UNIVERSITY OF TURKU

Large-scale integration of sequence and structure data in the study of carbonic anhydrase enzymes

Digital Health

Master's Degree Programme in Information and Communication Technology

Department of Computing, Faculty of Technology

Master of Science in Technology Thesis

Author:

Hosein Daneshpour

Supervisor:

Dr. Martti Tolvanen (Department of Computing)

December 2023

Master of Science in Technology Thesis
Department of Computing, Faculty of Technology
University of Turku

Subject: Digital Health
Programme: Master's Degree Programme in Information and Communication Technology
Author: Hosein Daneshpour
Title: Large-scale integration of sequence and structure data in the study of carbonic anhydrase enzymes
Number of pages: 64 pages, 39 appendix pages
Date: December 2023

In this discovery of carbonic anhydrases, several computational techniques have been used to examine different facets of these enzymes. Methods produced in this work aid for instance in the evaluation of the accuracy of Multiple Sequence Alignments and AlphaFold models, the retrieval of combinations of amino acids from targeted areas of Multiple Sequence Alignments, the evaluation of disulfides and metal-binding residues, the programmatic assessments of molecular geometry, and the production of molecular pictures. Therefore, by utilizing these data mining methods, we have learnt more about the composition, properties, and interactions of carbonic anhydrases, which will help us comprehend these enzymes and their function in biological systems better. Most importantly, the catalytic center and mechanism of the thus far poorly characterized delta carbonic anhydrases has been reliably established. Accordingly, this research offers novelty in several aspects such as combining large-scale data from AlphaFold models, analyzing the reproducibility of functional aspects in AlphaFold models, and formulating theories for the catalytic mechanism of a nearly unknown protein family using features of sequence conservation and structural evaluation.

Keywords: Bioinformatics, Data science, Digital health, Carbonic anhydrase, AlphaFold

Table of Contents

1 Introduction

In recent years the studies of carbonic anhydrases (CAs, EC 4.2.1.1) have significantly increased. Basically, these metalloenzymes, known as CAs, have a critical function in the process of synthesis of carbon dioxide and water to carbonic acid, which allows them to control the levels of CO2, HCO3-, and H+ (Campestre et al., 2021). Simply speaking, they directly affect our life and many biological functions of our body, such as the production of acid in the stomach lining, or controlling the water content of the cell. CAs are classified into six genetically distinct groups α, β, γ, δ, ζ and η. While α, β, and γ types of CA have different structures, and η has recently been discovered in Plasmodium (Supuran et al., 2014).
There is no considerable resemblance between the amino acid sequence of these CA families (Scozzafava et al., 2006). α-CA can be found in, algae, vertebrates and several bacteria. In mammals, 16 distinct α-CA isoforms exist (Supuran, 2008). The beta type is comprised of the majority of bacterial and plant chloroplast, as well (Sawaya, et al., 2006). Methanogens are the source of the gamma class, and also the delta type has been found in diatoms (Zimmerman and Ferry, 2008). Meanwhile, only a few chemolithotrophs and marine organisms have the zeta class, nevertheless, according to research, ζ-CA and β-CA have certain structural similarities (Sawaya et al., 2006). Since several of these CAs are drug discovery targets, there is a great opportunity to create medication compounds with novel mechanisms. Table 1 provides an overview of eight different classes of CA families.

Table 1. Different kinds of carbonic anhydrases (Jensen et al., 2020)

CA Class	Metal Cofactor	Organism(s)
α-CA	Zn^{2+}	Mammals, plants, algae, prokaryotes
β-CA	Zn^{2+}	Plants, algae, bacteria
γ-CA	Zn^{2+}, Fe^{2+}, Co^{2+}	Prokaryotes, plants, fungi, algae
δ-CA	Zn^{2+}, Co^{2+}	Marine phytoplankton
ζ-CA	Cd^{2+}, Zn^{2+}	Diatoms
η-CA	Zn^{2+}	*Plasmodium* sp
θ-CA	Zn^{2+}	Diatoms, green algae
ι-CA	Mn^{2+}	Marine phytoplankton

Delta CAs are an important part of the CO_2 concentrating mechanisms and thus photosynthesis. Because the group of diatoms accounts for approx. 20% of photosynthesis globally, this is also important for CO_2 fixation and climate change. Figure 1 shows a schematic view of different carbonic anhydrases.

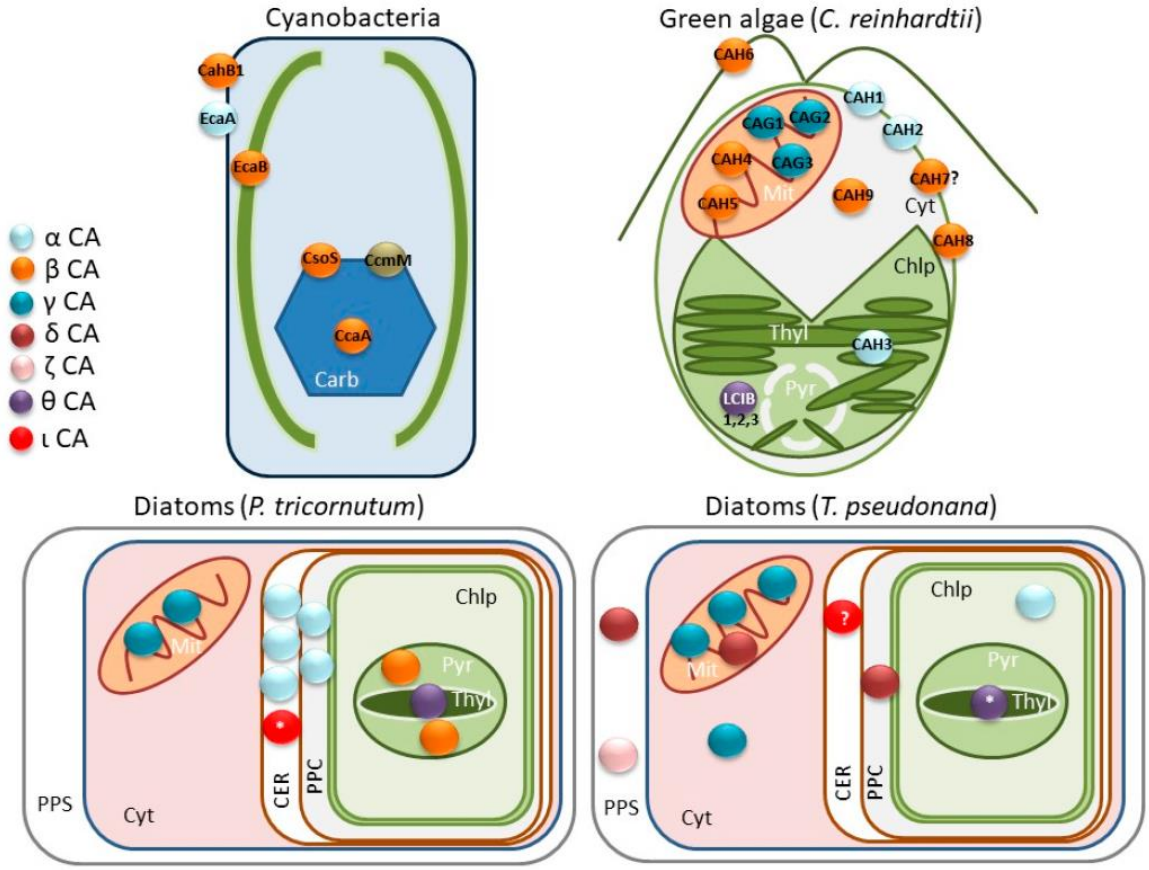


Figure 1. Different CA classes in a microalga (Jensen et al., 2020)

Accordingly, Jensen et al. (2020) argue that more study is required to fully comprehend the molecular mechanisms of CAs and their function in many physiological processes. The possibility of using CAs as biotechnological instruments for CO_2 fixation and the generation of biofuel is also suggested.

Although synthetic activity and natural sources create CO_2, cellular respiration is one of the main sources of CO_2 in nature. The main source of carbon for life on earth is then absorbed from CO_2. Through the process of photosynthesis, which utilises sunlight as an energy source, CO_2 as a source of carbon, and H_2O as a supply of electrons for the transmission chain for photosynthetic electrons, algae, plants, and cyanobacteria create oxygen and carbohydrates. Aerobic organisms produce CO_2, use oxygen for cellular metabolism, and then release CO_2

back into the atmosphere. As a result, aerobic organisms create large levels of CO_2, which need to be controlled at the cellular level. When CO_2 and H_2O combine, an unstable substance called carbonic acid (H_2CO_3) is created. This combination naturally divides into bicarbonate (HCO_3) and a proton (H+). Although the generation of H_2CO_3 is an efficient reaction at high pH, it is a weak process at neutral pH, which is frequently seen in most biological systems in all tissues and species. In this case, the hydration of CO_2 is catalyzed by a class of metalloenzymes called carbonic anhydrases (CAs) or carbonate dehydratases, which are found in all species. Based on variations in highly conserved amino acid sequences in their catalytic active sites, eight different evolutionary families of CAs have been found to date. The names of these enzyme are α-, β-, γ-, δ-, ζ-, η-, θ-, and ι-CAs. Algae, bacteria, vertebrates, protozoa, plants, and corals all contain α-CAs. Meanwhile, γ-CAs have been found in plants, archaea, and bacteria. Only a few marine diatoms have been found to have δ-CAs and ζ-CAs. η-CA has been determined at the malarial parasite Plasmodium sp. The marine diatom *Phaeodactylum tricornutum* has been shown to have θ-CA, while *Burkholderia territorii*, a Gram-negative bacteria, contains ι-CA. Although the major function of CA families is the reversible catalysis or hydration of CO_2 to H_2CO_3 preceding rapid dissociation to H+ and HCO_3, certain members of these enzymes also take part in the catalysis of CS_2 hydration, hydrolysis of a wide range of esters, and cyanamide hydration to urea. Additionally, CAs are essential for various metabolic processes, including bone resorption, calcification, ureagenesis, gluconeogenesis, and tumorigenicity in animals. Additionally, by acting a crucial part in the CO_2 concentration mechanism (CCM) and CO_2 fixation, they support photosynthesis in algae, green plants, and cyanobacteria, strengthening the potential of pathogenesis in parasites. The provision of CO_2 for photosynthesis closely depends on CCM and CO_2 fixation, particularly in aquatic settings where CO_2 levels are low and when phytoplankton is the dominant ecosystem. These significant photosynthetic bacteria are in charge of producing over half of the biosphere's initial production. As a result of the presence of metal ions as cofactors in their catalytic active site, carbonic anhydrase classes are categorized as metalloenzymes. All eight CA groups include Zn(II), and some families allow for the interchange of Zn(II) with other metal ions. Some α- and δ- CAs, may have Co(II) ions in place of Zn(II) in their catalytic active sites, while Fe(II) is probably present in γ-CAs under anaerobic conditions. Additionally, some ζ-CAs (cambialistic enzymes) may be activated with either Zn(II) or Cd(II) depending on environmental concentrations. A unique CA class that might activate lacking a metal ion in its active site has also just been discovered. Three highly conserved amino acid residues form up the catalytic core of CAs, and the

coordination sphere is often completed by a water molecule or hydroxide ion. Three His residues in the α-, γ-, and δ-CAs coordinate with the metal ion in the active site, whereas in the β- and ζ-CAs, one His and two Cys residues do similarly. The wide distribution of β-CA in several biology kingdoms, except in vertebrates, has been demonstrated by prior data showing evolutionary relationships between different CA families, as well as by similarities between coordination sites of amino acid residues in β- and ζ-CAs. However, diatoms, which contain ζ-CA, are not frequently described. These results prompted us to conduct the current research to determine the distribution of ζ-CA in other eukaryotic and prokaryotic microorganisms, investigate the evolutionary linkage between β- and ζ-CAs, and determine the likelihood that these two CA families would coexist in a microorganism utilizing in silico methods (Launay et al., 2020; Roberts et al., 1997; Shalileh et al., 2023). Figure 2 depicts an overview of the catalyst core of eight CAs.

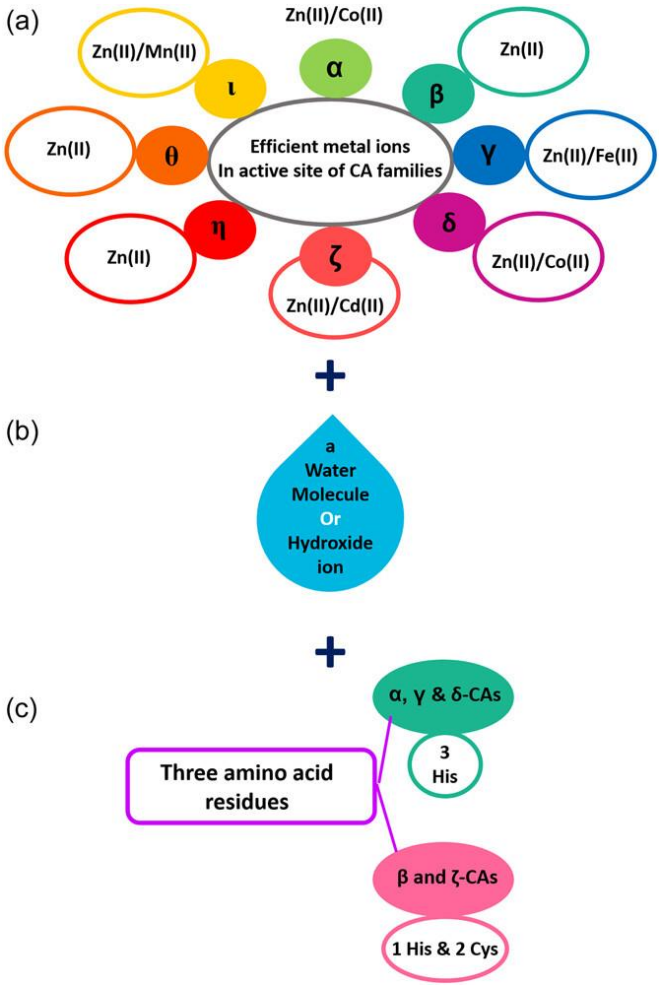


Figure 2. eight CA families (Shalileh, et al., 2023)

1.1 AlphaFold

AlphaFold is a novel artificial intelligence software that can predict 3D protein shapes and was created by DeepMind. Computationally, AlphaFold depends on a deep learning method, by the use of amino acid sequences. The initial release of AlphaFold was in the year 2018, which demonstrated its outstanding precision in predicting the protein structure. Given that this algorithm performs significantly better than the other rivals, it is named a “game-changer” (Callaway, 2020). Subsequently, the system has undergone updates and enhancements and is employed to forecast the structures of several proteins, such as disorders like Parkinson's. Furthermore, it lets us avoid using costly experimentation techniques like cryo-EM or X-ray crystallography. Accordingly, DeepMind and EMBL's European Bioinformatics Institute (EMBL-EBI) joined together to establish the AlphaFold database and offer the results of the predictions publicly accessible to scientists. The current version offers a comprehensive analysis of UniProt entries with more than 200 million data (alphafold, 2023).

The Critical Assessment of Structure Prediction (CASP) is a global competition that happens every two years for these structure predictions. Started in 1994 it is called the Olympics of protein folding challenge. Basically, CASP aims to assess the latest developments in protein structure prediction and pinpoint the gaps that require advancement. Almost 100 organizations worldwide presented more than 67,000 models in the most recent CASP round, CASP14. In 2020 AlphaFold2 overtook the former by presenting a much better performance. According to the Forbes “AlphaFold Is The Most Important Achievement In AI” (www.forbes.com, 2021).

2 Motivation of the thesis

Carbonic anhydrases are enzymes which are everywhere, because of the simple reaction which is needed for many things. Getting rid of carbon dioxide, retaining carbon dioxide for photosynthesis, pH regulation, etc. The reaction is so important that the “wheel has been reinvented”. We have CA activity in at least four completely different protein types (alpha, beta, gamma and iota CAs) and then some of these have been transformed during evolution so much that our sequence-based search tools (such as Blast) cannot recognize them. For example, the alphas are relatively close to eta CA (but still they don’t find each other in Blast) and very far from deltas. One purpose of this thesis is to show all the details of the structural similarities and differences between the alpha and delta families. The similarities are evidence of a common ancestor far back in evolution, and the differences are useful basic data for drug development. Another example of clearly similar protein structure but undetectable sequence similarity comes from the beta-like class, which contains beta, zeta and theta CAs.

This thesis also has methodological novelty. Until now no research has made comparisons of a large number of AlphaFold models to generalize structural features that are essential for a family of proteins. To use such extracted sets of structural features in comparisons between two classes of proteins is even more innovative.

Moreover, the present research explicitly demonstrates that the previous theories regarding the active site in DCAs have been incorrect, as my work shows that the active sites of ACAs and DCAs are structurally highly similar, despite quite different protein structures surrounding the centers, and indetectable levels of overall sequence similarity. This finding is notable as it suggests that the structural similarities between ACAs and DCAs may be indicative of the divergent evolution of a conserved catalytic centre.

3 Methods

Methodologically, in this research, a wide variety of tools and techniques have been implemented. In the review section, I have investigated the background and multidisciplinary nature of bioinformatics. Followings I have explained the main methods used in this research.

3.1 Data mining

In this research, I have implemented different data mining methods such as outliers' detection, different aggregation approaches, and normalizations (García et al., 2015). For example, in the model generation phase in section 5.6, for handling the MSA and pLDDT min-max scalier has been utilized to normalize the B factors values. In this method, the data is normalized using the min-max scaling, which scales the data to a predetermined range of values, often between 0 and 1, while each data point is subtracted from the minimum value and divided by the data range. This method was preferred to the other scaling techniques especially as MSA weights are in range 0 - 1, as well. Figure 3 depicts the large-scale data science and knowledge discovery workflow implemented in this research.

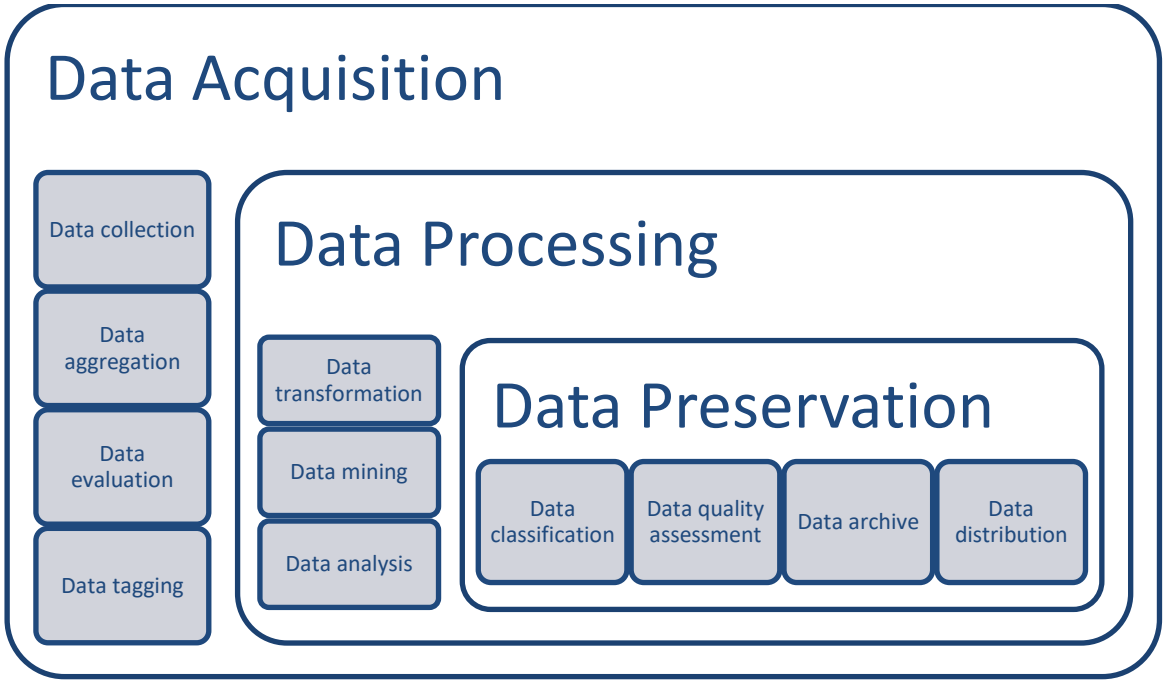


Figure 3. The process of bioinformatics large-scale data (adopted from Cullen & Garcia, 2021)

3.2 Biopython

Biopython programming language is a collection of open-source tools for biological analysis. It consists of a variety of modules that can analyze protein structures, sequencing data, population genetics, phylogenetic trees, and other things. Some of Biopython's significant advantages consist of parsers that support multiple bioinformatics file formats such as BLAST, FASTA, and Genbank; access to online databases such as NCBI, Expasy, and UniProt; interfaces to well-known applications such as Clustalw, DSSP, MSMS. Considering some of its minor weaknesses, as I experienced in this research as well, can be minor API bugs in mapping with UniProt, which perhaps will be fixed in future revisions (Hang et al., 2023).

The analysis in this study benefitted from the following key Python and Biopython libraries. The NumPy, which is an open-source project, was used to perform mathematical operations on arrays. Pandas for the data structuring and analysis. Matplotlib and Seaborn as some of the most popular libraries, which extensively help in the visualization process. Also, the Bio is the most comprehensive and popular Python bioinformatics package, and it includes several distinct sub-modules for frequently performed bioinformatics operations, including sequence analysis, phylogenetics, alignment, and database access, such as the AlignIO module to read and write MSA in various formats, such as Clustal, FASTA, PHYLIP. To cope with the protein 3D model, we used a module called PDBParser is included in Biopython's Bio.PDB sub-package. It enables us to extract data from Protein Data Bank (PDB) files and translate that data into a structure object. Thus, it allows us to access and study the protein structures and other features.

3.3 Multiple sequence alignment

In bioinformatics, the method of multiple sequence alignment (MSA) is implemented to find similar trends and evolutionary links among genes such as DNA, RNA, or protein sequences. Several analytical algorithms are used to produce and assess alignments to enable us in the inference of homology and the determination of the evolutionary linkages between the sequences. Clustal Omega, MAFFT, and WebPRANK are some of the tools for this purpose. The Uniport website dashboard provides an online tool for performing alignment with Clustal Omega, as well.

3.4 SeaView

Is a visualisation application for handling multiple sequence alignments. Technically, it can handle MSA files including protein, DNA sequences, and phylogenetic trees in different formats such as NEXUS, CLUSTAL, and FASTA (Gouy et al., 2010). Figure 4 presents a sample view of the MSA of 60 DCA alignments in the SeaView software.

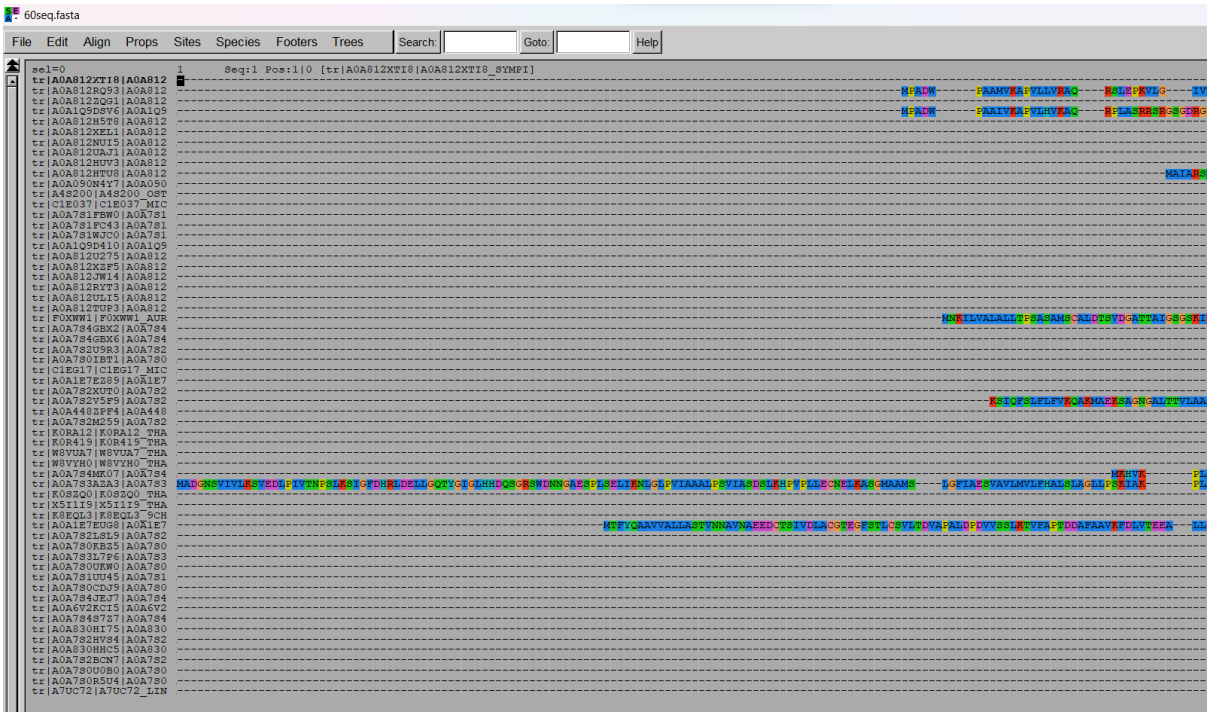


Figure 4. a Schematic view of Seaview with 60 Seq

3.5 Web logo

Web-logo is an online tool which can produce a graphical picture of sequence logos of amino acid or nucleic acid, by reading the MSA files. Accordingly, the sequence conservation can be seen as the total height of the stack, whereas the size of the symbols inside the stack indicates the relative abundance of each residue. Its key advantage in the bioinformatics analysis can be mentioned as being user-friendly-orientated (Crooks et al., 2004).

