a 3-D model of the target protein if one is available (Figure 12). Additionally, a linear figure can be produced showing graded conservation (Figure 12), which is especially useful if no model has been determined for the target protein. The downloadable version produces output files necessary to produce these outputs but it must be done manually.

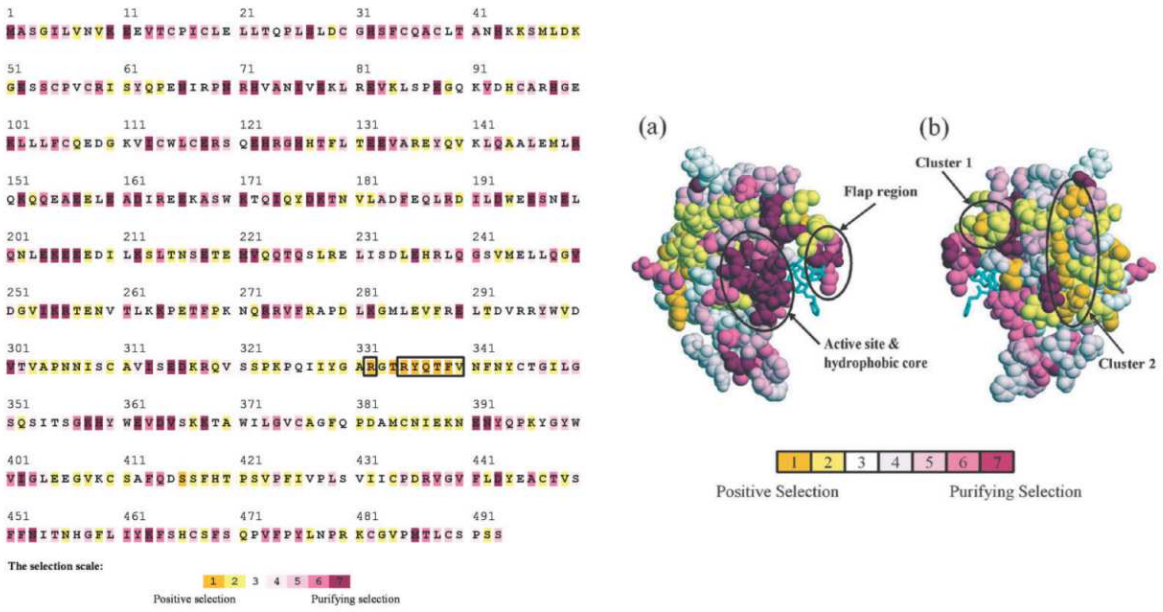


**Figure 12 – At left, a linear figure created by Selecton webserver, depicting results of Ka/Ks conservation analysis of a 493 amino acid protein. Ka/Ks values have been grouped to 7 categories (1-7) and each amino site colored accordingly to allow visual identification of regions of continuous positive or negative selection** (Stern, et al., 2007). **At right, 3D model produced by Selecton webserver. The protein sequences are analyzed for Ka/Ks values which are then mapped onto each amino acid in the protein. Ka/Ks values were divided into 7 categories each of which was assigned a color** (Doron-Faigenboim, Stern, Mayrose, Bacharach, & Pupko, 2005) (Selecton, 2013).

### *3.3.10.1 $K_a/K_s$*

While the Selecton program produces excellent images and models for understanding conservation within groups of homologous proteins, these are simply tools for interpreting the core output, which are $K_a/K_s$ values (identified as ω). In this type of conservation analysis, $K_a$ indicates the incidence of asynonymous mutation (where the subscript 'a' stands for asynonymous). In a codon where the three bases TTT would result in phenylalanine, and a mutation changes the third base to 'A' (TTA), the mutation is asynonymous as the new codon would now present a leucine. On the other hand, if mutation changes the third base to a 'C' (TTC), the redundancy of the codon table allows that this too will code for phenylalanine and as such the mutation is considered synonymous. Assuming that mutation events are essentially random, we would predict that if no selection was occurring then synonymous and asynonymous mutations would occur at equal frequency (when adjusting for the redundancy of the codon table). As a result, an amino acid site with a $K_a/K_s$ ratio that is higher than 1 has *more* mutation than expected by chance, which is considered evidence for positive selection. On the other hand an amino acid site with a $K_a/K_s$ ratio that is less than 1 has *less* mutation than expected by chance, which is considered evidence for 'purifying selection' (Stern, et al., 2007). The seeds for $K_a/K_s$ analysis were first sown by Kimura in 1968 when he proposed the Neutral Theory (Kimura, Evolutionary Rate at the Molecular Level, 1968). Counter to prevailing strictly Darwinian Theory of evolution at the time, the Neutral Theory proposed that the majority of

evolutionary change occurred as a result of random neutral mutations which can be fixed by random genetic drift (Kimura, The neutral theory of molecular evolution: A review of recent evidence, 1991). The Neutral Theory can easily exist alongside Darwinian Theory, at least in the mind of its author, and simply explains the more common, if less dramatic, evolutionary changes that life experiences.

The Selecton program determines ω at each site in a query sequence (e.g. Human CAII) by comparing codon-aligned DNA sequences of homologous proteins (e.g. Human CAII orthologs). The Selecton webserver documentation notes a requirement of 3 sequences with at least 10 being recommended for significant results (Selecton, 2013).

#### *3.3.10.2 Codon Evolutionary Models*

Evolutionary models whose resolution only extends to the level of amino acids can only reliably describe conservation. The benefit of determining these areas of conservation is that it allows identification of what residues might be of most significance within a protein. These residues might be involved in protein-protein interaction, ligand-binding, or be of structural significance. Conversely, residues and regions that are more actively mutated are regions of interest as well for some, but not all, of the same reasons. While a more variable residue is detrimental to internal structure, it can be useful for interacting with other proteins. The example query protein presented in (Stern, et al., 2007) is TRIM5α which, owing to an identified region of increased variability, is thought to confer resistance to HIV-1 in rhesus monkey cells. The benefit of extending evolutionary models to incorporate codon data, and determining ω, is that not only can conservation be determined in a more precise manner but also variation can be measured (Doron-Faigenboim & Pupko, A Combined Empirical and Mechanistic Codon Model, 2006). As there are 3 nucleotide positions per codon, and 4 possible nucleotides per position, the total number of possible codons is $4^3$, or 64. When discounting the three stop codons, there are 61 possible codon substitutions that can be considered; in one of the first codon models of evolution, by Goldman and Yang in 1994, this 61x61 matrix came into use (Zoller & Schneider, 2010). In this early model and those that have come since, a series of parameters are considered to be important to properly build the substitution matrix. Thus, the models have become known as 'parametric models'. The parameters commonly introduced in a number of these models are (Zoller & Schneider, 2010):

- ω – $K_a/K_s$ ratio
- π – Codon frequencies
- *K* – Relative rates of transition/tranversion
- *V* – Physical and chemical distances

The objective of parametric models was to emulate real world frequencies of substitution as closely as possible. In 2005 the first empirical model was created by Schneider et al. by making use of the recently widely expanded sequenced vertebrate genomic data available at Ensembl (Zoller & Schneider, 2010) (Schneider, Cannarozzi, & Gonnet, 2005). However, while this new method was an improvement over many parametric methods, the authors recognized that it was just starting point, and that additional analysis of specific sets of sequences could produce better models (Schneider, Cannarozzi, & Gonnet, 2005). In 2006, Doron-Faigenboim and Pupko created a codon model that was a hybrid of the parametric approach and empirical data; termed the Mechanistic Empirical Codon model (MEC), it is discussed more thoroughly in the next section (Doron-Faigenboim & Pupko, A Combined Empirical and Mechanistic Codon Model, 2006). In 2007, Kosiol et al. claimed their Empirical Codon Model (ECM) was the first empirical model, despite the existence of the 2005 model produced by Schneider et al. The primary difference touted by Kosiol et al. is the consideration of codon substitutions resulting from double and triple simultaneous mutations, instead of just single changes (Kosiol, Holmes, & Goldman, 2007).

| Model | Author | Type | Year |
|---|---|---|---|
| Codon-based Model of Nucleotide Substitution (Goldman & Yang, 1994) | Goldman and Yang | Parametric | 1994 |
| A Likelihood Approach for Comparing Synonymous and Nonsynonymous Nucleotide Substitution Rates (Muse & Gaut, 1994) | Muse and Gaut | Parametric | 1994 |
| Continuous Positive-Selection (Nielsen & Yang, 1998) | Nielsen and Yang | Parametric | 1998 |
| M0-M13 (Yang, Nielsen, Goldman, & Pedersen, Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites, 2000) | Yang et al. | Parametric | 2000 |
| Branch-Site (Yang & Nielsen, Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages, 2002) | Yang and Nielsen | Parametric | 2002 |
| Fixed-Site (Yang & Swanson, Codon-Substitution Models to Detect Adaptive Evolution that Account for Heterogeneous Selective Pressures Among Site Classes, 2002) | Yang and Swanson | Parametric | 2002 |
| Empirical Codon (E) (Schneider, Cannarozzi, & Gonnet, 2005) | Schneider et. al | Empirical | 2005 |
| Mechanistic Empirical Codon (MEC) (Doron-Faigenboim & Pupko, A Combined Empirical and Mechanistic Codon Model, 2006) | Doron-Faigenboim and Pupko | Parametric+Empirical | 2006 |
| Empirical Codon Model (ECM) (Kosiol, Holmes, & Goldman, 2007) | Kosiol, Holmes, and Goldman | Empirical | 2007 |

**Table 4 - Codon models for conservation analysis.**

Selecton provides a number of options of evolutionary models for determination of $K_a/K_s$ values (M5, M7, M8, M8a, and MEC). While the earlier strictly mechanistic models provide a less computationally intensive approach the results are also less accurate. It is for this reason that the MEC model was used.

The MEC model is the only codon model utilizing an empirical component available in the Selecton program, and was the only model used in this thesis research. In order to combine the strength of empirical codon models with that of parametric (mechanical) models, the MEC model uses three different empirical data sets. The first is the JTT data matrix so named for its creators Jones, Taylor, and Thornton which was compiled in 1992 from all proteins longer than 20 peptides within SWISS-PROT Release 15.0 (Jones, Taylor, & Thornton, 1992). The second is mtREV, which was derived from mitochondrial genes of 20 vertebrate species by Adachi et al. (Adachi & Hasegawa, 1996). Finally is cpREV, which was compiled from chloroplast containing organisms such as algae, diatom, plants etc. (Adachi, Waddell, & Hasegawa, 2000). Like the ECM model, the MEC model takes into account the possibility for substitutions at more than one codon position simultaneously (Doron-Faigenboim & Pupko, A Combined Empirical and Mechanistic Codon Model, 2006). The following equation describes the basis on which codon substitution values are weighted within the MEC model:

$$\psi_i \cdot A_{ij} = \sum_{\{l:\mathrm{aa}_l = i\}} \sum_{\{s:\mathrm{aa}_s = j\}} \pi_l \cdot Q^*_{ls},$$

$\psi_i$ = Frequency of amino acid *i* in empirical dataset (e.g. JTT matrix)

$A_{ij}$ = Substitution rate of amino acid *i* to amino acid *j*

$\Sigma_{\{l:\mathrm{aa}l = i\}}$ = Number of codons coding for amino acid *i*

$\Sigma_{\{s:\mathrm{aa}s = j\}}$ = Number of codons coding for amino acid *j*

$\pi_l$ = Frequency of codon *l*

$Q^*_{ls}$ = Substitution rate of codon *l* to codon *s*, accounting for any combination of transition or transversion across all three base positions within the codon.

Coupling the determined codon frequency with transition/transversion bias allows for a determination of likelihood of codon substitution. Within this equation are visible the two

elements that make this a combined Mechanistic and Empirical codon model. On the left side $\psi_i$ and $A_{ij}$ rely on previously defined large empirical datasets of protein/nucleotide sequences. On the right side $\pi_l$ and $Q^*_{ls}$ take into account codon frequencies and transitions/transversions to set mechanistic parameters from the target sequences. In the case of the MEC model these parameters are estimated using maximum likelihood, a statistical method for estimating parameters of a larger population from a subset of it, derived by R. A. Fisher in 1922 (Aldrich, 1997).

In their paper (Doron-Faigenboim & Pupko, A Combined Empirical and Mechanistic Codon Model, 2006), Doron-Faigenboim and Pupko compare their MEC model with M8 (Yang, Nielsen, Goldman, & Pedersen, Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites, 2000), M5 (Yang, Nielsen, Goldman, & Pedersen, Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites, 2000), and E (Schneider, Cannarozzi, & Gonnet, 2005) codon models using 13 data sets including nuclear, mitochondrial, chloroplast, and viral sequences. The MEC model outperformed the other models in 11 of the 13 sets, and scored relatively near the top model in the other two.

### 3.3.11 Chimera

Chimera is a molecular visualization software program produced by the University of California at San Francisco (UCSF); it is written primarily in Python and available for Windows, Linux, Mac OS X and other operating systems (Petterson, et al., 2004). Chimera accepts PDB files as input and possesses a variety of visualization options for displaying a protein in a simulation of three dimensions as depicted in Figure 13. In this thesis, script files (.cmd) controlling the Chimera software were automatically generated from Python scripts in order to create publication-quality images.

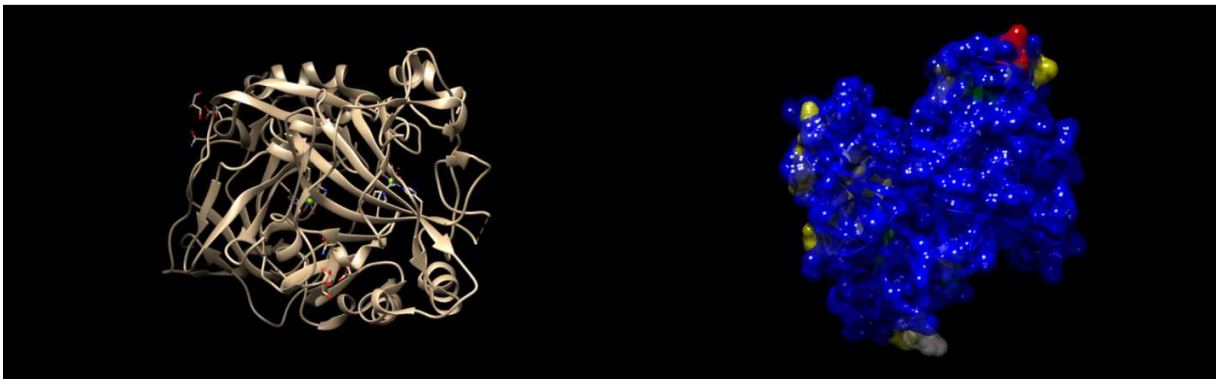

Figure 13 - PDB model 3FE4 for Human Carbonic Anhydrase VI, as visualized in Chimera. The image on the left is the initial view of pdb entry 3FE4, a dimer of CA-VI, and the image on the right is the same protein, as a monomer, after being altered via a Chimera script to highlight conservation.

## 4. Research Goals

The overall goal of this research is development of a pipeline for conservation analysis of a target multi-species protein. This goal is accomplished through a number of smaller tasks including:

- Automated retrieval of Ensembl orthologs and identification of incomplete protein sequences.
- Manual annotation and correction of incomplete sequences through gene prediction and homology analysis.
- $K_a/K_s$ conservation analysis, using the MEC codon model, resulting in generation of figures and 3D models highlighting contrasting conservation by individual residue.

The created pipeline should then be used to identify areas of positive and negative selection on the now expanded complete carbonic anhydrase IV orthologs. Through comparative analysis of varying branches of life it is the intention of this research to also identify amino acids and regions of interest that are unique, and functionally significant, to each. From this analysis it may be possible to provoke hypotheses about the structure and function of CAs and generate targets for future research.

# 5. Methods

## 5.1 Scripts

All scripts were created using Python in the Linux environment. In the following sections each script of major importance has been reduced to pseudo-code for readability.

### 5.1.1 Orthologer

The 'Orthologer' script was the most complex work created during this thesis research. The final version contains approximately 600 lines of python code, uses 20+ python modules, connects to the Ensembl database, and calls 6 external bioinformatics programs to process data.

*For each human α-CA gene in Ensembl:*

*Create output files for protein, CDS, and Complete sequences*
*Retrieve all orthologs from Ensembl (1to1, 1tomany, manytomany)*
*Write protein and CDS sequences for each ortholog to files*
*Create a distance table from a phylogeny based on F81 model from all sequences*
*Set INCOMPLETE = orthologs with protein sequences with lengths more than 5% longer or shorter than average for that CA gene*

*For each ortholog in INCOMPLETE:*

*Extract genomic sequence for ortholog from Ensembl, with additional 50% of length of gene added to each end*
*Set NEIGHBOR = closest related Complete protein sequence from distance table*
*Call Exonerate program to make prediction of gene from genomic sequence using NEIGHBOR as template*
*If any ESTs for ortholog:*

*Call Exonerate program to make prediction of gene from EST sequence(s) using NEIGHBOR as template*

*Create file containing ORTHOLOG, NEIGHBOR, and EST protein sequences*
*Call Clustal Omega to create alignment and output file*

### 5.1.2 SEQs2Categories

The purpose of the 'SEQs2Categories.py' script is to create categorized FASTA files from all complete sequences retrieved, or derived, from Ensembl. Five categories of organisms

(vertebrates, invertebrates, fishes, birds, and mammals) are separated so that each can have $K_a/K_s$ analysis completed independently to identify unique areas of conservation. Immediately preceding the use of this script is the only truly manual step in this created conservation analysis pipeline. In the manual step, protein sequences that have been predicted by Exonerate are compared in an alignment. The sequences determined to be good predictions are added to a FASTA file containing all complete protein sequences, while the same is done for the corresponding CDS sequences. This file is the primary input for this script. In advance of completion of this script, all species within the Ensembl database were individually placed into 17 distinct lists which were then referenced by this script.

*Open FASTA files of complete protein and CDS sequences*
*For each category (vertebrates, invertebrates, fishes, birds, and mammals):*
    *Create protein and CDS output file*
    *For each specie entry in protein FASTA file:*
        *Write specie entry to output file if contained in category list*
    *For each specie entry in CDS FASTA file:*
        *Write specie entry to output file if contained in category list*

### 5.1.3 Unaligned2KaKs

The primary task of this script is to call the $K_a/K_s$ program to complete conservation analysis on any of the category FASTA files created by the previous script. However, these files must first be codon aligned therefore Clustal Omega and Pal2Nal are called.

*Open protein and CDS FASTA files*
*For each specie entry CDS FASTA file:*
    *If entry contains 'homo sapiens' set KaKsQueryid to entry's id*
*Call Clustal Omega to create alignment file from protein FASTA file*
*Call Pal2Nal to create codon alignment file from aligned protein FASTA file and unaligned CDS FASTA file*
*Set Selecton parameters*
*Call Selecton to create conservation analysis from codon aligned file and KaKsQueryid using defined parameters*

### 5.1.4 RESparser – Histo+Line

First of the scripts designed to produce easily interpreted visuals from the raw $K_a/K_s$ data produced by the Selecton program; the output of this script is a colored histogram of the $K_a/K_s$ values associated to each amino acid. A line graph of the data is also produced.

*Open .res file*

*Parse entries for each amino acid*

*Set list ySeries = 1/(Exracted $K_a/K_s$ values)*

*Set list xSeries = Extracted amino acid #*

*Give each amino acid, for both .res files, a score of 1-7 based on $K_a/K_s$ value*

*Create colored histogram, and separate linegraph, across all amino acids based on level of conservation*

### 5.1.5 RESparser – Histo Compare

The objective of this script is to compare two .res conservation files of two different groups of species. In the resulting histogram variations in conservation values become easily identified, and the residue type and number easily obtained.

*Open .res files*

*Parse entries for each amino acid of both files*

*Set list ySeries = 1/(Exracted $K_a/K_s$ values)*

*Set list xSeries = Extracted amino acid #*

*Give each amino acid, for the primary .res file, a score of 1-7 based on $K_a/K_s$ value*

*Assign a unique color to each level of conservation*

*For the primary .res amino acids create a colored histogram across all amino acids based on level of conservation*

*For the secondary .res amino acids underlay a thinner black colored bar*

### 5.1.6 RESparser – PDB

These final two scripts also provide automated visualization of the conservation analysis. They take as input a single $K_a/K_s$ analysis .res file that is output by the Selecton program, and a PDB file. One of seven colors is mapped onto each amino acid, in the PDB model, that is not disordered with coloration based on level of conservation.

*Query for .res file*

*Open .res file*

*Parse entries for each amino acid*

*Set list ySeries = 1/(Exracted $K_a/K_s$ values)*

*Set list xSeries = Extracted amino acid #*

*Query for PDB model number, and model chain*

*Call DSSP program to extract DSSP data from full PDB model*

*Create new PDB file containing only the protein chain specified by user*

*Call DSSP program to extract DSSP data from single chain PDB model*

*Determine relative solvent accessibility (RSA) for each amino acid*
*Align protein sequence derived from PDB DSSP data with Ensembl sequence within .res file*
*Correlate Ensembl protein amino acid numbering with PDB protein model numbering*
*Give each amino acid a score of 1-7 based on $K_a/K_s$ value*
*Assign a unique color to each level of conservation*
*Create .cmd output file for chimera with each amino acid colored*

### 5.1.7 RESparser – PDB – Compare

This final, and most complex, visualization script takes as input two .res files of taxa between which a comparison should be made. The $K_a/K_s$ values at each residue are compared and those which vary by a significant amount are marked for coloration in the resulting Chimera .cmd file.

*Query for .res files to be compared*
*Open .res file*
*Parse entries for each amino acid*
*Set list ySeries = 1/(Exracted $K_a/K_s$ values)*
*Set list xSeries = Extracted amino acid #*
*Query for PDB model number, and model chain*
*Call DSSP program to extract DSSP data from full PDB model*
*Create new PDB file containing only the protein chain specified by user*
*Call DSSP program to extract DSSP data from single chain PDB model*
*Determine relative solvent accessibility (RSA) for each amino acid*
*Align protein sequence derived from PDB DSSP data with Ensembl sequence within .res file*
*Correlate Ensembl protein amino acid numbering with PDB protein model numbering*
*Give each amino acid, for both .res files, a score of 1-7 based on $K_a/K_s$ value*
*Compare each amino acid to equivalent in other .res file*
*Identify those sites where the score differs by 3 or more*
*Create .cmd output file for chimera with amino acids colored either red or blue if the site in the first .res file is 3 or more levels greater or if the amino acid in the second .res file is 3 or more levels greater, respectively.*

## 5.2 Manual Analysis

Three criteria were used to identify an 'Incomplete' sequence for the orthologs proteins retrieved from the Ensembl database:

1. Lack of initiating methionine 'M'.
2. Presence of 'X' in sequence; this indicates Ensembl homology prediction of sequence length but lack of supporting sequence data.

3. Length of sequence less than 95% of human sequence.

For each incomplete protein an alignment was created, using Clustal Omega, from the incomplete protein, Exonerate prediction from extragenomic data, and Exonerate prediction for EST data. Manual analysis of this alignment, in many cases, resulted in corrected or more complete sequences.

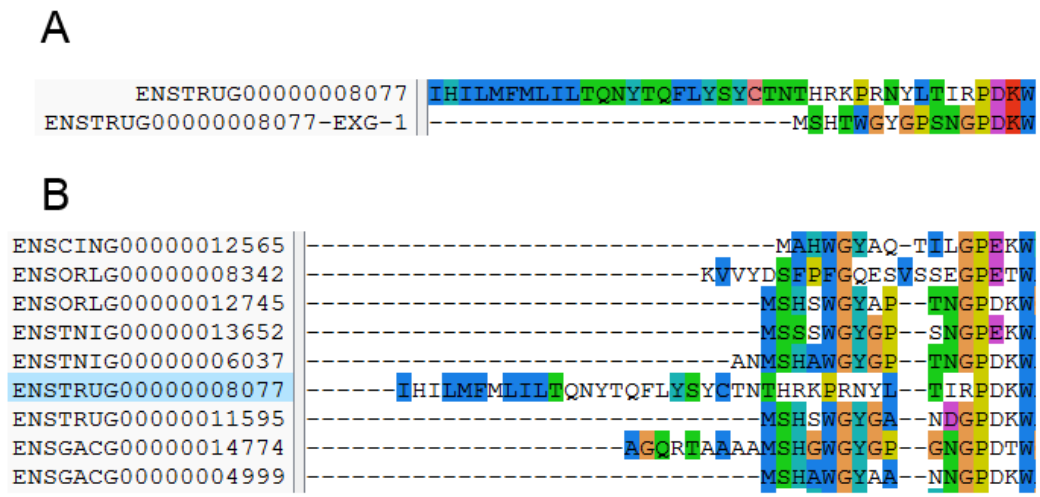


Figure 14 - A) Alignment of *incomplete* protein Takifugu rubripes CA-II, and protein predicted from T. rubripes extragenomic data using Danio rerio CAH-Z. B) Partial segment of alignment of all CA-II orthologs. Based on the Exonerate prediction and supporting evidence from the alignment, the beginning of the predicted sequence is taken as accurate and the corresponding portion of the Ensembl sequence is discarded.

Further work will be completed to conservatively automate as much of this manual analysis as possible, however this endeavor is beyond the scope of this thesis.