Figure 5 displays an alignment of several amino acid sequences. The letters are coloured in accordance with their chemical characteristics, and each row in the columns represents a distinct sequence. Acidic amino acids are represented by red, basic amino acids by blue, and hydrophobic amino acids by green. Also, the picture contains numbers identifying the positions of the amino acids in the sequence.

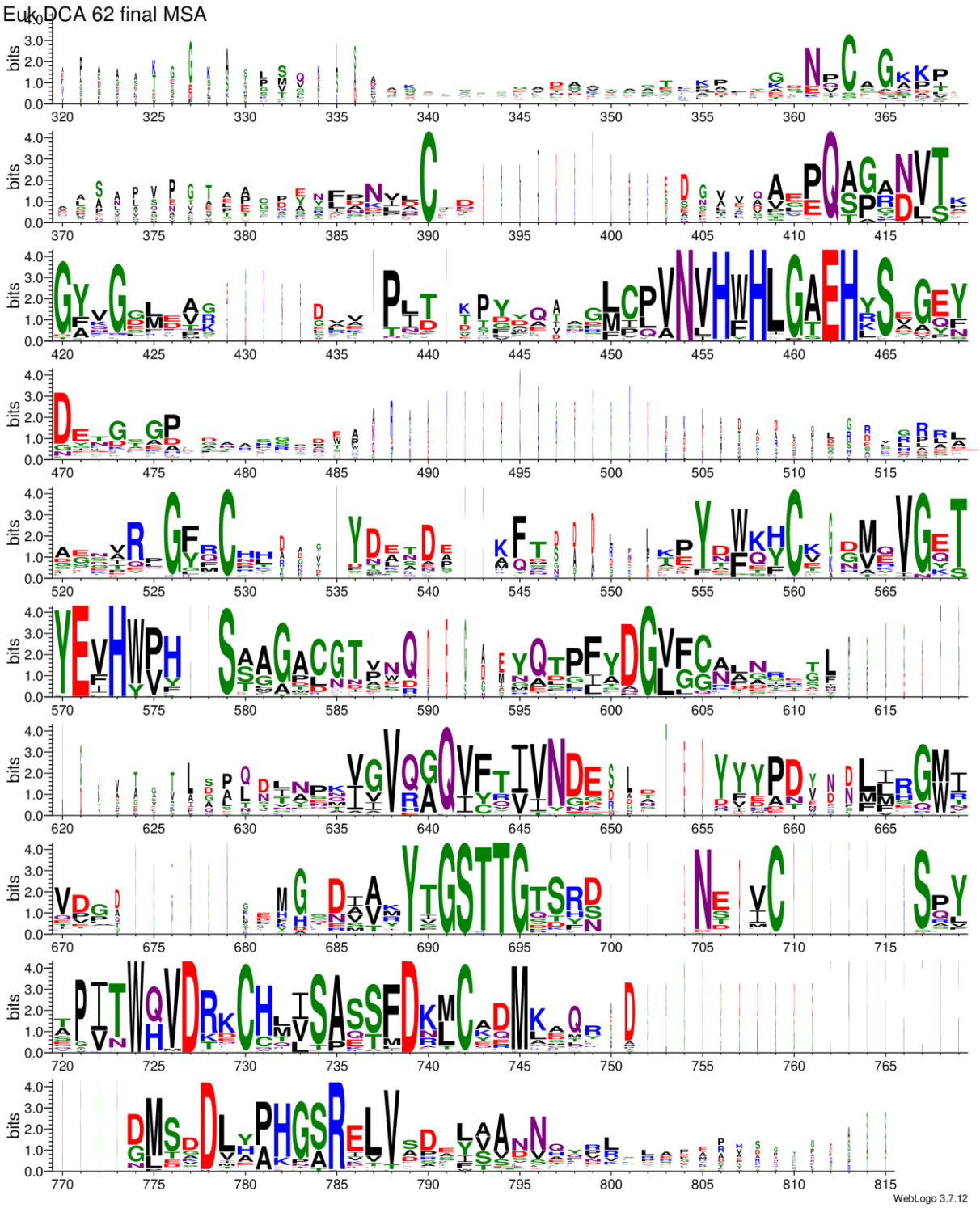


Figure 5. Weblogo of Delta CA consist of 62 sequences

3.6 UniProt

Uniport is one of the main sources we have utilized in this project, and basically for any bioinformatics research study. UniProt is an extensive repository of data about protein sequences and annotations, and it comprises the UniProt Knowledgebase (UniProtKB) as a primary location for gathering data on proteins' functions, the UniProt Archive (UniParc), and the UniProt Reference Clusters (UniRef).

3.7 RCSB Protein Data Bank

The RCSB Protein Data Bank (PDB) is a collection of proteins, nucleic acids, and other biological molecules that have had their 3D structures experimentally established. It is the most extensive repository of structural biology information available and is open to the general public, as well. The National Institutes of Health (NIH) and the Protein Data Bank in Europe (PDBe) founded the PDB in 1971. The Research Collaboratory for Structural Bioinformatics (RCSB), a nonprofit organization that also manages the RCSB PDB website, is currently in charge of its administration. Over 200,000 structures can be found in the PDB, which is continually expanding. Figure 6 presents the growing trend until now.

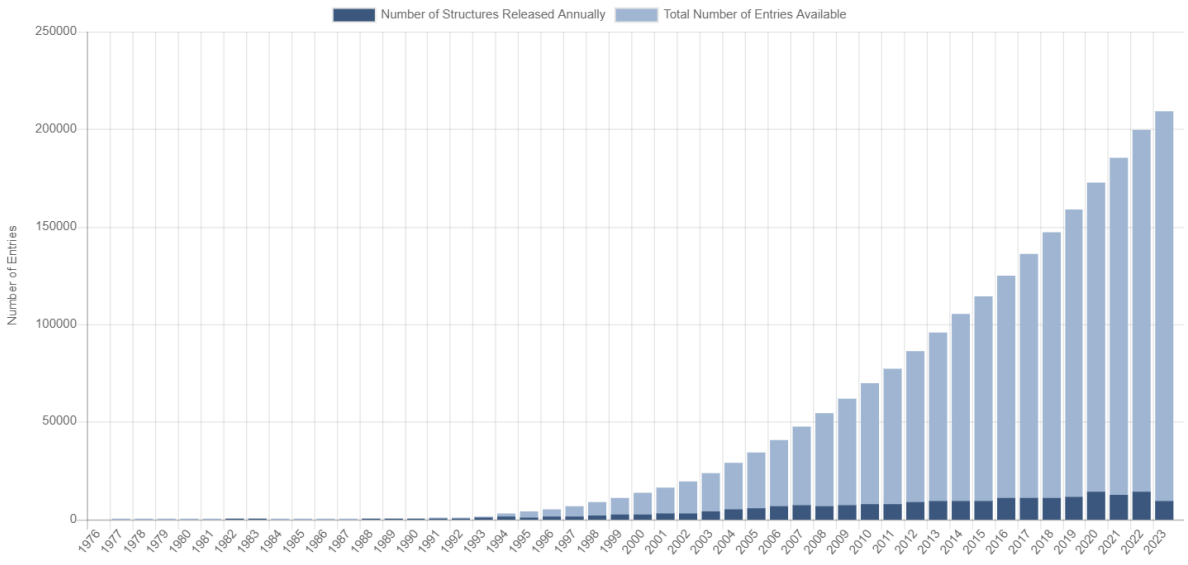


Figure 6. Overall statistics of released structures per year in PDB (PDB, 2023)

Technically, the RCSB verifies the structures before they are included in the database after they are submitted by researchers from all over the world. Research in a wide range of disciplines, including biochemistry, biophysics, molecular biology, and medicine, can benefit

greatly from the PDB. For the study of structural biology, the RCSB PDB is an essential tool. PDB can be employed to research the composition, behaviour, and utilization of proteins and other biological components, and consequently, new medications and treatments can be also designed and created using the PDB. Several tools and resources are available on the RCSB PDB website for accessing and studying PDB data. Generally, this resource consists of a search engine, a visualization tool, and an annotations database. Several instructional tools, such as tutorials and webinars, are also available on the website. Some of the great advantages offered by the RCSB Protein Data Bank can be summarised as: It has an extensive source of information about structural biology and the public has free access to it. Before they are uploaded to the database, the structures are checked by experts, as well. PDB data may be accessed and analyzed using several tools and resources available on the website, and there are several instructional tools available on the website.

Therefore, the current study on carbonic anhydrase enzymes can benefit from the RCSB Protein Data Bank in a variety of ways. Basically, we may look up the structures of carbonic anhydrase enzymes in the PDB, as extensive carbonic anhydrase enzyme structures from various species are presently available in the PDB. This provides us with large-scale data and enables us to compare and evaluate the structural features of several carbonic anhydrase enzymes. So, we may utilize the PDB to see how carbonic anhydrase enzyme structures are represented. A 3D viewer and a molecular surface viewer are some of the tools for viewing PDB data that are available on the PDB website. This enables us to rigorously visualize the enzyme's structure and recognize crucial elements like the active site. Tools for interpreting PDB data are available on the PDB website, including ones for detecting the conserved residues. We can learn more about the enzyme's role and its interactions with its surroundings as a result. Meanwhile, we may find a range of resources on the PDB website that might assist you in learning more about these enzymes, including literature citations and annotations.

3.8 EMBL-EBI

The European Bioinformatics Institute (EMBL-EBI) in Cambridge, United Kingdom is a research centre that offers a wide range of bioinformatics tools and services. The EMBL-EBI aids in the study of carbonic anhydrase enzymes in many ways, such as: Access to the RCSB Protein Data Bank, which has the structural information on enzymes. Protein sequence analysis tools like BLAST and Clustal Omega. Meanwhile, PyMOL and Chimera are two examples of the visualization tools offered by EMBL-EBI. Online courses and tutorials are

among their educational materials. In addition, the BioServices Platform, for example, offers a platform for academia to cooperate on bioinformatics initiatives.

3.9 Chimera

The "UCSF Chimera" software is used to interactively visualize and analyze molecular structures and associated data, such as density maps, trajectories, and sequence alignments. It is free to use for non-commercial purposes. Users are highly advised to try "ChimeraX", which is actively being developed and gives superior performance with huge structures and extra capabilities.

For example, Figure 7 depicts a CA model using the Chimera software. It can be recognized that the zinc cofactor is coordinated by water molecules and histidine residues (94, 96, and 119).

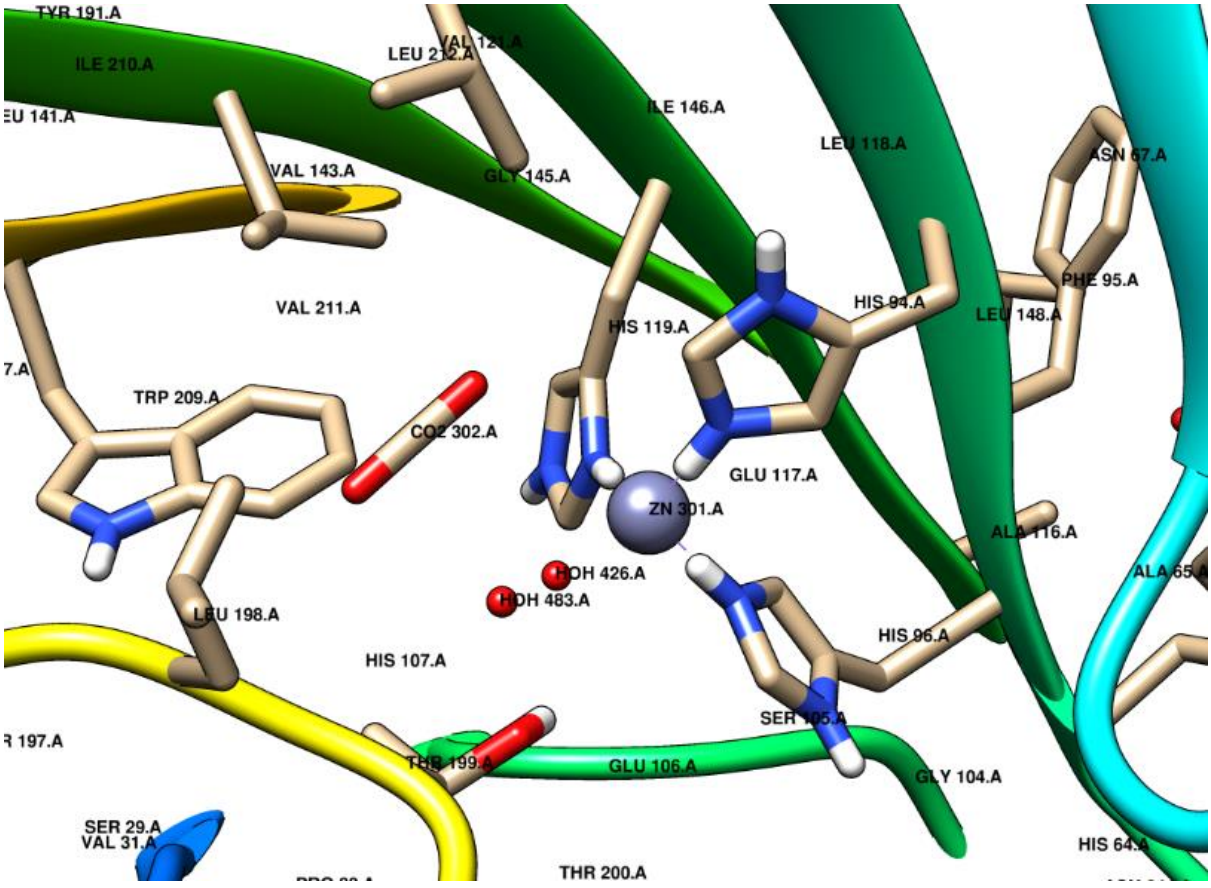


Figure 7. Human carbonic anhydrase II active site of PDB ID: 6LUX (Source: Author)

3.10 Percentage Identity Matrix

The percentage identity between each pair of sequences in an MSA can be displayed in a table called a percentage identity matrix (PIM). The PIM is determined by dividing the total number of aligned residues by the number of identical residues between two sequences and multiplying the result by 100. To display the level of similarity between each pair of sequences, the PIM is coloured while the degree of identification is greater with the darker colours. For instance, the sequence in the first row and column is the same as itself, hence the value in that cell is 100%. Figure 8 shows the sequence set of eukaryotic delta CAs containing 64 sequences. For example, the sequence A0A812XTI8 is 100% identical to itself, and 47.18% identical to A0A812RQ93, and so on. Thus, to find closely similar sequences in an MSA, we can utilize the PIM. The sequences with the highest levels of identity are more likely to be connected. The PIM may also be used to recognize conserved areas of sequences, which are probably crucial from a functional perspective.

A0A812XTI8|A0A812XTI8_SYMPI	100	47,18	48,8	48,97	48,97	44,01	42,6	41,59	42,27	42,27	41,56	41,14	27,4	30,27	33,46	32,35	32,35	32,77	3
A0A812RQ93|A0A812RQ93_9DINO	47,18	100	91,64	86,04	86,04	46,48	47,1	46,65	45,71	45,71	48,21	43,52	28,53	32,57	37,64	31,16	31,75	30,86	3
A0A812ZQG1|A0A812ZQG1_9DINO	48,8	91,64	100	92,24	92,24	46,83	47,1	46,01	45,08	45,08	47,23	44,75	29,62	32,18	37,27	35	35,67	35,4	3
A0A1Q9DSV6|A0A1Q9DSV6_SYMMI	48,97	86,04	92,24	100	100	47,18	46,57	45,22	44,62	44,62	47,23	42,64	27,94	31,42	36,53	30,29	30,88	29,97	3
A0A812LI67|A0A812LI67_9DINO	48,97	86,04	92,24	100	100	47,18	46,57	45,22	44,62	44,62	47,23	42,64	27,94	31,42	36,53	30,29	30,88	29,97	3
A0A812H5T8|A0A812H5T8_9DINO	44,01	46,48	46,83	47,18	47,18	100	47,39	46,44	46,6	46,6	46,42	45,39	31,44	32,56	36,16	36,8	36,43	37,89	3
A0A812XEL1|A0A812XEL1_9DINO	42,6	47,1	47,1	46,57	46,57	47,39	100	77	73,49	73,49	63,36	61,28	26,57	31,87	36,33	32,45	32,83	31,39	3
A0A812NUI5|A0A812NUI5_9DINO	41,59	46,65	46,01	45,22	45,22	46,44	77	100	88,76	88,76	70,21	65,67	27,09	32,73	36,67	32,89	33,22	32,89	3
A0A812UAJ1|A0A812UAJ1_SYMMI	42,27	45,71	45,08	44,62	44,62	46,6	73,49	88,76	100	100	69,28	65,38	27,24	33,82	37,69	33,77	33,77	33,77	3
A0A812VVC6|A0A812VVC6_9DINO	42,27	45,71	45,08	44,62	44,62	46,6	73,49	88,76	100	100	69,28	65,38	27,24	33,82	37,69	33,77	33,77	33,77	3
A0A812HUV3|A0A812HUV3_9DINO	41,56	48,21	47,23	47,23	47,23	46,42	63,36	70,21	69,28	69,28	100	73,89	30,61	33,21	36,67	35,33	35,67	35,03	3
A0A812HTU8|A0A812HTU8_9DINO	41,14	43,52	44,75	42,64	42,64	45,39	61,28	65,67	65,38	65,38	73,89	100	28,3	31,9	34,44	33,44	33,12	34,39	3
A0A090N4Y7|A0A090N4Y7_OSTTA	27,4	28,53	29,62	27,94	27,94	31,44	26,57	27,09	27,24	27,24	30,61	28,3	100	45,99	40,24	30,7	30,7	32,11	3
A4S200|A4S200_OSTLU	30,27	32,57	32,18	31,42	31,42	32,56	31,87	32,73	33,82	33,82	33,21	31,9	45,99	100	42,29	38,85	38,49	38,99	3
C1E037|C1E037_MICCC	33,46	37,64	37,27	36,53	36,53	36,16	36,33	36,67	37,69	37,69	36,67	34,44	40,24	42,29	100	45,9	46,27	47,24	4
A0A7S1FBW0|A0A7S1FBW0_NOCSC	32,35	31,16	35	30,29	30,29	36,8	32,45	32,89	33,77	33,77	35,33	33,44	30,7	38,85	45,9	100	97,61	94,68	9
A0A7S1FD63|A0A7S1FD63_NOCSC	32,35	31,75	35,67	30,88	30,88	36,43	32,83	33,22	33,77	33,77	35,67	33,12	30,7	38,49	46,27	97,61	100	92,16	9
A0A7S1FC43|A0A7S1FC43_NOCSC	32,77	30,86	35,4	29,97	29,97	37,89	31,39	32,89	33,77	33,77	35,03	34,39	32,11	38,99	47,24	94,68	92,16	100	9
A0A7S1FD71|A0A7S1FD71_NOCSC	32,26	31,16	35,53	30,29	30,29	36,53	31,75	32,27	32,81	32,81	34,63	32,83	31,66	37,98	46,27	92,47	94,89	97,54	
A0A7S1FDB9|A0A7S1FDB9_NOCSC	32,25	31,44	35,55	30,56	30,56	36,3	32,83	33,22	34,09	34,09	36	33,44	31,31	38,85	46,27	94,88	95,15	97,2	9
F0XWW1|F0XWW1_AURAN	26,07	26,25	29,15	26,61	26,61	30,04	24,82	25,59	26,01	26,01	26,35	20,99	28,25	29,74	32,7	27,64	27,95	27,64	2
A0A7S3ZV04|A0A7S3ZV04_9STRA	22,77	25,51	26,58	25,29	25,29	27,76	25,36	25,62	25,78	25,78	25	19,58	26,4	29,89	29,24	25,85	26,42	27,57	2
A0A8J2WXS8|A0A8J2WXS8_9STRA	24,34	28,14	28,14	27,95	27,95	27,76	25,82	25,83	26,58	26,58	25,5	20,6	28,77	30,22	29,24	28,43	29,1	29,45	2
A0A7S1RP87|A0A7S1RP87_ALECA	30,27	34,87	34,87	34,87	34,87	33,21	34,73	32,98	33,21	33,21	33,69	31,56	32,34	32,84	34,66	35,5	34,73	35,88	3
A0A7S1RQI3|A0A7S1RQI3_ALECA	29,5	34,1	34,1	34,1	34,1	33,58	34,35	32,27	32,5	32,5	32,98	30,9	31,23	31,73	33,47	34,73	33,97	35,11	3
A0A7S1WJC0|A0A7S1WJC0_ALECA	30,27	34,87	34,87	34,87	34,87	33,96	34,73	32,98	33,21	33,21	33,69	31,56	31,97	32,1	34,26	35,5	34,73	35,88	3
A0A812TUB8|A0A812TUB8_9DINO	33,7	33,58	33,96	33,58	33,58	34,93	32,57	30,9	32,06	32,06	32,65	30,57	30,63	30,88	34,77	35,07	34,7	34,73	3

Figure 8. a PIM view

3.11 Protein Structure Database

The website of the AlphaFold Protein Structure Database contains a database of protein structures predicted by the AlphaFold artificial intelligence system created by DeepMind.

With more than 200 million items, the database offers comprehensive coverage of UniProt. To predict a protein's 3D structure from its amino acid sequence, the AlphaFold method uses deep learning methods. It has been demonstrated to attain accuracy competitive with experimental techniques and has presented the power to completely change the way structural biology is studied. The website offers visitors a variety of resources, such as: a tool for discovering protein structure searches; a program for visualizing protein structures; and publications that have utilised the database.
Figure 9 presents the overall view of a model data, retrieved from an ALPHAFOLD MONOMER V2.0 PREDICTION FOR CARBONIC ANHYDRASE (A0A7V5EM53). The protein is a carbonic anhydrase from the Saprospriraceae bacteria. Utilizing the AlphaFold Monomer v2.0 process, the prediction was produced. The file includes details such as the protein's molecule, chain, and source.

3.12 MSA Entropy

Entropy is a notion in science that is generally related to chaos, unpredictability, or uncertainty in an ecosystem. In thermodynamics, it is a measure of the inability of a system's thermal energy to be converted into mechanical work. Entropy is a broad term for the quantity of energy that cannot be used to perform work or the variety of potential configurations that atoms can take in a system (Crooks, 2017).
In the field of bioinformatics, entropy is a unit used to describe the amount of variance or conservation at each place in a sequence alignment. It can be employed to find conserved areas in a sequence alignment and to guess about the evolution of the links between the sequences. A popular entropy measurement in bioinformatics is the Shannon entropy. It is determined by adding the negative products of the residue frequencies and their respective logarithms (Schneider & Stephens, 1990).

```
HEADER                                          01-JUN-22
TITLE     ALPHAFOLD MONOMER V2.0 PREDICTION FOR CARBONIC ANHYDRASE (A0A7V5EM53)
COMPND    MOL_ID: 1;
COMPND   2 MOLECULE: CARBONIC ANHYDRASE;
COMPND   3 CHAIN: A
SOURCE    MOL_ID: 1;
SOURCE   2 ORGANISM_SCIENTIFIC: SAPROSPIRACEAE BACTERIUM;
SOURCE   3 ORGANISM_TAXID: 2202734
REMARK   1
REMARK   1 REFERENCE 1
REMARK   1  AUTH   JOHN JUMPER, RICHARD EVANS, ALEXANDER PRITZEL, TIM GREEN,
REMARK   1  AUTH 2 MICHAEL FIGURNOV, OLAF RONNEBERGER, KATHRYN TUNYASUVUNAKOOL,
REMARK   1  AUTH 3 RUSS BATES, AUGUSTIN ZIDEK, ANNA POTAPENKO, ALEX BRIDGLAND,
REMARK   1  AUTH 4 CLEMENS MEYER, SIMON A A KOHL, ANDREW J BALLARD,
REMARK   1  AUTH 5 ANDREW COWIE, BERNARDINO ROMERA-PAREDES, STANISLAV NIKOLOV,
REMARK   1  AUTH 6 RISHUB JAIN, JONAS ADLER, TREVOR BACK, STIG PETERSEN,
REMARK   1  AUTH 7 DAVID REIMAN, ELLEN CLANCY, MICHAL ZIELINSKI,
REMARK   1  AUTH 8 MARTIN STEINEGGER, MICHALINA PACHOLSKA, TAMAS BERGHAMMER,
REMARK   1  AUTH 9 DAVID SILVER, ORIOL VINYALS, ANDREW W SENIOR,
REMARK   1  AUTH10 KORAY KAVUKCUOGLU, PUSHMEET KOHLI, DEMIS HASSABIS
REMARK   1  TITL   HIGHLY ACCURATE PROTEIN STRUCTURE PREDICTION WITH ALPHAFOLD
REMARK   1  REF    NATURE                        V. 596   583 2021
REMARK   1  REFN                   ISSN 0028-0836
REMARK   1  PMID   34265844
REMARK   1  DOI    10.1038/s41586-021-03819-2
REMARK   1
REMARK   1 DISCLAIMERS
REMARK   1 ALPHAFOLD DATA, COPYRIGHT (2021) DEEPMIND TECHNOLOGIES LIMITED. THE
REMARK   1 INFORMATION PROVIDED IS THEORETICAL MODELLING ONLY AND CAUTION SHOULD
REMARK   1 BE EXERCISED IN ITS USE. IT IS PROVIDED "AS-IS" WITHOUT ANY WARRANTY
REMARK   1 OF ANY KIND, WHETHER EXPRESSED OR IMPLIED. NO WARRANTY IS GIVEN THAT
REMARK   1 USE OF THE INFORMATION SHALL NOT INFRINGE THE RIGHTS OF ANY THIRD
REMARK   1 PARTY. THE INFORMATION IS NOT INTENDED TO BE A SUBSTITUTE FOR
REMARK   1 PROFESSIONAL MEDICAL ADVICE, DIAGNOSIS, OR TREATMENT, AND DOES NOT
REMARK   1 CONSTITUTE MEDICAL OR OTHER PROFESSIONAL ADVICE. IT IS AVAILABLE FOR
REMARK   1 ACADEMIC AND COMMERCIAL PURPOSES, UNDER CC-BY 4.0 LICENCE.
DBREF  XXXX A    1   269  UNP    A0A7V5EM53 A0A7V5EM53_9BACT     1    269
SEQRES   1 A  269  MET PRO PHE THR PHE ALA SER TYR PHE LEU ILE THR LYS
SEQRES   2 A  269  ILE SER ASN MET LYS LYS HIS ILE PHE PRO PHE LEU ALA
SEQRES   3 A  269  MET MET THR LEU LEU LEU ALA SER CYS HIS THR ALA LYS
SEQRES   4 A  269  GLU ALA HIS HIS HIS ALA HIS TRP SER TYR ALA GLY GLU
SEQRES   5 A  269  THR ASP PRO ALA HIS TRP ALA GLU LEU GLU LYS ASP ALA
SEQRES   6 A  269  GLN CYS ASP GLY LYS HIS GLN SER PRO ILE ASN ILE ILE
SEQRES   7 A  269  GLU LYS ASP VAL LYS PRO SER ASN ALA ASN ASN LEU VAL
SEQRES   8 A  269  PHE HIS TYR SER PRO GLN THR LYS ILE LYS ASP ALA VAL
SEQRES   9 A  269  ASN ASN GLY HIS SER ILE GLN PHE ASN PHE ASP GLU GLY
SEQRES  10 A  269  ASN PHE ILE TYR TYR ASN GLY LYS GLU TYR LYS LEU LYS
SEQRES  11 A  269  GLN LEU HIS PHE HIS GLU GLY SER GLU HIS THR VAL ASN
SEQRES  12 A  269  GLY ILE ARG TYR PRO ILE GLU MET HIS LEU VAL HIS VAL
SEQRES  13 A  269  SER ASP ASP GLY GLN ILE ALA VAL VAL GLY VAL PHE GLY
SEQRES  14 A  269  SER GLU GLY THR ASP SER GLN LEU PHE GLU PHE PHE ASP
SEQRES  15 A  269  LYS PHE LEU PRO ILE ALA VAL ASP GLU HIS LYS SER ILE
SEQRES  16 A  269  HIS GLN ALA LEU ASP LEU LYS SER PHE LEU PRO THR ASN
SEQRES  17 A  269  SER ALA TYR TYR SER TYR THR GLY SER LEU THR THR PRO

SEQRES  18 A  269  PRO CYS SER GLU ASN VAL ASN TRP ILE VAL TYR LYS MET
SEQRES  19 A  269  PRO ILE VAL LEU SER VAL ASP GLU VAL GLU GLN ILE ARG
SEQRES  20 A  269  LYS GLU LEU PRO ILE ARG ASN TYR ARG PRO THR GLN PRO
SEQRES  21 A  269  LEU ASN GLY ARG THR VAL TYR GLU ASN
CRYST1    1.000    1.000    1.000  90.00  90.00  90.00 P 1           1
ORIGX1      1.000000  0.000000  0.000000        0.00000
ORIGX2      0.000000  1.000000  0.000000        0.00000
ORIGX3      0.000000  0.000000  1.000000        0.00000
SCALE1      1.000000  0.000000  0.000000        0.00000
SCALE2      0.000000  1.000000  0.000000        0.00000
SCALE3      0.000000  0.000000  1.000000        0.00000
MODEL        1
ATOM      1  N   MET A   1     101.499   7.250 -15.196  1.00 30.80           N
ATOM      2  CA  MET A   1     100.308   8.111 -15.032  1.00 30.80           C
ATOM      3  C   MET A   1     100.037   8.797 -16.362  1.00 30.80           C
ATOM      4  CB  MET A   1     100.544   9.198 -13.965  1.00 30.80           C
ATOM      5  O   MET A   1     100.990   9.335 -16.908  1.00 30.80           O
ATOM      6  CG  MET A   1     100.383   8.713 -12.522  1.00 30.80           C
ATOM      7  SD  MET A   1     100.457  10.092 -11.351  1.00 30.80           S
ATOM      8  CE  MET A   1     100.094   9.251  -9.786  1.00 30.80           C
ATOM      9  N   PRO A   2      98.786   8.861 -16.834  1.00 47.88           N
ATOM     10  CA  PRO A   2      97.843   7.744 -17.035  1.00 47.88           C
ATOM     11  C   PRO A   2      97.448   7.629 -18.534  1.00 47.88           C
```

Figure 9. a sample AFmodel data used in this research

4 Results

4.1 Preprocessing phase

Primarily, for handling our large MSA we needed some codes to filter any sequence records which have a gap at any of the columns of interest. Basically, the code needed to check that none of the given columns have gaps, and if there are gaps, moves to the next record; if ok, add the record to a new array; at the end, the algorithm writes the new MSA into a new file for utilization in the next steps. In this process initially, the alignment that contains the MSA is loaded from a fasta file, and then a list of integers ("columns_of_interest") has been defined, where the indices of the MSA columns in the enquiry are contained in this list of particular significance. Then, a sequence from the MSA without gaps in the relevant columns will be stored in a new list. Accordingly, every sequence in the MSA is iterated over in the for loop. The loop examines each sequence to see whether there is a gap in any of the selected columns, and then it is added to the list if there is no gap in the sequence. Finally, we got a new fasta file by writing the recent list to it.
From the biological point of view, we need to be confident that we are only taking into consideration sequences with the right amino acids in the active site by choosing sequences that do not have any gaps in these columns. The active site of an enzyme is where the reaction occurs, hence this is a significant issue. This procedure can aid in locating areas of the sequences that have changed or been preserved over time. Thus, the evolution and operation of the enzyme may be better understood using this. The amino acids located in the enzyme's active site are essential to its operation. We may gain an idea of the potential modifications in the active site that can still retain the activity of the enzyme by listing all conceivable combinations of amino acids there. This can aid in our understanding of the evolution of carbonic anhydrase enzymes across time and their roles in various species.
Afterwards, I calculated the distances between the second and third metal-binding histidines. To verify all metal-binding triplets, I performed the identification of the amino acids in columns of interest. The distribution of distances can be utilized to examine the carbonic anhydrase enzyme's structural adaptability; while a large variety of distances indicates that the enzyme is more flexible, a small distribution of distances indicates that the enzyme is relatively rigid. Thus, this method can be used for learning about the composition and operation of carbonic anhydrase enzymes. While the identities of the amino acids can offer information about the interactions between the histidines, the distances between the metal-

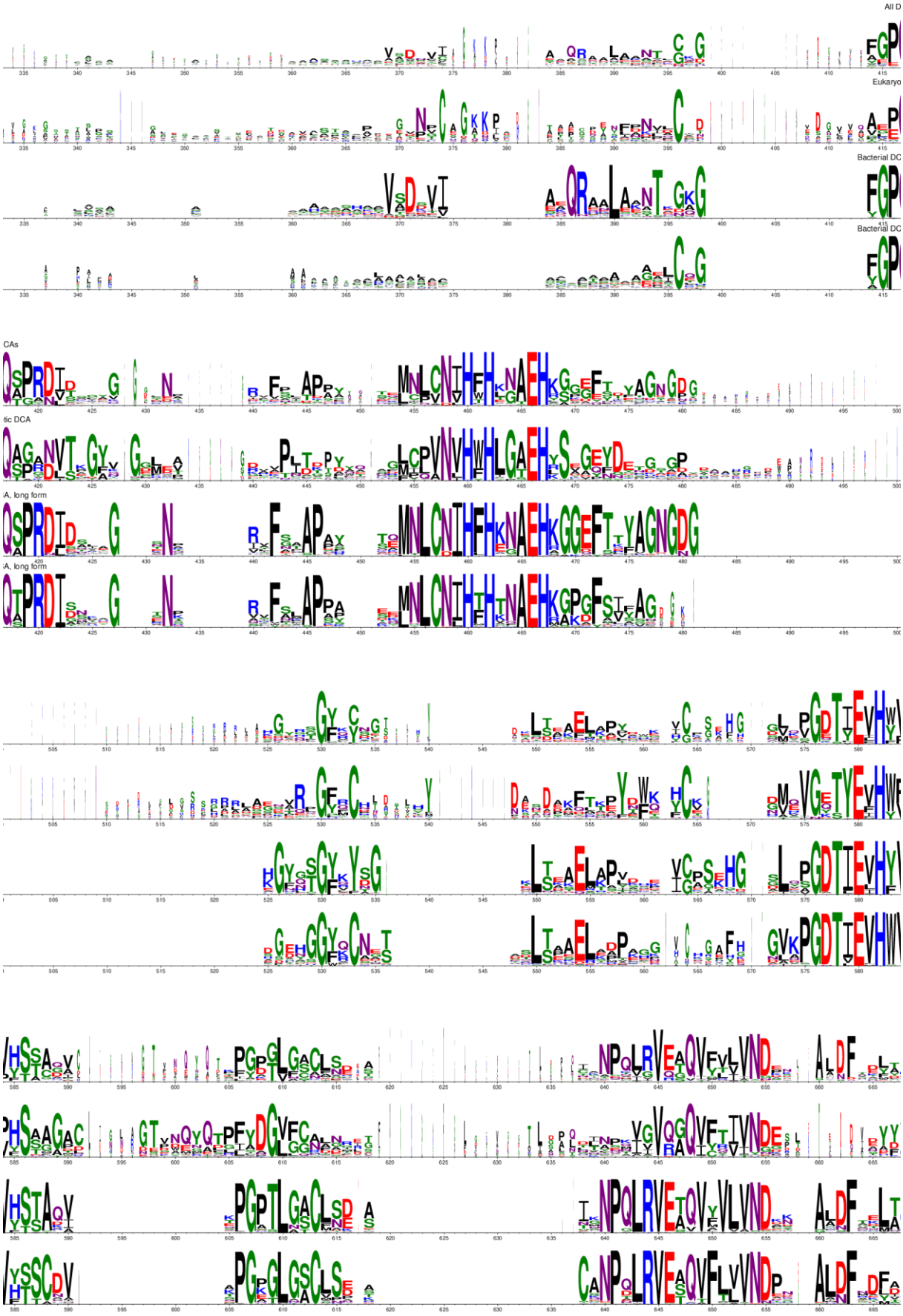
binding histidines can provide information about the conformation of the enzyme. For further readability, the codes for this section have been saved in the GitHub repository through the following link.

- https://github.com/hoseindaneshpour/Bioinformatics-Project---quality-control/blob/main/Bioinformatics%20Project.py

Meanwhile, I have done disulfide measurements. Generally, between two sulfur atoms, there is a powerful covalent link called the disulfide bond. Numerous proteins include it, notably carbonic anhydrase. Protein structure and stability may be studied using the disulfide bond's distance and dihedral angle. Thus, in the coding process, the vector subtraction operation was used to determine the separation between the two sulfur atoms. The vector subtraction determines the distance in space between two points. The two points in this instance represent the sulfur atoms that make up the disulfide bond. The code determines the dihedral angle between the four atoms.
Basically, the structure and stability of proteins depend on the disulfide bond's distance and dihedral angle. The two sulfur atoms should be separated by around 2.0 Å. Also, between 90 and 180 degrees should be the dihedral angle. Thus, the protein may become unstable if the distance or dihedral angle is outside of this range. Meanwhile, some of the main biological considerations are: Disulfide bonds play a crucial role in the stability and folding of proteins and in maintaining the protein's proper structure. Disulfide bonds can also be employed to connect several protein components together. Protein unfolding and denaturation can result from the destruction of disulfide links. Heat, pH fluctuations, or the presence of denaturing chemicals can cause this (Chichiarelli et al., 2022).
The codes for this section have been saved in the GitHub repository through the following link.

- https://github.com/hoseindaneshpour/disulfide-measurements/blob/main/disulfide%20measurements.py

Figure 10 depicts four DCAs as a one-liner logo file (which I have split into six pieces in a row) consisting of all DCAs, Eukaryitic DCAs, and long and short bacteria DCAs, where it is visible that five motifs are conserved.


All D
Eukaryo
Bacterial DC
Bacterial DC
CAs
tic DCA
A, long form
A, long form

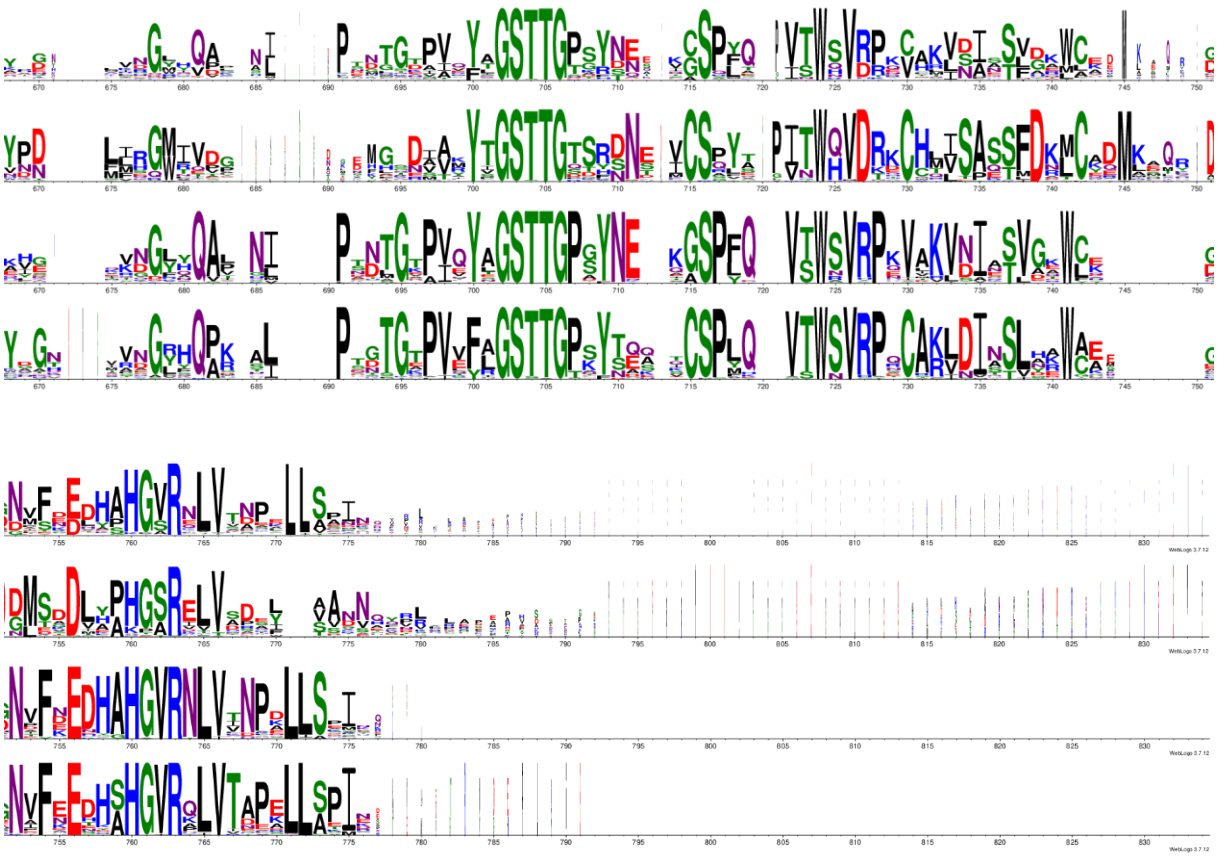


Figure 10. important motifs in DCA

Furthermore, the examination of significant conserved residues and their identities was done, by the same method, for all groups of DCAs and ACAs. A summary has been presented in Table 2, while the complete table of all residues is in Appendix 1.

Table 2. important conserved residues for all groups of DCAs and ACAs

	Metazoan ACA	count	Bacterial DCA short	count	Bacterial DCA long	count	Eukaryotic DCA	count	Bacterial ACA	count
0	[Q, H, H, E, H, T]	46.0	[Q, H, H, E, H, T]	49.0	[Q, H, H, E, H, T]	96.0	[Q, H, H, E, H, T]	59.0	[Q, H, H, E, H, T]	892
1	[Q, R, H, E, Q, T]	4.0	[Q, H, H, E, H, A]	1.0			[Q, R, H, E, H, T]	1.0	[E, H, H, E, H, T]	13
2	[Q, R, H, E, H, T]	3.0							[Q, H, H, E, H, S]	7
3	[Q, R, L, E, Q, S]	2.0							[Q, D, H, E, Q, T]	6
4	[H, H, H, E, H, T]	1.0							[Q, N, H, E, Q, T]	5
...	...	...	...	...	...	...	...	...	...	...

56	[Q, Q, H, E, Y, T]	1
57	[Q, R, K, E, Q, T]	1
58	[Q, S, H, L, V, P]	1
59	[Q, Y, K, E, Q, T]	1
60	[Q, Q, R, E, D, T]	1

As discussed, the active site residues need distances measured as well in figure 11:

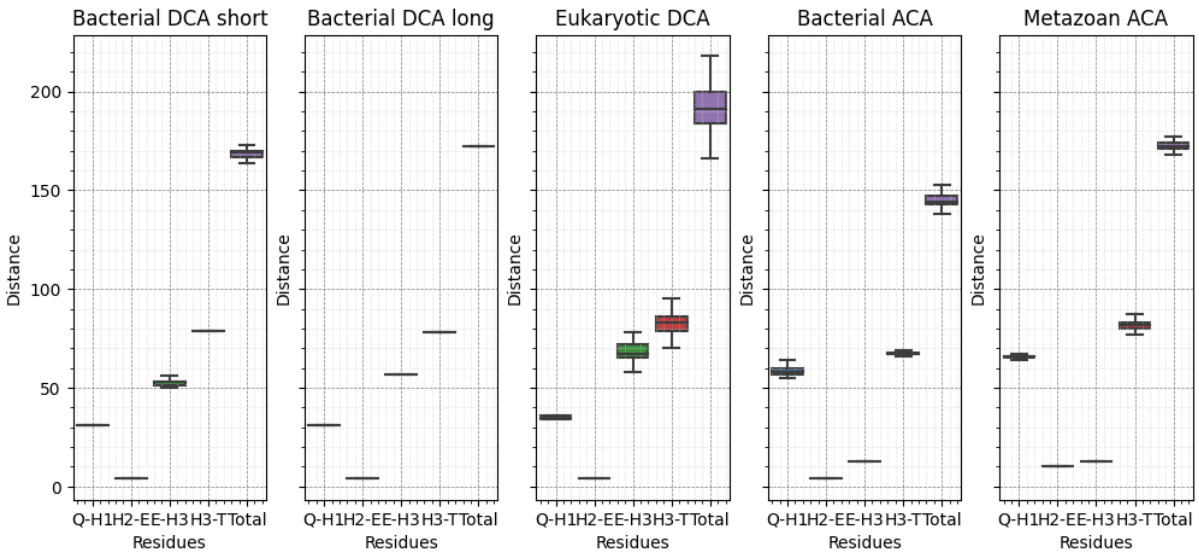


Figure 11. Distances between key residues in five groups of sequences

The GitHub links for figure 10 and 11 are as follows:

- https://github.com/hoseindaneshpour/Distances-between-key-residues-in-five-groups-of-sequences/blob/main/distance%20analysis.py
- https://github.com/hoseindaneshpour/important-conserved-residues-for-all-groups-of-DCAs-and-ACAs/blob/main/important%20conserved%20residues%20for%20all%20groups%20of%20DCAs%20and%20ACAs.py

4.2 Analysis of disulfides

As explained above, it is crucial to analyze disulfides (Cys-Cys pairs) in the context of carbonic anhydrase enzymes. It can aid in our comprehension of the structure of CAs. The

zinc ion is bound to the disulfide link which also aids in stabilizing the CA tetramer's structure. Also, disulfide analysis can benefit our comprehension of how CAs work. Likewise, identifying CA gene mutations linked to illness may be accomplished with the use of disulfide analysis. CAs' structure and function can be disrupted by mutations in the genes that encode them, which can cause some health issues (Campestre et al., 2021). Thus, disulfide analysis is an effective method for examining the composition, mechanism, and historical development of carbonic anhydrase enzymes. Understanding the function of disulfides in CAs will help us better comprehend these vital enzymes and how they affect human health. Accordingly, in this study, we will investigate the analysis of disulfides in the context of carbonic anhydrase enzymes to: design new medications that target CAs, advance our knowledge of CA inhibition mechanisms, and create new techniques for CA production development, and purification.

Subsequently, analyzing the existence of disulfide links in a protein alignment is facilitated by the code I developed. The algorithm is capable of determining whether sequences have particular disulfide bonds or gauging the alignment's overall disulfide bonds. As discussed, the covalent connections between two cysteine residues are known as disulfide bonds. They are widespread in proteins and are crucial for the stability and folding of proteins. Analyzing a protein's amino acid sequence can reveal if disulfide linkages are present. This method is employed by the Python codes to locate sequences that have particular disulfide bonds. Analyzing the protein structure can also reveal how many disulfide linkages are present in a protein.

The overall coding process can be determined as follows. The code outputs the sequence ID and the residues for each sequence in the designated columns. Then the number of sequences that include each residue at each column is determined as well. Then a data frame is created by combining the sequence ID data and residue data. To improve readability, the column names have been changed (C1 to C10). Next, the program verifies that each sequence has a cysteine residue at each of the predetermined columns. Hereby, the sequences of 62 proteins are displayed in a table, along with the locations of 10 conserved residues, while the letter "C" designates the conserved Cys residues. Then, the code gives the cell in the result data frame a value of 1 if a sequence has a cysteine residue at a certain column, and a value of 0 if a sequence at a given column does not contain a cysteine residue. The written code checks to see if the two cysteine residues required for the disulfide bond are both present in the sequence before deciding whether or not it exists. The code assigns a value of 1 to the

associated cell in the DataFrame if both cysteine residues are present and assigns a value of 0 to the relevant cell if either one or both cysteine residues are absent (Figure 12).

This link to the Python codes in GitHub:

- https://github.com/hoseindaneshpour/Analysis-of-disulfides-Cys-Cys-pairs-/blob/main/Thesis%20ideas%20-%20Analysis%20of%20disulfides%20(Cys-Cys%20pairs).py

ID	DS_A	DS_B	DS_C	DS_D	DS_E	Row Sum
A0A812XTI8	0.0	1.0	0.0	1.0	1.0	3.0
A0A812RQ93	0.0	1.0	0.0	1.0	1.0	3.0
A0A812ZQG1	0.0	1.0	0.0	1.0	1.0	3.0
A0A1Q9DSV6	0.0	1.0	0.0	1.0	1.0	3.0
A0A812H5T8	0.0	1.0	0.0	1.0	1.0	3.0
A0A812XEL1	0.5	1.0	0.0	1.0	0.5	3.0
A0A812NUI5	0.0	1.0	0.0	1.0	1.0	3.0
A0A812UAJ1	0.0	1.0	0.0	1.0	1.0	3.0
A0A812HUV3	0.0	1.0	0.0	1.0	1.0	3.0
A0A812HTU8	0.0	1.0	0.0	1.0	1.0	3.0
A0A090N4Y7	0.0	1.0	0.0	1.0	1.0	3.0
A4S200	0.0	1.0	0.0	1.0	1.0	3.0
C1E037	0.0	1.0	0.0	1.0	1.0	3.0
A0A7S1FBW0	0.0	1.0	0.0	1.0	1.0	3.0
A0A7S1FC43	0.0	1.0	0.0	1.0	1.0	3.0
F0XWW1	1.0	1.0	1.0	1.0	1.0	5.0
A0A7S1WJC0	0.5	1.0	0.0	1.0	1.0	3.5
A0A1Q9D410	0.5	1.0	0.0	1.0	1.0	3.5
A0A812U275	0.5	1.0	0.0	1.0	1.0	3.5
A0A812XZF5	0.5	1.0	0.0	1.0	1.0	3.5
A0A812RYT3	0.5	1.0	0.0	1.0	1.0	3.5
A0A812JW14	0.5	1.0	0.0	1.0	1.0	3.5
A0A812ULI5	0.5	1.0	0.0	1.0	1.0	3.5
A0A812TUP3	0.5	1.0	0.0	1.0	1.0	3.5
R1FQT9	0.5	0.5	1.0	1.0	0.5	3.5
K0SCN2	1.0	1.0	1.0	0.0	1.0	4.0
A0A7S4GBX2	1.0	1.0	1.0	1.0	0.5	4.5
A0A7S4GBX6	1.0	1.0	1.0	1.0	0.5	4.5
A0A7S2U9R3	1.0	1.0	1.0	1.0	1.0	5.0
A0A7S0IBT1	1.0	1.0	1.0	1.0	1.0	5.0
C1EG17	1.0	1.0	1.0	1.0	1.0	5.0
A0A1E7EZ89	1.0	1.0	1.0	1.0	1.0	5.0
A0A7S2XUT0	1.0	1.0	1.0	1.0	1.0	5.0
A0A7S2V5F9	1.0	1.0	1.0	1.0	1.0	5.0
A0A448ZPF4	1.0	1.0	1.0	1.0	1.0	5.0
A0A7S2M259	1.0	1.0	1.0	1.0	1.0	5.0
K0RA12	1.0	1.0	1.0	1.0	1.0	5.0
K0R419	1.0	1.0	1.0	1.0	1.0	5.0
W8VUA7	1.0	1.0	1.0	1.0	1.0	5.0
W8VYH0	1.0	1.0	1.0	1.0	1.0	5.0
K0SZQ0	1.0	1.0	1.0	1.0	1.0	5.0
A0A7S4MK07	1.0	1.0	1.0	1.0	1.0	5.0
A0A7S3AZA3	1.0	1.0	1.0	1.0	1.0	5.0
K8EQL3	1.0	1.0	1.0	1.0	1.0	5.0
X5I1I9	1.0	1.0	1.0	1.0	1.0	5.0
A0A1E7EUG8	1.0	1.0	1.0	1.0	1.0	5.0
A0A7S2LSL9	1.0	1.0	1.0	1.0	1.0	5.0
A0A7S3L7P6	1.0	1.0	1.0	1.0	1.0	5.0
A0A7S0KBZ5	1.0	1.0	1.0	1.0	1.0	5.0
A0A7S1UU45	1.0	1.0	1.0	1.0	1.0	5.0
A0A7S0UKW0	1.0	1.0	1.0	1.0	1.0	5.0
A0A7S0CDJ9	1.0	1.0	1.0	1.0	1.0	5.0
A0A7S4JEJ7	1.0	1.0	1.0	1.0	1.0	5.0
A0A6V2KCI5	1.0	1.0	1.0	1.0	1.0	5.0
A0A7S4S7Z7	1.0	1.0	1.0	1.0	1.0	5.0
A0A830HI75	1.0	1.0	1.0	1.0	1.0	5.0
A0A7S2HVS4	1.0	1.0	1.0	1.0	1.0	5.0
A0A830HHC5	1.0	1.0	1.0	1.0	1.0	5.0
A0A7S2BCN7	1.0	1.0	1.0	1.0	1.0	5.0
A0A7S0U0B0	1.0	1.0	1.0	1.0	1.0	5.0
A0A7S0R5U4	1.0	1.0	1.0	1.0	1.0	5.0
A7UC72	0.5	1.0	1.0	1.0	1.0	4.5
NaN	127.5	184.5	117.0	183.0	180.0	NaN

Figure 12. Overall evaluation of the 10 Cyc in the alignment

4.3 Blast search and sequence selection

4.3.1 Eukaryotic DCAs

>AAQ56178.1 intracellular carbonic anhydrase [Conticribra weissflogii] 1-275
was used as a query sequence. C-terminal KHEHNHSHGHSHVRGHQHHQWF was removed because it does not align with the DCA domain but does find other histidine-rich regions in Blast.