# 6. Results

## 6.1 Manual Analysis

The orthologer script extracted 32 *complete* and 33 *incomplete* orthologs, to human CA-IV, from the Ensembl database. Manual analysis of the *incomplete* orthologs resulted in modification of 7 of the sequences. A total of 79 amino acid changes were made to the 7 sequences, by either removal or addition. Another 7 sequences were sufficiently long to merit inclusion into the *complete* group, for a total of 46 *complete* proteins. The final collection of *complete* CA-IV proteins for $K_a/K_s$ analysis included the following numbers of proteins within respective groupings: 33 vertebrates, 13 invertebrates, 17 mammals, 13 fishes, 2 birds, 1 lizard. An extended study was made through manual analysis of the cytoplasmic CAs (I, II, III, VII, and XIII). In these five isozyme groups, there were 90 proteins determined to be incomplete. An additional 420 residue changes were made among 50 of these proteins.

| CA # | Ensembl ID and Description | Does not start with Met | Short (SP) | Short (anterior) | Long (anterior) | short (posterior) | long (posterior) | Contains 'X' | ACTION | STATUS | starting length | Finishing length | Residues Changed |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CA1 | >ENSCPOG00000012870 Cavia_porcellus-CA1 | x | | | | x | | | Removed 'I' | Complete | 261 | 260 | 1 |
| CA1 | >ENSGACG00000014774 Gasterosteus_aculeatus-CA1 | x | | | | x | | | Removed “AGQRTAAAA” | Complete | 269 | 260 | 9 |
| CA1 | >ENSGGOG00000008746 Gorilla_gorilla-CA1 | x | x | | | | | | NO CHANGE | Complete | 257 | 257 | 0 |
| CA1 | >ENSMICG00000000860 Microcebus_murinus-CA1 | x | x | x | | | | x | NO CHANGE | INCOMPLETE | 183 | 183 | 0 |
| CA1 | >ENSOPRG00000014213 Ochotona_princeps-CA1 | x | x | | | | | | Added exonerate supported anterior “G” | Complete | 247 | 248 | 1 |
| CA1 | >ENSOCUG00000005039 Oryctolagus_cuniculus-CA1 | | | | | x | | | NO CHANGE | INCOMPLETE | 227 | 227 | 0 |
| CA1 | >ENSORLG00000008342 Oryzias_latipes-CA1 | x | | | | | | | NO CHANGE | Complete | 266 | 266 | 0 |
| CA1 | >ENSPMAG00000008967 Petromyzon_marinus-CA1 | x | | | x | | | | NO CHANGE | Complete | 261 | 261 | 0 |
| CA1 | >ENSSARG00000001803 Sorex_araneus-CA1 | | | | | x | | | NO CHANGE | INCOMPLETE | 241 | 241 | 0 |
| CA1 | >ENSTRUG00000008077 Takifugu_rubripes-CA1 | x | | | x | x | | | Replace anterior “IHILMFMLILTQNYTQFLYSYCTNTHRKPRNYLTIR” with “MSHTWGYGPSNGP” | INCOMPLETE | 261 | 222 | 39 |
| CA1 | >ENSTSYG00000006199 Tarsius_syrichta-CA1 | x | x | | | | | | Added exonerate supported anterior “G” | Complete | 248 | 249 | 1 |
| CA1 | >ENSTNIG00000006037 Tetraodon_nigroviridis-CA1 | | | | x | | | | Removed anterior “AG” | Complete | 260 | 258 | 2 |
| CA1 | >ENSTBEG00000004632 Tupaia_belangeri-CA1 | | | | | x | | x | NO CHANGE | INCOMPLETE | 223 | 223 | 0 |
| CA2 | >ENSACAG00000009431 Anolis_carolinensis-CA2 | x | | | x | | | | Removed anterior “FT” | Complete | 259 | 257 | 2 |
| CA2 | >ENSBTAG00000017733 Bos_taurus-CA2 | x | x | | | | | | NO CHANGE | Complete | 249 | 249 | 0 |
| CA2 | >ENSCJAG00000000351 Callithrix_jacchus-CA2 | x | | | x | | | | Removed anterior “AT” | Complete | 262 | 260 | 2 |
| CA2 | >ENSCHOG00000000722 Choloepus_hoffmanni-CA2 | x | x | x | | | | | NO CHANGE | INCOMPLETE | 182 | 182 | 0 |
| CA2 | >ENSECAG00000016228 Equus_caballus-CA2 | x | | | x | | | | Removed anterior “T” | Complete | 261 | 260 | 1 |
| CA2 | >ENSFCAG00000005137 Felis_catus-CA2 | x | x | | | | | | Added exonerate supported anterior “G” | Complete | 248 | 249 | 1 |
| CA2 | >ENSGACG00000014774 Gasterosteus_aculeatus-CA2 | x | | | x | | | | Removed anterior “AGQRTAAAA” | Complete | 269 | 260 | 9 |
| CA2 | >ENSMODG00000025608 Monodelphis_domestica-CA2 | x | | | | | | | NO CHANGE | Complete | 264 | 264 | 0 |
| CA2 | >ENSMODG00000015855 Monodelphis_domestica-CA2 | x | | | x | | | | Removed anterior “TFRSAVQSTTEKLKDFKN” | Complete | 282 | 254 | 28 |
| CA2 | >ENSMLUG00000022410 Myotis_lucifugus-CA2 | x | | | | | | | NEEDS REVIEW | NEEDS REVIEW | 266 | 266 | 0 |
| CA2 | >ENSOANG00000002036 Ornithorhynchus_anatinus-CA2 | x | x | | | | | | NEEDS REVIEW | NEEDS REVIEW | 259 | 259 | 0 |
| CA2 | >ENSORLG00000008342 Oryzias_latipes-CA2 | x | | | | | | | Removed anterior “KVVYDSFPFGQESVS” | Complete | 266 | 251 | 15 |
| CA2 | >ENSPMAG00000008967 Petromyzon_marinus-CA2 | x | | | | | | | Added anterior exonerate supported “K” | Complete | 261 | 262 | 1 |
| CA2 | >ENSPVAG00000002689 Pteropus_vampyrus-CA2 | x | x | | | | | | Added exonerate supported anterior “HKG” | Complete | 248 | 251 | 3 |
| CA2 | >ENSSARG00000001849 Sorex_araneus-CA2 | x | x | x | | | | | NO CHANGE | INCOMPLETE | 182 | 182 | 0 |
| CA2 | >ENSSSCG00000006140 Sus_scrofa-CA2 | x | | | | | | | Removed anterior “TT” | Complete | 262 | 260 | 2 |
| CA2 | >ENSTGUG00000011728 Taeniopygia_guttata-CA2 | x | | | | | | | Removed anterior “T” | Complete | 261 | 260 | 1 |
| CA2 | >ENSTRUG00000008077 Takifugu_rubripes-CA2 | x | | | | | | | NEEDS REVIEW | NEEDS REVIEW | 247 | 247 | 0 |
| CA2 | >ENSTNIG00000006037 Tetraodon_nigroviridis-CA2 | x | | | | | | | Removed anterior "AN" | Complete | 260 | 258 | 2 |
| CA2 | >ENSTBEG00000010555 Tupaia_belangeri-CA2 | | | | | x | | | NO CHANGE | INCOMPLETE | 220 | 220 | 0 |
| | | | | | | | | | | | | | |
| CA3 | >ENSACAG00000024792 Anolis_carolinensis-CA3 | x | | | | | | | NO CHANGE | Complete | 248 | 248 | 0 |
| CA3 | >ENSDORG00000001002 Dipodomys_ordii-CA3 | | | | | | | x | NO CHANGE | INCOMPLETE | 260 | 264 | 4 |
| CA3 | >ENSETEG00000007613 Echinops_telfairi-CA3 | | | | | | | x | NO CHANGE | INCOMPLETE | 260 | 260 | 0 |
| CA3 | >ENSEEUG00000007144 Erinaceus_europaeus-CA3 | | | | | x | | | NO CHANGE | INCOMPLETE | 221 | 221 | 0 |
| CA3 | >ENSFCAG00000008051 Felis_catus-CA3 | | | | | x | | x | NO CHANGE | INCOMPLETE | 219 | 219 | 0 |
| CA3 | >ENSGALG00000019454 Gallus_gallus-CA3 | x | | | x | | | | Removed anterior “GTGAPCRRSRGELRCRAERST” | Complete | 283 | 262 | 21 |
| CA3 | >ENSGACG00000014774 Gasterosteus_aculeatus-CA3 | x | | | x | | | | Removed anterior “AGQRTAAAA” | Complete | 269 | 260 | 9 |
| CA3 | >ENSGGOG00000010673 Gorilla_gorilla-CA3 | x | | | x | | | | Removed anterior “T” | Complete | 262 | 261 | 1 |
| CA3 | >ENSLAFG00000004404 Loxodonta_africana-CA3 | x | | | x | x | | | Removed anterior “T”. Added Posterior “K”. | Complete | 260 | 260 | 2 |
| CA3 | >ENSMEUG00000000900 Macropus_eugenii-CA3 | x | | x | | | | | NO CHANGE | Complete | 248 | 248 | 0 |

| CA # | Ensembl ID and Description | Does not start with Met | Short (SP) | Short (anterior) | Long (anterior) | short (posterior) | long (posterior) | Contains 'X' | ACTION | STATUS | starting length | Finishing length | Residues Changed |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CA3 | >ENSMLUG00000016286 Myotis_lucifugus-CA3 | x | | x | | | | | Removed anterior "ELTSLSG" | INCOMPLETE | 192 | 185 | 7 |
| CA3 | >ENSOANG00000002034 Ornithorhynchus_anatinus-CA3 | x | | | x | | | | Removed anterior "AGRQGIPEERAGGQSRSRRPRDGERRSRGRKEI". | Complete | 293 | 260 | 33 |
| CA3 | >ENSORLG00000008342 Oryzias_latipes-CA3 | x | | | | | | | Removed anterior "KVVYDSFPFGQESVS". | Complete | 266 | 251 | 15 |
| CA3 | >ENSPMAG00000008967 Petromyzon_marinus-CA3 | x | | | | | | | Removed anterior "LSSRIEDNDDRA" | Complete | 261 | 249 | 12 |
| CA3 | >ENSPVAG00000007169 Pteropus_vampyrus-CA3 | | | | | | | x | NO CHANGE | INCOMPLETE | 258 | 258 | 0 |
| CA3 | >ENSTGUG00000018183 Taeniopygia_guttata-CA3 | x | | | | | | | NO CHANGE | Complete | 258 | 258 | 0 |
| CA3 | >ENSTRUG00000008077 Takifugu_rubripes-CA3 | x | | | x | x | | | Replaced anterior "IHILMFMLILTQNYTQFLYSYCTNTHRKPRNYLTIR" with "MSHTWGYGPSNG" | INCOMPLETE | 247 | 226 | 21 |
| CA3 | >ENSTSYG00000010505 Tarsius_syrichta-CA3 | | | | | x | | | Added "RSLFASAENEPPVPLVGNWRPP" intron. | Complete | 238 | 260 | 22 |
| CA3 | >ENSTNIG00000006037 Tetraodon_nigroviridis-CA3 | x | | | x | | | | Removed anterior "AN" | Complete | 261 | 259 | 2 |
| CA3 | >ENSTTRG00000005803 Tursiops_truncatus-CA3 | x | | x | | | | | NO CHANGE | Complete | 248 | 248 | 0 |
| CA3 | >ENSVPAG00000003347 Vicugna_pacos-CA3 | x | | x | | | | | NO CHANGE | Complete | 248 | 248 | 0 |
| | | | | | | | | | | | | | |
| CA4 | >ENSACAG00000015167 Anolis_carolinensis-CA4 | x | x | | | | | | NO CHANGE | Complete | 318 | 318 | 0 |
| CA4 | >K05G3.3 Caenorhabditis_elegans-CA4 | | | x | | x | | | NO CHANGE | Complete | 246 | 246 | 0 |
| CA4 | >ENSCJAG00000001097 Callithrix_jacchus-CA4 | | | | | | | x | NO CHANGE | NEEDS REVIEW | 256 | 256 | 0 |
| CA4 | >ENSCING00000010731 Ciona_intestinalis-CA4 | | | | | X | | | NO CHANGE | NEEDS REVIEW | 258 | 258 | 0 |
| CA4 | >ENSDNOG00000010140 Dasypus_novemcinctus-CA4 | x | | | | | | x | Removed anterior "GR" | INCOMPLETE | 311 | 309 | 2 |
| CA4 | >ENSDORG00000007143 Dipodomys_ordii-CA4 | x | x | | | X | | | NO CHANGE | INCOMPLETE | 225 | 225 | 0 |
| CA4 | >FBgn0039235 Drosophila_melanogaster-CA4 | | | | | x | | | NO CHANGE | Complete | 279 | 279 | 0 |
| CA4 | >FBgn0040628 Drosophila_melanogaster-CA4 | | | | | x | | | NO CHANGE | Complete | 284 | 284 | 0 |
| CA4 | >ENSETEG00000017554 Echinops_telfairi-CA4 | x | x | x | | | | | NO CHANGE | INCOMPLETE | 218 | 218 | 0 |
| CA4 | >ENSEEUG00000013445 Erinaceus_europaeus-CA4 | x | x | x | | x | | | NO CHANGE | INCOMPLETE | 208 | 208 | 0 |
| CA4 | >ENSGACG00000020471 Gasterosteus_aculeatus-CA4 | x | x | | | x | | | NO CHANGE | INCOMPLETE | 282 | 282 | 0 |
| CA4 | >ENSGACG00000011230 Gasterosteus_aculeatus-CA4 | x | | | | | | | NO CHANGE | Complete | 317 | 317 | 0 |
| CA4 | >ENSMICG00000013382 Microcebus_murinus-CA4 | x | | | | | | x | NO CHANGE | INCOMPLETE | 313 | 313 | 0 |
| CA4 | >ENSMODG00000014115 Monodelphis_domestica-CA4 | x | x | x | | x | | | NO CHANGE | INCOMPLETE | 225 | 225 | 0 |
| CA4 | >ENSMODG00000003392 Monodelphis_domestica-CA4 | x | x | | | | | | NO CHANGE | Complete | 292 | 292 | 0 |
| CA4 | >ENSMLUG00000024982 Myotis_lucifugus-CA4 | x | | | | | | | NEEDS REVIEW | INCOMPLETE | 302 | 302 | 0 |
| CA4 | >ENSOPRG00000017181 Ochotona_princeps-CA4 | x | x | x | | | | | NO CHANGE | INCOMPLETE | 270 | 270 | 0 |
| CA4 | >ENSOANG00000004545 Ornithorhynchus_anatinus-CA4 | x | | | | x | | | NO CHANGE | INCOMPLETE | 286 | 286 | 0 |
| CA4 | >ENSOCUG00000009399 Oryctolagus_cuniculus-CA4 | x | | | | | | | NO CHANGE | Complete | 293 | 293 | 0 |
| CA4 | >ENSORLG00000002507 Oryzias_latipes-CA4 | x | | | | | | | NO CHANGE | Complete | 309 | 309 | 0 |
| CA4 | >ENSPCAG00000005108 Procavia_capensis-CA4 | | | | | | | x | NO CHANGE | INCOMPLETE | 306 | 306 | 0 |
| CA4 | >ENSTRUG00000012774 Takifugu_rubripes-CA4 | x | | | | | | | NO CHANGE | Complete | 316 | 316 | 0 |
| CA4 | >ENSTRUG00000006382 Takifugu_rubripes-CA4 | x | | | | | | | Removed anterior "APLPQTLTANFDVTFISTLK" | Complete | 304 | 284 | 20 |
| CA4 | >ENSTRUG00000007204 Takifugu_rubripes-CA4 | x | | | | | | | Removed anterior "RSILPSTL" | Complete | 311 | 303 | 8 |
| CA4 | >ENSTRUG00000007249 Takifugu_rubripes-CA4 | x | | | | | | | Removed anterior "TL" | Complete | 313 | 311 | 2 |
| CA4 | >ENSTNIG00000008558 Tetraodon_nigroviridis-CA4 | | | | | x | | | Added posterior "GKPMVGTFRPVQPLNGRQVFHSGAAAALTSSA" | Complete | 260 | 292 | 32 |
| CA4 | >ENSTNIG00000016135 Tetraodon_nigroviridis-CA4 | x | | | | | | | NO CHANGE | Complete | 309 | 309 | 0 |
| CA4 | >ENSTBEG00000014112 Tupaia_belangeri-CA4 | | | | | | | x | NO CHANGE | INCOMPLETE | 311 | 311 | 0 |
| CA4 | >ENSTTRG00000008780 Tursiops_truncatus-CA4 | x | | | | | | x | Added anterior "MPIAGTPEEVGLG" | INCOMPLETE | 290 | 304 | 14 |
| CA4 | >ENSVPAG00000009703 Vicugna_pacos-CA4 | x | x | x | | | | | Added anterior "V" | INCOMPLETE | 219 | 220 | 1 |
| CA4 | >ENSXETG00000013489 Xenopus_tropicalis-CA4 | x | | | | x | | | NO CHANGE | INCOMPLETE | 282 | 282 | 0 |

| CA # | Ensembl ID and Description | Does not start with Met | Short (SP) | Short (anterior) | Long (anterior) | short (posterior) | long (posterior) | Contains 'X' | ACTION | STATUS | starting length | Finishing length | Residues Changed |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| CA7 | >ENSCPOG00000024957 Cavia_porcellus-CA7 | x | | | | | | | Removed anterior "SSG" | Complete | 264 | 261 | 3 |
| CA7 | >ENSDNOG00000019175 Dasypus_novemcinctus-CA7 | x | | x | | | | x | NO CHANGE | INCOMPLETE | 263 | 263 | 0 |
| CA7 | >ENSDORG00000013532 Dipodomys_ordii-CA7 | | | | | | | | NO CHANGE | Complete | 250 | 250 | 0 |
| CA7 | >ENSETEG00000011874 Echinops_telfairi-CA7 | | | | | | | x | NO CHANGE | INCOMPLETE | 263 | 263 | 0 |
| CA7 | >ENSECAG00000024943 Equus_caballus-CA7 | x | | | | | | | Added anterior "GHHGLGAYQNDGPSEWHKLYPIAQ" | Complete | 238 | 262 | 24 |
| CA7 | >ENSEEUG00000011058 Erinaceus_europaeus-CA7 | x | | x | | | | | Added anterior "TV" | INCOMPLETE | 184 | 186 | 2 |
| CA7 | >ENSFCAG00000012773 Felis_catus-CA7 | x | | x | | | | x | NO CHANGE | INCOMPLETE | 250 | 250 | 0 |
| CA7 | >ENSMEUG00000007011 Macropus_eugenii-CA7 | | | | | x | | | NO CHANGE | INCOMPLETE | 224 | 224 | 0 |
| CA7 | >ENSOANG00000012209 Ornithorhynchus_anatinus-CA7 | x | | x | | | | | NO CHANGE | Complete | 250 | 250 | 0 |
| CA7 | >ENSPVAG00000000385 Pteropus_vampyrus-CA7 | | | | | | | x | NO CHANGE | INCOMPLETE | 262 | 262 | 0 |
| CA7 | >ENSSARG00000011702 Sorex_araneus-CA7 | x | | x | | | | | NO CHANGE | INCOMPLETE | 224 | 224 | 0 |
| CA7 | >ENSSSCG00000022094 Sus_scrofa-CA7 | | | | | | | x | Replaced anterior intron of "APXSDERSHSYFSGERQRRPPITNQYRIQPSCVLTX" with "GPSXXXXXXXXXXGDRQSPINIVSSQAVYSPSLKPL". | INCOMPLETE | 269 | 269 | 36 |
| CA7 | >ENSTSYG00000001353 Tarsius_syrichta-CA7 | x | | x | | | | | NO CHANGE | INCOMPLETE | 183 | 183 | 0 |
| CA7 | >ENSTNIG00000009424 Tetraodon_nigroviridis-CA7 | x | | | x | | | | Removed anterior "GK" | Complete | 263 | 261 | 2 |
| CA7 | >ENSTBEG00000010634 Tupaia_belangeri-CA7 | x | | x | | | | | NO CHANGE | INCOMPLETE | 177 | 177 | 0 |
| | | | | | | | | | | | | | |
| CA13 | >ENSACAG00000014147 Anolis_carolinensis-CA13 | x | | | | | | | Removed anterior "PA" | Complete | 262 | 260 | 2 |
| CA13 | >ENSDORG00000004299 Dipodomys_ordii-CA13 | x | | x | | | | | NO CHANGE | Complete | 250 | 250 | 0 |
| CA13 | >ENSETEG00000015776 Echinops_telfairi-CA13 | x | | x | | | | x | Added anterior "G" | INCOMPLETE | 250 | 251 | 1 |
| CA13 | >ENSECAG00000013368 Equus_caballus-CA13 | x | | | | | | | NO CHANGE | Complete | 251 | 251 | 0 |
| CA13 | >ENSEEUG00000014581 Erinaceus_europaeus-CA13 | x | | | | | | x | Added anterior "G" | INCOMPLETE | 250 | 251 | 1 |
| CA13 | >ENSFCAG00000008686 Felis_catus-CA13 | | | | | | | x | NO CHANGE | INCOMPLETE | 263 | 263 | 0 |
| CA13 | >ENSGALG00000015820 Gallus_gallus-CA13 | x | | | | | | | Removed anterior "A" | Complete | 260 | 259 | 1 |
| CA13 | >ENSGACG00000014774 Gasterosteus_aculeatus-CA13 | x | | | x | | | | Removed anterior "AGQRTAAAA" | Complete | 269 | 260 | 9 |
| CA13 | >ENSMEUG00000002395 Macropus_eugenii-CA13 | x | | | | x | | x | NO CHANGE | INCOMPLETE | 210 | 210 | 0 |
| CA13 | >ENSMLUG00000006164 Myotis_lucifugus-CA13 | x | | | | | | | NO CHANGE | Complete | 263 | 263 | 0 |
| CA13 | >ENSOPRG00000014191 Ochotona_princeps-CA13 | | | | | | | x | NO CHANGE | INCOMPLETE | 257 | 257 | 0 |
| CA13 | >ENSOANG00000002031 Ornithorhynchus_anatinus-CA13 | x | | | | | | | Removed anterior "YSELGWGFL" | Complete | 259 | 250 | 9 |
| CA13 | >ENSORLG00000008342 Oryzias_latipes-CA13 | x | | | | | | | Removed anterior "KVVYDSFPFGQESVS". | Complete | 266 | 251 | 15 |
| CA13 | >ENSPMAG00000008967 Petromyzon_marinus-CA13 | x | | | | | | | Removed anterior "LSSRIEDNDDRA". Added posterior "K". | Complete | 261 | 250 | 13 |
| CA13 | >ENSPVAG00000005236 Pteropus_vampyrus-CA13 | x | | | | | | | Added anterior "G" | Complete | 250 | 251 | 1 |
| CA13 | >ENSTGUG00000011722 Taeniopygia_guttata-CA13 | x | | | | | | | Removed anterior "A" | Complete | 259 | 258 | 1 |
| CA13 | >ENSTRUG00000008077 Takifugu_rubripes-CA13 | x | | | | x | | | Removed anterior "IHILMFMLILTQNYTQFLYSYCTNTHRKPRNYLTIRPDKWAKDFPIADGSRQSPINIVPM EAQYDPSLKPLKLNYNQSNAKGILNNGHSFQVLGENKSA", replaced with "MSHTWGYGPSNGPDKWAKDFPIADGSRQSPINIVPM EAQYDPSLKPLKLNYNQSNAKGILNNGHSFQVDFVDDADSST" | INCOMPLETE | 247 | 222 | 25 |
| CA13 | >ENSTSYG00000008066 Tarsius_syrichta-CA13 | | | | | x | | | NO CHANGE | INCOMPLETE | 223 | 223 | 0 |
| CA13 | >ENSTNIG00000006037 Tetraodon_nigroviridis-CA13 | x | | | | | | | Removed anterior "AN" | Complete | 260 | 258 | 2 |
| CA13 | >ENSTBEG00000004374 Tupaia_belangeri-CA13 | | | | | | | x | NO CHANGE | INCOMPLETE | 262 | 262 | 0 |
| CA13 | >ENSVPAG00000003345 Vicugna_pacos-CA13 | | | | | | | x | NO CHANGE | INCOMPLETE | 264 | 264 | 0 |
| CA13 | >ENSXETG00000026010 Xenopus_tropicalis-CA13 | | | | | | | | Removed anterior "TSNM" | Complete | 266 | 262 | 4 |
| | | | | | | | | | | | | | 499 |

**Table 5 - Manual changes made to CA-I, CA-II, CA-III, CA-IV, CA-VII and CA-XIII proteins identified as 'incomplete' by Ortholoģer script. Changes were based on evidence from protein sequences predicted from genomic and EST data.**

## 6.2 Histograms

After manual gene annotation was performed and a Selecton $K_a/K_s$ analysis performed on the resulting updated orthologs, figures were created using data parsed from the Selecton (.res) results file. There are four scripts which parse the .res file containing $K_a/K_s$ analysis and produce varying types of output. The RESparser-histo script is responsible for creating histograms depicting the inverse of $K_a/K_s$ values at each amino acid position. As lower $K_a/K_s$ values indicate greater conservation the inverse value is used for greater ease of interpretation in graphic depictions. Histograms were created for $K_a/K_s$ analysis of all CA-IV orthologs (Figures 15-19).

If the *RESparser-histo-compare* script is provided with two Selecton results files, for example conservation analysis files for both mammal and non-mammal species, the script will produce a histogram where the conservation value for each group at each position is superimposed for comparison. The results from the comparison of mammals/non-mammals are included as Figures 20-24.

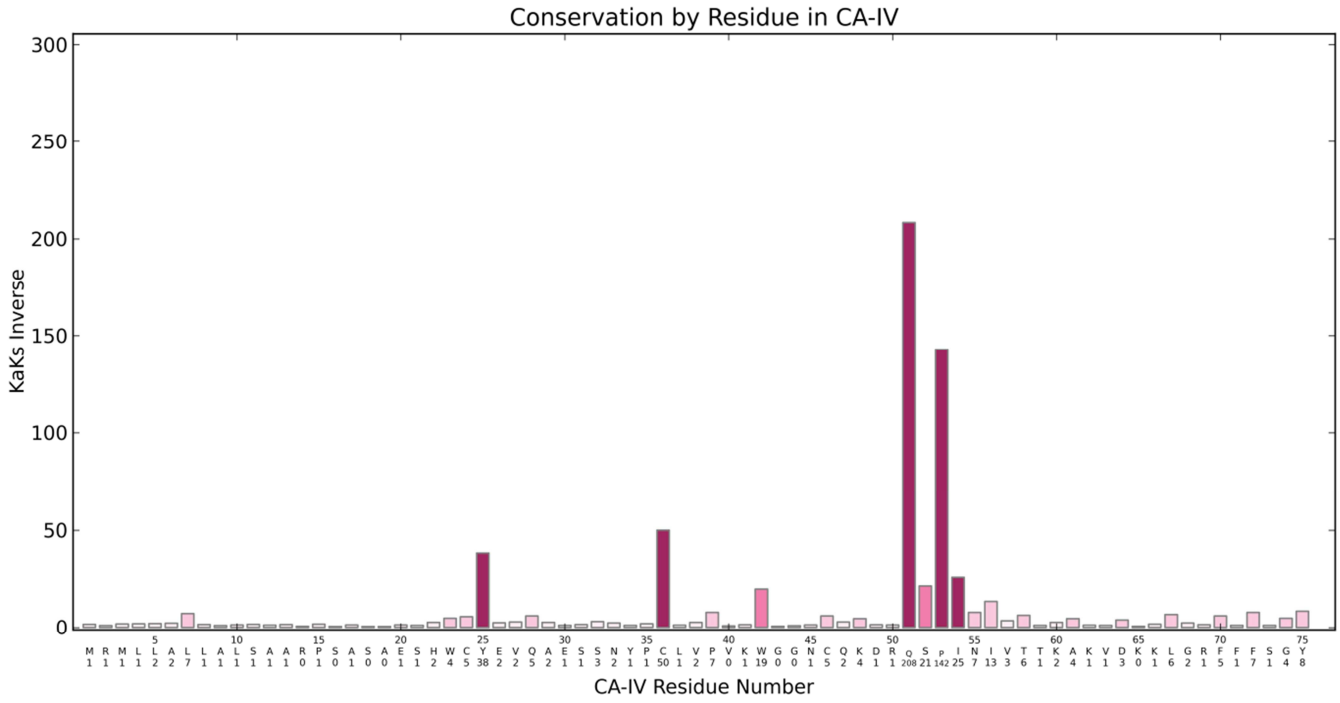


Figure 15 – Histogram of Ka/Ks values for residues 1-75 of all complete CA-IV sequences. Created using Python (Python.org) and Matplotlib (Hunter, 2007).

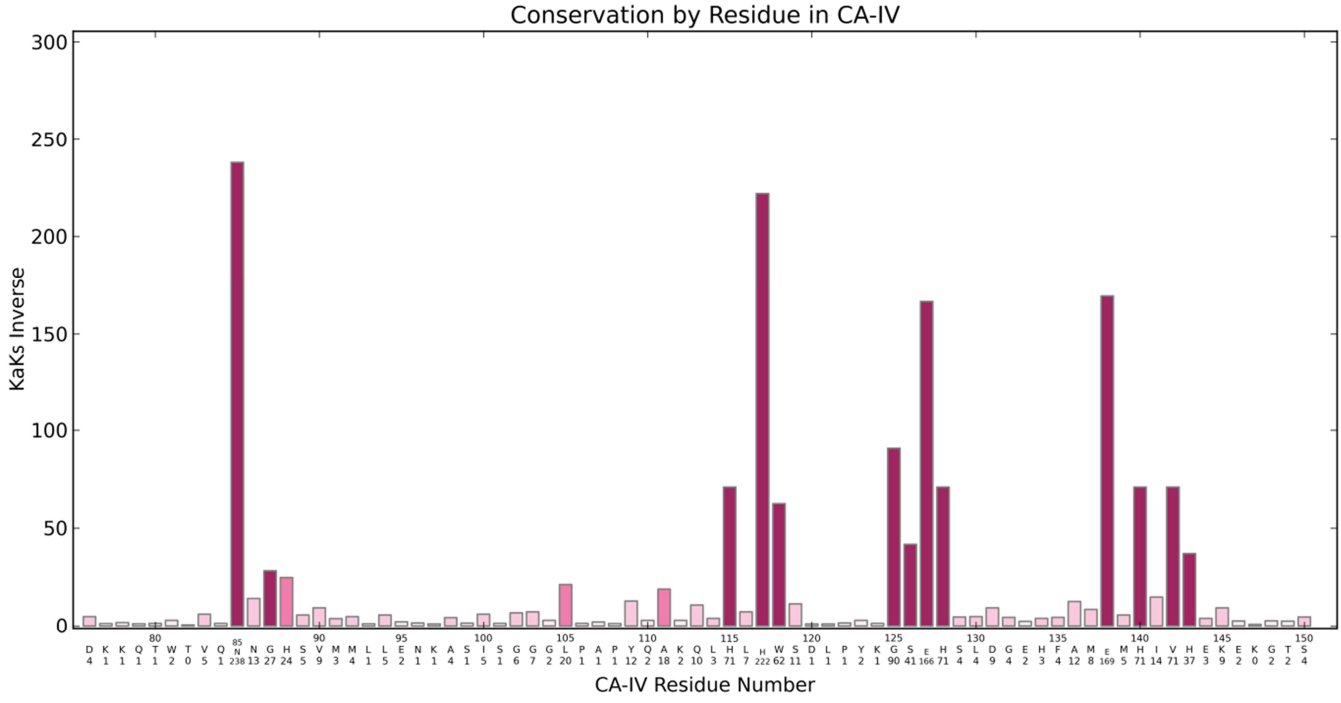


Figure 16 - Histogram of Ka/Ks values for residues 76-150 of all complete CA-IV sequences. Created using Python (Python.org) and Matplotlib (Hunter, 2007).

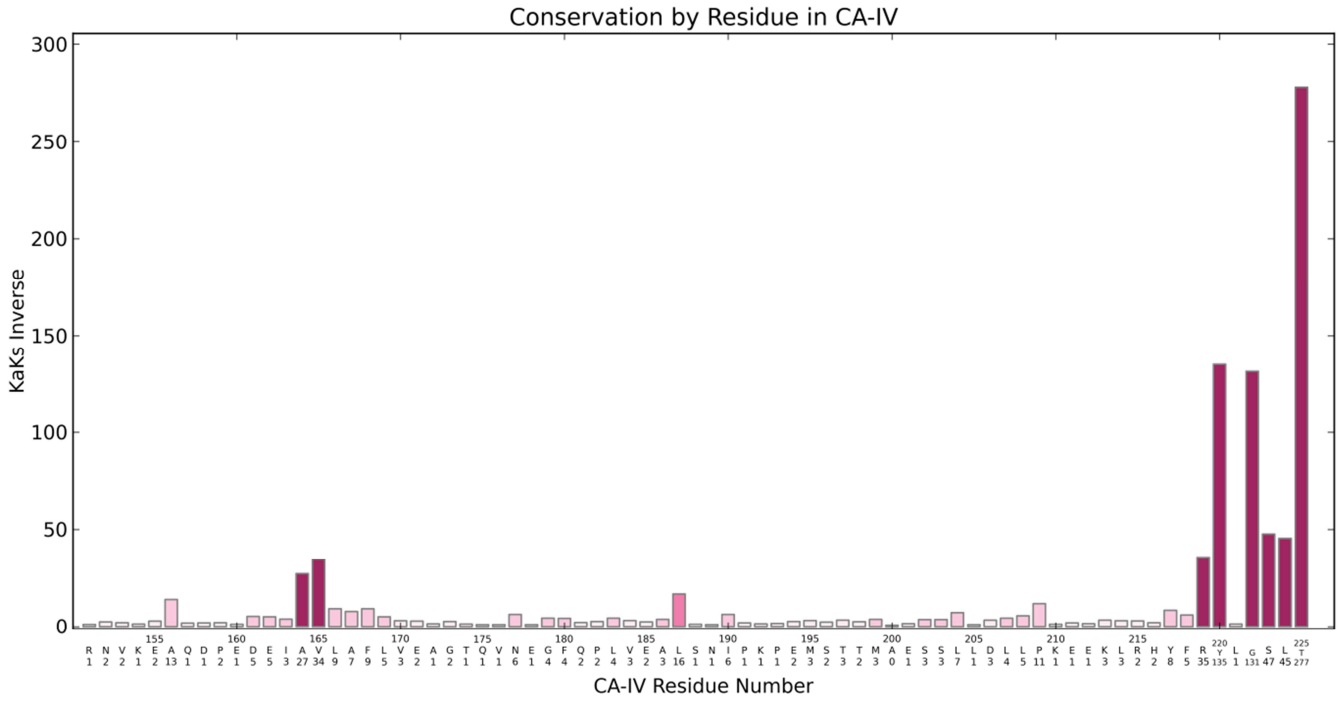


Figure 17 - Histogram of Ka/Ks values for residues 151-225 of all complete CA-IV sequences. Created using Python (Python.org) and Matplotlib (Hunter, 2007).

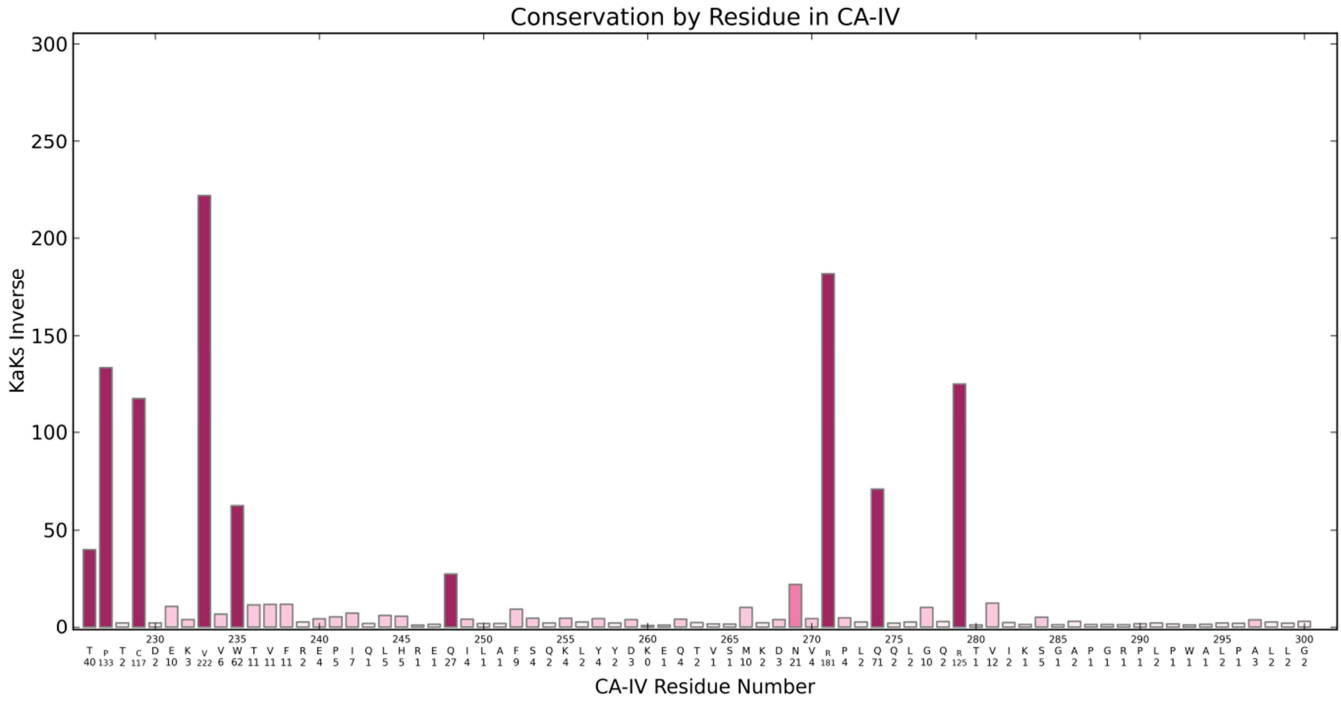


Figure 18 - Histogram of Ka/Ks values for residues 226-300 of all complete CA-IV sequences. Created using Python (Python.org) and Matplotlib (Hunter, 2007).

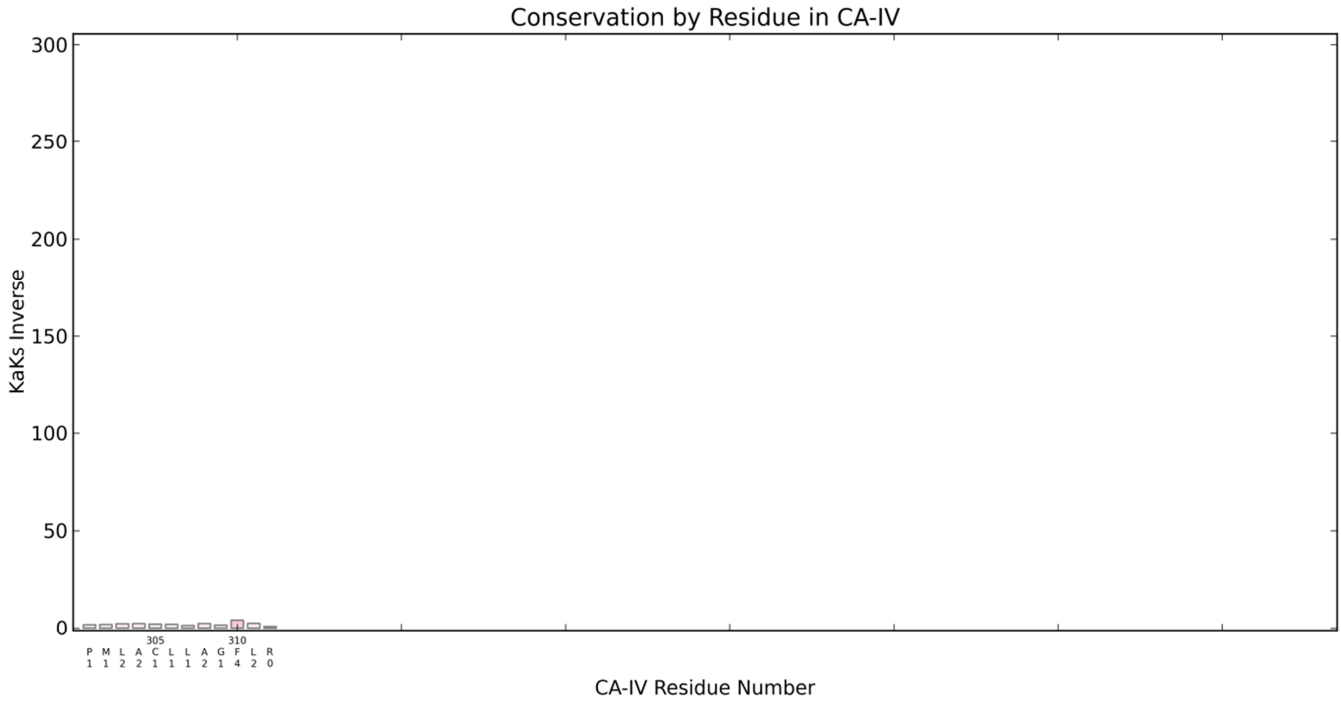


Figure 19 - Histogram of Ka/Ks values for residues 301-312 of all complete CA-IV sequences. Created using Python (Python.org) and Matplotlib (Hunter, 2007).

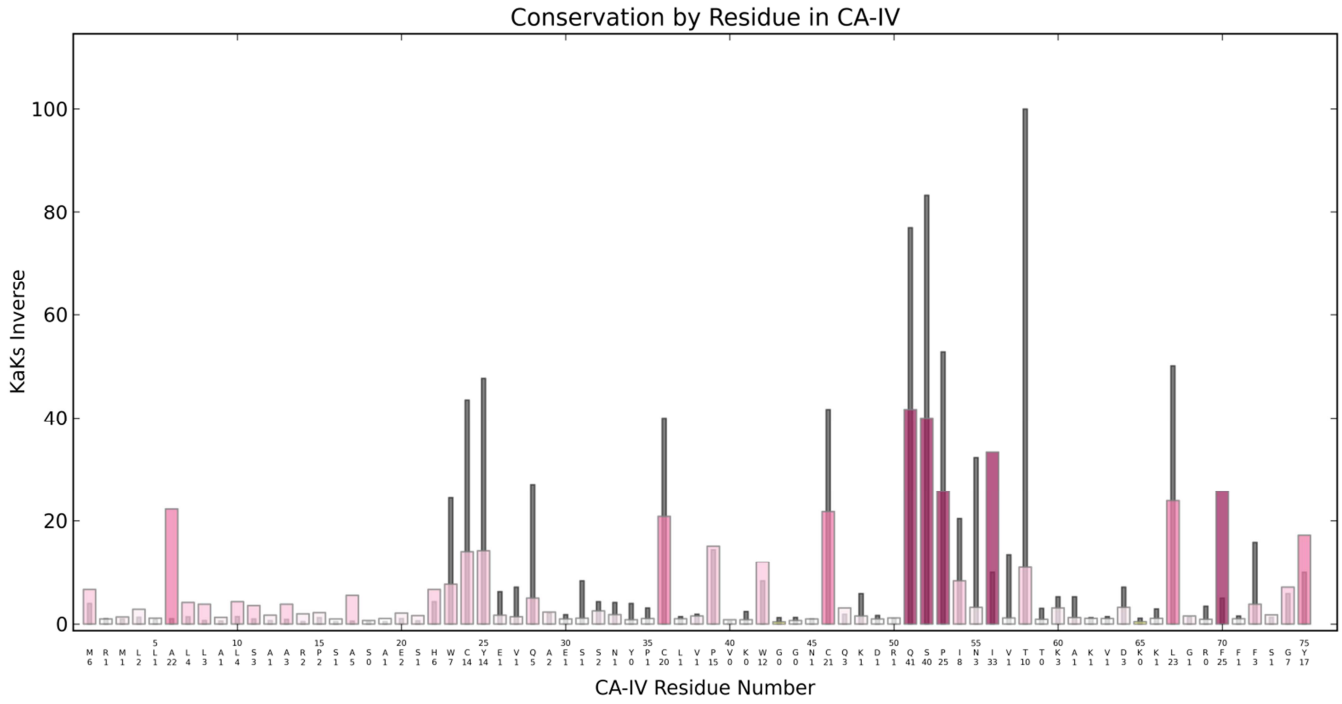


Figure 20 - Histogram of compared Ka/Ks values, for complete CA-IV sequences from mammals (thick bars) and non-mammals (thin bars), across residues 1-75. Created using Python (Python.org) and Matplotlib (Hunter, 2007).

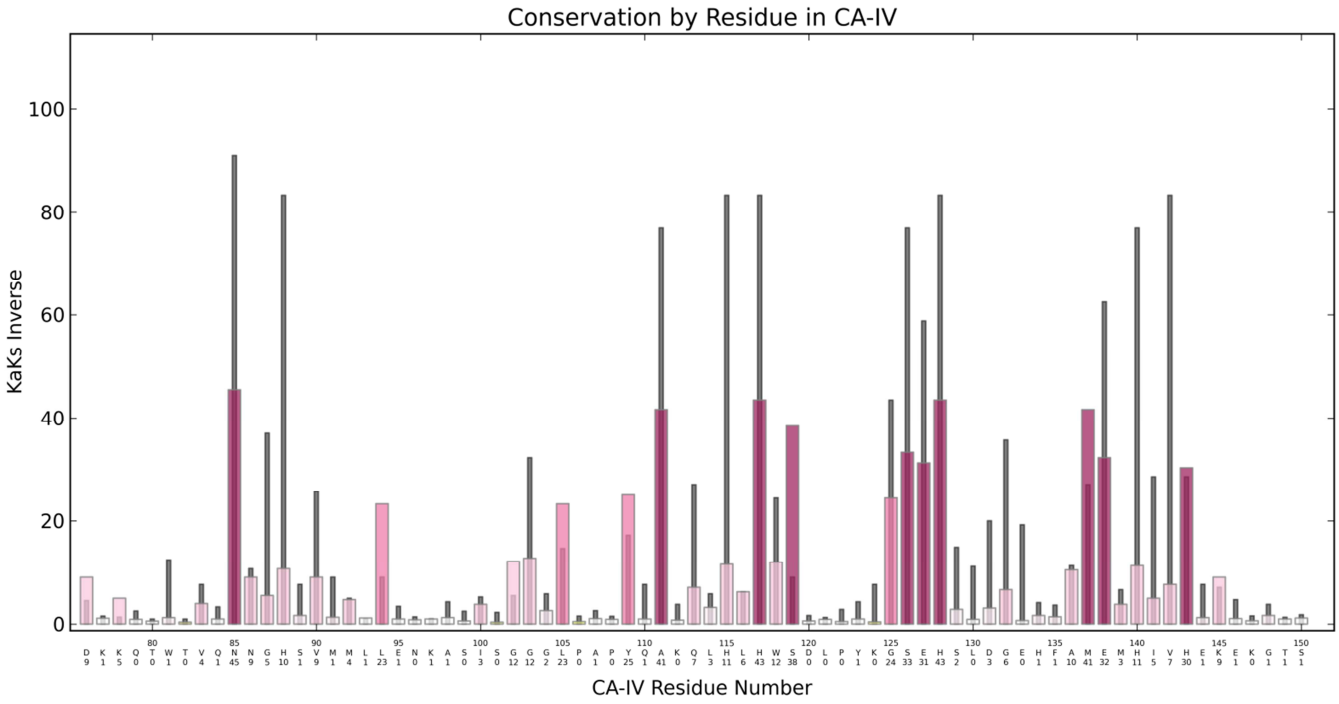


Figure 21 - Histogram of compared Ka/Ks values, for complete CA-IV sequences from mammals (thick bars) and non-mammals (thin bars), across residues 76-150. Created using Python (Python.org) and Matplotlib (Hunter, 2007).

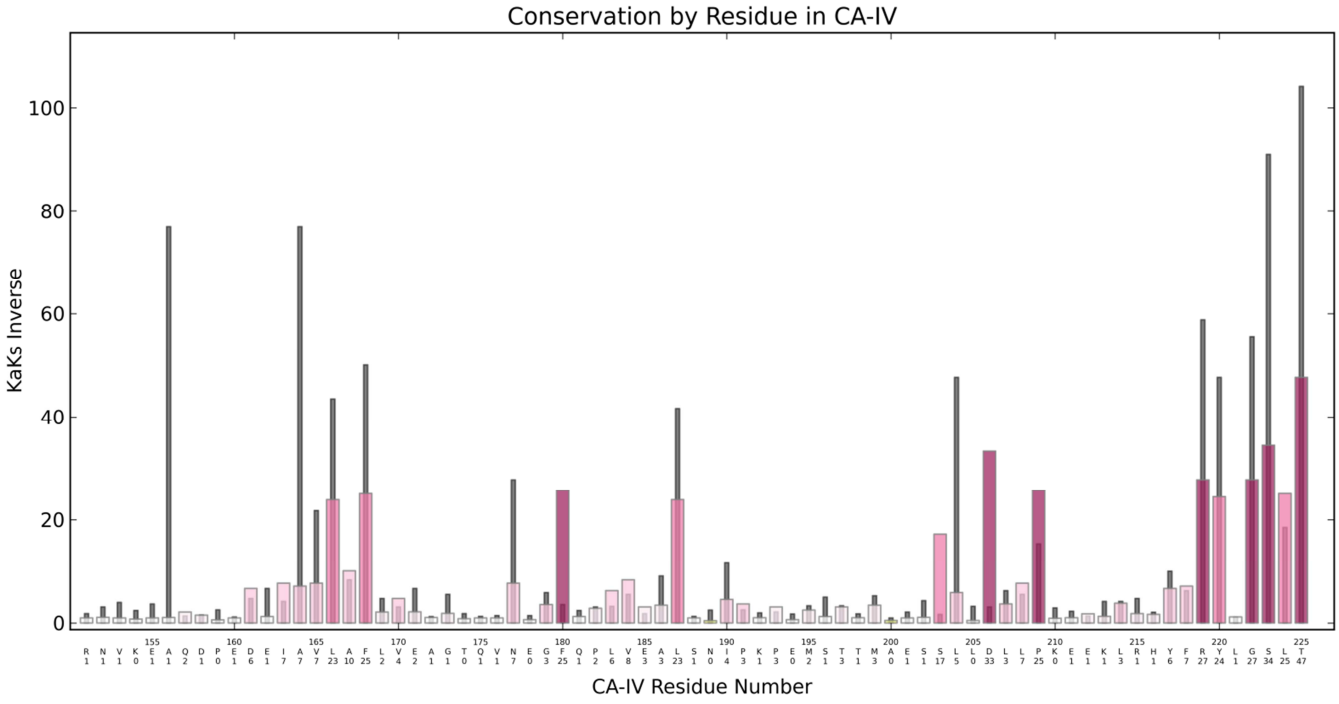


Figure 22 - Histogram of compared Ka/Ks values, for complete CA-IV sequences from mammals (thick bars) and non-mammals (thin bars), across residues 151-225. Created using Python (Python.org) and Matplotlib (Hunter, 2007).