This is the delta CA which was first reported in the literature:

```
 TITLE     Carbonic anhydrase in the marine diatom Thalassiosira
weissflogii
            (Bacillariophyceae)
  JOURNAL   J. Phycol. 33 (5), 845-850 (1997)
  AUTHORS   Roberts,S.B., Lane,T.W. and Morel,F.M.M.
```

The scientific name of the organism was different at the time.

Blast at UniProt, target database UniProt KB, restricted to Eukaryota, E threshold 0.0001, 1000 hits gives 181 hits.

Sequence length:

1 - 200 (29)

201 - 400 (105)

401 - 600 (27)

601 - 800 (11)

>= 801 (9)

Lengths 201-600 will be taken to MSA and further analysis. The shortest sequences cannot cover the required 80% of the MSA length and the longer sequences will contain long, superfluous sequences or multiple DCA domains which will misguide the sequence alignment.

The bracket 201-400 (105 sequences) was taken to preliminary MSA (clustalo of SeaView) and then visual selection. The sequences that did not have anything aligning with the following conserved/important residues were deleted: C preceding the “QSP” motif; the “QSP” motif; HxHxxxEH of the active site; the “EVHW” with the third metal-binding His; “DRKC” motif.

Sequences with up to 40 residues missing from the C terminus were accepted, and so were sequences missing the hydrophobic regions which many sequences carry within their first ~20 to 90 residues. These latter regions are predicted to be signal peptides or transmembrane helices, which will be missing or separate from the DCA domains. By preliminary

examination of some of the AlphaFold models these regions have been mainly seen to be predictions of low confidence, in extended conformation outside of the globular, confidently predicted domain (“DCA domain”). Finally, sequences with unique insertions with lengths of more than 20 residues were removed (two sequences). These criteria leave us with 62 sequences. The longer bracket 401-600 contained less of truncated sequences as could be expected. Only one sequence was rejected because of C-terminal truncation, and one more because of an unreasonable insertion. Accordingly, keeping 25 of 27 sequences.
The next step was joining the two sets. After alignment in SeaView, one of the longer sets was seen to miss the “DRKC” motif and was removed. Realigning with ClustalO at EBI. Iterations were selected as shown in Figure 13:

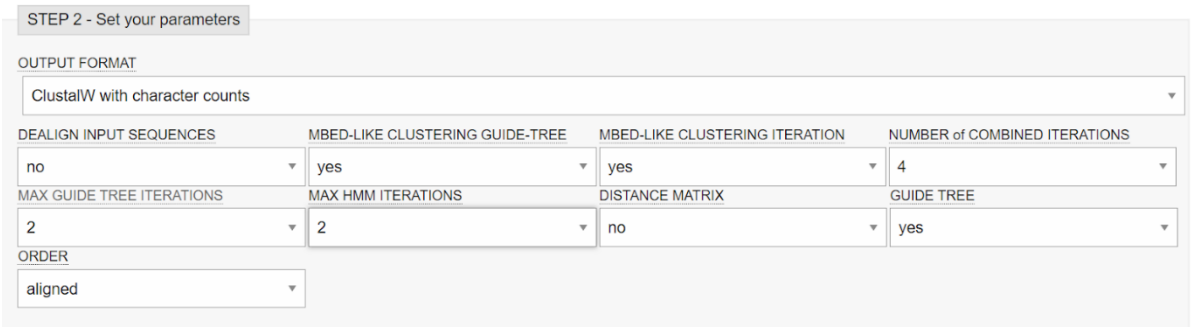


Figure 13. parameter setting in EBI

As SeaView cannot handle numbered Clustal format well, colours get misplaced from their letters, so I repeated the process but clustal without numbers as output (Figure 14).

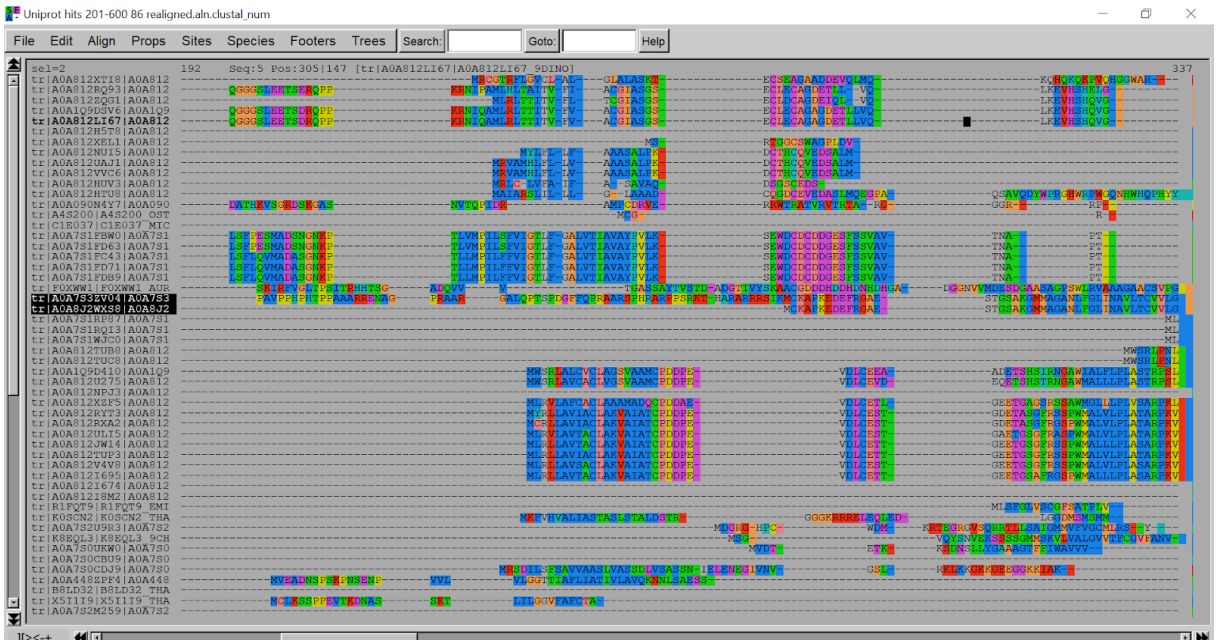


Figure 14. produced Seaview

After realignment, we have several pairs of 100% identical sequences as seen in the PIM matrix and some groups of near-identity. This leaves us 65 sequences, another evaluation with clustalO at EBI. After that, the active sites are split into two places. It is A0A812TUB8 which has two active sites. So, the final sequence set of eukaryotic delta CAs contains 64 sequences as an aln formatted file.

4.3.2 Bacterial DCAs

Starting Blast with A0A545U8G7 in UniProt against bacterial reference proteomes I get 167 hits. Six of them are of length 125 or less, they are discarded. The set of 161 is aligned. Visual inspection reveals three sequences truncated in N-terminus (not having the QSP site) and two with long, unique insertions. The new set of 156 sequences is realigned. However, 10 of 156 AF models were not found. Then I realigned the set of 146 sequences and made a new logo file.

4.4 The process of generation of MSA weights and entropy

The final sequence set of eukaryotic delta CAs contains 64 sequences. Their identity profile, taken from the final ClustalO at EBI, same parameters are above, is seen coloured in the PIM matrix. Nevertheless, four alignments were also removed as the AF model for A0A8J2WXS8 and A7UC73 does not exist, and also two other sequences with suspicious insertions which yield 60 sequences. Next, the retrieved sequences consisting of UniProt IDs have been realigned through the “https://www.ebi.ac.uk/Tools/msa/clustalo/” service of the Clustal Omega webpage. Figure 15 shows the results page of the aligned sequences.

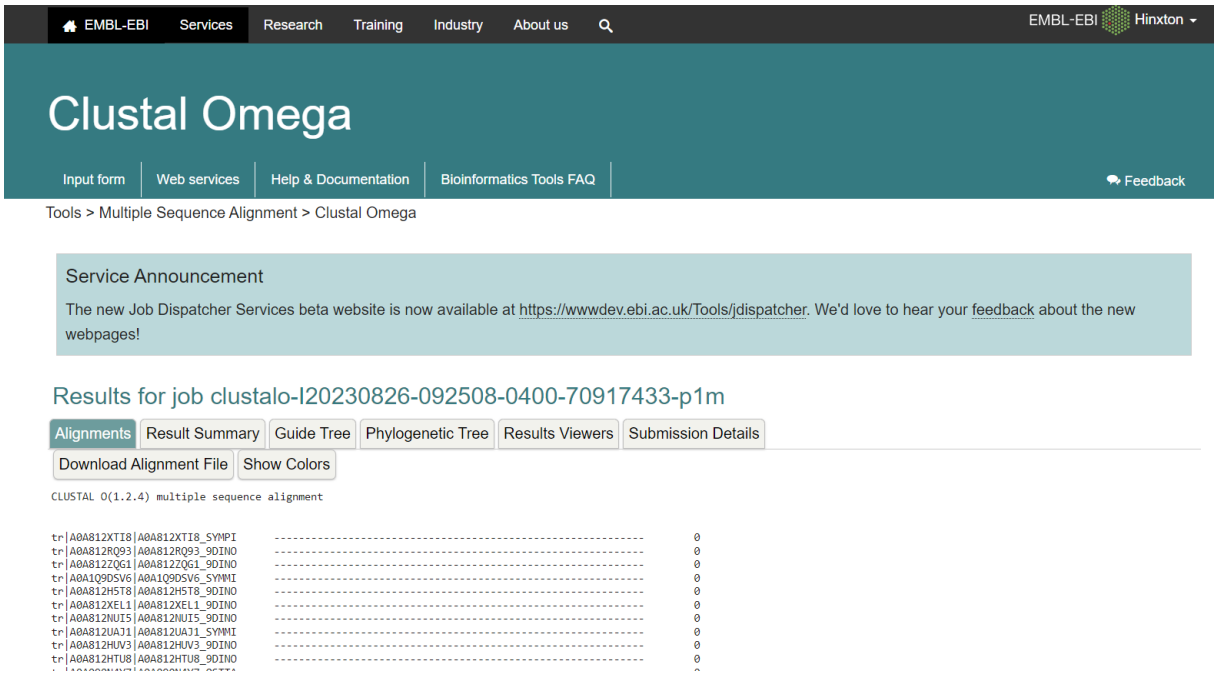


Figure 15. A summary view of multiple sequence alignment 60 seq

Then, the final alignment was uploaded to https://weblogo.threeplusone.com/create.cgi. With the selectable output of plain text. The main steps are depicted in Figure 16.

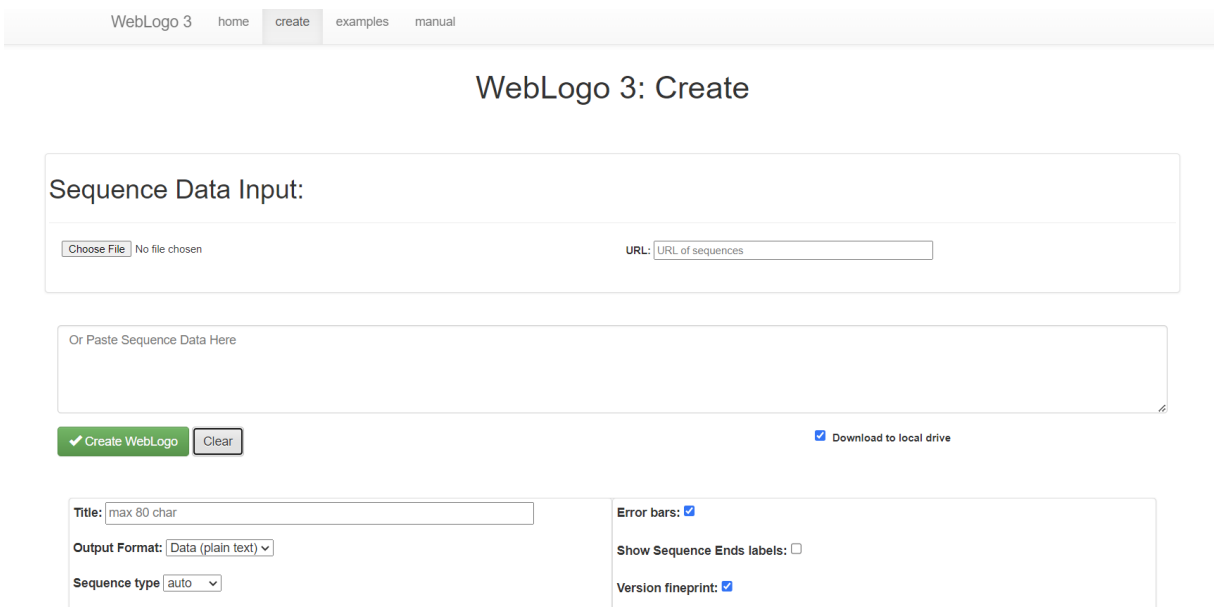
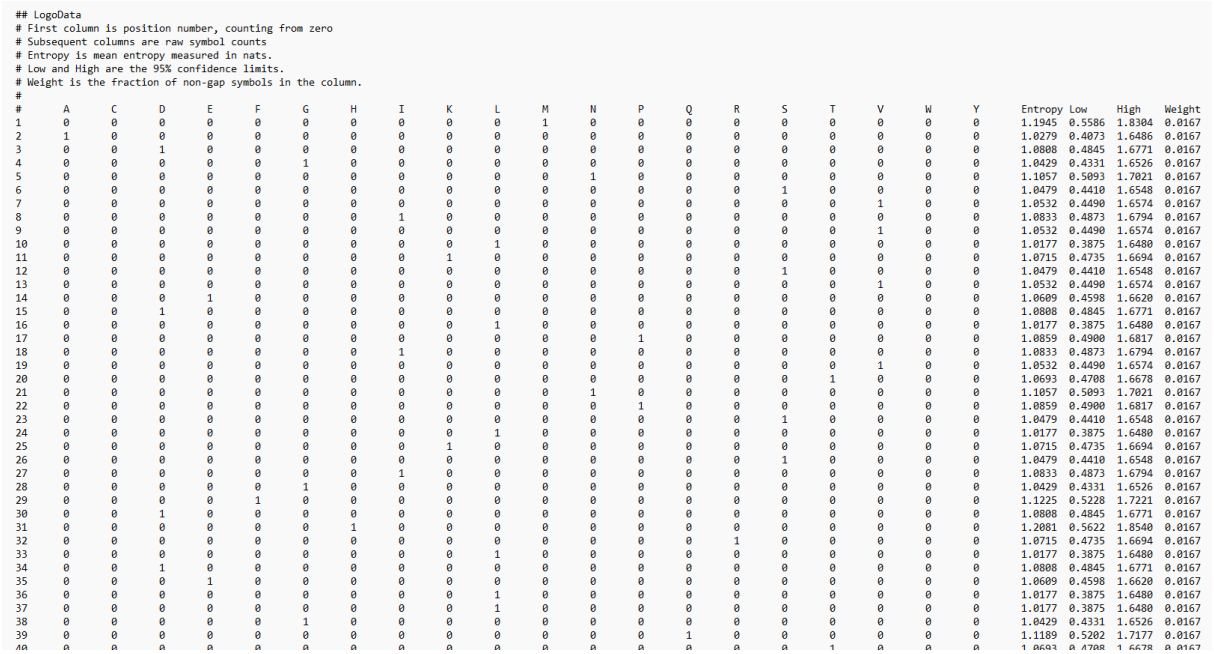


Figure 16. Weblogo weight and entropy retrieval process

Finally, the output looks like a text file presented in Figure 17. Which consists of 814 rows of data with the headers partially visible in the image below. Therefore, the information for each position in the sequence alignment is contained in each line. The first number on each line

indicates the location, which is followed by the number of times each amino acid symbol appears there. The entropy and weight at that point are the last two columns in the file. The produced text file is a key input to the result analysis of the Python codes.



```
## LogoData
# First column is position number, counting from zero
# Subsequent columns are raw symbol counts
# Entropy is mean entropy measured in nats.
# Low and High are the 95% confidence limits.
# Weight is the fraction of non-gap symbols in the column.
#
#      A     C     D     E     F     G     H     I     K     L     M     N     P     Q     R     S     T     V     W     Y     Entropy Low     High    Weight
1      0     0     0     0     0     0     0     0     0     0     1     0     0     0     0     0     0     0     0     0     1.1945  0.5586  1.8304  0.0167
2      1     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     1.0279  0.4073  1.6486  0.0167
3      0     0     1     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     1.0808  0.4845  1.6771  0.0167
4      0     0     0     0     0     1     0     0     0     0     0     0     0     0     0     0     0     0     0     0     1.0429  0.4331  1.6526  0.0167
5      0     0     0     0     0     0     0     0     0     0     0     1     0     0     0     0     0     0     0     0     1.1057  0.5093  1.7021  0.0167
6      0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     1     0     0     0     0     1.0479  0.4410  1.6548  0.0167
7      0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     1     0     0     1.0532  0.4490  1.6574  0.0167
8      0     0     0     0     0     0     0     1     0     0     0     0     0     0     0     0     0     0     0     0     1.0833  0.4873  1.6794  0.0167
9      0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     1     0     0     1.0532  0.4490  1.6574  0.0167
10     0     0     0     0     0     0     0     0     0     1     0     0     0     0     0     0     0     0     0     0     1.0177  0.3875  1.6480  0.0167
11     0     0     0     0     0     0     0     0     1     0     0     0     0     0     0     0     0     0     0     0     1.0715  0.4735  1.6694  0.0167
12     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     1     0     0     0     0     1.0479  0.4410  1.6548  0.0167
13     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     1     0     0     1.0532  0.4490  1.6574  0.0167
14     0     0     0     1     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     1.0609  0.4598  1.6620  0.0167
15     0     0     1     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     1.0808  0.4845  1.6771  0.0167
16     0     0     0     0     0     0     0     0     0     1     0     0     0     0     0     0     0     0     0     0     1.0177  0.3875  1.6480  0.0167
17     0     0     0     0     0     0     0     0     0     0     0     0     0     1     0     0     0     0     0     0     1.0859  0.4900  1.6817  0.0167
18     0     0     0     0     0     0     0     1     0     0     0     0     0     0     0     0     0     0     0     0     1.0833  0.4873  1.6794  0.0167
19     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     1     0     0     1.0532  0.4490  1.6574  0.0167
20     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     1     0     0     1.0693  0.4708  1.6678  0.0167
21     0     0     0     0     0     0     0     0     0     0     0     1     0     0     0     0     0     0     0     0     1.1057  0.5093  1.7021  0.0167
22     0     0     0     0     0     0     0     0     0     0     0     0     0     1     0     0     0     0     0     0     1.0859  0.4900  1.6817  0.0167
23     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     1     0     0     0     0     1.0479  0.4410  1.6548  0.0167
24     0     0     0     0     0     0     0     0     0     1     0     0     0     0     0     0     0     0     0     0     1.0177  0.3875  1.6480  0.0167
25     0     0     0     0     0     0     0     0     1     0     0     0     0     0     0     0     0     0     0     0     1.0715  0.4735  1.6694  0.0167
26     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     1     0     0     0     0     1.0479  0.4410  1.6548  0.0167
27     0     0     0     0     0     0     0     1     0     0     0     0     0     0     0     0     0     0     0     0     1.0833  0.4873  1.6794  0.0167
28     0     0     0     0     0     1     0     0     0     0     0     0     0     0     0     0     0     0     0     0     1.0429  0.4331  1.6526  0.0167
29     0     0     0     0     1     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     1.1225  0.5228  1.7221  0.0167
30     0     0     1     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     1.0808  0.4845  1.6771  0.0167
31     0     0     0     0     0     0     1     0     0     0     0     0     0     0     0     0     0     0     0     0     1.2081  0.5622  1.8540  0.0167
32     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     1     0     0     0     1.0715  0.4735  1.6694  0.0167
33     0     0     0     0     0     0     0     0     0     1     0     0     0     0     0     0     0     0     0     0     1.0177  0.3875  1.6480  0.0167
34     0     0     1     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     1.0808  0.4845  1.6771  0.0167
35     0     0     0     1     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     1.0609  0.4598  1.6620  0.0167
36     0     0     0     0     0     0     0     0     0     1     0     0     0     0     0     0     0     0     0     0     1.0177  0.3875  1.6480  0.0167
37     0     0     0     0     0     0     0     0     0     1     0     0     0     0     0     0     0     0     0     0     1.0177  0.3875  1.6480  0.0167
38     0     0     0     0     0     1     0     0     0     0     0     0     0     0     0     0     0     0     0     0     1.0429  0.4331  1.6526  0.0167
39     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     1     0     0     0     0     1.1189  0.5202  1.7177  0.0167
40     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     0     1     0     0     1.0693  0.4708  1.6678  0.0167
```

Figure 17. A sample view of the WebLogo produced for analysis of DCA 60 sequences

4.5 Analysis of DCA

In this part, I have analysed the MSA consisting of:

```
Length of alignment file 60 records
Unique seq lengths in MSA: {814}
MSA columns in logofile 814
```

Initially, a logo file in text format and an MSA file are parsed by the Python codes. The MSA file in Clustal format is read by the 'AlignIO.read()' module of the Biopython library, which then returns an alignment object. The code then determines if each sequence in the alignment is the same length and if not, it raises an exception message. Using the pandas package, the logo file is read in and a logo plot is made by using the yielded data frame. If there is a discrepancy between the length of the MSA and the number of columns in the logo file, the code produces an error message. Next, the algorithm creates the carbonic anhydrase logo plot. The panda's package function reads a file in tab-separated format and produces a pandas data frame. The entropy and weight of each point in the MSA are displayed in Figure 18, as a graph of entropy and weight against position. The x-axis is the 814 positions, and the y-axis is

entropy and weight, while the blue line represents the entropy, and the red line represents the weight. The plot shows that entropy and weight are almost presenting the same trend after 360. The plot also shows that the entropy and weight values are highest at positions 360 -750.