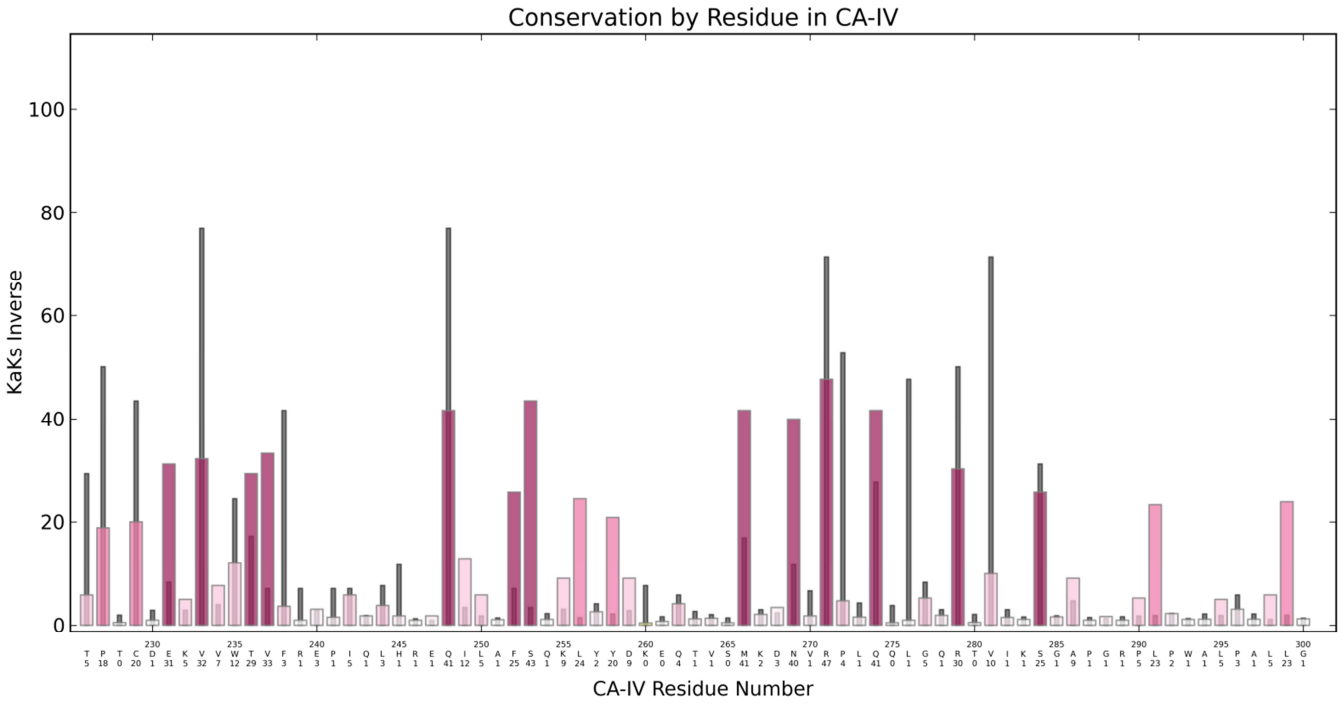


Figure 23 - Histogram of compared Ka/Ks values, for complete CA-IV sequences from mammals (thick bars) and non-mammals (thin bars), across residues 226-300. Created using Python (Python.org) and Matplotlib (Hunter, 2007).

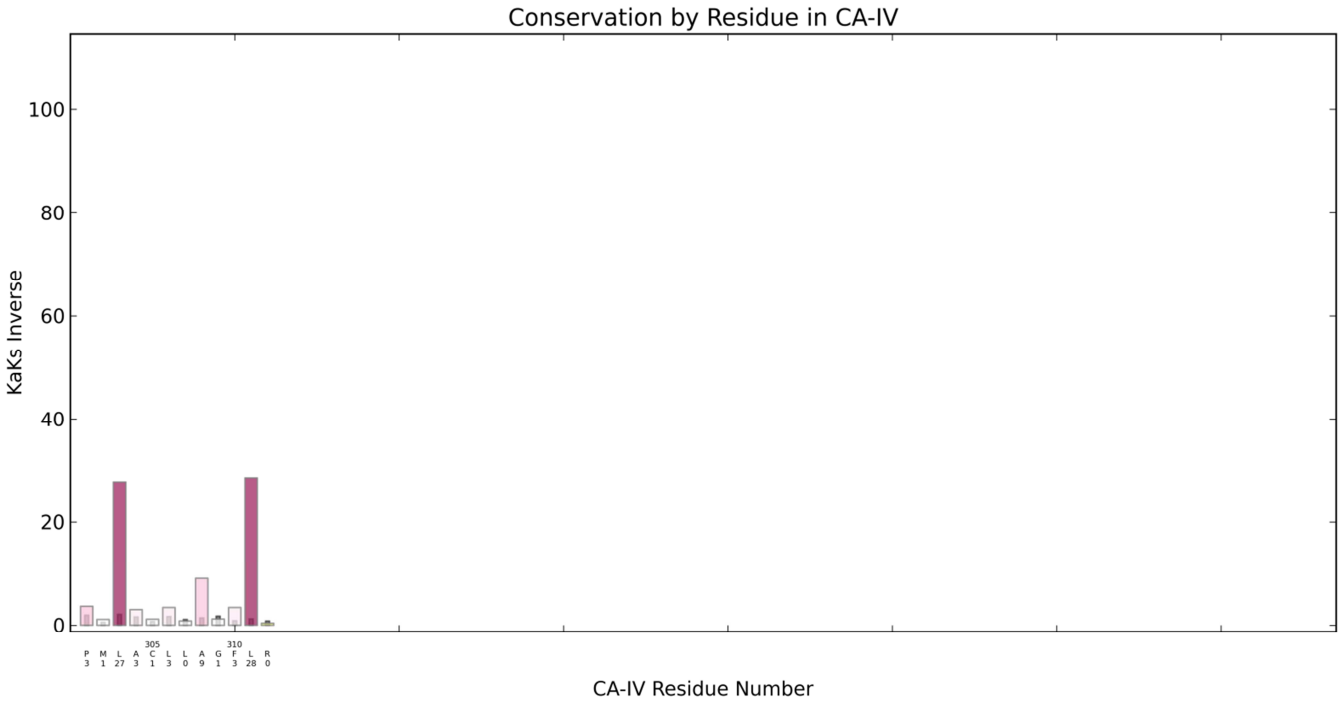


Figure 24 - Histogram of compared Ka/Ks values, for complete CA-IV sequences from mammals (thick bars) and non-mammals (thin bars), across residues 301-312. Created using Python (Python.org) and Matplotlib (Hunter, 2007).

## 6.3 3D Protein Models

The results of the *RESparser-PDB* script are seen in the depiction of CA-IV conservation in mammal and non-mammal groups in Figure 25. Two different view types are produced (columns), each with 3 different orientations (rows) produced by 120 degree rotations around the y-axis. The first view, 'protein specific' (column 1 (from left to right) for mammals and column 3 for non-mammals), depicts a coloration scheme showing degree of conservation for each amino in comparison to other amino acids within the protein. The second view, 'selecton' (column 2 for mammals and column 4 for non-mammals) is based on the Selecton model which has predetermined values for levels of conservation which are static and independent of which protein is being studied. The output of *REparser-PDB-compare* is seen in the farthest right column, in which 3 views of a 3D model were created comparing mammal and non-mammal species conservation values. There were 17 residues identified as having a conservation level score difference of 3 or higher between the two groups. These residues which were more conserved in mammals were Asp183 and Ser227 (marked as red). The residues more conserved in non-mammals were Asn32, Thr35, His64, His94, Lys103, Glu112, His119, Val121, Ala134, Ala142, Ser197, Thr199, Lys233, Leu251, and Val256 (marked as blue). Those most likely contributing to structure of the protein are those which have a RSA less than .20, qualifying as 'buried' (His94, His119, Val121, Ala142, Ser197, Thr199, ), while those with an RSA greater than, or equal to, .20 qualifying as 'surface' (Asn32, Thr35, His64, Lys103, Glu112, Ala134, Asp183, Ser227, Lys233, Leu251, Val256). There are some previously identified residues strongly correlated with the active site and catalytic activity, such as His64, Thr199, His94, and His 199; the remaining residues are previously unidentified as of significant value in the CA proteins, in general or to mammal or non-mammal groups.

The same analysis was performed for vertebrates and invertebrates and the results visualized in Figure 26. In this comparison the following amino acids were considerably more conserved in vertebrates: Trp5, Cys6, Gln10, Ala11, Cys23, Ser29, Thr35, Leu44, Asn61, Ala90, Ser105, His107, Met116, Leu144, Phe146, Arg193, Ser197, Thr199, Gln222, and Arg246. Those amino acids considerably more conserved in invertebrates: Ile31, Val143, and Thr200.

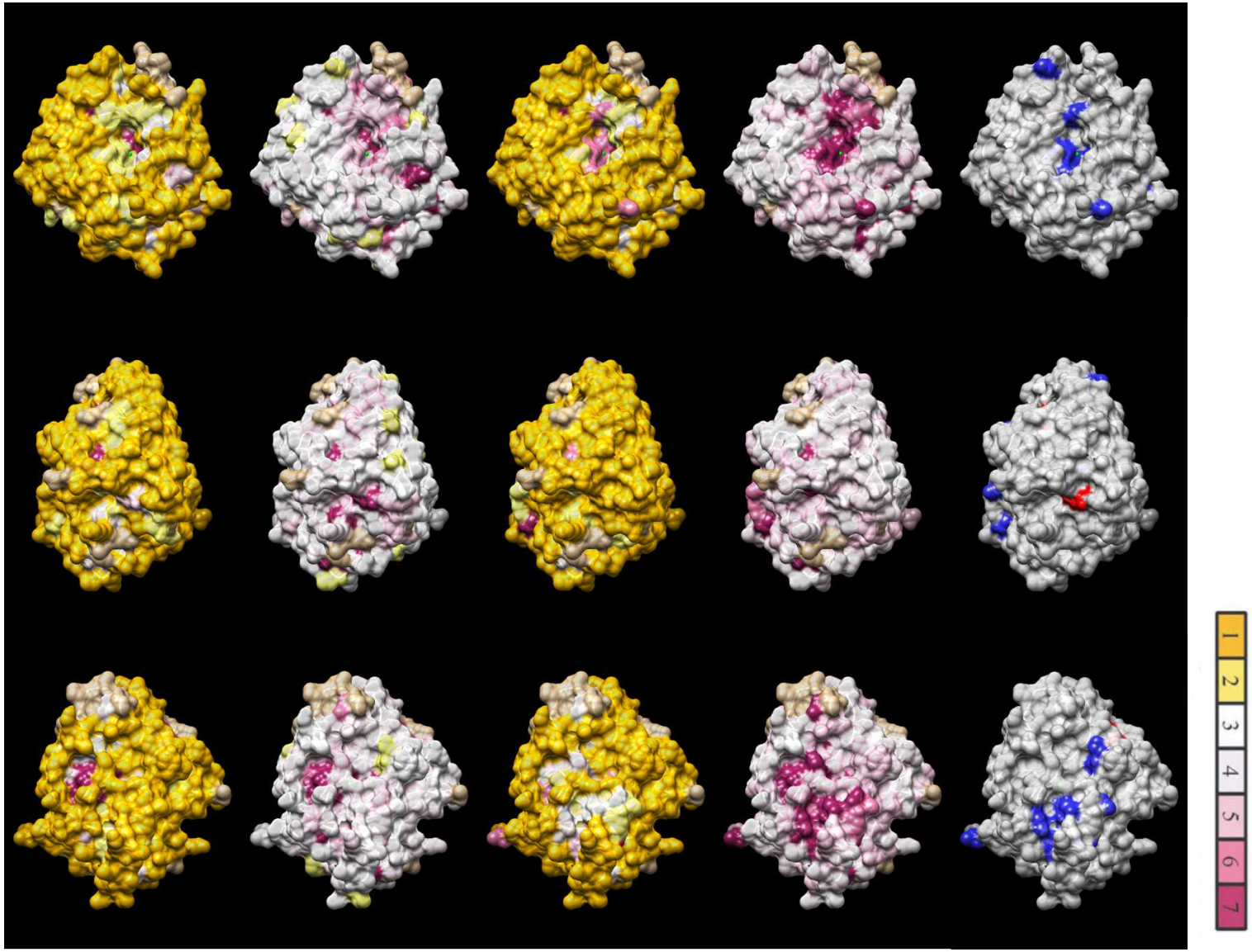


**Figure 25 – Panel showing mammal (columns 1,2) and non-mammal (columns 3,4) Ka/Ks conservation mapped onto PDB model 3FW3 Chain B** (Vernier, et al., 2010)**. In the protein specific models (columns 1,3) the colors are determined by taking the top conservation score for the whole protein and dividing by 7. Scores for each amino acid are colored according to which of the seven levels they fall within with 1 being the least conserved and 7 the most conserved. This provides an understanding of which amino acids are most important within this protein alone. In the Selecton specific models (columns 2,4) the colors are based on the Selecton model where yellow/orange colors represent positive selection, near white colors represent mostly neutral selection, and increasingly intense red colors represent conservation. The scores resulting in the Selecton color scheme are more general thus providing a comparison of conservation to all proteins. In column 5 is a comparison of mammals and non-mammals. Amino acids which are significantly more conserved in mammals are marked in red, while those significantly more conserved in non-mammals are marked in blue.**

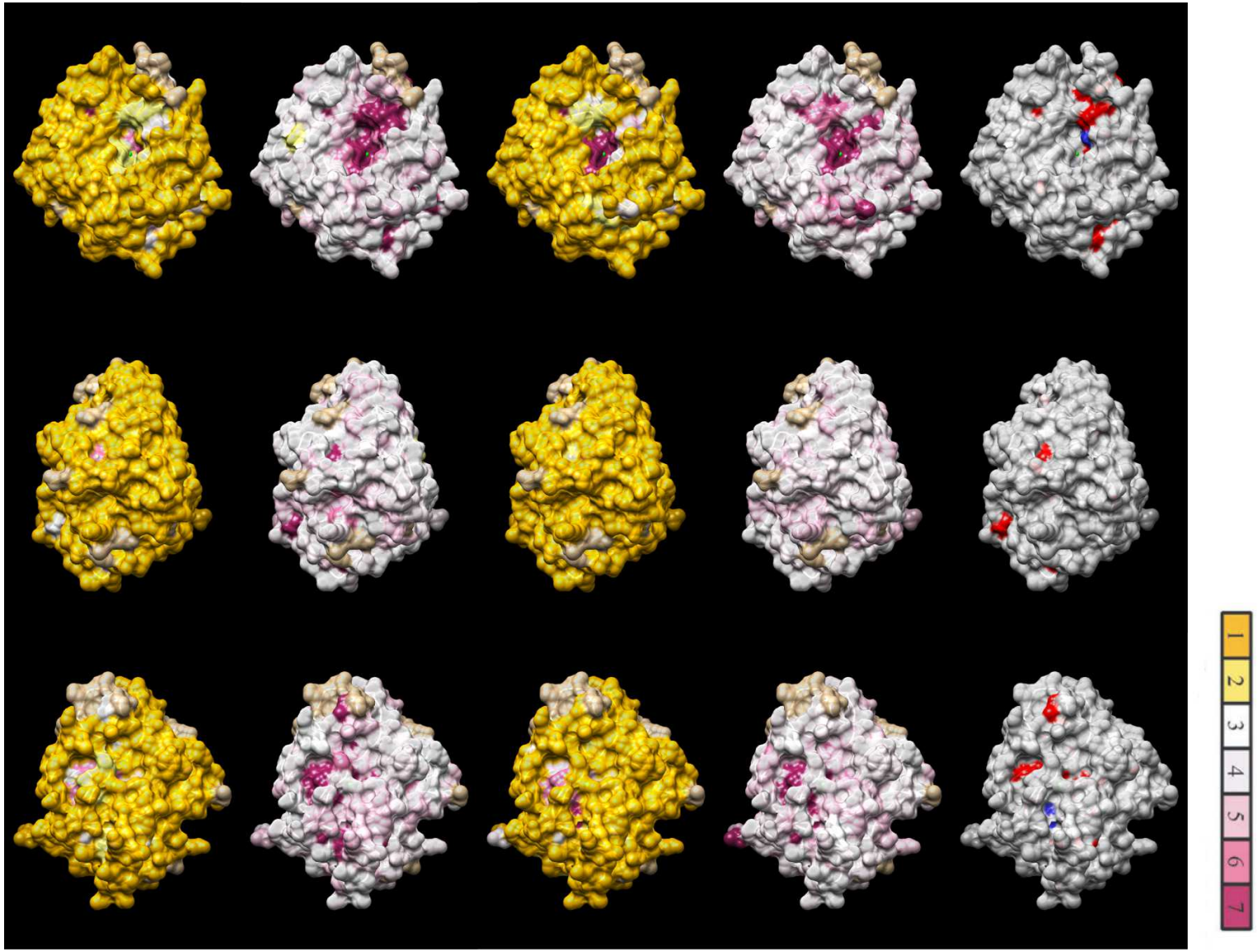


Figure 26 - Panel showing vertebrate (columns 1,2) and invertebrate (columns 3,4) Ka/Ks conservation mapped onto PDB model 3FW3 Chain B (Vernier, et al., 2010). In the protein specific models (columns 1,3) the colors are determined by taking the top conservation score for the whole protein and dividing by 7. Scores for each amino acid are colored according to which of the seven levels they fall within with 1 being the least conserved and 7 the most conserved. This provides an understanding of which amino acids are most important within this protein alone. In the Selecton specific models (columns 2,4) the colors are based on the Selecton model where yellow/orange colors represent positive selection, near white colors represent mostly neutral selection, and increasingly intense red colors represent conservation. The scores resulting in the Selecton color scheme are more general thus providing a comparison of conservation to all proteins. In column 5 is a comparison of mammals and non-mammals. Amino acids which are significantly more conserved in vertebrates are marked in red, while those significantly more conserved in invertebrates are marked in blue.

# 7. Discussion

If the utilization of tools can be regarded a hallmark of intelligence in animals, what conclusions can we make about the cell's use of proteins? Whether the function regards DNA replication, cell maintenance, or disabling pathogens there are certainly proteins involved. While the beginnings of life on earth could function without the wide variety of proteins available to modern organisms, attempting to build and maintain a modern cell without them is analogous to building a modern computer from raw materials and with no tools.

Understanding a protein's structure and function is largely advanced through identification of amino acids which have been conserved. As is evident on the larger scale of physical attributes of the body, conservation of an unimportant feature is evolutionarily costly; therefore, only those attributes of value are conserved. An experimental approach to determining the importance of a particular amino acid is to use site specific mutation to create a protein altered at that single point. However, this process can take time, especially when considering a complete analysis of a whole protein which may possess many hundreds of amino acids. A conservation analysis approach thus saves significant time in the analysis of proteins. At the very least it allows for protein specific identification of subsets of amino acids that could then be addressed experimentally. At best, analysis of a large numbers of proteins can quickly identify recurring amino acids critical to structure and function.

## 7.1 Analysis of Output

### 7.1.1 Sequences

The process created in this thesis work suggests 499 residue changes in 123 carbonic anhydrase proteins (groups CA-I, CA-II, CA-III, CA-IV, CA-VII, CA-XIII), and identified 17 key amino acid residues that vary most significantly between mammal and non-mammal vertebrates. A total of 23 such amino acid residues were found to differ between vertebrates and invertebrates. The approach can be applied to any protein which exists within the Ensembl database and has a sufficient number of identified orthologs. Comparisons can be easily made among any number of pairs created from predefined or custom groupings (mammals, fishes, birds, in/vertebrates, etc.) by creation of a histogram or 3D model provided there is a PDB structure available.

### 7.1.2 Models

A comparison made in this study highlighted a region of high conservation in non-mammal groups not as strongly conserved in mammals. The residues comprising this region are Asn32, Thr35, Lys103, Glu112, Leu251, and Val256. A previous study of N-glycosylation sites in CA isoforms IV, XV, and XVII, showed the same region uniquely devoid of the N-glycosylation motif

(Tolvanen, et al., 2012). Tolvanen et al. proposes that this region is likely devoid of N-glycosylation in order to allow binding with an ion channel (Tolvanen, et al., 2012). If this supposition is correct, then the differences between mammal and non-mammal CA-IV, in this region, could be explained by the fact that mammal ion channel genes differ significantly from those in fish (Jegla, Zmasek, Batalov, & Nayak, 2009), and coevolution of each taxon's CA-IV and distinctive ion transporter resulted in a different mode of binding. In humans, the anion exchange 1 (AE1) protein has been suggested to bind to CA-IV (Sterling, Alvarez, & Casey, 2002); this research proposes that CA-IV binds with extracellular loop 4 of AE1 while cytoplasmic CA-II binds with AE1 on the internal surface of the cell membrane. AE1 matches tissue expression of CA-IV closely, with a presence in erythrocytes, kidney, retina, and heart (Sterling, Alvarez, & Casey, 2002) (Wu, et al., 2009). BioGPS Microarray data analysis of AE1 shows strongest expression in early erythrocytes, and while there is significant expression in fetal lung, there is not in mature lung (Wu, et al., 2009). Future research regarding CA-IV ion channel binding should focus on the identified residues as potential binding points for AE1.

## 7.2 Sources of Error

There are sources of error, inherent to this approach, which are not present in experimental approaches. Perhaps most significant are the errors introduced by an incorrect alignment. In this thesis work an alignment of 46 CA-IV sequences was created, each sequence comprised of approximately 300 amino acids. Any misalignments, at any of the 300 positions, result in an altered score of conservation. However, given the similarity of these sequences, and the similarity inherent in any group of true orthologs, this concern is reduced.

In building the alignments for $K_a/K_s$ analysis, a target sequence must be included to serve as the template which $K_a/K_s$ values for the whole alignment are mapped onto. With the goal of mapping two distinct taxa onto single protein, and making a one to one comparison, the target sequence for all $K_a/K_s$ analyses was human CA-IV. This in turn requires that human CA-IV be included in the alignment for each of the taxa being compared, even if it was not a part of that taxon (e.g. non-mammals, fishes, and invertebrates). Subsequently, the $K_a/K_s$ analysis for that taxon becomes partially 'contaminated' by the presence of the human sequence. A potential solution for this would be $K_a/K_s$ analysis of each pure taxon with choice of target sequence going to the two sequences most similar between the two taxa. After analysis, the target sequences, and their associated conservation values, could be aligned together. A PDB protein sequence would need to be included into the alignment to allow for display of results on its structure.

An issue related to alignment, is the mapping of conservation values onto a protein which is potentially not part of the taxa being analyzed. The PDB model 3FW3 chain B of human CA-IV

was used for all visualizations. Thus, generating a 3D model for non-mammal or invertebrate conservation using 3WF3 as a scaffold is not going to provide as good a presentation of protein configuration as mapping those values onto a species more representative of that group. However, for consistency, and making 1 to 1 comparisons, the same model is used.

Another source of error is the conservation analysis. While it appears that the MEC model is the most comprehensive and accurate of all codon models, by definition these models are only attempting to approximate the rates of substitution that occur in the natural world; at the very least some level of inaccuracy is expected. Yet there is still much to be gleaned from even the broad strokes painted by this method. Codons representing stronger signals of conservation present important opportunities for prediction of protein function and interaction.

## 7.3 Future Research

A primary goal of this thesis work was automation of a process to produce visualization of protein conservation. The majority of the steps were translated to python scripting. However, the gene annotation step remained strictly manual. A significant number of times, analysis of predicted sequence alignments proved challenging. Therefore, creation and utilization of a strict set of rules would prove beneficial here. This final step of automation would open up much larger venues of analysis. However, there are a significant number of possible alignment and prediction scenarios to account for. In this attempt it would likely be beneficial to use a combined empirical and mechanistic approach similar in concept to the chosen MEC codon model. Analysis of the deficiencies of the sequence targeted will provide the parameters for its attempted repair, while the alignment of all good orthologs will provide the empirical data for comparison.

Larger scale analysis of many proteins within a family (e.g. all CA proteins), or similarly folded proteins, could provide insight as to which amino acids are most often present at critical fold points, provide interaction with other proteins, or are necessary for catalytic activity.

Perhaps most obviously, the comparative analyses that this pipeline produces can help to uncover the activity of proteins whose function is only partially known, or only known in one species; distinct differences in conservation allude to a modified purpose specific to that organism. The CA family is a perfect target for this approach as it exists in many forms, both catalytically active and inactive, across many varied species. In the case of disease, identification of conserved amino acids and regions is important for the creation of inhibitors or treatments. Where these regions differ between species is important to notice when using animal models to test treatment methods.

# 8. Conclusions

The goal of this thesis was to create a pipeline for analysis of the carbonic anhydrase protein family with particular emphasis on CA-IV. With the aid of the tools developed it is now possible to start with an Ensembl protein id and subsequently: retrieve all orthologs, make predictions for incomplete orthologs, derive conservation values across all amino acid sites for complete sequences, compare conservation among different taxa, produce a 3D model highlighting conservation values, and produce publication-quality images. Through utilization of this pipeline it was possible to identify a number of CA-IV amino acids uniquely important within different taxa.

## Works Cited


*FigTree*. (2013). Retrieved from http://tree.bio.ed.ac.uk/software/figtree/

*HIV Databases*. (2013). Retrieved from http://www.hiv.lanl.gov

*Selecton*. (2013). Retrieved 2013, from http://selecton.tau.ac.il/faq.html

Adachi, J., & Hasegawa, M. (1996). Model of amino acid substitution in proteins encoded by mitochondrial DNA. *Journal of Molecular Evolution, 42*(4), 459-468.

Adachi, J., Waddell, P., & Hasegawa, M. (2000). Plastid Genome Phylogeny and a Model of Amino Acid Substitution for Proteins Encoded by Chloroplast DNA. *Molecular Evolution, 50*, 348-358.

Aldrich, J. (1997). R.A. Fisher and the Making of Maximum Likelihood 1912-1922. *Statistical Science, 12*(3), 162-176.

Alterio, V., & al., e. (2009). Crystal structure of the catalytic domain of the tumor-associated human carbonic anhydrase IX. *PNAS, 106*(38), 16233-16238.

An, H., Tu, C., Duda, D., Montanez-Clemente, I., Math, K., Laipis, P., . . . Silverman, D. (2002). Chemical Rescue in Catalysis by Human Carbonic Anhydrases II and III. *Biochemistry, 41*, 3235-3242.

Becker, H. M., Klier, M., Schuler, C., McKenna, R., & Deitmer, J. W. (2011). Intramolecular proton shuttle supports no only catalytic but also noncatalytic function carbonic anhydrase II. *PNAS, 108*(7), 3071-3076.

Behravan, G., Jonsson, B.-H., & Lindskog, S. (1991). Fine tuning of the catalytic properties of human carbonic anhydrase II. *European Journal of Biochemistry, 195*, 393-396.

Bernstein, F., Koetzle, T., Williams, G., Meyer, E., Brice, M., Rodgers, J., . . . Tasumi, M. (1977). The Protein Data Bank: A Computer-based Archival File For Macromolecular Structures. *Molecular Biology, 112*(535).

Birney, E., Clamp, M., & Durbin, R. (2004). GeneWise and Genomewise. *Genome Research, 14*, 988-995.

Brinkman, R., Margaria, R., Meldrum, D., & al, e. (1932). The CO2 catalyst present in blood. *Proceedings of the Physiological Society*(March), 3-4.

Chandrashekar, J., Yarmolinsky, D., Buchholtz, L. v., Oka, Y., Sly, W., Ryba, N. J., & Zuker, C. S. (2009). The Taste of Carbonation. *Science, 326*, 443-445.

Chiang, W.-L., Chu, S.-C., Yang, S.-S., & al, e. (2002). The aberrant expression of cytosolic carbonic anhydrase and its clinical significance in human non-small cell lung cancer. *Cancer Letters*(188), 199-205.

Chien, M.-H., Ying, T.-H., Hsieh, Y.-H., Lin, C.-H., Shih, C.-H., Wei, L.-H., & Fang, S.-F. (2012). Tumor-associated carbonic anhydrase XII is linked to the growth of primary oral squamous cell carcinoma and its poor prognosis. *Oral Oncology, 48*, 417-423.

Cock, P., Antao, T., Chang, J., Chapman, B., Cox, C., Dalke, A., . . . de Hoon, M. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics, 25*(11), 1422-1423.

Cukierman, S. (2006). Et tu, Grotthuss! and other unfinished stories. *Biochimica et Biophysica*, 876-885.

Culp, D., Robinson, B., Parkkila, S., Pan, P.-w., Cash, M., Truong, H., . . . Gullett, S. (2011). Oral colonization by Streptococcus mutans and caries development is reduced upon deletion of carbonic anhydrase VI expression in saliva. *Biochemica et Biophysica Acta - Molecular Basis of Disease, 1812*(12), 1567-1576.

Delport, W., Poon, A., Frost, S., & Pond, S. (2010). Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. *Bioinformatics, 26*(19), 2455-2457.

Di Fiore, A., & al., e. (2009). Crystal structure of human carbonic anhydrase XIII and its complex with the inhibitor acetazolamide. *Proteins, 74*(1), 164-175.

Doron-Faigenboim, A., & Pupko, T. (2006). A Combined Empirical and Mechanistic Codon Model. *Molecular Biology and Evolution, 24*(2), 388-397.

Doron-Faigenboim, A., Stern, A., Mayrose, I., Bacharach, E., & Pupko, T. (2005). Selecton: a server for detecting evolutionary forces at a single amino-acid site. *Bioinformatics, 21*(9), 2101-2103.

Dungwa, J., Hunt, L., Ramani, P., & al, e. (2012). Carbonic anhydrase IX up-regulation is associated with adverse clinicopathologic and biologic factors in neuroblastomas. *Human Pathology, 43*, 1651-1660.

Edgar, R. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research, 32*(5), 1792-1797.

Elder, I., & al., e. (2007). Structural and kinetic analysis of proton shuttle residues in the active site of human carbonic anhydrase III. *Proteins, 68*(1), 337-343.

Entrez. (n.d.). *Entrez - CA1.* Retrieved October 2012, from http://www.ncbi.nlm.nih.gov/sites/entrez?db=gene&cmd=Retrieve&dopt=full_report&list_uids=759

Fisher, S. e. (2006). X-ray crystallographic studies reveal that the incorporation of spacer groups in carbonic anhydrase inhibitors causes alternate binding modes. *Acta Crystallographica, 62*, 618-622.

Fisher, Z., Hernandez, P., Tu, C., Duda, D., Yoshioka, C., An, H., . . . McKenna, R. (2005). Structural and Kinetic Characterization of Active-Site Histidine as a Proton Shuttle in Catalysis by Human Carbonic Anhydrase II. *Biochemistry, 44*, 1097-1105.

Fisher, Z., Tu, C., Bhatt, D., Lakshmanan, G., Agbandje-McKenna, M., McKenna, R., & Silverman, D. (2007). Speeding Up Proton Transfer in a Fast Enzyme: Kinetic and Crystallographic Studies on the Effect of Hydrophobic Amino Acid Substitutions in the Active Site of Human Carbonic Anhydrase II. *Biochemistry, 45*, 3803-3813.

Flicek, P., Amode, M. R., Barrell, D., & al., e. (2012). Ensembl 2012. *Nucleic Acids Research, 40*(Database Issue D84-D90).

Gilmour, K. (2010). Perspectives on carbonic anhydrase. *Comparative Biochemistry and Physiology, A*(157), 193-197.

Goldman, N., & Yang, Z. (1994). A Codon-based Model of Nucleotide Substitution for Protein-coding DNA Sequences. *Molecular Biology and Evolution, 11*(5), 725-736.

Gu, X. (1997). The Age of the Common Ancestor of Eukaryotes and Prokaryotes: Statistical Inferences. *Molecular Biology and Evolution, 14*(8), 861-866.

Guindon, S., & Gascuel, O. (2003). A Simple, Fast, and Accurate Algorithm to Estimate Large Phylogenies by Maximum Likelihood. *Systematic Biology, 52*(5), 696-704.

Hassan, I., Shajee, B., Waheed, A., Ahmad, F., & Sly, W. (2013). Structure, function and applications of carbonic anhydrase isozymes. *Bioorganic & Medicinal Chemistry, 21*(6), 1570-1582.

Hilvo, M., Baranauskien, L., Salzano, A. M., Scaloni, A., Matulis, D., Innocenti, A., . . . al., e. (2008). Biochemical Characterization of CA IX, One of the Most Active Carbonic Anhydrase Isozymes. *Biological Chemistry, 283*(41), 27799-27809.

Hsieh, M.-J., Chen, K.-S., Chiou, H.-L., & al, e. (2010). Carbonic anhydrase XII promotes invasion and migration ability of MDA-MB-231 breast cancer cells through the p38 MAPK signaling pathway. *European Journal of Cell Biology, 89*, 598-606.

Hunter, J. (2007). Matplotlib is a 2D graphics package used for Python for application development, interactive scripting, and publication-quality image generation across user interfaces and operating systems. *Computing in Science*, 90-95.

Hynninen, P. P., Huhtala, H., Pastorekova, S., Pastorek, J., Waheed, A., Sly, W., & Tomas, E. (2011). Carbonic anhydrase isozymes II, IX, and XII in uterine tumors. *ACTA Pathologica, Microbiologica, 120*, 117-129.

Jegla, T., Zmasek, C., Batalov, S., & Nayak, S. (2009). Evolution of the Human Ion Channel Set. *Combinatorial Chemistry & High Throughput Screening, 12*, 2-23.

Jones, D., Taylor, W., & Thornton, J. (1992). The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences, 8*(3), 275-282.

Joosten, R., te Beek, T., Krieger, E., Hekkelman, M., Hooft, R., Schneider, R., . . . Vriend, G. (2010). A Series of PDB related databases for everyday needs. *Nucleic Acids Research*, 1-9.

Jude, K., & al., e. (2002). Crystal structure of F65A/Y131C-methylimidazole carbonic anhydrase V reveals architectural features of an engineered proton shuttle. *Biochemistry, 41*(8), 2485-2491.

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

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers, 22*, 2577-2673.

Kale, S., Herfeld, J., Dai, S., & Blank, M. (2012). Lewis-inspired representation of dissociable water in clusters and Grotthuss chains. *Journal of Biological Physics, 38*, 49-59.

Kaunisto, K., Parkkila, S., Rajaniemi, H., Waheed, A., Grubb, J., & Sly, W. (2002). Carbonic anhydrase XIV: Luminal expression suggests key role in renal acidification. *Kidney International, 61*, 2111-2118.

Kimoto, M., Kishino, M., Yura, Y., & Ogawa, Y. (2006). A role of salivary carbonic anhydrase VI in dental plaque. *Oral Biology, 51*, 117-122.

Kimura, M. (1968). Evolutionary Rate at the Molecular Level. *Nature, 217*, 625-626.

Kimura, M. (1991). The neutral theory of molecular evolution: A review of recent evidence. *Genetics, 66*, 367-386.

Kivela, A., Knuuttila, A., Sihvo, E., & al, e. (2012). Carbonic anhydrase IX in malignant pleural mesotheliomas: A potential target for anti-cancer therapy. *Bioorganic & Medicinal Chemistry*.

Knight, R., Maxwell, P., Birmingham, A., Carnes, J., Caporaso, J. G., Easton, B., . . . Smit, S. (2007). PyCogent: a toolkit for making sense from sequence. *Genome Biology, 8*(R171).

Korber, B. (2000). HIV Signature and Sequence Variation Analysis. In *Computational Analysis of HIV Molecular Sequences* (pp. 55-72). Dordrecht, Netherlands: Kluwer Academic Publishers.

Kosiol, C., Holmes, I., & Goldman, N. (2007). An Empirical Codon Model for Protein Sequence Evolution. *Molecular Biology and Evolution, 24*(9).

Krogh, A. (1998). An Introduction to Hidden Markov Models for Biological Sequences. *Computational Methods in Molecular Biology*, 45-63.

Kummola, L., Hamalainen, J., Kivela, J., Kivela, A., Saarnio, J., Karttunen, T., & Parkkila, S. (2005). Expression of a novel carbonic anhydrase, CA XIII, in normal and neoplastic colorectal mucosa. *BMC Cancer, 5*(41).

Kupriyanova, E., & Pronina, N. (2011). Carbonic Anhydrase: Enzyme That Has Transformed the Biosphere. *Russian Journal of Plant Physiology, 58*(2), 197-209.

Lee, Y. S., Kim, T.-H., Kang, T.-W., Chung, W.-H., & Shin, G.-S. (2008). WSPMaker: a web tool for calculating selection pressure in proteins and domains using window-sliding. *BMC Bioinformatics, 9*(13).

Leggat, W., Dixon, R., Saleh, S., & al, e. (2005). A novel carbonic anhydrase from the giant clam Tridacna gigas contains two carbonic anhydrase domains. *FEBS, 272*(13), 3297-3305.

Lehtonen, J., Shen, B., Vihenen, M., Casini, A., Scozzafava, A., Supuran, C., . . . Parkkila, S. (2004). Characterization of CA XIII, a Novel Member of the Carbonic Anhydrase Isozyme Family. *Biological Chemistry, 279*(4), 2719-2727.

Liang, Z., Xue, Y., Behravan, G., Jonsson, B.-H., & Lindskog, S. (1993). Importance of the conserved active-site residues Tyr7, Glu106 and Thr199 for the catalytic function of human carbonic anhydrase II. *European Journal of Biochemistry, 211*(3), 821-827.

Liao, S.-Y., Darcy, K., Randall, L., & al, e. (2010). Prognostic Relevance of Carbonic Anhydrase-IX in High-Risk, Early-Stage Cervical Cancer: A Gynecologic Oncology Group Study. *Gynecol Oncol., 116*(3), 1-16.

Lindskog, S. (1997). Structure and Mechanism of Carbonic Anhydrase. *Pharmacology & Therapeutics , 74*(1), 1-20.

Liu, C.-M., Lin, Y.-M., Yeh, K.-T., Chen, M.-K., Chang, J.-H., Chen, C.-J., . . . Chien, M.-H. (2012). Expression of carbonic anhydrases I/II and the correlation to clinical aspects of oral squamous cell carcinoma analyzed using tissue microarray. *Oral Pathology & Medicine, 41*, 533-539.

Masayuki Nakao, G. I. (2009). Prognostic Significance of Carbonic Anhydrase IX Expression by Cancer-associated Fibroblasts in Lung Adenocarcinoma. *Cancer, 115*(12), 2732-2743.

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

Maupin, M., & Voth, G. (2010). Proton transport in carbonic anhydrase, Insights from molecular simulation. *Biochemica et Biophysica Acta*, 332-341.

Meyer, P.-A. (2009). Stochastic Processes from 1950 to the Present. *Electronic Journal for History of Probability and Statistics, 5*(1), 813-848.

Mikulski, R., Avvaru, B., Tu, C., Case, N., McKenna, R., & Silverman, D. (2011). Kinetic and Crystallographic Studies of the Role of Tyrosine 7 in the Active Site of Human Carbonic Anhydrase II. *Arch Biochem Biophys., 506*(2), 181-187.

Mikulski, R., West, D., Sippel, K., Avvaru, B. S., Aggarwal, M., Tu, C., . . . Silverman, D. N. (2012). Water Networks in Fast Proton Transfer during Catalysis by Human Carbonic Anhydrase II. *BioChemistry, 52*, 125-131.

Momen-Roknabadi, A., Sadeghi, M., Pezeshk, H., & Marashi, S.-A. (2008). Impact of residue accessible surface area on the prediction of protein secondary structures. *BMC Bioinformatics, 9*(357).

Muse, S., & Gaut, B. (1994). A Likelihood Approach for Comparing Synonymous and Nonsynonymous Nucleotide Substitution Rates, with Application to the Chloroplast Genome. *Molecular Biology and Evolution, 11*(5), 715-724.

Nagle, J., & Morowitz, H. (1978). Molecular mechanisms for proton transport in membranes. *PNAS, 75*(1), 298-302.

Needleman, S. B., & Wunsch, C. D. (1970). A General Method Applicable to the Search for Similarities in the Amino Acid Sequence of Two Proteins. *Journal of Molecular Biology, 48*, 443-453.

Nielsen, R., & Yang, Z. (1998). Likelihood Models for Detecting Positively Selected Amino Acid Sites and Applications to the HIV-1 Envelope Gene. *Genetics, 148*, 929-936.

Nishimori, I., Vullo, D., Minakuchi, T., Scozzafava, A., Capasso, C., & Supuran, C. T. (2012). Restoring catalytic activity to the human carbonic anhydrase (CA) related proteins VIII, X and XI affords isoforms with high catalytic efficiency and susceptibility to anion inhibition. *Bioorganic & Medicinal Chemistry Letters, 23*, 256-260.

Odcikin, E., Ozdemir, H., Ciftci, M., & al, e. (2002). INVESTIGATION OF RED BLOOD CELL CARBONIC ANHYDRASE, GLUCOSE 6-PHOSPHATE DEHYDROGENASE, HEXOKINASE ENZYME ACTIVITIES, AND ZINC CONCENTRATION IN PATIENTS WITH HYPERTHYROID DISEASES. *Endocrine Research, 28*(1,2), 61-68.

Okuyama, T., Waheed, A., Kusumoto, W., Zhu, X. L., & Sly, W. (1995). Carbonic Anhydrase IV: Role of Removal of C-terminal Domain in Glycosylphosphatidylinositol Anchoring and Realization of Enzyme Activity. *Archives of Biochemistry and Biophysics, 320*(2), 315-322.

Orlean, P., & Menon, A. (2007). GPI anchoring of protein in yeast and mammalian cells, or: how we learned to stop worrying and love glycophospholipids. *Lipid Research, 48*, 993-1011.

Ozensoy, O., Kockar, F., Arslan, O., & al, e. (2006). An evaluation of cytosolic erythrocyte carbonic anhydrase and catalase in carcinoma patients: An elevation of carbonic anhydrase activity. *Clinical BioChemistry*(39), 804-809.

Petterson, E., Goddard, T., Huang, C., Couch, G., Greenblatt, D., Meng, E., & Ferrin, T. (2004). UCSF Chimera—A Visualization System for Exploratory. *Computational Chemistry, 25*(13), 1605-1612.

Picaud, S., & al., e. (2009). Crystal structure of human carbonic anhydrase-related protein VIII reveals the basis for catalytic silencing. *Proteins, 76*(2), 507-511.

Pike, L. (2009). The challenge of lipid rafts. *Lipid Research, April*(Supplement), 323-328.

Pilka, E., & al., e. (n.d.). To be published.

Pond, S., Frost, S., & Muse, S. (2005). HyPhy: hypothesis testing using phyogenies. *Bioinformatics, 21*(5), 676-679.

*Python.org*. (n.d.). Retrieved 2013, from http://www.python.org/

Rabiner, L. (1989). A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IEEE, 77*(2).

Rasheed, S., Harris, A., & Tekkis, P. (2008). Assessment ofmicrovesseldensityandcarbonicanhydrase-9(CA-9) expression in rectal cancer. *Pathology Research and Practice*(205), 1-9.

Riccardi, D., Konig, P., Prat-Resina, X., Yu, H., Elstner, M., Frauenheim, T., & Cui, Q. (2006). "Proton holes" in long-range proton transfer reactions in solution and enzymes: A theoretical analysis. *Journal of the American Chemistry Society, 128*(50), 16302-16311.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D., Darling, A., Hohna, S., . . . Huelsenbeck, J. P. (2011). MrBayes 3.2: Efficient Bayesian Phylogenetic Inference and Model Choice Across a Large Model Space. *Systematic Biology, 61*(3), 539-542.

Scherrer, M., Meyer, A., & Wilke, C. (2012). Modeling coding-sequence evolution within the context of residue solvent accessibiility. *BMC Evolutionary Biology, 12*(179).

Schneider, A., Cannarozzi, G., & Gonnet, G. (2005). Empirical codon substitution matrix. *BMC Bioinformatics, 6*(134).

Schneider, H.-P., Alt, M., Klier, M., Spiess, A., Andes, F., Waheed, A., . . . Deitmer, J. (2013). GPI-anchored carbonic anhydrase IV displays both intra- and extracellular activity in cRNA-injected oocytes and in mouse neurons. *PNAS, 10*(4), 1494-1499.

Shah, G., Hewett-Emett, D., Grubb, J., Migas, M., Fleming, R., Waheed, A., & Sly, W. (2000). Mitochondrial carbonic anhydrase CA VB: Differences in tissue distribution and pattern of evolution from those of CA VA suggest distinct physiological roles. *PNAS, 97*(4), 1677-1682.

Shakin-Eshleman, S. H., Spitalnik, S. L., & Kasturi, L. (1996). The amino acid following an asn-X-Ser/Thr sequon is an important determinant of N-linked core glycosylation efficiency. *Biological Chemistry, 271*(11), 6363-6366.

Sherwood, B. T., Colquhoun, A., & D., R. (2007). Carbonic Anhydrase IX Expression and Outcome after Radiotherapy for Muscle-invasive Bladder Cancer. *Clinical Oncology*(19), 777-783.

Sievers, F., Wilm, A., Dineen, D., Gibson, T. J., Karplus, K., Li, W., . . . Higgins, D. G. (2011). Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology, 7*(539).

Slater, G. S., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics, 6*(31).

Sonnhammer, E., Heijne, G., & Krogh, A. (1998). A hidden Markov model for predicting transmembrane helices in protein sequences. *Proceedings of Sixth International Conference on Intelligent Systems for Molecular Biology*, 175-182.

Srivastava, D. e. (2007). Structural Analysis of Charge Discrimination in the Binding of Inhibitors to Human Carbonic Anhydrases I and II. *American Chemistry Society, 129*(17), 5528-5537.

Stams, T., & al., e. (1996). Crystal structure of the secretory form of membrane-associated human carbonic anhydrase IV at 2.8-A resolution. *PNAS, 93*(24), 13589-13594.

Stelzer, G., Dalah, I., Stein, I., & al, e. (2011). In-silico Human Genomics with GeneCards. *Human Genomics, 5*(6), 709-717. Retrieved from www.genecards.org

Sterling, D., Alvarez, B., & Casey, J. (2002). The extracellular Component of a Transport Metabolon. *Biological Chemistry, 277*(28), 25239-25246.

Stern, A., Doron-Faigenboim, A., Erez, E., Martz, E., Bacharach, E., & Pupko, T. (2007). Selecton 2007: advanced models for detecting positive and purifying selection using a Bayesian inference approach. *Nucleic Acids Research, 35*(Web Server).

Su, A., Wiltshire, T., Batalov, S., Lapp, H., & al., e. (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. *Proceedings of the National Academy of Science, 16*(6062-6067), 101.

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Research, 34*(Web Server), 609-612.

Tashian, R. (1989). The Carbonic Anhydrases: Widening Perspectives on Their Evolution, Expression and Function. *BioEssays, 10*(6), 186-192.

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

Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., M, F., . . . Ponten, F. (2010). Towards a knowledge-based Human Protein Atlas. *Nature Biotechnology, 28*(12), 1248-1250.

Vernier, W., Chong, W., Rewolinski, D., Greasley, S., Pauly, T., Shaw, M., . . . Reyner, E. (2010). Thioether benzenesulfonamide inhibitors of carbonic anhydrases II and IV: Structure-based drug design, synthesis, and biological evaluation. *Bioorganic & Medicinal Chemistry, 18*(19), 3307-3319.

Whittington, D., & al., e. (2001). Crystal structure of the dimeric extracellular domain of human carbonic anhydrase XII, a bitopic membrane protein overexpressed in certain cancer tumor cells. *PNAS, 98*(17), 9545-9550.

Whittington, D., & al., e. (2004). Expression, assay, and structure of the extracellular domain of murine carbonic anhydrase XIV: implications for selective inhibition of membrane-associated isozymes. *Biological Chemistry, 279*(8), 7223-7228.

Wistrand, P., Carter, N., Conroy, C., & Maheiu, I. (1999). Carbonic anhydrase IV activity is localized on the exterior surface of human erythrocytes. *Acta Physiologica Scandinavica, 165*(2), 211-218.

Wraight, C. (2006). Chance and design - Proton transfer in water, channels and bioenergetic proteins. *Biochimica et Biophysica Acta, 1757*, 886-912.

Wu, C., Orozco, C., Boyer, J., Leglise, M., Goodale, J., Batalov, S., . . . Su, A. (2009). BioGPS: an extensible and customizable portal for querying and organizing gene annotation resources. *Genome Biology, 10*(11).

Yang, Z. (1997). *Phylogenetic Analysis by Maximum Likelihood (PAML).* http://abacus.gene.ucl.ac.uk/software/paml.html).

Yang, Z., & Nielsen, R. (2002). Codon-Substitution Models for Detecting Molecular Adaptation at Individual Sites Along Specific Lineages. *Molecular Biology and Evolution, 19*(6), 908-917.

Yang, Z., & Swanson, W. (2002). Codon-Substitution Models to Detect Adaptive Evolution that Account for Heterogeneous Selective Pressures Among Site Classes. *Molecular Biology and Evolution, 19*(1), 49-57.

Yang, Z., Nielsen, R., Goldman, N., & Pedersen, A.-M. K. (2000). Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites. *Genetics, 155*, 431-449.

Yang, Z., Nielsen, R., Goldman, N., & Pedersen, A.-M. K. (2000). Codon-Substitution Models for Heterogeneous Selection Pressure at Amino Acid Sites. *Genetics, 155*, 431-449.

Zhang, C., Wang, J., Long, M., & Fan, C. (2013). gKaKs: the pipeline for genome-level Ka/Ks calculation. *Bioinformatics, 29*(5), 645-646.

Zheng, J., Avvaru, B. S., Tu, C. M., & Silverman, D. (2008). Role of Hydrophilic Residues in Proton Transfer during Catalysis by Human Carbonic Anhydrase II. *Biochemistry, 47*, 12028-12036.

Zhu, X. L., & Sly, W. (1990). Carbonic Anhydrase IV from Human Lung. *Biological Chemistry, 265*(15), 8795-8801.

Zoller, S., & Schneider, A. (2010). Empirical Analysis of the Most Relevant Parameters of Codon Substitution Models. *Jounal of Molecular Evolution, 70*, 605-612.