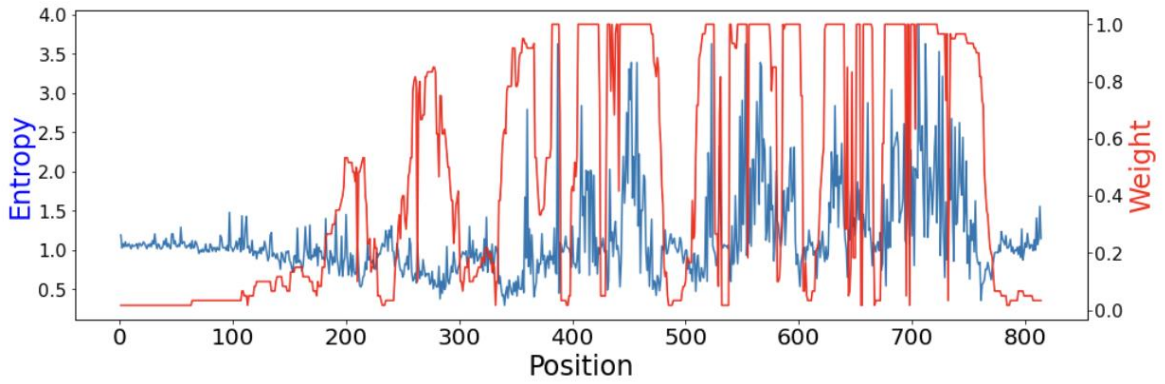


Figure 18. entropy in blue and weight in red

Next, the 'Weight' column of the data frame is used by the algorithm to conduct change point detection using the 'ruptures' library. With a model of "l1" and a penalty value of 2, it specifically applies the "Pelt" method to find change spots. The 'rpt.display()' method from the 'ruptures' package is then used to plot the resulting change points. Similarly, the plot's y-axis denotes weight, while the x-axis shows position. The line graph in Figure 19 depicts the relationship between position and weight.

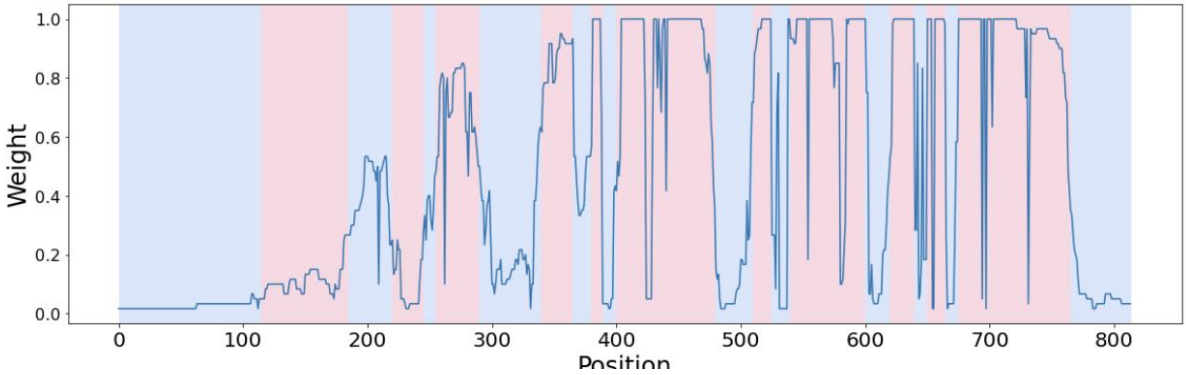


Figure 19. Change point detection in MSA

Then, I downloaded AlphaFold models for each sequence in the MSA file by using the codes. A written function utilizes the 'requests' library to download files from a URL and save them in a predetermined location in my computer drive. The path in which the downloaded models will be saved is indicated by the 'model_dir' variable. A list of UniProt IDs taken from the

MSA file makes up the 'uniprot_ids' variable. The program then iterates over each UniProt ID and determines whether the relevant model file is already present in my directory folder, and the file is downloaded from a URL and stored in the "model_dir" if it is not already there. The GitHub link to the python codes for AFmodels collection:

- https://github.com/hoseindaneshpour/AF-model-download-link-creator/blob/main/AF%20model%20download%20link%20creator.py

Then, a novel method for the creation of the main data frame has been presented. This algorithm reads the MSA file in Clustal format and generates a pandas data frame including the UniProt ID, residue, and location of each non-gap character in the MSA. Each character in each MSA sequence is iterated by using the 'enumerate()' function, and gap characters are excluded using the 'if' condition. To create three-letter codes from one-letter amino acid codes, utilize a dictionary. Then the weights are mapped to the positions in the MSA by the code, which creates a new data frame. The weights are mapped to the places using the produced data frame ('dfweights'), and the final data frame is saved in 'new_df'. Then, 'df1' and 'new_df' are combined using the 'pd.merge()' function based on their indexes to create a new data frame that includes the UniProt ID, residue, location, and weight for each non-gap character in the MSA. The method then extracts the B-factor values for each residue in each model by reading the AlphaFold models for each sequence in the MSA file. Each residue in each model is identified by its UniProt ID, residue ID, and B-factor in the final data frame (Table 3).

Table 3. Summary of the outcome data frame for DCA

	Uniprot ID	Residue	Residue ID	B Factor	position	weights_collection
0	A0A812XTI8	MET	1	34.22	256	0.4833
1	A0A812XTI8	ARG	2	27.67	257	0.5333
2	A0A812XTI8	CYS	3	25.52	258	0.5333
3	A0A812XTI8	GLY	4	27.34	259	0.7667
4	A0A812XTI8	THR	5	33.56	260	0.8000
...	...	...	...	...	...	...
22660	A7UC72	LEU	282	93.25	757	0.9000
22661	A7UC72	VAL	283	89.06	758	0.9000

	Uniprot ID	Residue	Residue ID	B Factor	position	weights_collection
22662	A7UC72	ARG	284	64.25	759	0.9000
22663	A7UC72	ARG	285	56.75	760	0.8167
22664	A7UC72	VAL	286	43.16	761	0.8167

22665 rows × 6 columns

Having the data frame ready in hand, I have made a code that generates a scatter plot of the B-factor values for each residue in each model versus the MSA weights for each residue. The plot was made using the 'sns.regplot()' function from the Seaborn library, and a red colour was added using the 'plt.fill()' function. The resultant plot reveals a positive link between the MSA weights and B-factor values, with greater MSA weights correlating with higher B-factor values. However, due to the availability of signal peptides and short and long sequences, the correlation measure guided us for future studies, as well. The MSA weights vs. pLDDT scatter plot is shown in Figure 20, while MSA weights are on the y-axis, while pLDDT is on the x-axis. Through the plot, there is a yellow regression line with a positive slope. In the plot's upper left corner, a group of so-called Redbox data points can be seen. The 'pLDDT' score is a metric of the projected structure's confidence, with higher values suggesting more certainty. Higher weights indicate better conservation, and the "MSA weights" are a measure of how well-conserved each site in the multiple sequence alignment is. The pink data points in the plot's upper-left corner correspond to highly conserved places where the projected structure is low confident.

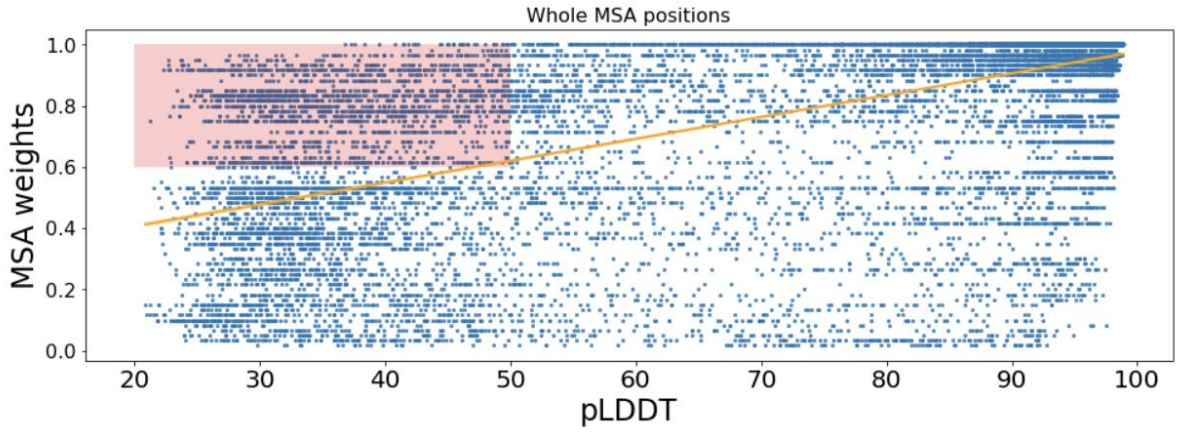


Figure 20. pLDDT and MSA weights for the 60 sequences

Figure 21 is a histogram of the complete MSA locations' B Factor distribution. The "B Factor" in the x-axis and the "Count" in the y-axis are labelled.

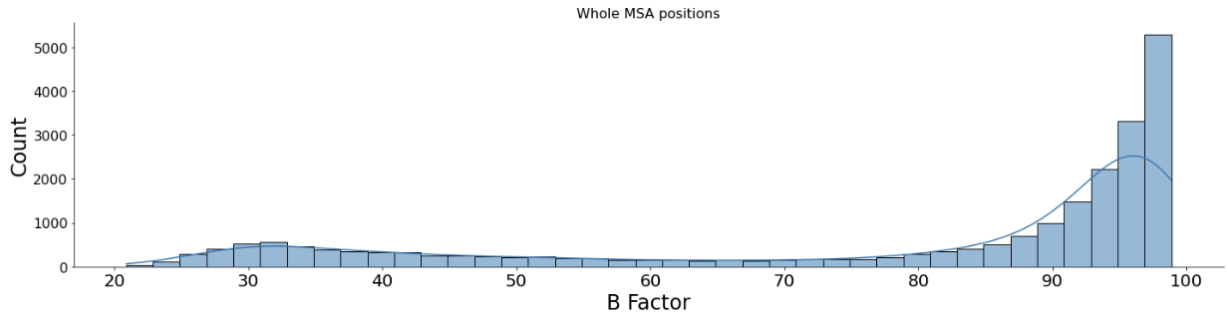
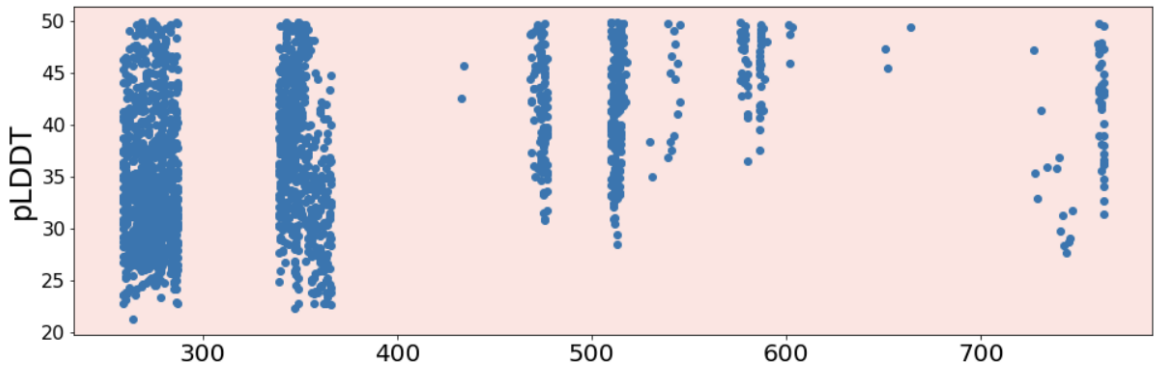


Figure 21. B Factor distribution of whole MSA

Accordingly, for the whole MSA, there is a 0.64 correlation between the B-factor values and MSA weights, which suggests a moderately positive connection between the two variables. In addition, I interpreted the Redbox further. For places with an MSA weight of more than 0.6 and a B-factor value lower than 50, the first plot in figure 22 is a scatter plot of the B-factor values versus the position in the MSA. Also, the histogram of the position in the MSA for places with an MSA weight of more than 0.6 and a B-factor value lower than 50 is shown in the second plot.



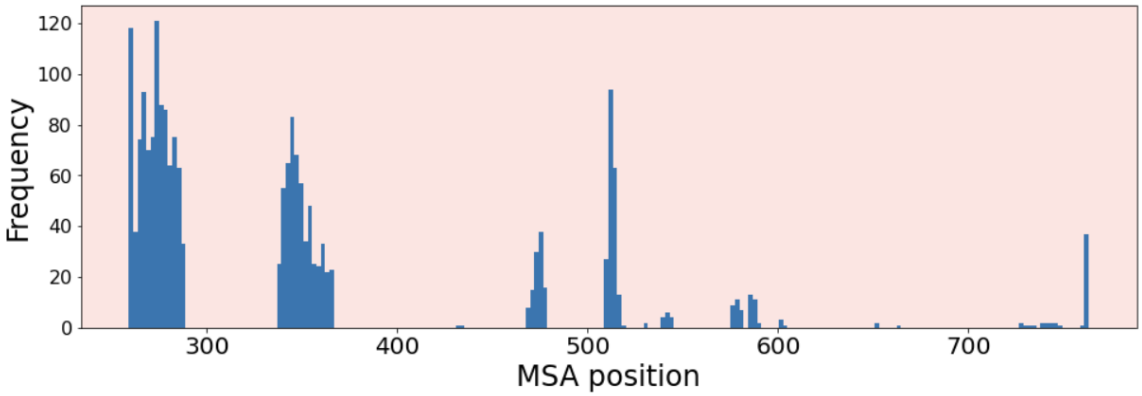


Figure 22. Residues at high MSA weight with low model quality

Then the code generates a scatter plot of the B-factor values vs. the MSA weights for all MSA positions that lie between the ranges of domain start and end point ('dca_start' and 'dca_end'). Figure 23 demonstrates that at sites in this range, there is a positive connection between the B-factor values and MSA weights.

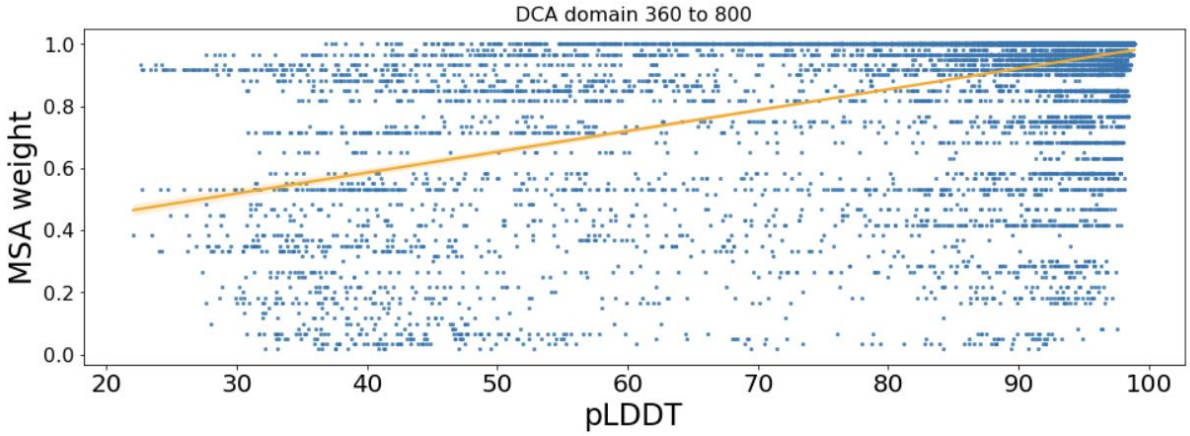


Figure 23. MSA versus pLDDT in domain 360 to 800

Likewise, the histogram in figure 24 shows the frequency of B-factor values for MSA points within the range of "dca_start" to "dca_end." The B-factor values are shown on the x-axis, while the frequency of occurrence is shown on the y-axis.

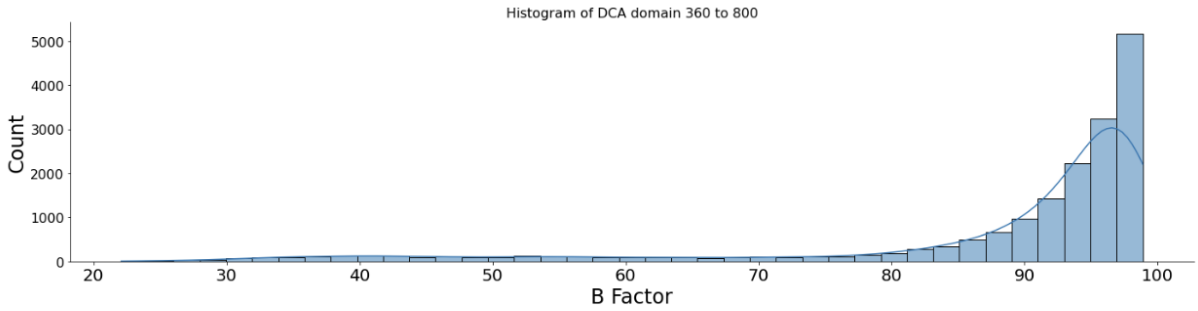


Figure 24. histogram of pLDDT in DCA domain 360 to 800

Meanwhile, for MSA locations that are between 360 and 800, there is a 0.53 correlation between the B-factor values and the MSA weights. So, the two variables appear to have a moderately positive correlation.

The mean B-factor value for each place in the MSA is displayed in figure 25. Position in the MSA is shown by the x-axis, while the mean B-factor value is represented by the y-axis. The graphic demonstrates that the mean B-factor values have many peaks.

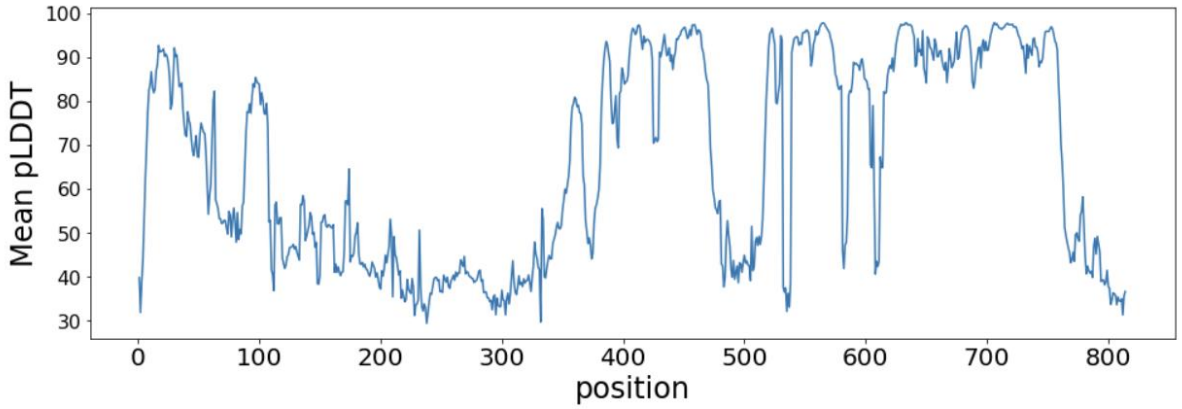


Figure 25. the mean B-factor value for 814 positions in the MSA

This line graph in figure 26 displays a protein sequence's location between 330 and 710 concerning the MSA weight and pLDDT. The MSA weight and pLDDT are represented on the y-axis, while the location within the protein sequence is represented on the x-axis. The blue line shows the pLDDT, while the red line shows the MSA weight. The MSA weight is plotted in the range of 0 and 1.2, and 0 to 100 is the mean pLDDT.

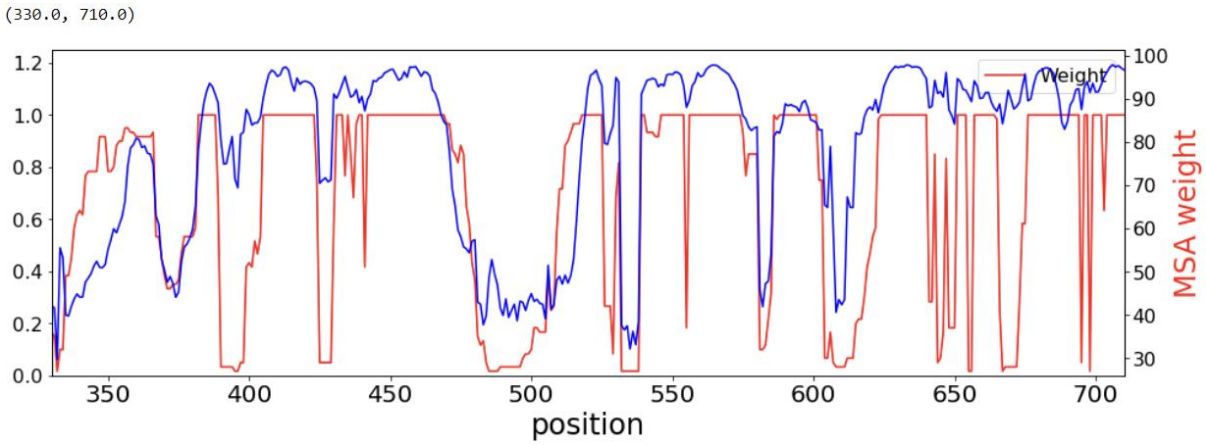


Figure 26. the MSA weight and pLDDT between 330 and 710 positions

Then, I determined the statistics of good-quality residues in AF models. Therefore, there are 319 sites with B-factor values higher than 70 for the domain limitations. There are 204 positions with B-factor values more than 90, and 122 places with B-factor values less than 70 within the range of positions between 360 and 800. Also, supergood residues may be found at 99 locations with B-factor values higher than 95.

Furthermore, for each Uniprot ID, the statistics of high-quality residues are computed using Python codes. Therefore, a data frame is produced while the column values for each Uniprot ID that satisfy specific requirements are contained in the data frame: Len of Column Values no. of > 70; 80>Len of Column Values no. of > 70; 90>Len of Column Values no. of > 80; Len of Column Values no. of > 90; Len of Column Values no. of < 70 (Table 4). Meanwhile, the HTML Styles CSS is used to modify the data frame's 'style' attribute dependent on the pLDDT colouring style.