# Appendix A

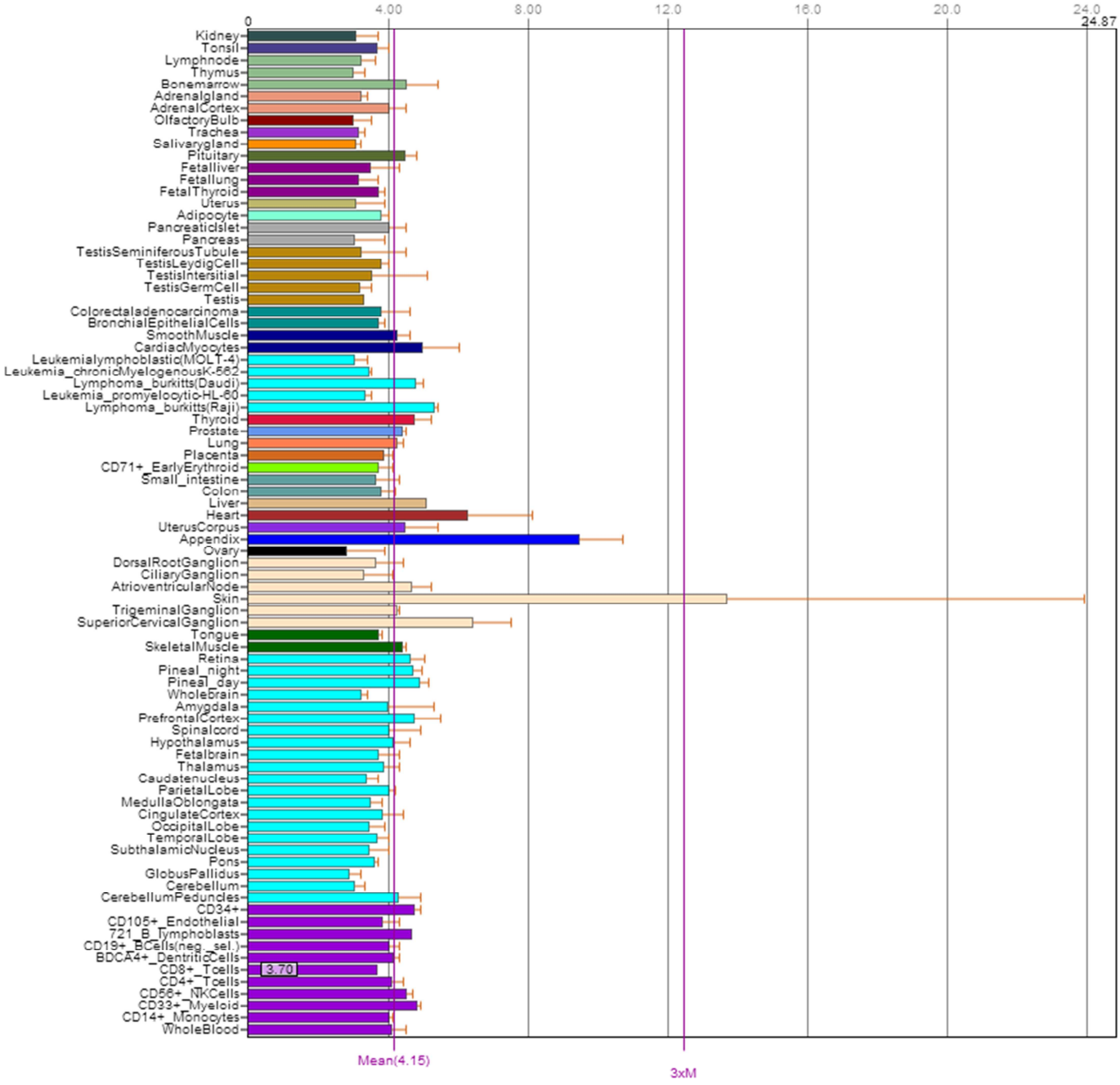


Figure 27 – CA1 expression by tissue (Wu, et al., 2009).

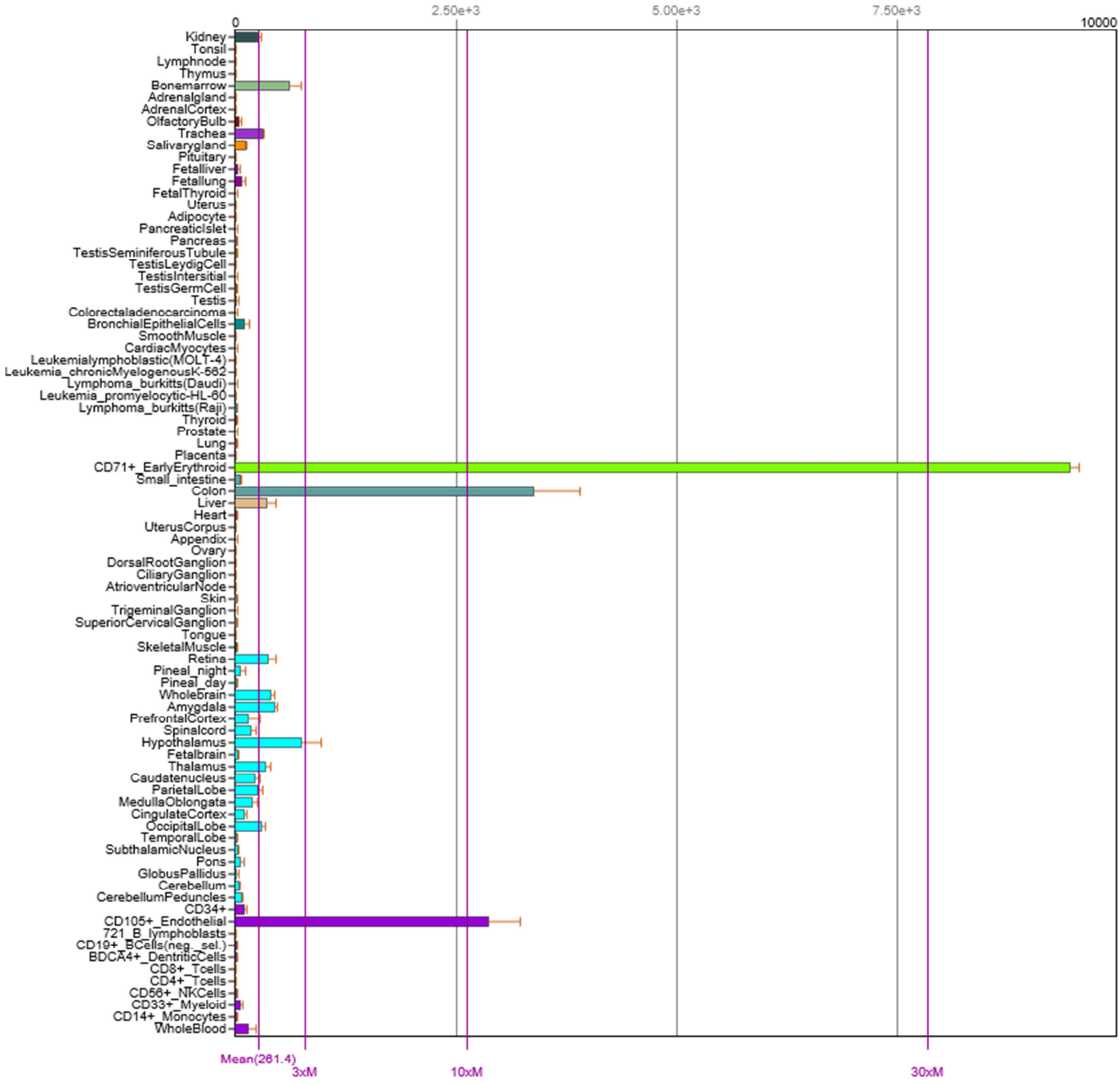


Figure 28 – CA2 expression by tissue (Wu, et al., 2009).

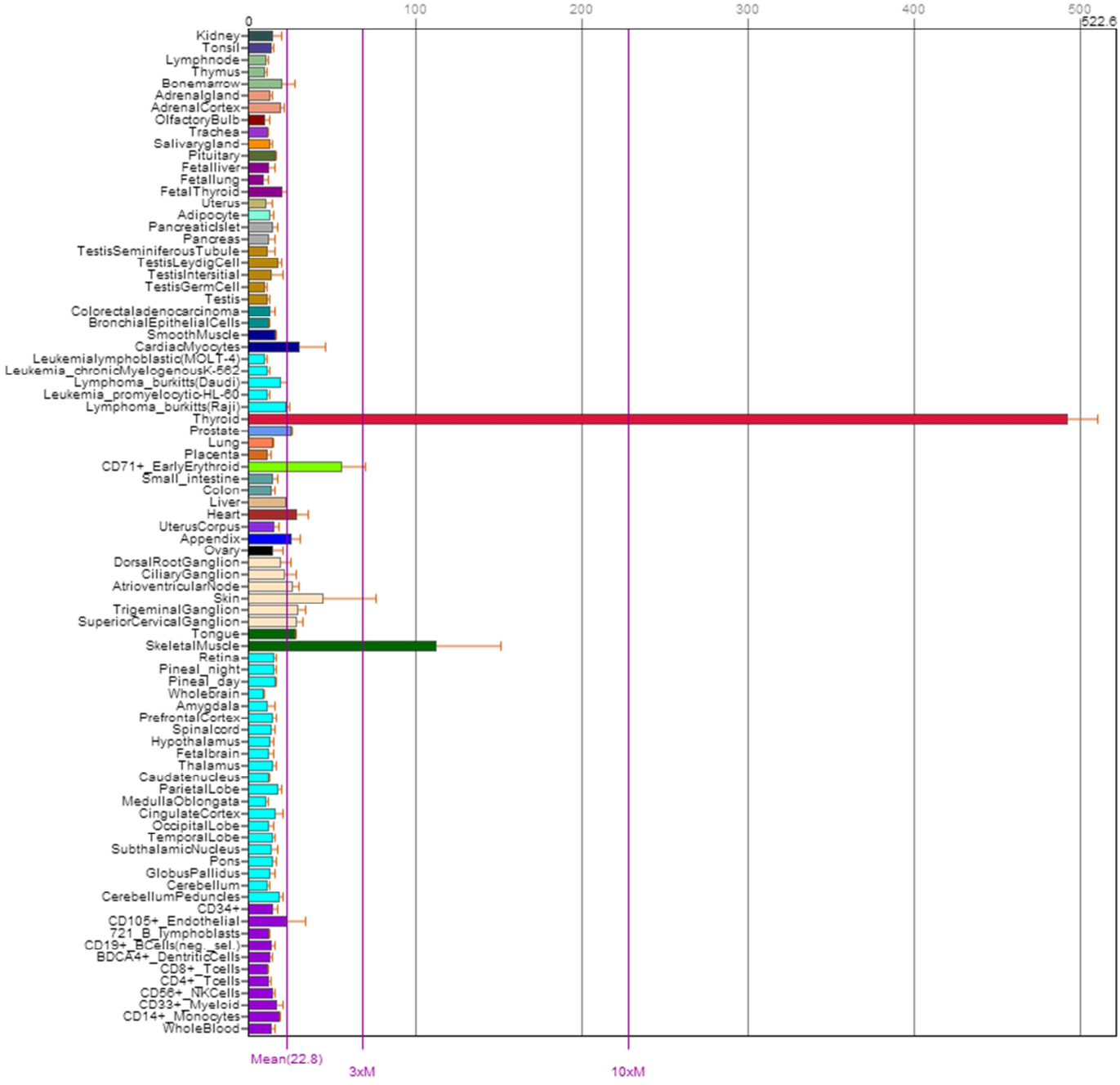


Figure 29 – CA3 expression by tissue (Wu, et al., 2009).

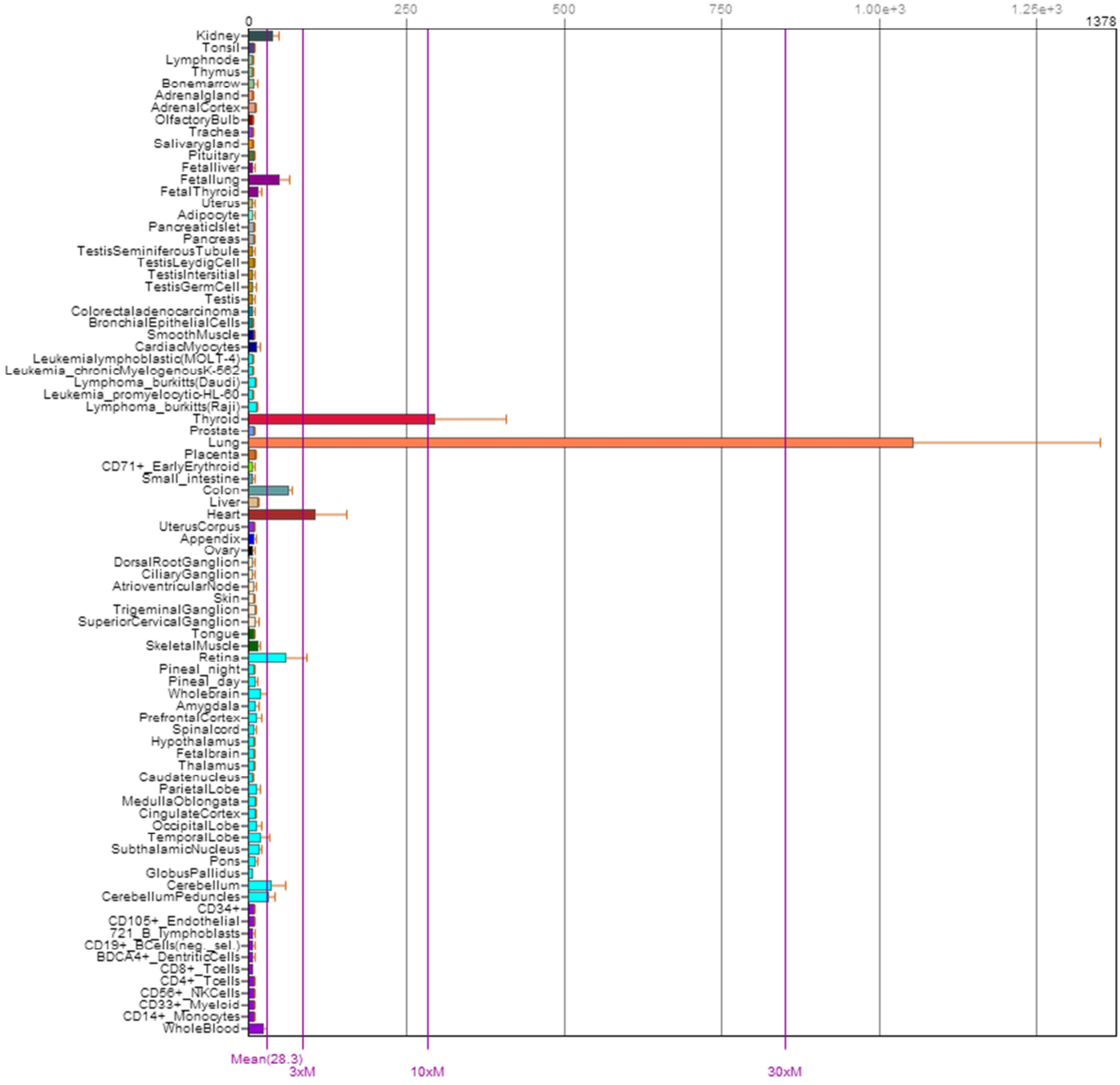


Figure 30 – CA4 expression by tissue (Wu, et al., 2009).

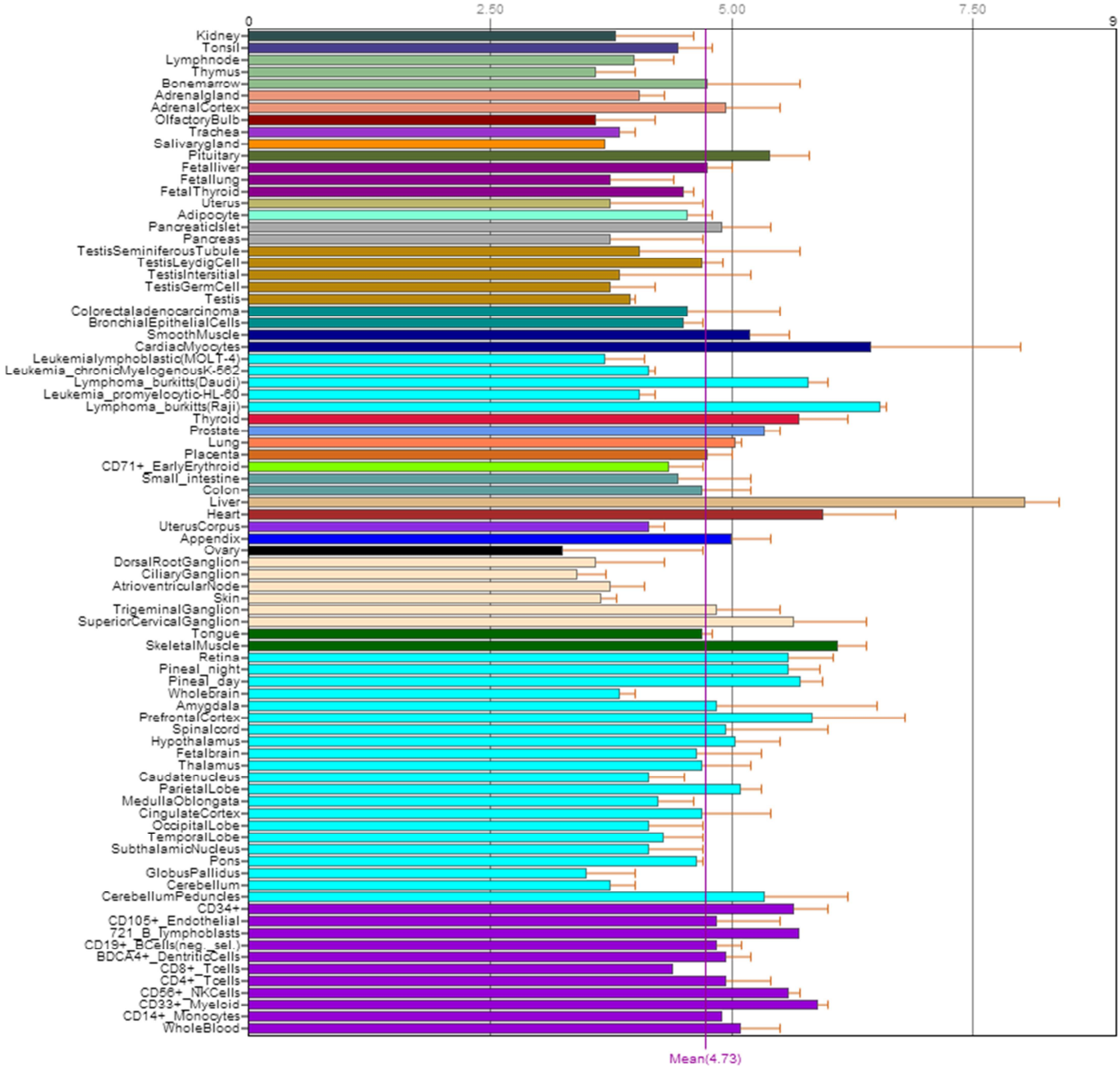


Figure 31 – CA5a expression by tissue (Wu, et al., 2009).

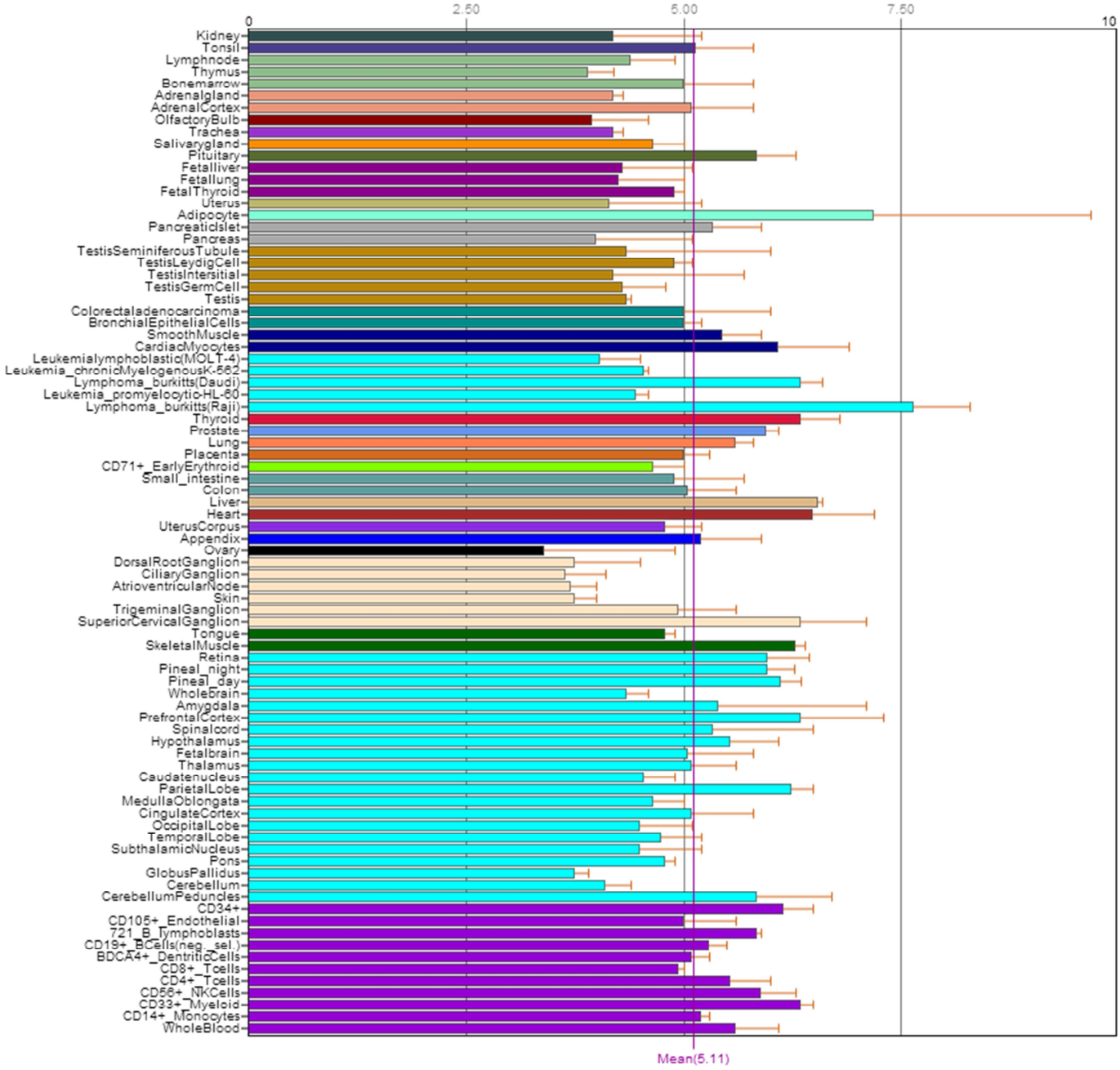


Figure 32 – CA5b expression by tissue (Wu, et al., 2009).

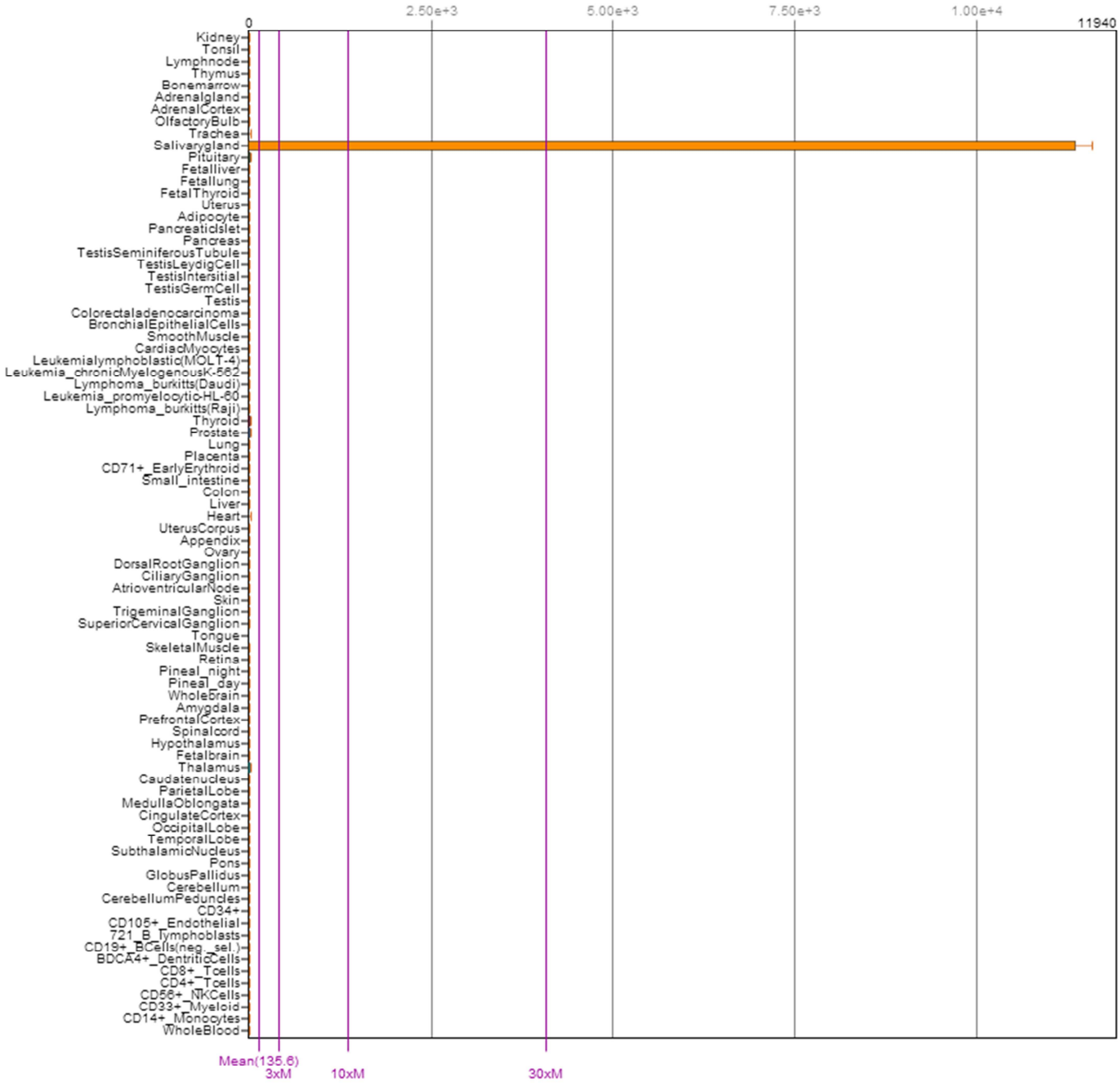


Figure 33 – CA6 expression by tissue (Wu, et al., 2009).