Table 4. statistics of quality residues in each 60 sequence

	Uniprot ID	360 < Residue Count < 800	Len of Column Values no. of > 70	80>Len of Column Values no. of > 70	90>Len of Column Values no. of > 80	Len of Column Values no. of > 90	Column Values	Residue	Len of Column Values no. of < 70
0	A0A812XTI8	297.0	240	17	54	169	[360, 361, 362, 363, 364, 365, 366, 367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 467, 468, 469, 470, 471, 472, 473, 474, 475, 476, 477, 478, 479, 480, 506, 509, 510, 511, 512, 513, 514, 515, 516, 658, 659, 660, 661, 762, 763, 764, 765, 766, 767, 768]	[GLY, TRP, ALA, ARG, PRO, THR, LYS, ALA, ASP, PRO, GLU, PRO, ARG, ALA, PRO, THR, MET, MET, GLN, THR, GLY, ARG, GLN, PRO, GLY, GLY, PRO, THR, ALA, ARG, ASP, ALA, GLU, ILE, GLN, SER, VAL, ILE, PRO, ALA, GLN, PRO, GLN, PRO, SER, SER, ILE, GLN, GLN, TRP, VAL, SER, PRO, PRO, ALA, THR, ARG]	57
1	A0A812RQ93	296.0	233	16	71	146	[360, 361, 362, 363, 364, 365, 366, 367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 467, 468, 469, 470, 471, 472, 473, 474, 475, 476, 477, 478, 479, 480, 506, 509, 510, 511, 512, 513, 514, 515, 516, 517, 586, 587, 588, 589, 593, 663, 664, 665, 676, 677, 762, 763, 764, 765, 766, 767]	[HIS, GLU, LEU, GLY, LYS, SER, ASN, THR, SER, ALA, ALA, VAL, GLN, GLU, VAL, GLU, LEU, GLU, GLY, SER, GLU, ASN, ASP, VAL, GLY, LYS, HIS, LEU, HIS, TYR, ARG, HIS, GLY, VAL, PRO, GLU, TRP, ASP, GLY, HIS, PRO, LEU, VAL, PRO, GLU, GLY, THR, ARG, GLU, PHE, ARG, GLY, GLU, PRO, TYR, PHE, GLY, ASP, GLU, ASP, MET, ARG, HIS]	63
2	A0A812ZQG1	296.0	224	18	63	143	[360, 361, 362, 363, 364, 365, 366, 367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 467, 468, 469, 470, 471, 472, 473, 474, 475, 476, 477, 478, 479, 480, 506, 509, 510, 511, 512, 513, 514, 515, 516, 585, 586, 587, 588, 589, 590, 591, 592, 593, 594, 595, 660, 661, 662, 663, 664, 665, 676, 677, 761, 762, 763, 764, 765, 766, 767]	[HIS, GLN, VAL, GLY, LYS, SER, ASN, ALA, ASN, ALA, ALA, THR, GLN, GLU, VAL, GLU, LEU, GLU, GLY, SER, GLU, ASN, ASP, VAL, GLY, LYS, HIS, LEU, HIS, TYR, ARG, HIS, GLY, VAL, PRO, GLU, TRP, ASP, GLY, HIS, PRO, LEU, VAL, PRO, GLU, GLY, GLU, ARG, GLU, PHE, ARG, ILE, PHE, ASN, GLY, LEU, GLY, GLN, TRP, ARG, GLU, PRO, TYR, LEU, GLY, ASN, ASP, GLU, ASP, MET, ARG, HIS]	72
3	A0A1Q9DSV6	296.0	235	17	60	158	[360, 361, 362, 363, 364, 365, 366, 367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 467, 468, 469, 470, 471, 472, 473, 474, 475, 476, 477, 478, 479, 480, 506, 509, 510, 511, 512, 513, 514, 515, 516, 526, 586, 587, 588, 664, 665, 676, 677, 761, 762, 763, 764, 765, 766, 767]	[HIS, GLN, VAL, GLY, LYS, SER, ASN, THR, SER, ALA, ALA, VAL, GLN, GLU, VAL, GLU, LEU, GLU, GLY, SER, GLU, ASN, ASP, VAL, GLY, LYS, HIS, LEU, HIS, TYR, ARG, HIS, GLY, VAL, PRO, GLU, TRP, ASP, GLY, HIS, PRO, LEU, VAL, PRO, GLU, GLY, ARG, ARG, GLU, PHE, PRO, TYR, PHE, GLY, ASP, ASP, GLU, ASP, MET, ARG, HIS]	61
4	A0A812H5T8	300.0	233	15	61	157	[366, 367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 573, 574, 575, 576, 577, 578, 579, 580, 581, 582, 583, 584, 585, 586, 587, 588, 589, 590, 591, 592, 593, 594, 595, 596, 597, 598, 599, 600, 601, 623, 643, 646, 647, 648, 649, 650, 651, 652, 653, 654, 762, 763, 764, 765, 766, 767, 768, 769, 770, 771]	[MET, GLN, LEU, SER, LYS, LEU, PHE, PHE, CYS, CYS, LEU, VAL, TRP, LYS, ALA, ARG, GLY, GLY, PRO, LYS, SER, ASP, GLY, LEU, GLN, GLU, ASP, GLY, ASP, GLU, GLY, GLN, LEU, GLY, ILE, THR, ALA, GLY, LEU, GLY, GLY, ALA, PHE, THR, ARG, THR, ASN, ASN, ASP, TYR, VAL, ASP, ALA, LYS, ASN, ARG, THR, PRO, GLN, GLY, ALA, GLY, ILE, VAL, GLN, GLN, GLY]	67
5	A0A812XEL1	275.0	201	25	94	82	[360, 361, 362, 363, 364, 365, 366, 367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 383, 438, 577, 578, 579, 580, 581, 582, 583, 584, 585, 586, 587, 588, 589, 590, 591, 592, 593, 594, 595, 596, 659, 660, 661, 662, 663, 664, 665, 676, 677, 688, 689, 690, 724, 725, 726, 727, 728, 729, 731, 734, 739, 740, 741, 742, 743, 744, 745, 746, 747]	[CYS, GLN, VAL, GLU, ASP, SER, ALA, LEU, MET, GLN, LEU, LEU, SER, ASN, SER, THR, GLU, THR, SER, ASN, GLU, THR, ALA, LYS, GLY, VAL, GLY, LEU, ARG, ASP, GLU, GLY, ALA, GLU, GLY, GLU, LEU, GLY, ILE, THR, ASP, GLY, LEU, ALA, GLY, GLN, GLN, TRP, ARG, GLN, PRO, PRO, LEU, GLY, GLY, THR, ASP, PRO, VAL, PRO, VAL, LEU, ALA, GLY, GLY, ARG, PHE, PRO, ARG, GLY, SER, SER, ARG, SER]	74
6	A0A812NUI5	302.0	249	11	57	181	[360, 361, 362, 363, 364, 365, 366, 367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 383, 577, 578, 579, 580, 581, 582, 583, 584, 585, 586, 587, 588, 589, 590, 591, 592, 593, 594, 595, 596, 762, 763, 764, 765, 766, 767, 768, 769, 770]	[SER, SER, SER, LEU, PRO, GLN, PHE, SER, PRO, ALA, LEU, LEU, SER, ASN, ALA, ALA, GLU, THR, SER, ASN, GLU, THR, ALA, LYS, GLU, GLY, LEU, ARG, ASP, GLU, GLY, ALA, GLU, GLY, GLN, LEU, GLY, ILE, THR, ASP, GLY, LEU, GLY, GLY, PRO, SER, PRO, GLU, PRO, GLU, SER, GLU, LEU]	53
7	A0A812UAJ1	301.0	235	19	57	159	[360, 361, 362, 363, 364, 365, 366, 367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 383, 384, 478, 479, 480, 506, 509, 512, 527, 577, 578, 579, 580, 581, 582, 583, 584, 585, 586, 587, 588, 589, 591, 592, 593, 594, 595, 660, 661, 662, 663, 664, 665, 676, 763, 764, 765, 766, 767, 768, 769, 770, 771]	[SER, SER, SER, LEU, PRO, GLN, PHE, SER, PRO, SER, LEU, LEU, SER, ASN, ALA, THR, GLU, PRO, SER, ASN, GLU, THR, ALA, PRO, VAL, ALA, THR, PRO, LEU, VAL, ARG, ASP, ILE, GLY, LEU, ARG, ASP, GLU, GLY, ALA, GLU, GLY, GLN, LEU, GLY, THR, ASP, GLY, LEU, GLY, GLN, TRP, ARG, GLN, PRO, PRO, LEU, SER, PRO, GLU, PRO, GLU, SER, VAL, GLU, LEU]	66
8	A0A812HUV3	307.0	250	13	65	172	[360, 361, 362, 363, 364, 365, 366, 367, 368, 369, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 383, 509, 510, 511, 512, 513, 514, 515, 577, 578, 579, 580, 581, 582, 583, 584, 585, 586, 587, 588, 589, 593, 594, 765, 766, 767, 768, 769,	[SER, THR, ALA, LEU, PRO, GLN, PHE, ASN, PRO, VAL, ALA, LEU, ALA, ALA, THR, ALA, LYS, ARG, GLU, THR, PRO, VAL, ARG, PRO, GLU, ASN, ALA, SER, ILE, GLY, LEU, ARG, ASP, GLU, GLY, GLY, GLU, GLY, ILE, LEU, GLY, GLY, LEU, LEU,	57

	Uniprot ID	360 < Residue Count < 800	Len of Column Values no. of > 70	80>Len of Column Values no. of > 70	90>Len of Column Values no. of > 80	Len of Column Values no. of > 90	Column Values	Residue	Len of Column Values no. of < 70
							770, 771, 772, 773, 774, 775, 776, 777, 778]	PRO, PRO, ALA, PRO, VAL, PRO, THR, SER, THR, SER, THR, PRO, ARG]	
9	A0A812HTU8	289.0	237	22	63	152	[374, 375, 376, 377, 378, 379, 380, 381, 382, 383, 384, 479, 480, 511, 512, 513, 514, 515, 516, 577, 578, 579, 580, 581, 582, 583, 584, 585, 586, 587, 588, 589, 593, 594, 595, 661, 662, 663, 664, 665, 676, 677, 762, 763, 764, 765, 766, 767, 768, 769, 770, 771]	[THR, SER, GLY, ALA, PRO, ALA, PRO, PRO, PRO, SER, ILE, GLU, PRO, PRO, GLU, ASN, ALA, SER, ASP, ILE, GLY, ILE, ARG, ASP, GLU, SER, GLY, GLU, GLY, ILE, LEU, GLY, GLY, LEU, GLY, TRP, ARG, GLN, PRO, PRO, GLU, VAL, VAL, SER, PRO, LEU, PRO, PRO, SER, THR, PRO, PRO]	52
10	A0A090N4Y7	274.0	213	8	56	149	[360, 361, 362, 363, 364, 365, 366, 367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 424, 431, 432, 433, 434, 435, 436, 467, 468, 469, 518, 526, 527, 528, 539, 540, 541, 542, 543, 544, 545, 546, 547, 548, 639, 640, 641, 642, 643, 649, 650, 651, 652, 653, 654, 763, 764, 765]	[ARG, ALA, ILE, VAL, ALA, VAL, VAL, ARG, ALA, ILE, VAL, ALA, GLY, VAL, GLY, ALA, THR, ALA, ALA, ARG, ALA, ASN, GLU, SER, GLU, THR, TRP, LYS, ARG, ALA, GLU, VAL, SER, GLU, ALA, ALA, THR, LEU, GLY, SER, ASP, GLY, GLU, ALA, SER, ASN, ASP, GLY, GLY, ASP, GLY, ALA, ASP, ASP, ASP, ALA, LYS, SER, PRO, PHE, PHE]	61
11	A4S200	276.0	211	12	81	118	[360, 361, 362, 363, 364, 365, 366, 367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 383, 420, 421, 422, 423, 424, 431, 432, 433, 435, 436, 437, 438, 439, 468, 469, 526, 527, 528, 529, 530, 531, 532, 533, 534, 535, 536, 537, 538, 539, 540, 541, 542, 543, 544, 545, 546, 547, 761, 762, 763, 764]	[ALA, VAL, THR, TYR, PRO, VAL, LEU, ARG, ASP, GLU, TRP, PHE, GLY, PRO, GLY, ASP, ASP, GLU, THR, GLY, ALA, ARG, ALA, ALA, SER, GLY, ALA, SER, ILE, LEU, PRO, THR, ASN, ALA, ILE, LEU, ARG, ALA, SER, ALA, SER, ASP, ALA, THR, GLY, THR, ALA, ARG, VAL, VAL, ALA, THR, ASN, GLY, ASP, ASP, VAL, SER, ASP, ALA, ARG, SER, THR, PHE, TYR]	65
12	C1E037	287.0	250	10	46	194	[375, 376, 377, 378, 379, 380, 381, 382, 383, 466, 467, 468, 469, 470, 471, 472, 473, 474, 475, 476, 477, 478, 479, 480, 506, 508, 509, 510, 511, 512, 513, 514, 515, 516, 517, 518, 594]	[MET, THR, GLU, VAL, GLU, GLU, ALA, PRO, ALA, THR, LYS, PRO, SER, ASP, ASP, THR, GLU, ALA, ARG, ARG, ARG, LEU, ARG, ARG, LEU, ASP, SER, GLY, ASP, PRO, ALA, LEU, PRO, LEU, GLU, THR, LEU]	37
13	A0A7S1FBW0	277.0	235	10	41	184	[367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 432, 433, 435, 436, 437, 439, 440, 441, 442, 469, 470, 471, 472, 473, 512, 513, 514, 515, 516, 517, 518, 586, 587, 588, 768, 769]	[VAL, VAL, ALA, THR, ASN, ALA, PRO, THR, VAL, ALA, PRO, THR, ALA, THR, SER, SER, SER, ASN, LEU, ASP, VAL, SER, SER, ASP, HIS, GLU, ASP, SER, SER, PHE, ASP, ALA, ALA, VAL, THR, ASP, GLU, ASN, GLU, ASN, ASP, PRO]	42
14	A0A7S1FC43	269.0	231	10	41	180	[360, 361, 362, 363, 364, 365, 366, 367, 368, 369, 370, 371, 372, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 440, 468, 469, 470, 471, 472, 473, 512, 513, 514, 515, 516, 517, 518, 519]	[VAL, ALA, THR, ASN, ALA, PRO, THR, ALA, ALA, ALA, THR, SER, ALA, PRO, THR, VAL, ALA, PRO, THR, ALA, ALA, SER, SER, SER, MET, GLU, ASP, SER, SER, PHE, ASP, ALA, ALA, VAL, THR, ASP, GLU, TYR]	38
15	A0A7S1WJC0	289.0	276	12	47	217	[376, 377, 378, 379, 380, 381, 476, 510, 511, 512, 513, 514, 515]	[PRO, ALA, CYS, PRO, TYR, SER, GLY, ARG, ARG, LEU, ALA, SER, GLY]	13
16	A0A1Q9D410	287.0	254	13	44	197	[369, 370, 371, 373, 374, 375, 376, 377, 378, 379, 380, 381, 475, 476, 510, 511, 512, 513, 514, 515, 575, 576, 577, 578, 579, 580, 585, 586, 587, 588, 589, 600, 602]	[SER, ALA, PRO, PRO, GLY, THR, ALA, SER, CYS, PRO, TYR, SER, SER, SER, GLY, ARG, ARG, LEU, ALA, SER, ASP, ASN, ASN, PRO, ASP, ALA, ILE, ASN, ALA, ASP, LEU, GLY, GLY]	33
17	A0A812U275	288.0	250	17	37	196	[369, 370, 371, 374, 375, 376, 377, 378, 379, 380, 381, 382, 475, 476, 510, 511, 512, 513, 514, 515, 574, 575, 576, 577, 578, 579, 580, 584, 585, 586, 587, 588, 595, 596, 599, 600, 601, 602]	[SER, ALA, PRO, GLY, THR, ALA, ALA, CYS, PRO, TYR, SER, PHE, SER, SER, GLY, ARG, ARG, LEU, ALA, GLY, MET, ASP, ASN, ASN, ALA, THR, ASP, ASP, MET, ASN, ALA, ASP, GLY, GLY, ASN, GLY, ARG, GLY]	38
18	A0A812XZF5	285.0	253	14	33	206	[369, 370, 371, 372, 373, 377, 378, 379, 380, 381, 475, 476, 510, 511, 512, 513, 514, 515, 575, 576, 577, 578, 579, 580, 584, 585, 586, 587, 588, 589, 600, 601]	[SER, VAL, THR, PRO, PRO, GLY, CYS, PRO, TYR, ASN, SER, SER, GLY, ARG, ARG, LEU, ALA, GLU, ASP, ALA, ASP, PRO, THR, ASP, ALA, VAL, ASN, ALA, ASP, LEU, GLY, ARG]	32
19	A0A812JW14	288.0	255	17	30	208	[371, 373, 374, 375, 376, 377, 378, 379, 380, 381, 475, 476, 510, 511, 512, 513, 514, 515, 575, 576, 577, 578, 579, 580, 584, 585, 586, 587, 588, 589, 600, 601, 602]	[PRO, PRO, GLY, THR, ALA, THR, CYS, PRO, TYR, ASN, SER, SER, GLY, ARG, ARG, LEU, ALA, SER, ASP, HIS, ASN, ALA, SER, ASP, ASP, MET, ASN, ALA, ASP, LEU, GLY, ARG, GLY]	33
20	A0A812RYT3	288.0	251	17	33	201	[364, 369, 370, 371, 373, 374, 375, 376, 377, 378, 379, 380, 381, 382, 475, 476, 510, 511, 512, 513, 514, 515, 574, 575, 576, 577, 578, 579, 580, 584, 585, 586, 587, 588, 600, 601, 602]	[PRO, SER, ASN, PRO, PRO, GLY, THR, PRO, ALA, CYS, PRO, TYR, ASN, PHE, SER, SER, GLY, ARG, ARG, LEU, ALA, GLY, SER, ASP, TYR, ASN, ALA, SER, ASP, ASP, ILE, THR, ALA, ASP, GLY, ARG, GLY]	37
21	A0A812ULI5	288.0	248	20	33	195	[369, 370, 371, 373, 374, 375, 376, 377, 378, 379, 380, 381, 475, 476, 510, 511, 512, 513, 514, 515, 574, 575, 576, 577, 578, 579, 580, 584, 585, 586, 587, 588, 589, 595, 599, 600, 601, 602, 603, 623]	[SER, ASN, PRO, PRO, GLY, THR, PRO, ALA, CYS, PRO, TYR, ASN, SER, SER, GLY, ARG, ARG, LEU, ALA, GLY, THR, ASP, ASN, ASP, ALA, SER, ASP, ASN, MET, ASN, ALA, ASP, LEU, GLY, ASN, GLY, ARG, GLY, LEU, LEU]	40
22	A0A812TUP3	288.0	253	14	36	203	[369, 370, 371, 373, 374, 375, 376, 377, 378, 379, 380, 381, 475, 476, 510, 511,	[SER, ASN, PRO, PRO, GLY, SER, PRO, ALA, CYS, PRO, TYR, ASN, SER, SER,	35

	Uniprot ID	360 < Residue Count < 800	Len of Column Values no. of > 70	80>Len of Column Values no. of > 70	90>Len of Column Values no. of > 80	Len of Column Values no. of > 90	Column Values	Residue	Len of Column Values no. of < 70
							512, 513, 514, 515, 574, 575, 576, 577, 578, 579, 580, 584, 585, 586, 587, 588, 589, 600, 602]	GLY, ARG, ARG, LEU, ALA, SER, THR, ASP, ASN, ASP, ALA, SER, ASP, GLY, MET, ASN, ALA, ASP, LEU, GLY, GLY]	
23	F0XWW1	315.0	286	11	64	211	[475, 476, 477, 478, 479, 480, 481, 498, 499, 500, 501, 502, 503, 504, 505, 506, 507, 508, 509, 510, 511, 512, 513, 514, 515, 516, 517, 689, 690]	[HIS, ASP, ASP, ASP, ALA, HIS, ARG, LYS, LEU, ALA, TRP, THR, ASN, GLY, LYS, GLN, ARG, PRO, ASN, GLY, ARG, ARG, LYS, LEU, ALA, GLY, GLY, ASP, ALA]	29
24	A0A7S4GBX2	240.0	232	12	38	182	[472, 473, 474, 515, 516, 517, 518, 689]	[THR, THR, VAL, SER, SER, GLU, GLN, SER]	8
25	A0A7S4GBX6	240.0	230	7	50	173	[410, 472, 473, 474, 515, 516, 517, 518, 602, 689]	[GLY, THR, THR, VAL, SER, SER, GLU, GLN, GLY, SER]	10
26	A0A7S2U9R3	271.0	258	8	28	222	[403, 404, 405, 406, 473, 474, 475, 514, 515, 516, 517, 518, 519]	[ARG, ASP, VAL, SER, PRO, PRO, PRO, GLY, ILE, SER, ALA, GLY, SER]	13
27	A0A7S0IBT1	352.0	278	27	88	163	[425, 426, 428, 429, 477, 478, 479, 480, 501, 502, 503, 504, 505, 506, 507, 508, 509, 510, 511, 512, 513, 514, 515, 516, 517, 518, 599, 600, 601, 602, 603, 604, 605, 606, 607, 608, 609, 610, 611, 612, 613, 614, 615, 616, 617, 618, 619, 620, 621, 622, 623, 761, 762, 763, 764, 765, 766, 767, 768, 769, 770, 771, 772, 773, 774, 775, 776, 777, 778, 779, 780, 781, 782, 783]	[GLU, GLY, ASN, SER, THR, ALA, THR, TYR, ASP, ALA, ARG, GLN, PRO, ALA, GLY, ARG, LYS, LEU, LEU, ALA, SER, GLY, GLY, THR, THR, TYR, GLY, GLU, GLU, ALA, TYR, PRO, GLY, LYS, GLY, ALA, PHE, GLY, LEU, GLY, LYS, VAL, GLY, GLU, THR, ALA, ASP, PRO, LEU, SER, LEU, GLY, PRO, ASN, GLY, ALA, THR, VAL, SER, GLY, THR, GLY, SER, ALA, SER, SER, ALA, ASN, VAL, TYR, ASN, ASP, ARG, GLN]	74
28	C1EG17	351.0	261	13	98	150	[423, 424, 425, 426, 427, 428, 429, 430, 431, 466, 467, 468, 469, 470, 471, 472, 473, 474, 475, 476, 477, 478, 479, 480, 501, 502, 503, 504, 505, 506, 507, 508, 509, 510, 511, 512, 513, 514, 515, 516, 517, 518, 519, 599, 600, 601, 602, 603, 604, 605, 606, 607, 608, 609, 610, 611, 612, 613, 614, 615, 616, 617, 618, 619, 620, 621, 622, 761, 762, 763, 764, 765, 766, 767, 768, 769, 770, 771, 772, 773, 774, 775, 776, 777, 778, 779, 780, 781, 782, 783]	[MET, LYS, GLN, GLY, GLN, THR, THR, LEU, ASN, ALA, HIS, THR, ALA, ASP, THR, ALA, ASN, GLY, PRO, THR, GLN, PRO, HIS, ALA, ASP, ALA, ARG, LEU, PRO, ALA, GLY, ARG, LYS, LEU, LEU, ALA, ALA, GLY, GLY, THR, THR, TYR, LEU, GLY, GLU, GLU, ALA, TYR, PRO, GLY, LYS, GLY, LYS, PHE, GLY, LEU, GLY, THR, VAL, GLY, GLY, VAL, GLU, GLN, PRO, LEU, SER, GLY, PRO, ASN, GLY, ALA, THR, VAL, SER, GLY, THR, GLY, ASP, ALA, SER, THR, ALA, ASN, VAL, TYR, ASN, ASP, ARG, GLN]	90
29	A0A1E7EZ89	330.0	295	10	61	224	[473, 474, 475, 476, 477, 489, 490, 491, 492, 493, 494, 495, 496, 497, 498, 499, 500, 501, 502, 503, 504, 505, 506, 507, 508, 509, 510, 511, 512, 513, 514, 515, 516, 517, 763]	[THR, GLY, LYS, GLY, THR, GLY, ARG, ARG, VAL, LEU, SER, LYS, TYR, GLY, LEU, LEU, ASP, LYS, LYS, ALA, ASP, GLY, THR, HIS, ARG, GLY, LEU, SER, ASP, ASP, ASP, GLY, PRO, ASP, ASN]	35
30	A0A7S2XUT0	292.0	267	7	16	244	[472, 473, 474, 475, 478, 510, 511, 512, 513, 514, 515, 516, 517, 601, 603, 621, 623, 760, 761, 762, 763, 764, 765, 766, 767]	[SER, ASP, TYR, ARG, ARG, ARG, LEU, ALA, GLU, SER, GLU, ILE, ILE, ASP, SER, ASN, GLY, ARG, LEU, ARG, HIS, PHE, GLU, ARG, VAL]	25
31	A0A7S2V5F9	291.0	262	4	21	237	[472, 473, 474, 475, 476, 477, 478, 510, 511, 512, 513, 514, 515, 516, 517, 600, 601, 602, 603, 621, 622, 623, 759, 760, 761, 762, 763, 764, 765]	[ALA, SER, THR, THR, GLU, THR, HIS, ARG, LEU, LEU, ALA, GLY, ASP, SER, THR, ARG, ASP, LEU, SER, ASN, LEU, GLY, ARG, ARG, LEU, ARG, PHE, VAL, GLU]	29
32	A0A448ZPF4	342.0	271	6	45	220	[402, 405, 475, 476, 477, 478, 479, 480, 481, 482, 483, 484, 485, 486, 487, 488, 489, 490, 491, 492, 493, 494, 495, 496, 497, 498, 499, 500, 501, 502, 503, 504, 505, 506, 507, 508, 509, 510, 511, 512, 513, 514, 515, 516, 759, 760, 761, 762, 763, 764, 765, 766, 767, 768, 769, 770, 771, 787, 788, 789, 790, 791, 792, 793, 794, 795, 796, 797, 798, 799, 800]	[VAL, PRO, GLY, LEU, GLU, PRO, GLU, SER, GLY, ASP, ALA, GLU, SER, TYR, HIS, ASP, GLY, HIS, THR, GLU, GLY, VAL, LYS, ASP, PHE, ALA, ASN, MET, THR, ASP, ASP, GLU, VAL, ASP, SER, ARG, PHE, LEU, ALA, ALA, GLU, ASP, HIS, GLU, ARG, GLN, LEU, HIS, GLU, HIS, GLU, PRO, VAL, THR, GLY, GLY, ASP, GLU, SER, GLU, LEU, THR, ALA, GLU, GLN, LEU, ARG, HIS, LEU, ARG, THR]	71
33	A0A7S2M259	336.0	288	16	56	216	[474, 475, 476, 477, 478, 479, 480, 481, 482, 483, 509, 510, 512, 513, 514, 515, 516, 517, 518, 519, 773, 774, 775, 776, 777, 778, 779, 780, 781, 782, 783, 784, 785, 786, 787, 788, 789, 790, 791, 792, 793, 794, 795, 796, 797, 798, 799, 800]	[ALA, GLU, ARG, PRO, GLU, TRP, ALA, ASN, ARG, ASP, LEU, ALA, ALA, ALA, GLY, GLU, LYS, VAL, ARG, GLY, PRO, ALA, ALA, GLU, ALA, ALA, PRO, GLU, GLU, ALA, PRO, ALA, ALA, GLU, GLU, ALA, PRO, ALA, ALA, GLU, ASP, VAL, ALA, ALA, GLU, VAL, ALA, ARG]	48
34	K0RA12	295.0	273	3	20	250	[475, 476, 477, 478, 479, 480, 481, 482, 483, 509, 510, 512, 513, 514, 515, 516, 517, 518, 761, 762, 763, 764]	[PRO, ARG, PRO, ASP, TRP, ALA, GLN, ARG, GLU, ARG, ASP, LEU, ALA, GLU, GLY, LYS, VAL, ARG, ALA, THR, ARG, GLU]	22
35	K0R419	297.0	272	3	21	248	[475, 476, 477, 478, 479, 480, 481, 482, 483, 509, 510, 512, 513, 514, 515, 516, 517, 518, 760, 761, 762, 763, 764, 765, 766]	[PRO, ARG, PRO, ASP, TRP, ALA, GLN, ARG, GLU, ARG, ASP, LEU, ALA, GLU, GLY, LYS, VAL, ARG, SER, ARG, LYS, ASN, LEU, ARG, ALA]	25
36	W8VUA7	294.0	273	5	18	250	[477, 478, 480, 481, 482, 483, 509, 510, 511, 512, 513, 514, 515, 516, 517, 518, 519, 760, 761, 762, 763]	[PRO, SER, ALA, ALA, ARG, TYR, LEU, THR, ASP, ALA, SER, ALA, GLU, GLU, GLU, GLN, PRO, LEU, ARG, ALA, MET]	21
37	W8VYH0	293.0	280	12	33	235	[480, 481, 482, 483, 509, 510, 512, 513, 514, 515, 516, 517, 518]	[ALA, ASN, ARG, ASP, LEU, ALA, GLY, ALA, GLY, GLU, SER, VAL, PRO]	13