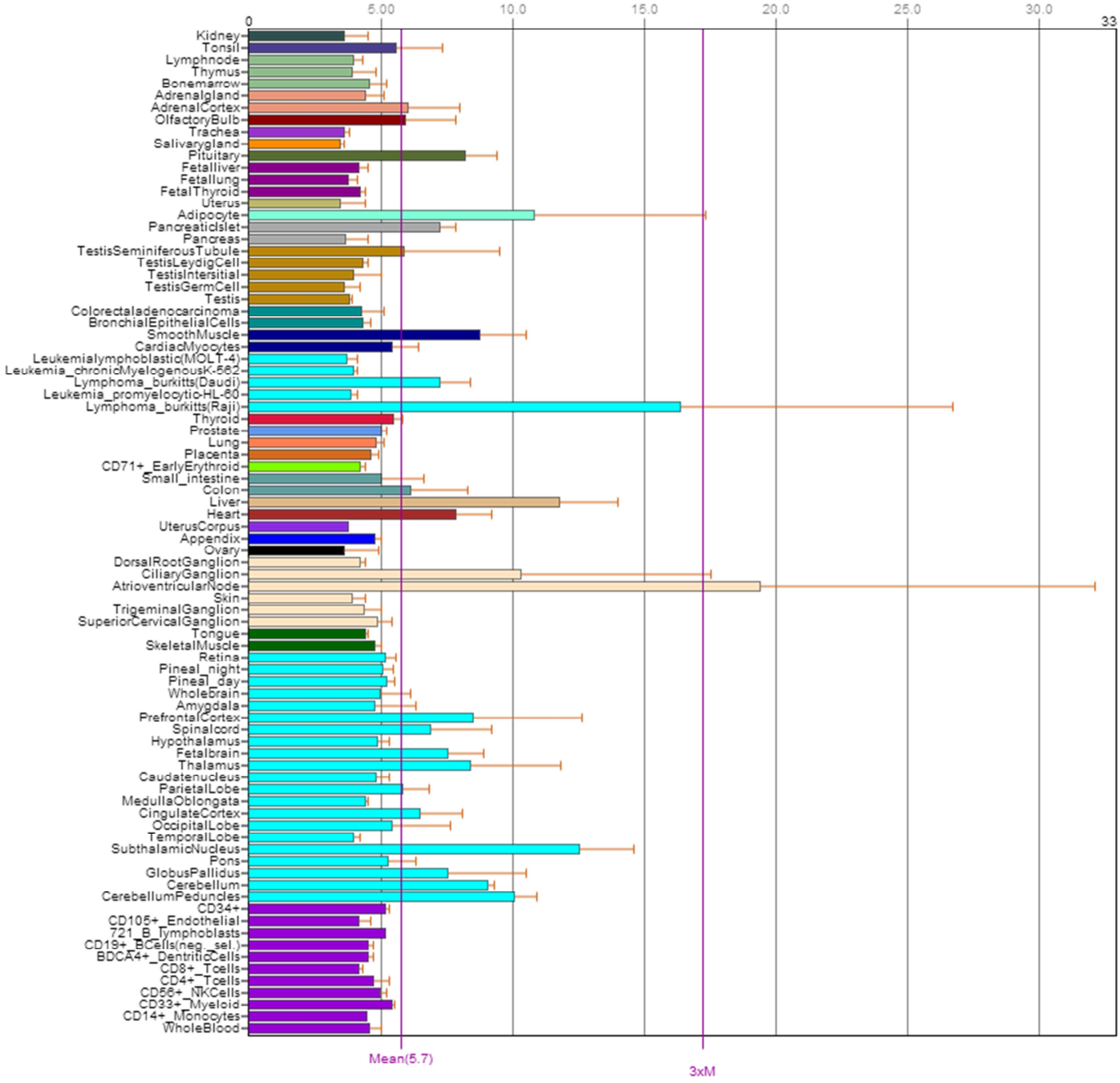


Figure 34 – CA7 expression by tissue (Wu, et al., 2009).

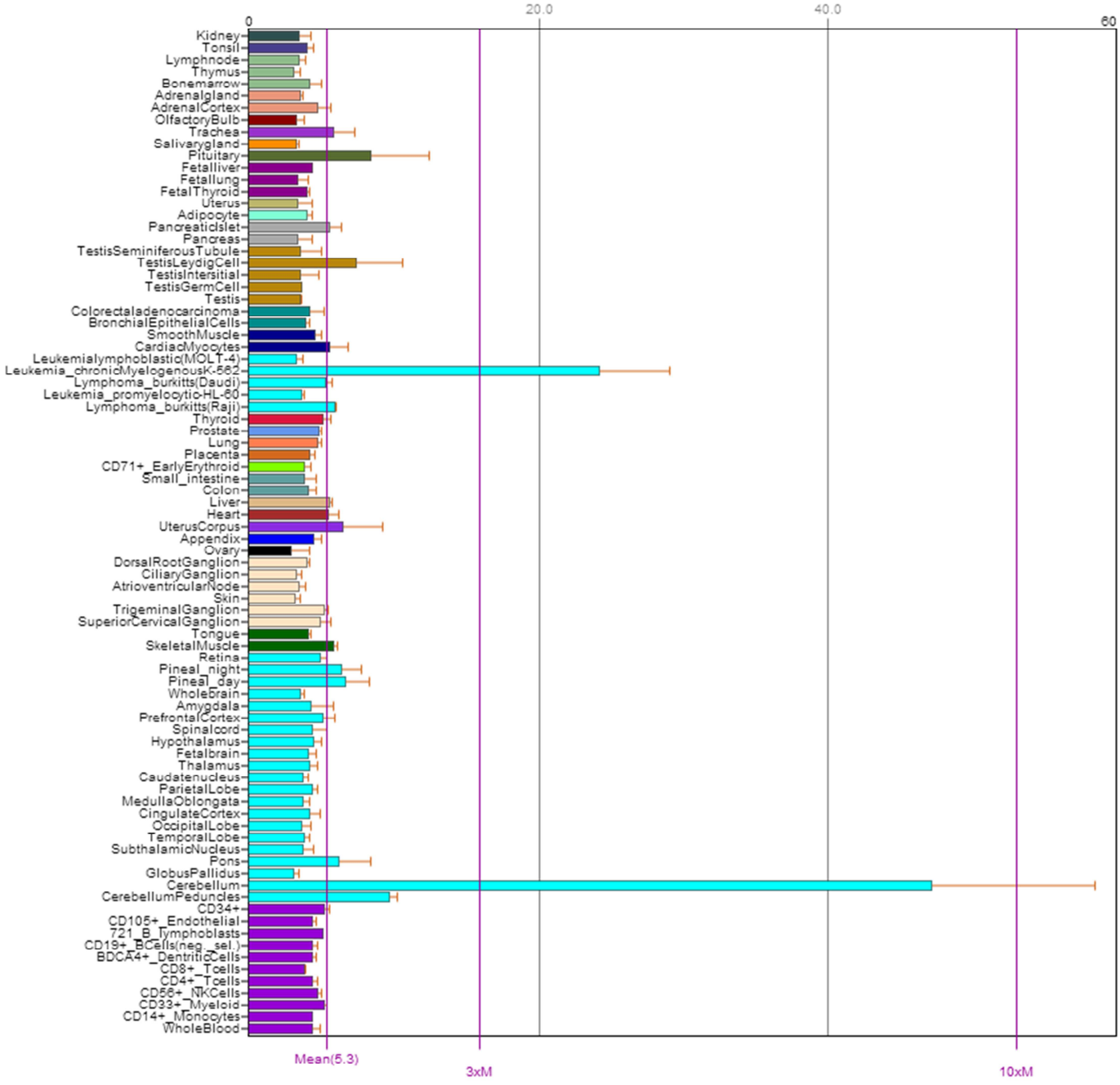


Figure 35 – CA8 expression by tissue (Wu, et al., 2009).

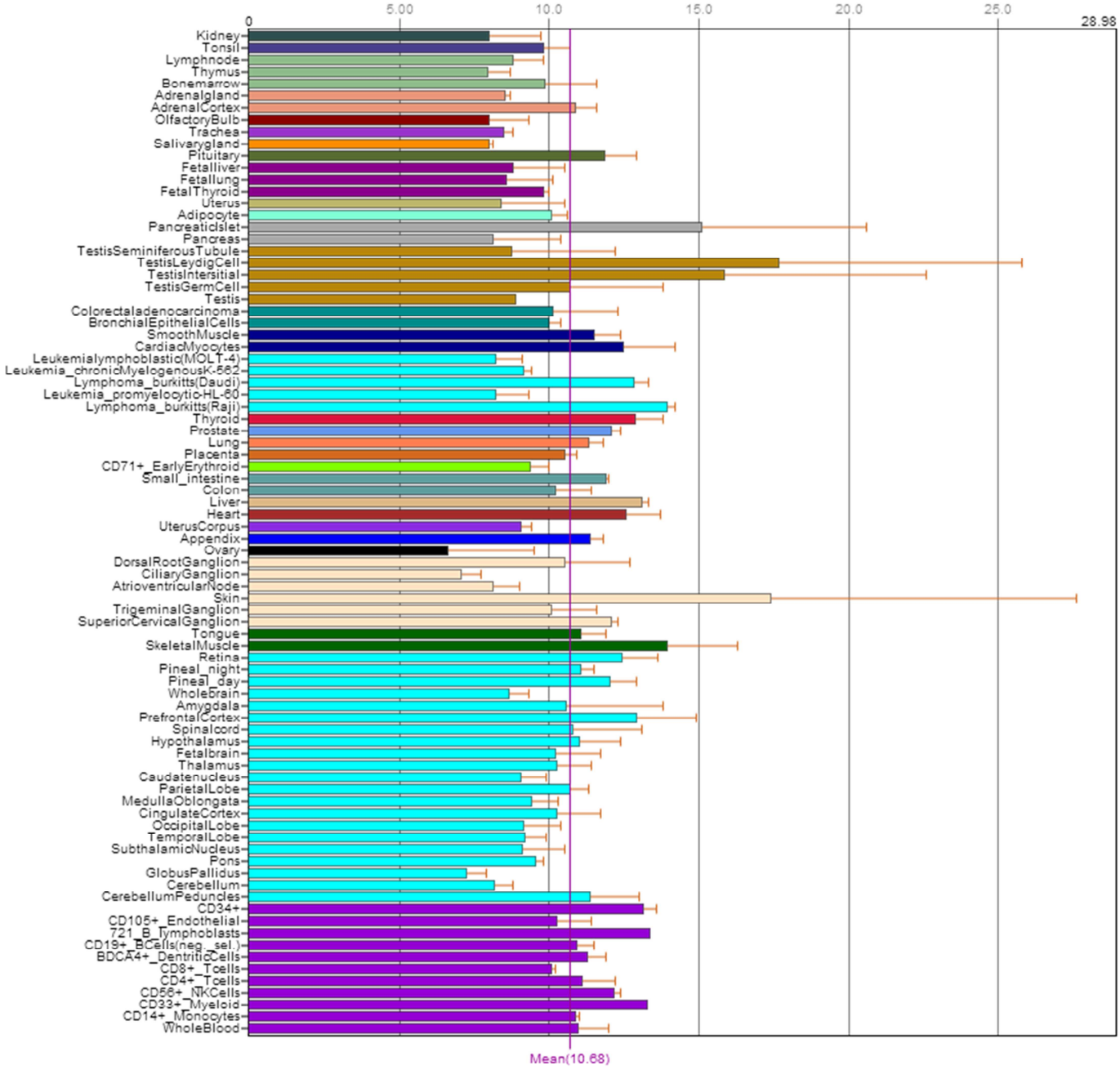


Figure 36 – CA9 expression by tissue (Wu, et al., 2009).

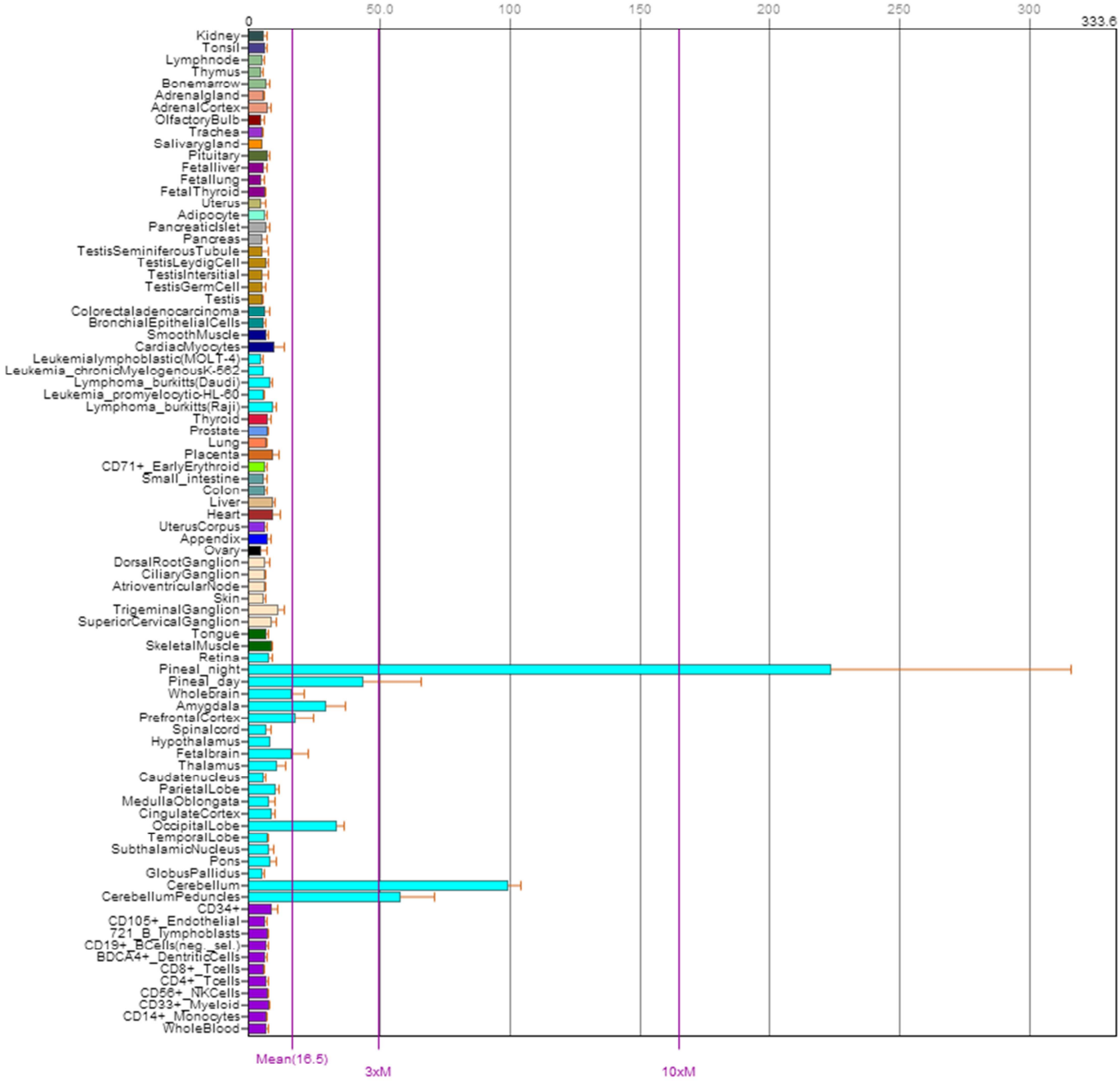


Figure 37 – CA10 expression by tissue (Wu, et al., 2009).

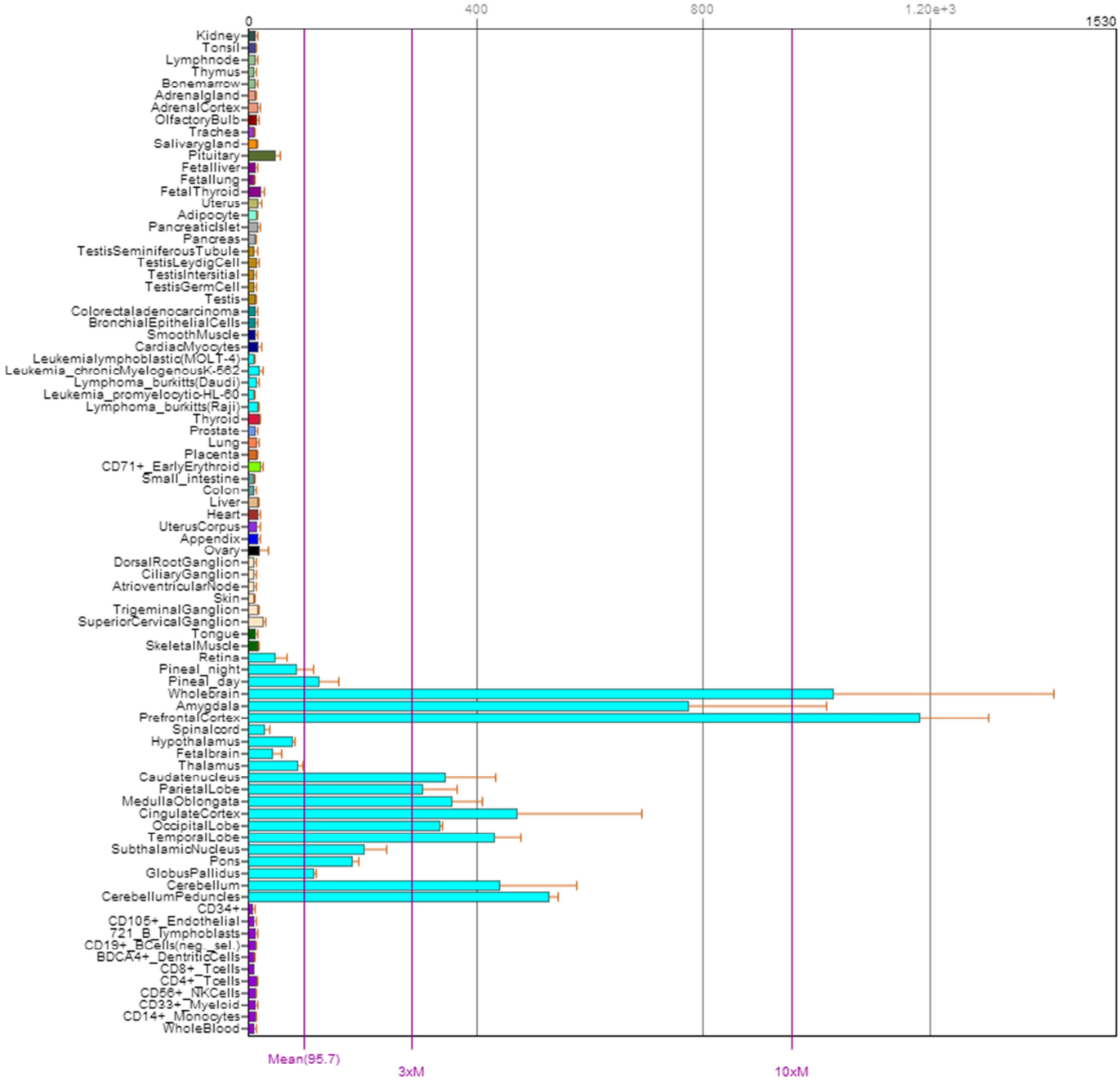


Figure 38 – CA11 expression by tissue (Wu, et al., 2009).

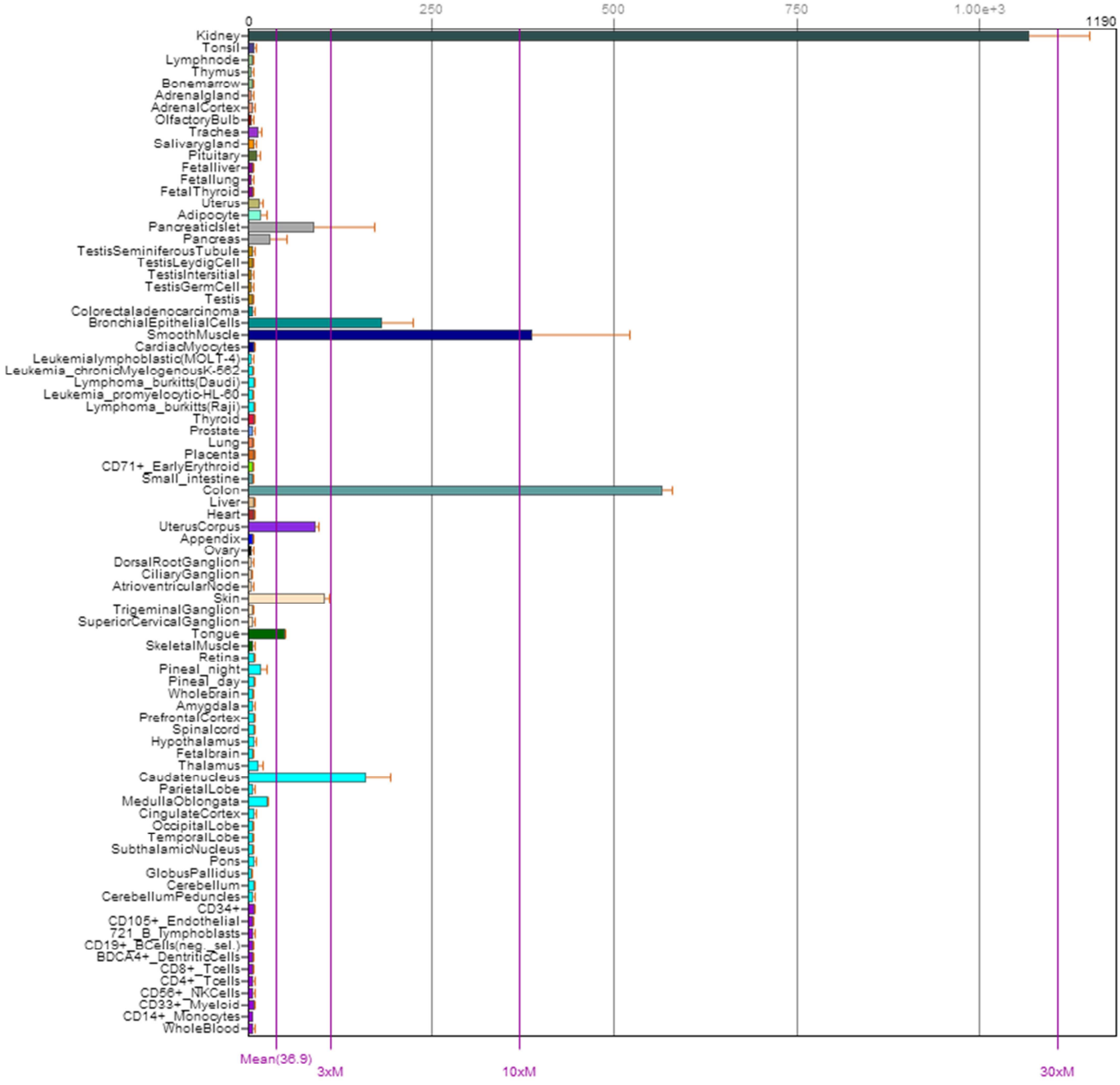


Figure 39 – CA12 expression by tissue (Wu, et al., 2009).

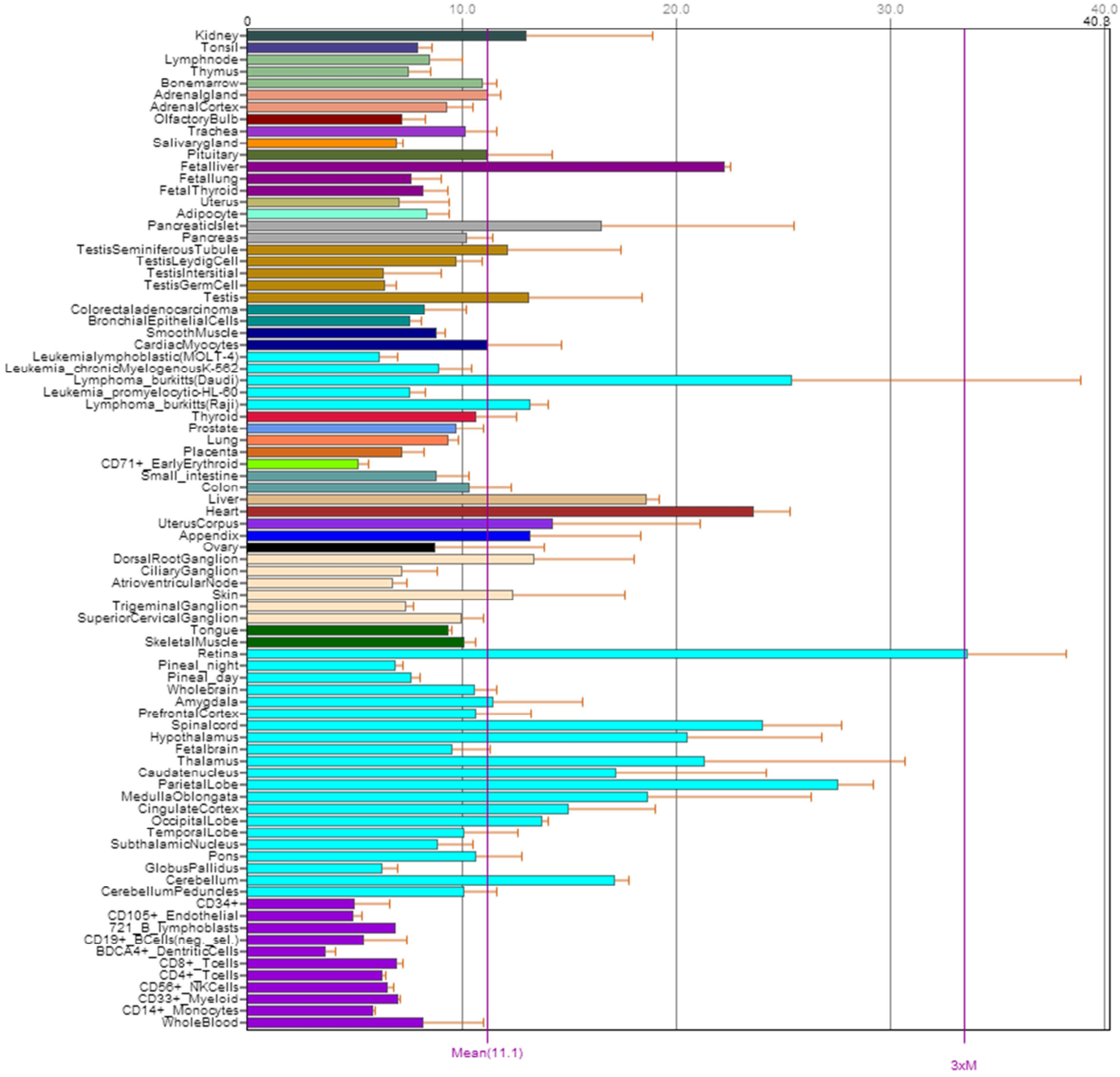


Figure 40 – CA14 expression by tissue (Wu, et al., 2009).

# Appendix B

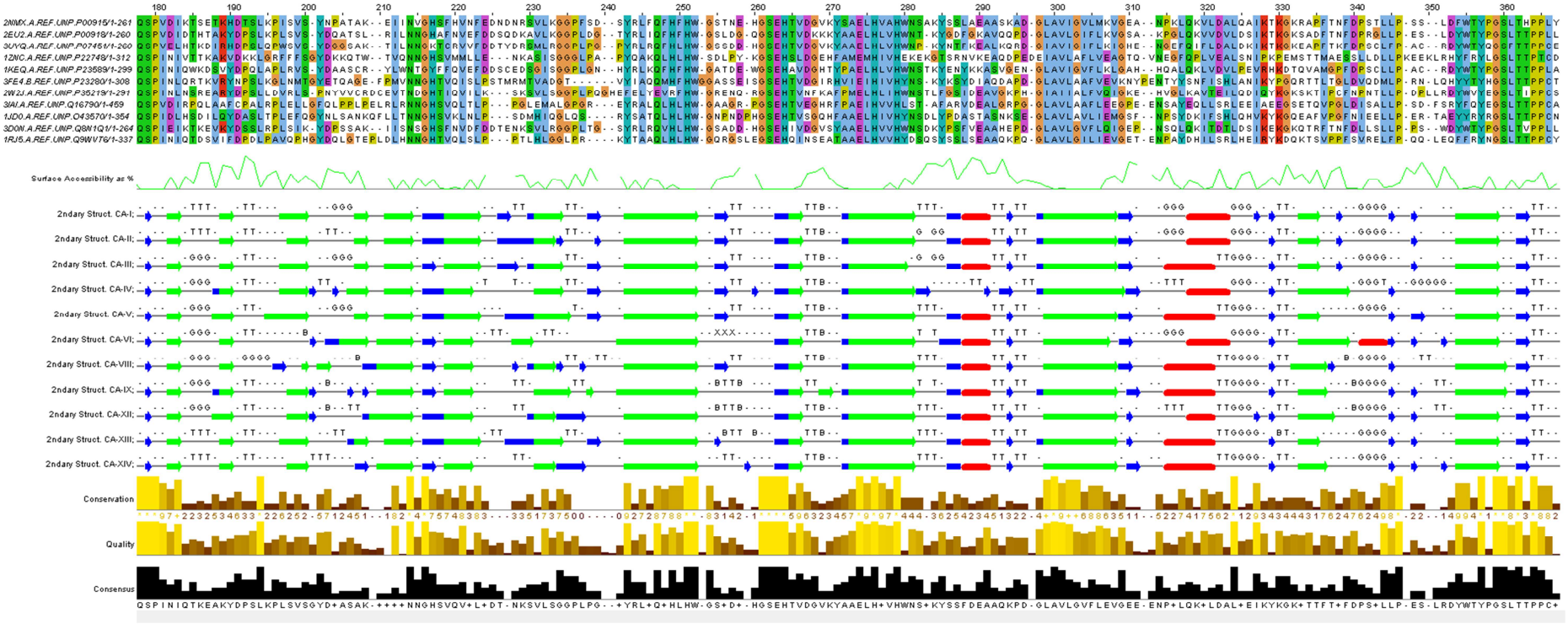


Table 6 – At top, a MUSCLE (Edgar, 2004) based alignment of the catalytic domain of carbonic anhydrases (I, II, III, IV, V, VI, VIII, IX, XIII, XIV) as retrieved from PDB (Bernstein, et al., 1977) entries (2NMX.A (Srivastava, 2007), 2EU2.A (Fisher S. e., 2006), 3UYQ.A (Elder & al., 2007), 1ZNC.A (Stams & al., 1996), 1KEQ.A (Jude & al., 2002), 3FE4.B (Pilka & al.), 2W2J.A (Picaud & al., 2009), 3IAI.A (Alterio & al., 2009), 1JD0.A (Whittington & al., 2001), 3D0N.A (Di Fiore & al., 2009), 1RJ5.A (Whittington & al., 2004)). RSA of CA-IV amino acids are presented as a line graph. At middle, secondary structure data created using DSSP is displayed in alignment for all CAs present; helices are shown as red tubes, sheets are green arrows, and bends as blue arrows. Image is presented as created in Jalview, using annotations generated using DSSP and python scripting.

# Appendix C

| CA-IV ENS_POS | ENS | PDB | PDB_POS | RSA | LOC | | CA-IV ENS_POS | ENS | PDB | PDB_POS | RSA | LOC | | CA-IV ENS_POS | ENS | PDB | PDB_POS | RSA | LOC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | M | - | - | NA | NA | | 73 | S | S | 50 | 0.2 | Surface | | 145 | K | K | 124 | 0.3 | Surface |
| 2 | R | - | - | NA | NA | | 74 | G | G | 50 | 0.2 | Surface | | 146 | E | E | 125 | 0.64 | Surface |
| 3 | M | - | - | NA | NA | | 75 | Y | Y | 51 | 0.01 | Buried | | 147 | K | - | - | NA | NA |
| 4 | L | - | - | NA | NA | | 76 | D | D | 52 | 0.72 | Surface | | 148 | G | - | - | NA | NA |
| 5 | L | - | - | NA | NA | | 77 | K | K | 53 | 0.38 | Surface | | 149 | T | - | - | NA | NA |
| 6 | A | - | - | NA | NA | | 78 | K | K | 54 | 0.48 | Surface | | 150 | S | - | - | NA | NA |
| 7 | L | - | - | NA | NA | | 79 | Q | Q | 55 | 0.22 | Surface | | 151 | R | - | - | NA | NA |
| 8 | L | - | - | NA | NA | | 80 | T | T | 56 | 0.63 | Surface | | 152 | N | - | - | NA | NA |
| 9 | A | - | - | NA | NA | | 81 | W | W | 57 | 0.01 | Buried | | 153 | V | - | - | NA | NA |
| 10 | L | - | - | NA | NA | | 82 | T | T | 58 | 0.3 | Surface | | 154 | K | - | - | NA | NA |
| 11 | S | - | - | NA | NA | | 83 | V | V | 59 | 0.01 | Buried | | 155 | E | E | 133 | 0.8 | Surface |
| 12 | A | - | - | NA | NA | | 84 | Q | Q | 60 | 0.26 | Surface | | 156 | A | A | 134 | 1.03 | Surface |
| 13 | A | - | - | NA | NA | | 85 | N | N | 61 | 0.01 | Buried | | 157 | Q | Q | 135 | 0.34 | Surface |
| 14 | R | - | - | NA | NA | | 86 | N | N | 62 | 0.36 | Surface | | 158 | D | D | 136 | 0.47 | Surface |
| 15 | P | - | - | NA | NA | | 87 | G | G | 63 | 0.13 | Buried | | 159 | P | P | 137 | 0.63 | Surface |
| 16 | S | - | - | NA | NA | | 88 | H | H | 64 | 0.27 | Surface | | 160 | E | E | 138 | 0.39 | Surface |
| 17 | A | - | - | NA | NA | | 89 | S | S | 65 | 0.04 | Buried | | 161 | D | D | 139 | 0.1 | Buried |
| 18 | S | - | - | NA | NA | | 90 | V | V | 66 | 0.01 | Buried | | 162 | E | E | 140 | 0.23 | Surface |
| 19 | A | - | - | NA | NA | | 91 | M | M | 67 | 0.12 | Buried | | 163 | I | I | 141 | 0.09 | Buried |
| 20 | E | - | - | NA | NA | | 92 | M | M | 68 | 0.01 | Buried | | 164 | A | A | 142 | 0 | Buried |
| 21 | S | - | - | NA | NA | | 93 | L | L | 69 | 0.26 | Surface | | 165 | V | V | 143 | 0.04 | Buried |
| 22 | H | H | 4 | 1.07 | Surface | | 94 | L | L | 70 | 0 | Buried | | 166 | L | L | 144 | 0 | Buried |
| 23 | W | W | 5 | 0.24 | Surface | | 95 | E | E | 71 | 0.48 | Surface | | 167 | A | A | 145 | 0 | Buried |
| 24 | C | C | 6 | 0.01 | Buried | | 96 | N | N | 72 | 0.47 | Surface | | 168 | F | F | 146 | 0 | Buried |
| 25 | Y | Y | 7 | 0.01 | Buried | | 97 | K | K | 76 | 0.48 | Surface | | 169 | L | L | 147 | 0 | Buried |
| 26 | E | E | 8 | 0.43 | Surface | | 98 | A | A | 77 | 0.04 | Buried | | 170 | V | V | 148 | 0 | Buried |
| 27 | V | V | 9 | 0.27 | Surface | | 99 | S | S | 78 | 0.2 | Buried | | 171 | E | E | 149 | 0.35 | Surface |
| 28 | Q | Q | 10 | 0.11 | Buried | | 100 | I | I | 79 | 0.01 | Buried | | 172 | A | A | 150 | 0.5 | Surface |
| 29 | A | A | 11 | 0.51 | Surface | | 101 | S | S | 80 | 0.35 | Surface | | 173 | G | G | 151 | 1.02 | Surface |
| 30 | E | E | 11 | 0.51 | Surface | | 102 | G | G | 81 | 0.3 | Surface | | 174 | T | T | 151 | 1.02 | Surface |
| 31 | S | S | 11 | 0.51 | Surface | | 103 | G | G | 82 | 0 | Buried | | 175 | Q | Q | 152 | 0.34 | Surface |
| 32 | S | S | 11 | 0.51 | Surface | | 104 | G | G | 83 | 0.49 | Surface | | 176 | V | V | 153 | 0.43 | Surface |
| 33 | N | N | 11 | 0.51 | Surface | | 105 | L | L | 84 | 0.11 | Buried | | 177 | N | N | 154 | 0.24 | Surface |
| 34 | Y | Y | 11 | 0.51 | Surface | | 106 | P | P | 85 | 0.72 | Surface | | 178 | E | E | 155 | 0.71 | Surface |
| 35 | P | P | 11 | 0.51 | Surface | | 107 | A | A | 86 | 0.25 | Surface | | 179 | G | G | 156 | 0.19 | Buried |
| 36 | C | C | 11 | 0.51 | Surface | | 108 | P | P | 87 | 0.39 | Surface | | 180 | F | F | 157 | 0 | Buried |
| 37 | L | L | 11 | 0.51 | Surface | | 109 | Y | Y | 88 | 0 | Buried | | 181 | Q | Q | 158 | 0.21 | Surface |
| 38 | V | V | 12 | 0.2 | Surface | | 110 | Q | Q | 89 | 0.27 | Surface | | 182 | P | P | 159 | 0.16 | Buried |
| 39 | P | P | 13 | 0.07 | Buried | | 111 | A | A | 90 | 0 | Buried | | 183 | L | L | 160 | 0.01 | Buried |
| 40 | V | V | 14 | 0.82 | Surface | | 112 | K | K | 91 | 0.36 | Surface | | 184 | V | V | 161 | 0.06 | Buried |
| 41 | K | K | 15 | 0.73 | Surface | | 113 | Q | Q | 92 | 0.21 | Surface | | 185 | E | E | 162 | 0.54 | Surface |
| 42 | W | W | 16 | 0.01 | Buried | | 114 | L | L | 93 | 0.01 | Buried | | 186 | A | A | 163 | 0.07 | Buried |
| 43 | G | G | 20 | 0.52 | Surface | | 115 | H | H | 94 | 0.15 | Buried | | 187 | L | L | 164 | 0 | Buried |
| 44 | G | G | 21 | 0.79 | Surface | | 116 | L | L | 95 | 0.01 | Buried | | 188 | S | S | 165 | 0.53 | Surface |
| 45 | N | N | 22 | 0.39 | Surface | | 117 | H | H | 96 | 0.01 | Buried | | 189 | N | N | 166 | 0.57 | Surface |
| 46 | C | C | 23 | 0 | Buried | | 118 | W | W | 97 | 0 | Buried | | 190 | I | I | 167 | 0 | Buried |
| 47 | Q | Q | 24 | 0.6 | Surface | | 119 | S | S | 98 | 0.01 | Buried | | 191 | P | P | 168 | 0.17 | Buried |
| 48 | K | K | 25 | 0.35 | Surface | | 120 | D | D | 99 | 0.31 | Surface | | 192 | K | K | 169 | 0.34 | Surface |
| 49 | D | D | 26 | 0.87 | Surface | | 121 | L | L | 100 | 0.4 | Surface | | 193 | P | P | 170 | 0.18 | Buried |
| 50 | R | R | 27 | 0.29 | Surface | | 122 | P | P | 101 | 0.57 | Surface | | 194 | E | E | 171 | 0.91 | Surface |
| 51 | Q | Q | 28 | 0.01 | Buried | | 123 | Y | Y | 102 | 0.64 | Surface | | 195 | M | M | 172 | 0.31 | Surface |
| 52 | S | S | 29 | 0 | Buried | | 124 | K | K | 103 | 0.42 | Surface | | 196 | S | S | 173 | 0.47 | Surface |
| 53 | P | P | 30 | 0 | Buried | | 125 | G | G | 104 | 0 | Buried | | 197 | T | T | 174 | 0.18 | Buried |
| 54 | I | I | 31 | 0.01 | Buried | | 126 | S | S | 105 | 0 | Buried | | 198 | T | T | 175 | 0.61 | Surface |
| 55 | N | N | 32 | 0.36 | Surface | | 127 | E | E | 106 | 0.01 | Buried | | 199 | M | M | 176 | 0.01 | Buried |
| 56 | I | I | 33 | 0.01 | Buried | | 128 | H | H | 107 | 0 | Buried | | 200 | A | A | 177 | 0.55 | Surface |
| 57 | V | V | 34 | 0.37 | Surface | | 129 | S | S | 108 | 0.01 | Buried | | 201 | E | E | 178 | 0.69 | Surface |
| 58 | T | T | 35 | 0.2 | Surface | | 130 | L | L | 109 | 0.1 | Buried | | 202 | S | S | 179 | 0.01 | Buried |
| 59 | T | T | 36 | 0.85 | Surface | | 131 | D | D | 110 | 0.5 | Surface | | 203 | S | S | 180 | 0.03 | Buried |
| 60 | K | K | 37 | 0.82 | Surface | | 132 | G | G | 111 | 0.54 | Surface | | 204 | L | L | 181 | 0 | Buried |
| 61 | A | A | 38 | 0.05 | Buried | | 133 | E | E | 112 | 0.36 | Surface | | 205 | L | L | 182 | 0.3 | Surface |
| 62 | K | K | 39 | 0.39 | Surface | | 134 | H | H | 113 | 0.38 | Surface | | 206 | D | D | 183 | 0.28 | Surface |
| 63 | V | V | 40 | 0.64 | Surface | | 135 | F | F | 114 | 0.15 | Buried | | 207 | L | L | 184 | 0.01 | Buried |
| 64 | D | D | 41 | 0.34 | Surface | | 136 | A | A | 115 | 0.06 | Buried | | 208 | L | L | 185 | 0.1 | Buried |
| 65 | K | K | 42 | 0.55 | Surface | | 137 | M | M | 116 | 0 | Buried | | 209 | P | P | 186 | 0.15 | Buried |
| 66 | K | K | 43 | 0.65 | Surface | | 138 | E | E | 117 | 0 | Buried | | 210 | K | K | 187 | 0.44 | Surface |
| 67 | L | L | 44 | 0.07 | Buried | | 139 | M | M | 118 | 0 | Buried | | 211 | E | E | 187 | 0.44 | Surface |
| 68 | G | G | 45 | 0.43 | Surface | | 140 | H | H | 119 | 0.01 | Buried | | 212 | E | E | 187 | 0.44 | Surface |
| 69 | R | R | 46 | 0.52 | Surface | | 141 | I | I | 120 | 0.01 | Buried | | 213 | K | K | 188 | 0.21 | Surface |
| 70 | F | F | 47 | 0.03 | Buried | | 142 | V | V | 121 | 0.13 | Buried | | 214 | L | L | 189 | 0.27 | Surface |
| 71 | F | F | 48 | 0.54 | Surface | | 143 | H | H | 122 | 0.01 | Buried | | 215 | R | R | 189 | 0.27 | Surface |
| 72 | F | F | 49 | 0.2 | Surface | | 144 | E | E | 123 | 0.38 | Surface | | 216 | H | H | 190 | 0.47 | Surface |

| CA-IV ENS_POS | ENS | PDB | PDB_POS | RSA | LOC | CA-IV ENS_POS | ENS | PDB | PDB_POS | RSA | LOC |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 217 | Y | Y | 191 | 0.01 | Buried | 289 | R | - | - | NA | NA |
| 218 | F | F | 192 | 0.03 | Buried | 290 | P | - | - | NA | NA |
| 219 | R | R | 193 | 0.07 | Buried | 291 | L | - | - | NA | NA |
| 220 | Y | Y | 194 | 0.02 | Buried | 292 | P | - | - | NA | NA |
| 221 | L | L | 195 | 0.35 | Surface | 293 | W | - | - | NA | NA |
| 222 | G | G | 196 | 0.06 | Buried | 294 | A | - | - | NA | NA |
| 223 | S | S | 197 | 0 | Buried | 295 | L | - | - | NA | NA |
| 224 | L | L | 198 | 0.26 | Surface | 296 | P | - | - | NA | NA |
| 225 | T | T | 199 | 0.06 | Buried | 297 | A | - | - | NA | NA |
| 226 | T | T | 200 | 0.27 | Surface | 298 | L | - | - | NA | NA |
| 227 | P | P | 201 | 0.24 | Surface | 299 | L | - | - | NA | NA |
| 228 | T | T | 202 | 0.53 | Surface | 300 | G | - | - | NA | NA |
| 229 | C | C | 203 | 0 | Buried | 301 | P | - | - | NA | NA |
| 230 | D | D | 204 | 0.26 | Surface | 302 | M | - | - | NA | NA |
| 231 | E | E | 205 | 0.26 | Surface | 303 | L | - | - | NA | NA |
| 232 | K | K | 206 | 0.2 | Surface | 304 | A | - | - | NA | NA |
| 233 | V | V | 207 | 0 | Buried | 305 | C | - | - | NA | NA |
| 234 | V | V | 208 | 0.06 | Buried | 306 | L | - | - | NA | NA |
| 235 | W | W | 209 | 0.02 | Buried | 307 | L | - | - | NA | NA |
| 236 | T | T | 210 | 0.01 | Buried | 308 | A | - | - | NA | NA |
| 237 | V | V | 211 | 0 | Buried | 309 | G | - | - | NA | NA |
| 238 | F | F | 212 | 0 | Buried | 310 | F | - | - | NA | NA |
| 239 | R | R | 213 | 0.51 | Surface | 311 | L | - | - | NA | NA |
| 240 | E | E | 214 | 0.41 | Surface | 312 | R | - | - | NA | NA |
| 241 | P | P | 215 | 0.19 | Buried | | | | | | |
| 242 | I | I | 216 | 0.03 | Buried | | | | | | |
| 243 | Q | Q | 217 | 0.51 | Surface | | | | | | |
| 244 | L | L | 218 | 0 | Buried | | | | | | |
| 245 | H | H | 219 | 0.33 | Surface | | | | | | |
| 246 | R | R | 220 | 0.58 | Surface | | | | | | |
| 247 | E | E | 221 | 0.36 | Surface | | | | | | |
| 248 | Q | Q | 222 | 0.02 | Buried | | | | | | |
| 249 | I | I | 223 | 0.08 | Buried | | | | | | |
| 250 | L | L | 224 | 0.16 | Buried | | | | | | |
| 251 | A | A | 225 | 0.06 | Buried | | | | | | |
| 252 | F | F | 226 | 0 | Buried | | | | | | |
| 253 | S | S | 227 | 0.66 | Surface | | | | | | |
| 254 | Q | Q | 227 | 0.66 | Surface | | | | | | |
| 255 | K | K | 228 | 0.29 | Surface | | | | | | |
| 256 | L | L | 229 | 0 | Buried | | | | | | |
| 257 | Y | Y | 230 | 0.28 | Surface | | | | | | |
| 258 | Y | Y | 231 | 0.05 | Buried | | | | | | |
| 259 | D | D | 232 | 0.16 | Buried | | | | | | |
| 260 | K | K | 233 | 0.39 | Surface | | | | | | |
| 261 | E | E | 236 | 0.73 | Surface | | | | | | |
| 262 | Q | Q | 237 | 0.36 | Surface | | | | | | |
| 263 | T | T | 238 | 0.73 | Surface | | | | | | |
| 264 | V | V | 239 | 0.11 | Buried | | | | | | |
| 265 | S | S | 240 | 0.37 | Surface | | | | | | |
| 266 | M | M | 241 | 0 | Buried | | | | | | |
| 267 | K | K | 242 | 0.28 | Surface | | | | | | |
| 268 | D | D | 243 | 0.13 | Buried | | | | | | |
| 269 | N | N | 244 | 0 | Buried | | | | | | |
| 270 | V | V | 245 | 0.23 | Surface | | | | | | |
| 271 | R | R | 246 | 0.02 | Buried | | | | | | |
| 272 | P | P | 247 | 0.46 | Surface | | | | | | |
| 273 | L | L | 248 | 0.4 | Surface | | | | | | |
| 274 | Q | Q | 249 | 0.24 | Surface | | | | | | |
| 275 | Q | Q | 250 | 0.81 | Surface | | | | | | |
| 276 | L | L | 251 | 0.34 | Surface | | | | | | |
| 277 | G | G | 252 | 0.6 | Surface | | | | | | |
| 278 | Q | Q | 253 | 0.94 | Surface | | | | | | |
| 279 | R | R | 254 | 0.06 | Buried | | | | | | |
| 280 | T | T | 255 | 0.75 | Surface | | | | | | |
| 281 | V | V | 256 | 0.08 | Buried | | | | | | |
| 282 | I | I | 257 | 0.25 | Surface | | | | | | |
| 283 | K | K | 258 | 0.25 | Surface | | | | | | |
| 284 | S | S | 259 | 0.27 | Surface | | | | | | |
| 285 | G | - | - | NA | NA | | | | | | |
| 286 | A | - | - | NA | NA | | | | | | |
| 287 | P | - | - | NA | NA | | | | | | |
| 288 | G | - | - | NA | NA | | | | | | |

**Table 7 - Alignment of Ensembl entry ENSG00000167434 (Flicek, Amode, Barrell, & al., 2012) CA-IV protein sequence (ENS) with PDB entry 3FW3 (Bernstein, et al., 1977) (Vernier, et al., 2010) chain B protein sequence (PDB). Associated RSA values as determined by DSSP (Kabsch & Sander, 1983). The alignment allows for the association of Ensembl sequence amino acid position Ka/Ks values to be translated to PDB amino acid positions and therefore mapped onto the PDB structure.**