	Uniprot ID	360 < Residue Count < 800	Len of Column Values no. of > 70	80>Len of Column Values no. of > 70	90>Len of Column Values no. of > 80	Len of Column Values no. of > 90	Column Values	Residue	Len of Column Values no. of < 70
38	A0A7S4MK07	307.0	281	4	28	249	[377, 472, 473, 474, 475, 476, 477, 478, 479, 501, 502, 503, 504, 505, 506, 507, 508, 509, 510, 511, 512, 513, 514, 515, 516, 761]	[PRO, GLN, THR, ASN, SER, SER, SER, SER, SER, SER, SER, MET, SER, GLY, ARG, ARG, GLU, LEU, ALA, GLY, ASP, TYR, ALA, GLU, LYS]	26
39	A0A7S3AZA3	303.0	273	15	31	227	[367, 368, 376, 377, 378, 379, 472, 473, 474, 475, 476, 477, 478, 479, 501, 502, 507, 508, 509, 510, 511, 512, 513, 514, 515, 516, 619, 759, 760, 761]	[ASP, ALA, GLY, PRO, GLY, ASP, ASP, GLY, GLY, SER, SER, HIS, GLY, GLY, ASP, SER, SER, ALA, ARG, ARG, GLN, LEU, ALA, GLY, SER, ASP, ALA, SER, ARG, LYS]	30
40	K0SZQ0	297.0	274	11	37	226	[478, 479, 500, 501, 502, 503, 504, 505, 506, 507, 508, 509, 510, 511, 512, 513, 514, 515, 516, 517, 763, 764, 765]	[GLU, TRP, ALA, THR, ARG, ASP, LEU, ALA, GLU, SER, ASP, ASP, ASN, ASP, ASP, ASP, ASP, GLU, ASN, VAL, ARG, PHE, ALA]	23
41	X5I1I9	300.0	270	1	7	262	[472, 473, 474, 475, 476, 477, 512, 513, 514, 515, 760, 761, 762, 763, 764, 765, 766, 767, 768, 769, 770, 771, 772, 794, 795, 796, 797, 798, 799, 800]	[SER, VAL, ASN, ASN, ASN, LEU, PRO, GLN, ASN, GLN, ARG, ARG, LEU, GLY, GLY, HIS, ASP, HIS, HIS, HIS, HIS, HIS, HIS, GLY, HIS, ASP, HIS, ALA, ASP, HIS]	30
42	K8EQL3	298.0	284	3	14	267	[472, 473, 474, 475, 476, 511, 512, 513, 514, 515, 516, 517, 763, 764]	[SER, ALA, SER, HIS, ARG, LYS, LEU, LEU, ALA, GLU, GLY, ALA, ALA, PRO]	14
43	A0A1E7EUG8	320.0	278	16	99	163	[476, 477, 478, 479, 480, 481, 483, 484, 497, 498, 499, 500, 501, 502, 503, 504, 505, 506, 507, 508, 509, 510, 511, 512, 513, 514, 515, 516, 517, 760, 761, 762, 763, 764, 765, 766, 767, 768, 769, 770, 771, 772]	[ALA, ALA, ASP, ASP, ALA, ALA, HIS, ASP, ASP, ASP, GLY, HIS, THR, ASP, ASP, ALA, GLY, GLU, GLY, ASP, SER, ARG, ARG, GLN, LEU, ALA, GLY, ASP, ALA, ARG, GLY, ARG, GLY, LEU, ARG, LEU, ARG, LYS, ASN, ASN, LYS, ASN]	42
44	A0A7S2LSL9	290.0	272	5	15	252	[472, 473, 474, 475, 476, 477, 509, 510, 511, 512, 513, 514, 515, 516, 760, 761, 762, 763]	[ASP, ASP, HIS, GLU, GLY, HIS, ASP, HIS, ARG, ALA, LEU, ALA, GLU, ASP, ARG, ASP, ARG, TYR]	18
45	A0A7S0KBZ5	296.0	272	1	6	265	[472, 511, 512, 513, 514, 515, 516, 761, 762, 763, 764, 765, 766, 767, 768, 769, 770, 771, 772, 794, 795, 796, 797, 798]	[GLU, ARG, ARG, LEU, ALA, GLU, GLY, ARG, LYS, THR, GLU, GLU, LYS, THR, ASP, LYS, ARG, VAL, PHE, ARG, MET, VAL, TYR, ASP]	24
46	A0A7S3L7P6	308.0	265	4	12	249	[471, 472, 473, 474, 475, 476, 477, 478, 479, 480, 481, 482, 483, 484, 497, 498, 499, 500, 501, 502, 503, 504, 505, 506, 507, 508, 509, 510, 511, 512, 513, 514, 515, 516, 760, 761, 762, 763, 764, 765, 766, 767, 768]	[VAL, ALA, GLU, VAL, HIS, GLU, GLY, GLU, GLU, GLY, ASN, THR, ASN, ALA, VAL, GLU, ASP, THR, GLN, VAL, ALA, ASN, THR, GLY, HIS, ARG, LYS, LEU, SER, GLY, GLY, ASP, ASP, GLU, ARG, ARG, LEU, ARG, ARG, HIS, ALA, TYR, ASP]	43
47	A0A7S0UKW0	313.0	278	11	15	252	[392, 396, 397, 399, 400, 467, 468, 469, 470, 471, 472, 473, 474, 475, 476, 477, 478, 479, 501, 503, 505, 506, 507, 508, 509, 510, 511, 512, 513, 514, 515, 516, 517, 761, 762]	[LEU, ALA, ASN, THR, GLU, ALA, SER, HIS, ASP, ASP, HIS, ARG, THR, LEU, MET, GLU, GLU, SER, MET, GLY, THR, PRO, ILE, HIS, ASP, ARG, ARG, LYS, LEU, ALA, GLY, LYS, ALA, ARG, ASN]	35
48	A0A7S1UU45	279.0	270	5	6	259	[402, 472, 473, 474, 475, 476, 513, 514, 515]	[ILE, VAL, GLY, ALA, ASP, THR, LEU, ASP, GLY]	9
49	A0A7S0CDJ9	280.0	268	2	17	249	[472, 473, 474, 475, 476, 511, 512, 513, 514, 515, 516, 517]	[GLU, ILE, ALA, HIS, ARG, ARG, LYS, LEU, ALA, GLY, LYS, ALA]	12
50	A0A7S4JEJ7	296.0	277	4	5	268	[389, 399, 400, 401, 402, 403, 473, 474, 475, 476, 477, 511, 512, 513, 514, 515, 516, 762, 763]	[ILE, VAL, ASP, GLY, VAL, ALA, HIS, SER, SER, HIS, GLY, ARG, ARG, MET, ALA, ASP, ASP, GLY, ARG]	19
51	A0A6V2KCI5	290.0	271	2	22	247	[472, 473, 474, 475, 476, 477, 478, 510, 511, 512, 513, 514, 515, 516, 759, 760, 761, 762, 763]	[ALA, ASP, GLU, LYS, ASP, ASP, ASP, HIS, ARG, ARG, LEU, ALA, GLU, GLU, ARG, LEU, LEU, ARG, ALA]	19
52	A0A7S4S7Z7	290.0	270	1	11	258	[471, 472, 473, 474, 475, 476, 477, 478, 510, 511, 512, 513, 514, 515, 516, 759, 760, 761, 762, 763]	[ALA, ALA, ASP, GLU, LYS, ASP, ASP, ASP, HIS, ARG, ARG, LEU, ALA, GLU, GLU, ARG, LEU, LEU, ARG, ALA]	20
53	A0A830HI75	303.0	281	2	6	273	[471, 472, 473, 474, 475, 476, 477, 478, 508, 509, 510, 511, 512, 513, 514, 515, 760, 761, 762, 763, 764, 765]	[LEU, MET, TYR, ARG, LYS, ALA, ARG, ARG, SER, LEU, ASN, ALA, ALA, ALA, GLY, ALA, LYS, LEU, ILE, GLU, ASP, SER]	22
54	A0A7S2HVS4	288.0	267	2	22	243	[472, 473, 474, 475, 476, 477, 478, 509, 510, 511, 512, 513, 514, 515, 516, 759, 760, 761, 762, 763, 764]	[LEU, ASN, ALA, ASN, VAL, ASP, ASN, LEU, ASP, ARG, ARG, LEU, GLU, LEU, ASN, LEU, ARG, ALA, LEU, ARG, ALA]	21
55	A0A830HHC5	298.0	277	4	14	259	[472, 473, 474, 475, 476, 477, 478, 506, 507, 508, 509, 510, 511, 512, 513, 514, 515, 516, 759, 760, 761]	[THR, THR, ARG, ARG, ARG, ARG, ARG, SER, ARG, SER, MET, LEU, ALA, TYR, ASP, ASP, GLU, ASN, ARG, LEU, HIS]	21
56	A0A7S2BCN7	296.0	276	5	20	251	[471, 472, 473, 474, 475, 476, 477, 478, 506, 507, 508, 509, 512, 513, 514, 515, 516, 759, 760, 761]	[LYS, THR, THR, ARG, ARG, ARG, ARG, ARG, SER, ARG, SER, MET, LEU, ALA, GLU, LYS, GLU, ARG, LEU, HIS]	20
57	A0A7S0U0B0	260.0	251	3	23	225	[472, 473, 474, 475, 476, 514, 515, 516, 517]	[PRO, ALA, LYS, GLU, TRP, ASP, GLY, LYS, GLY]	9
58	A0A7S0R5U4	299.0	286	9	18	259	[472, 473, 474, 475, 476, 477, 512, 513, 514, 515, 516, 617, 618]	[ALA, HIS, ASP, ASP, SER, HIS, ASP, ASP, SER, HIS, SER, PRO, LEU]	13
59	A7UC72	286.0	264	2	17	245	[382, 472, 473, 474, 475, 476, 477, 478, 506, 507, 508, 509, 510, 511, 512, 513, 514, 515, 516, 759, 760, 761]	[PHE, ASN, SER, ALA, GLU, GLU, ASP, ASP, GLU, GLY, GLU, ALA, ASP, SER, ARG, ARG, LEU, ALA, LYS, ARG, ARG, VAL]	22

The following step generates a line graph of all models using the 'plotly.express' module. The link between the location and B Factor for various Uniprot IDs is depicted in Graph 27. To

distinguish between the various Uniprot IDs, the graph has distinct colours. The graph comprises 60 lines, each of which represents a distinct Uniprot ID.

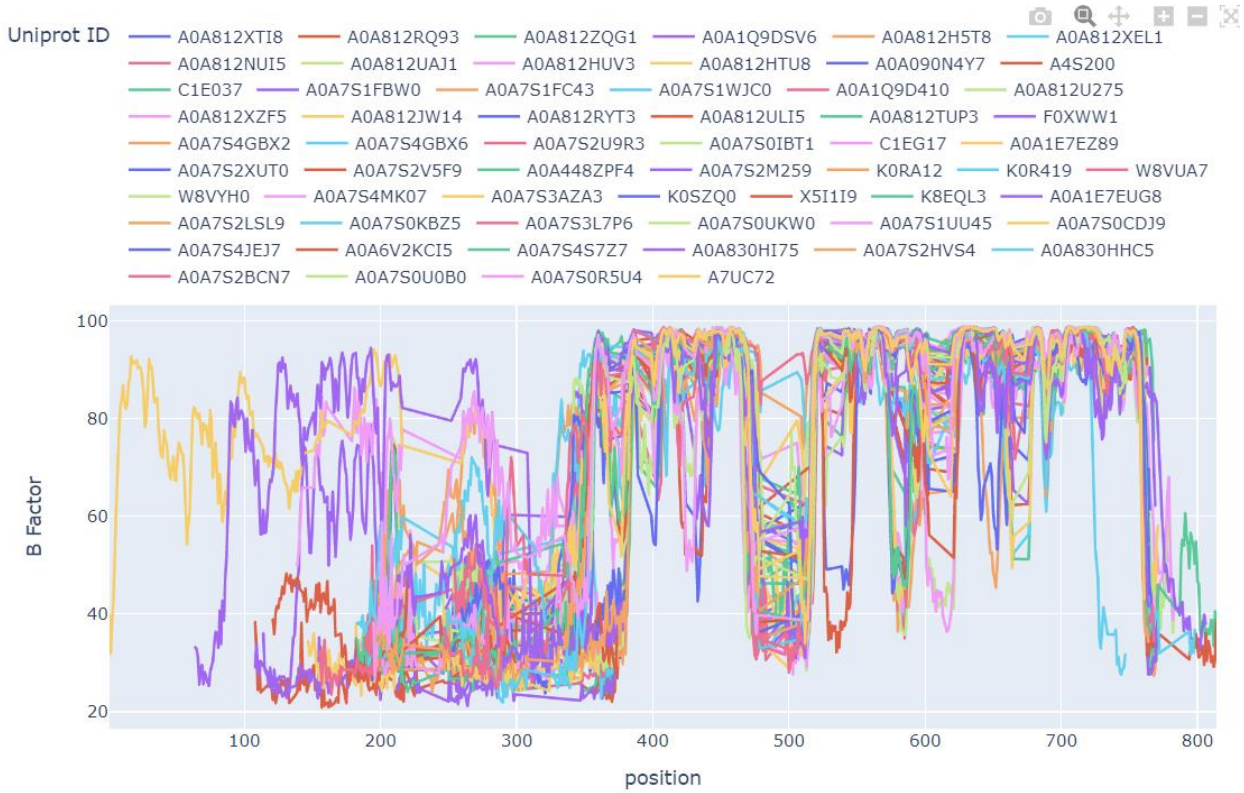
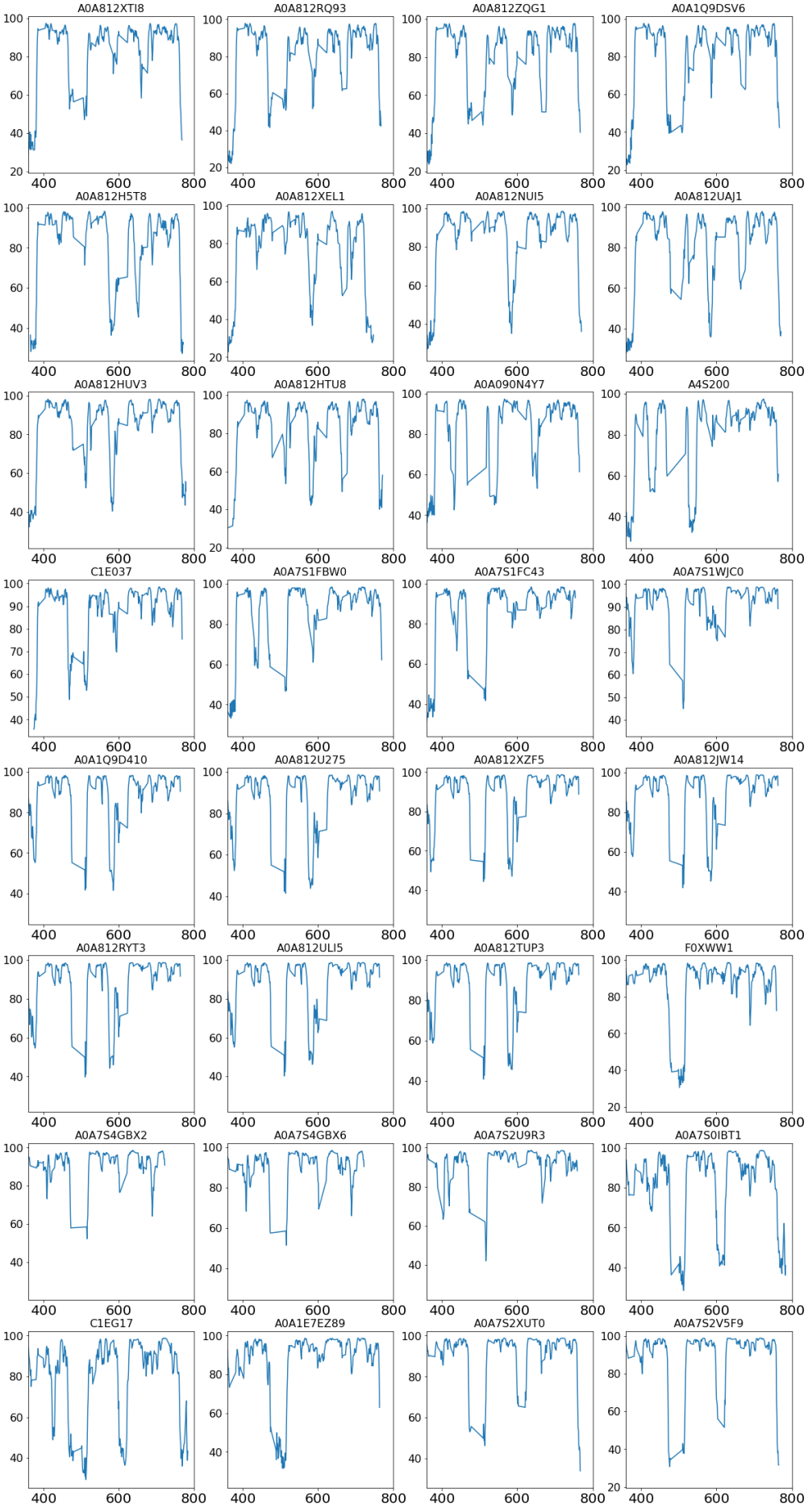


Figure 27. line graph of 60 DCA sequences

Finally, using the 'matplotlib' package, the code generates a set of line graphs organized in a 16x4 grid as illustrated in Figure 28. The B Factor of a particular Uniprot ID is plotted on each graph. Each graph has an x-axis with a range of 360 to 800.


A0A812XTI8
A0A812RQ93
A0A812ZQG1
A0A1Q9DSV6
A0A812H5T8
A0A812XEL1
A0A812NUI5
A0A812UAJ1
A0A812HUV3
A0A812HTU8
A0A090N4Y7
A4S200
C1E037
A0A7S1FBW0
A0A7S1FC43
A0A7S1WJC0
A0A1Q9D410
A0A812U275
A0A812XZF5
A0A812JW14
A0A812RYT3
A0A812ULI5
A0A812TUP3
F0XWW1
A0A7S4GBX2
A0A7S4GBX6
A0A7S2U9R3
A0A7S0IBT1
C1EG17
A0A1E7EZ89
A0A7S2XUT0
A0A7S2V5F9

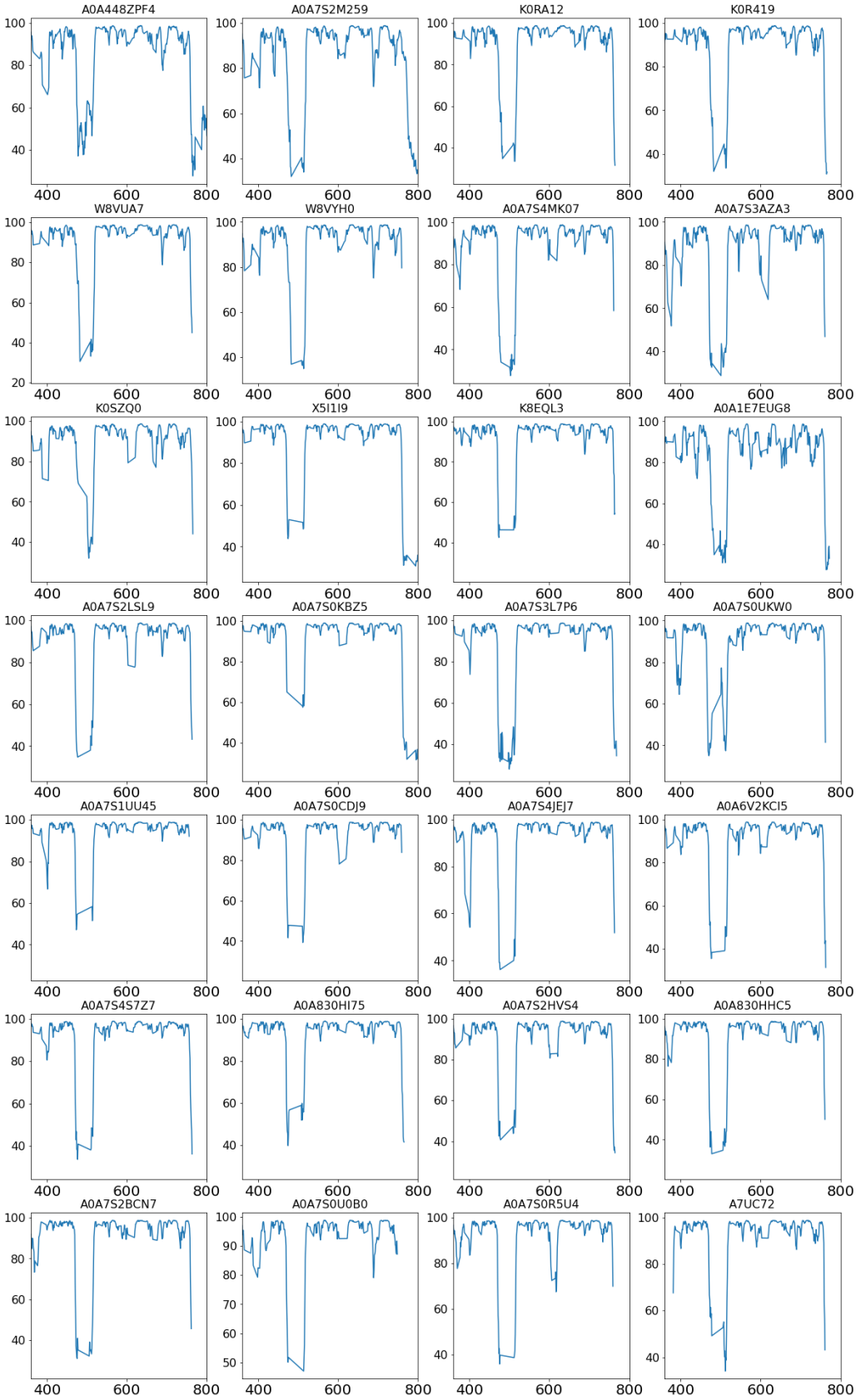


Figure 28. separate plot for each 60 DCA

The code for this section is in the Appendix 2. The same computing process has been implemented in the sections 5.6 and 5.7, as well.

4.6 Analysis of ACA

Subsequently, the analysis narrows down to the bacteria ACA from 590 to 494 (& 466) with AF models available realigned, as:

```
Length of alignment file 466 records
Unique seq lengths in MSA: {739}
MSA columns in logofile 739
```

I reduced the 494 set to 466 by taking out the sequences corresponding to the 20 worst models (% bad > 9.6) and 8 more which showed unusual features in the MSA and the fraction of bad residues in the range 6 to 9 %. I realigned the edited MSA which was shortened more in the MSA step than in removing gaps left by removed insertions. Then I made the logo and a new notebook to process this.

Most of the sequence abnormalities seem to have deletions near the start of the ACA domain, and when looking at the bad sequence graphics (Red box contents) in the 494 and 466 figures 29 and 30, the unedited set of models has more dots of pLDDT <50 in that region of the MSA.

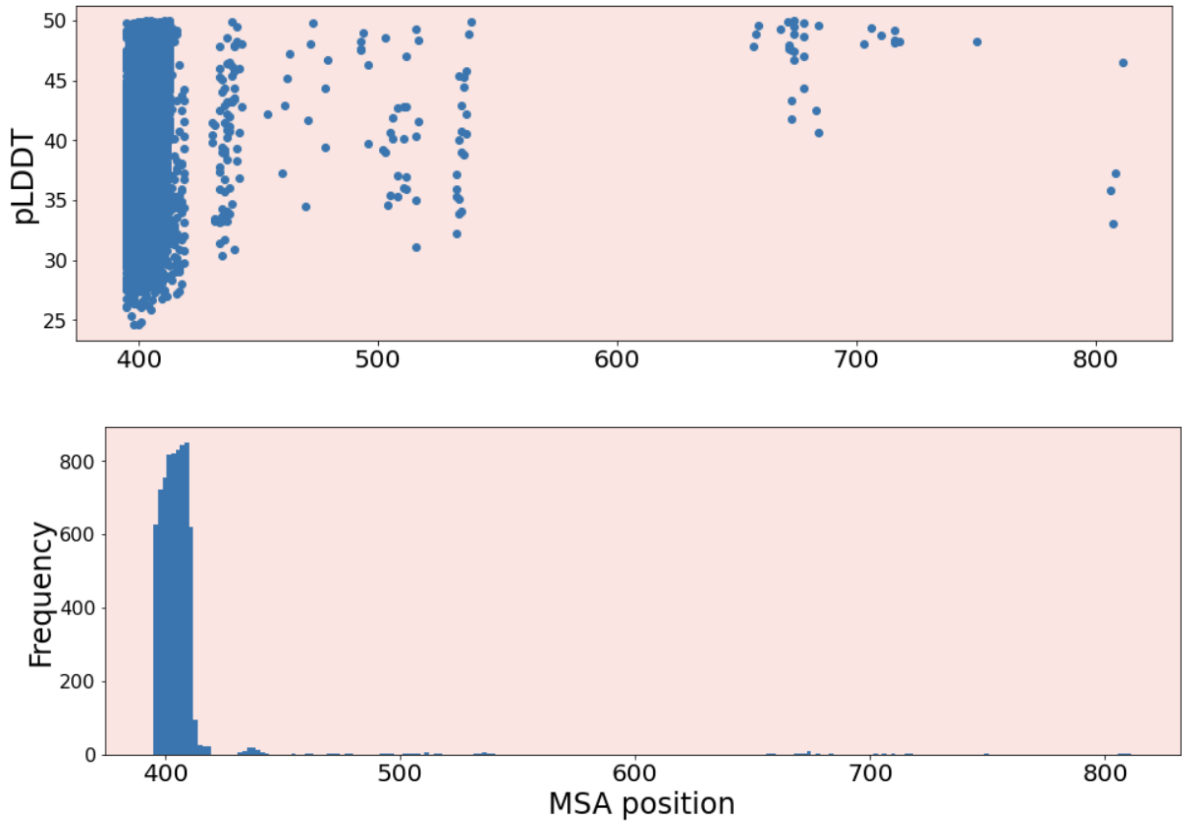


Figure 29. ACA 494

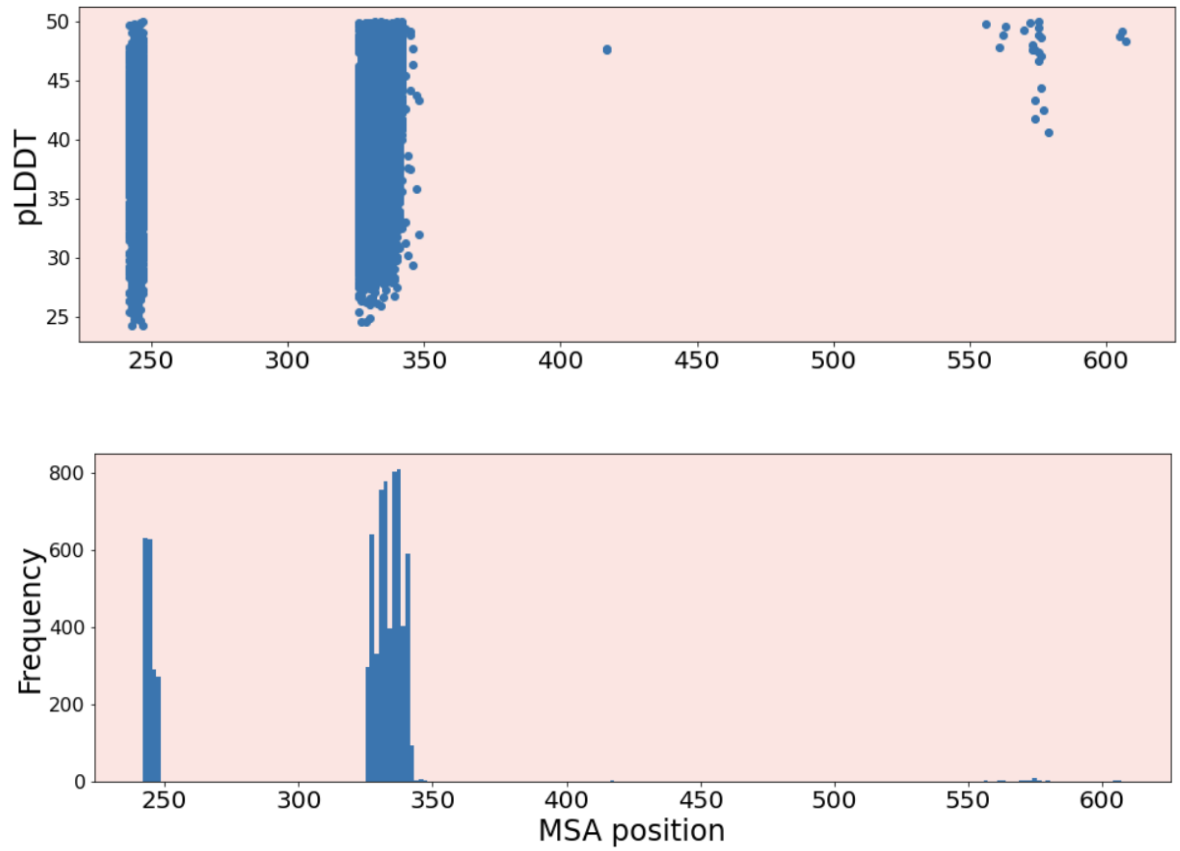


Figure 30. ACA 466

We can also see improvements in the mean pLDDT curve and the MSA quality, in weight curves in figure 31.

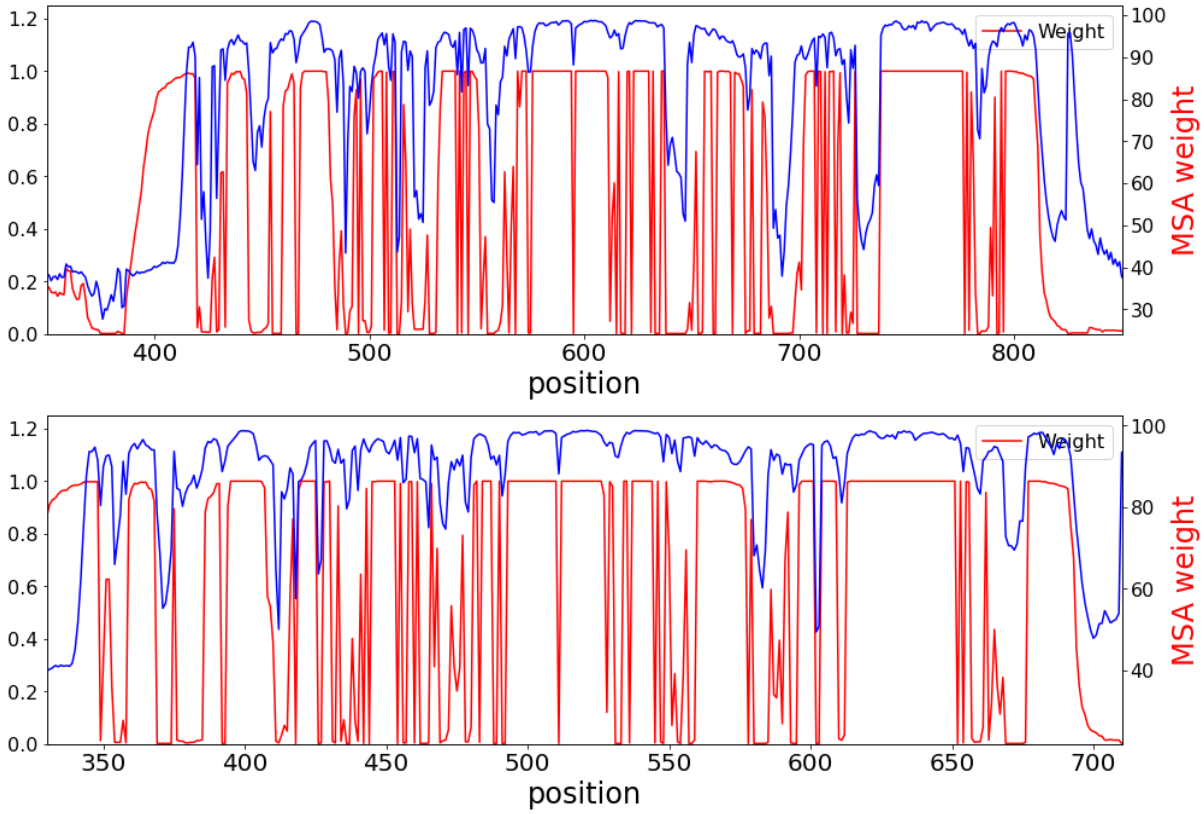


Figure 31. ACA 494 (top) and ACA 466 (bottom)

Likewise, the hairball diagram of all pLDDT curves in one graph in figure 32.

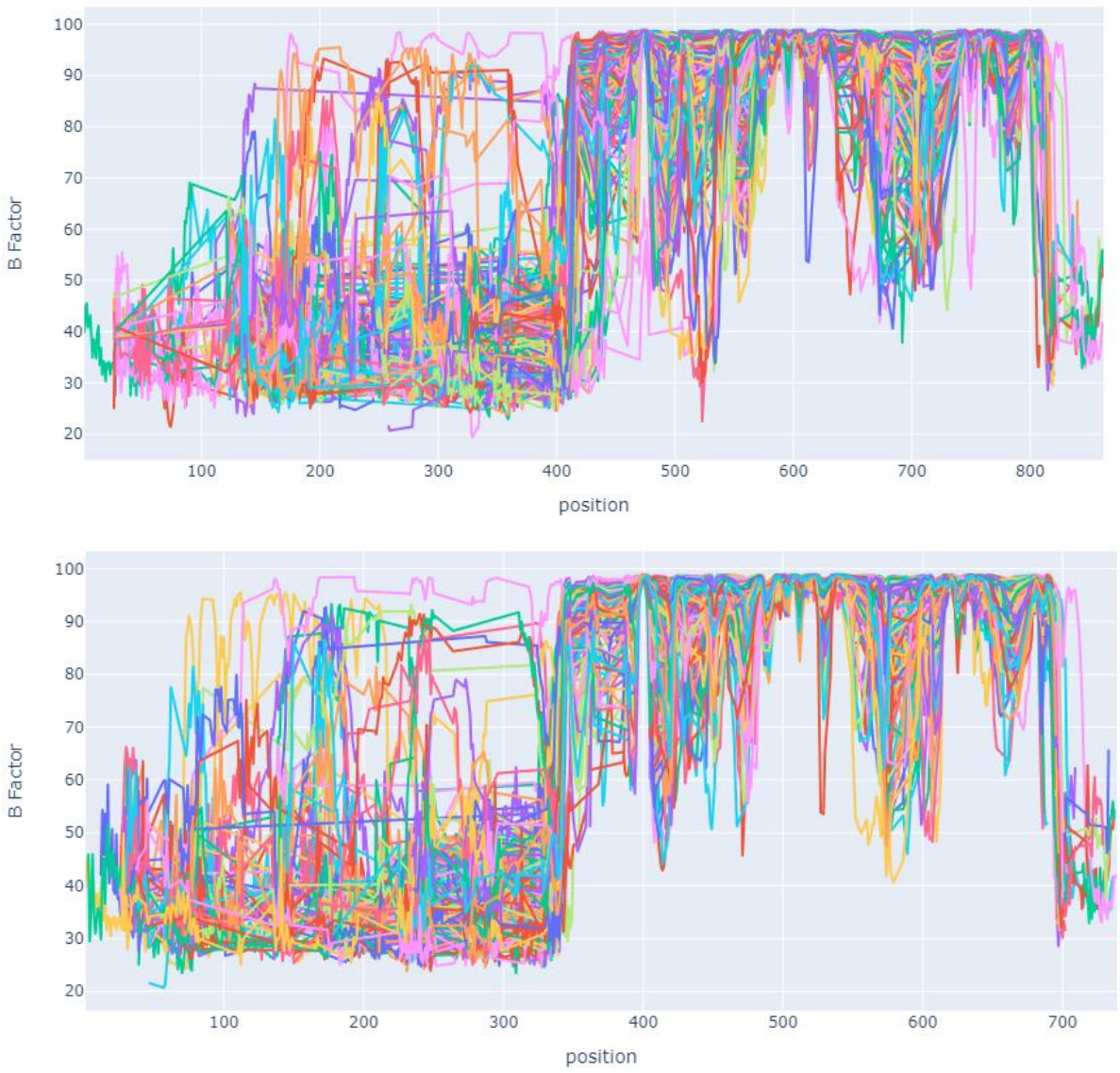


Figure 32. ACA 494 (top) and ACA 466 (bottom)

The codes for this section have been saved in the GitHub repository through the following link:

- https://github.com/hoseindaneshpour/ACA-466-MT/blob/main/ACA%20466%20MT%20(plots%20fine-tuned).py
- https://github.com/hoseindaneshpour/ACA-494-MT/blob/main/ACA%20494%20MT%20(plots%20fine-tuned).py

4.7 Analysis of DCA bacteria

The analysis of this section relies on 146 proteomes of UniProt, with the start domain dca_start = 83. Plot 33 depicts the Histogram of pLDDT of the whole MSA with length 373.

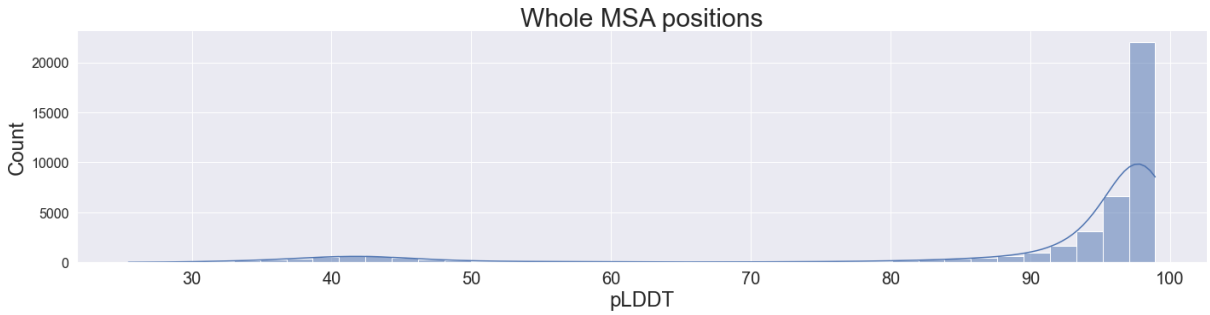


Figure 33. pLDDT of the whole MSA

Plot 34 illustrates the weighs more than 0.7 considering the whole MSA.

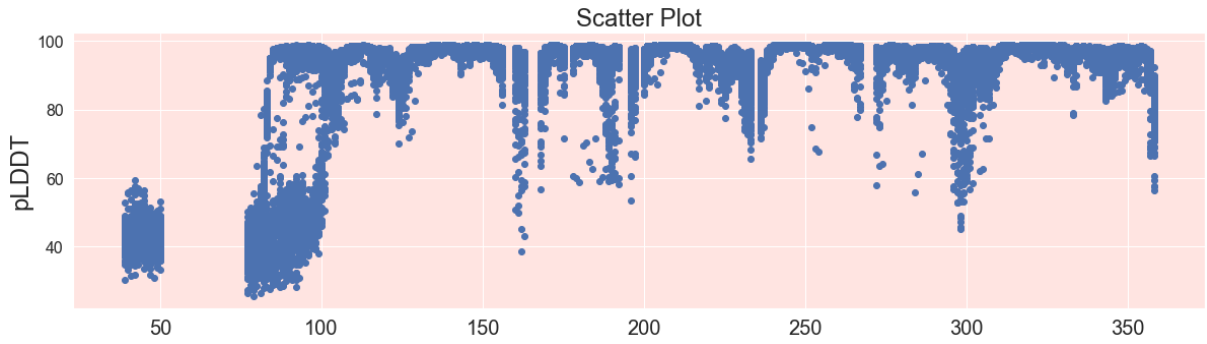


Figure 34. weights more than 0.7 with whole B Factor

The overlapping plot of MSA and pLDDT is in figure 35.

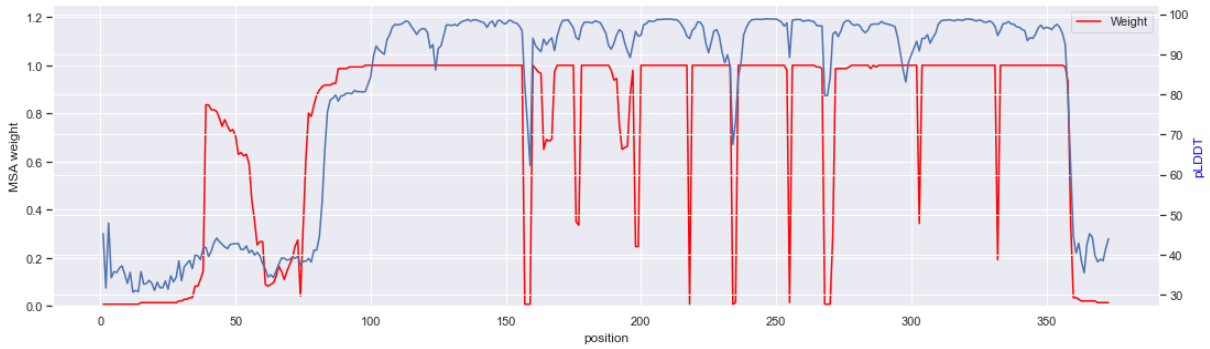


Figure 35. Mean pLDDT versus MSA weight

Figure 36 depicts the line plot of 146 sequences.

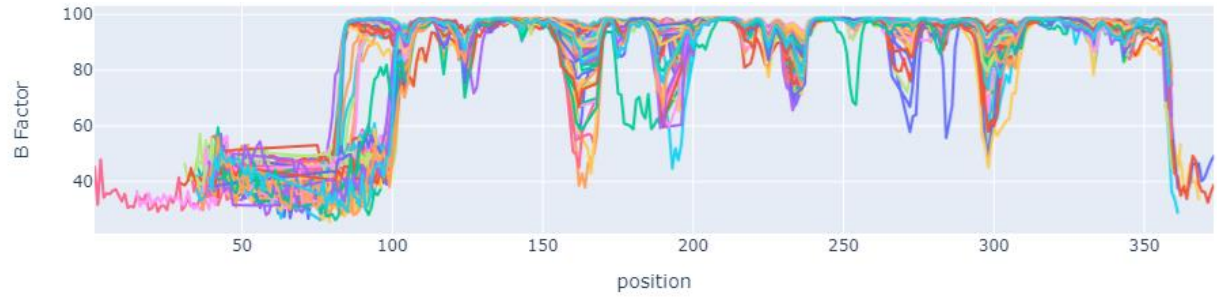


Figure 36. DCA bacteria146 sequences

The codes for this section have been saved in the GitHub repository through the following link.

- https://github.com/hoseindaneshpour/DCAb5-MT/blob/main/DCAb5%20MT.py

By exploring the bacterial DCA it was discovered that its disulfides are three of the same as in eukaryotes (B, D, E) but it was missing two. Table 5 is the Disulfides analysis of 146 sequences.

Table 5. Disulfides of 146 sequences

	ID	DS_B	DS_D	DS_E	DS_K	DS_L	No. of Disulfides
0	A0A1Y0CV22	0.0	0.0	1.0	1.0	0.0	2.0
1	E1SMY4	0.0	0.0	1.0	1.0	0.0	2.0
2	A0A1M5Z5J9	0.0	0.0	1.0	1.0	0.0	2.0
3	A0A6G8D3H9	0.0	0.0	0.0	1.0	0.0	1.0
4	U3A2Q2	0.0	0.0	0.0	1.0	0.0	1.0
...	...	...	...	...	...	...	...
142	A0A327YRR2	1.0	0.2	0.0	1.0	1.0	3.2
143	A0A1G8T7Z6	1.0	1.0	0.0	1.0	1.0	4.0
144	A0A1G7GQA1	1.0	1.0	0.0	1.0	1.0	4.0
145	A0A1P8UZE2	1.0	1.0	0.0	1.0	1.0	4.0
Column Sum		50.0	49.2	73.0	146.0	47.0	

Next, I developed some code to count the number of Signal Peptides and to explore the other features as well, such as disordered parts. The results are: Signal peptide in 121 proteins, 17 with no features, and 129 with some features. As a result, all 8 of the compositional bias features overlap with a disordered region. The compositional bias features are either of type Polar residues (6) or Basic and acidic residues (2). Most disordered regions are right after the signal peptide.

```
UniProt ID: A0A1G7GQA1, [('Signal', 1, 21), ('Disordered', 24, 78), ('CB', 'Polar residues', 63, 77)]
UniProt ID: A0A1G7P8L1, [('Signal', 1, 21), ('Disordered', 44, 67), ('CB', 'Polar residues', 49, 67)
UniProt ID: A0A4U0ZVH1, [('Signal', 1, 21), ('Disordered', 24, 68), ('CB', 'Polar residues', 24, 42)]
UniProt ID: A0A6I1EDA4, [('Signal', 1, 24), ('Disordered', 48, 75), ('CB', 'Polar residues', 61, 75)]
```

As present in figure 37, in the top left box with sequences with high mutual similarity, all sequences have 1 or 2 disulfides, the middle box has mainly 5 disulfides and the third box (bottom right) has mainly 4 disulfides.

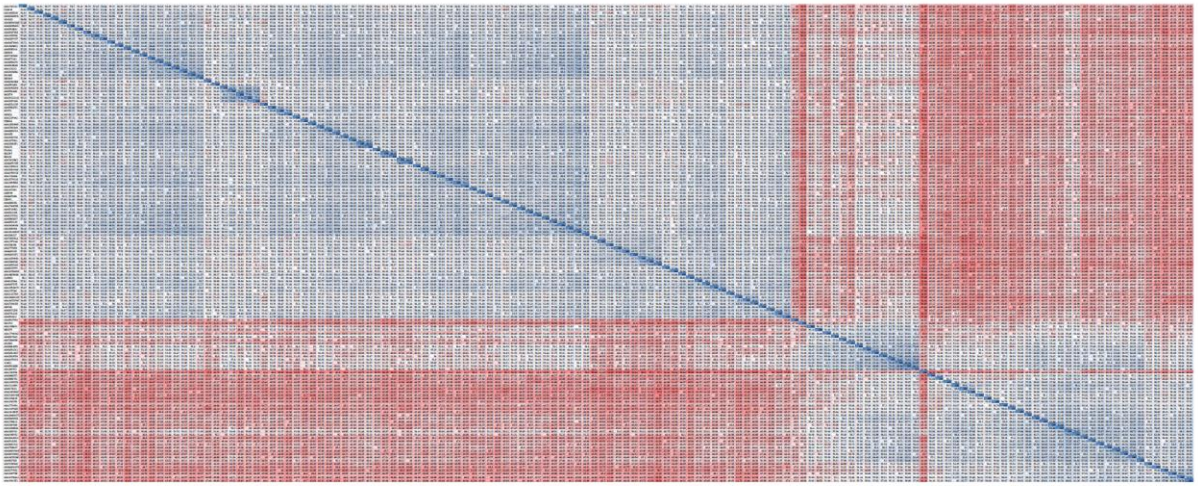


Figure 37. Overall view of PIM of 146

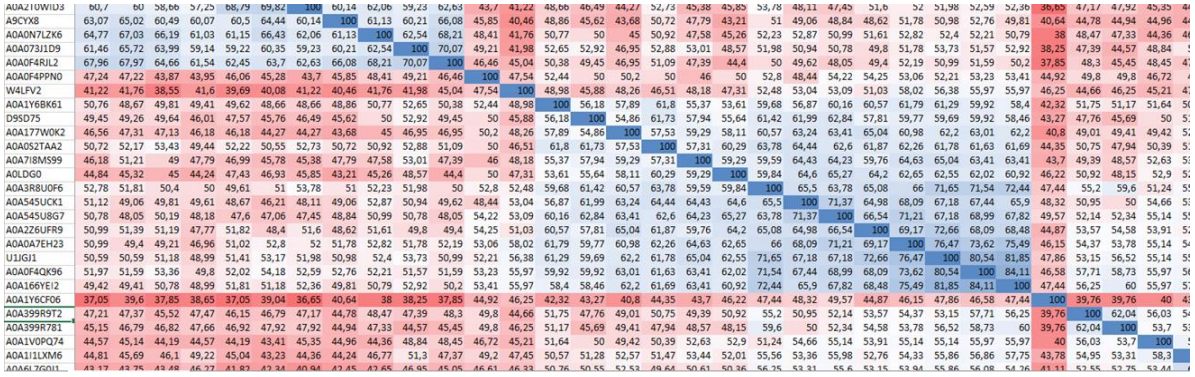


Figure 38. End of the first box, and the start of the third box of PIM

Based on figure 38 and table 6, the disulfide table is divided close to the way the excel PIM table was divided.

Table 6. End of the first box, and the start of the third box of PIM

A0A0N7LZK6	0.0	0.0	0.0	1.0	0.0	1.0
A0A073J1D9	0.0	0.0	0.0	1.0	0.0	1.0
A0A0F4RJL2	0.0	0.0	0.0	1.0	0.0	1.0
A0A0F4PPN0	1.0	1.0	1.0	1.0	0.0	4.0
W4LFV2	1.0	1.0	0.2	1.0	1.0	4.2
A0A1Y6BK61	1.0	1.0	1.0	1.0	1.0	5.0
D9SD75	1.0	1.0	1.0	1.0	0.0	4.0
A0A177W0K2	1.0	1.0	1.0	1.0	1.0	5.0
A0A0S2TAA2	1.0	1.0	1.0	1.0	0.0	4.0
A0A7I8MS99	1.0	1.0	1.0	1.0	1.0	5.0
A0LDG0	1.0	1.0	1.0	1.0	1.0	5.0
A0A3R8U0F6	1.0	1.0	1.0	1.0	1.0	5.0
A0A545UCK1	1.0	1.0	1.0	1.0	1.0	5.0
A0A545U8G7	1.0	1.0	1.0	1.0	1.0	5.0
A0A2Z6UFR9	1.0	1.0	1.0	1.0	1.0	5.0
A0A0A7EH23	1.0	1.0	1.0	1.0	1.0	5.0
U1JGJ1	1.0	1.0	1.0	1.0	1.0	5.0
A0A0F4QK96	1.0	1.0	1.0	1.0	1.0	5.0
A0A166YEI2	1.0	1.0	1.0	1.0	1.0	5.0
A0A1Y6CF06	1.0	1.0	0.2	1.0	1.0	4.2
A0A399R9T2	1.0	1.0	0.0	1.0	1.0	4.0
A0A399R781	1.0	1.0	0.0	1.0	1.0	4.0

4.8 Visualization of DCA short vs. long

Accordingly, the DCA long A0A1T4W5S3 and DCA short A0A0P1G5I0 were superimposed with ChimeraX, by deleting low-quality residues in the N-terminus: 1-43 of the short one and: 1-26 or the long one. Superimposition, MatchMaker with 3.0 Å threshold, and RMSD between 204 pruned atom pairs is 0.864 angstroms (across all 229 pairs: 2.945). Sequence lengths after deletion is: 239 short, and 259 long, and the percentage of aligned residues is (by

shorter) 85%. The N-terminal extra in the long DCA is 26-42 (by the domain definition based on the mean pLDDT of each residue). Looking at the sequence logos, the region in which the short DCAs have their specific conservation pattern is shown in figure 39.

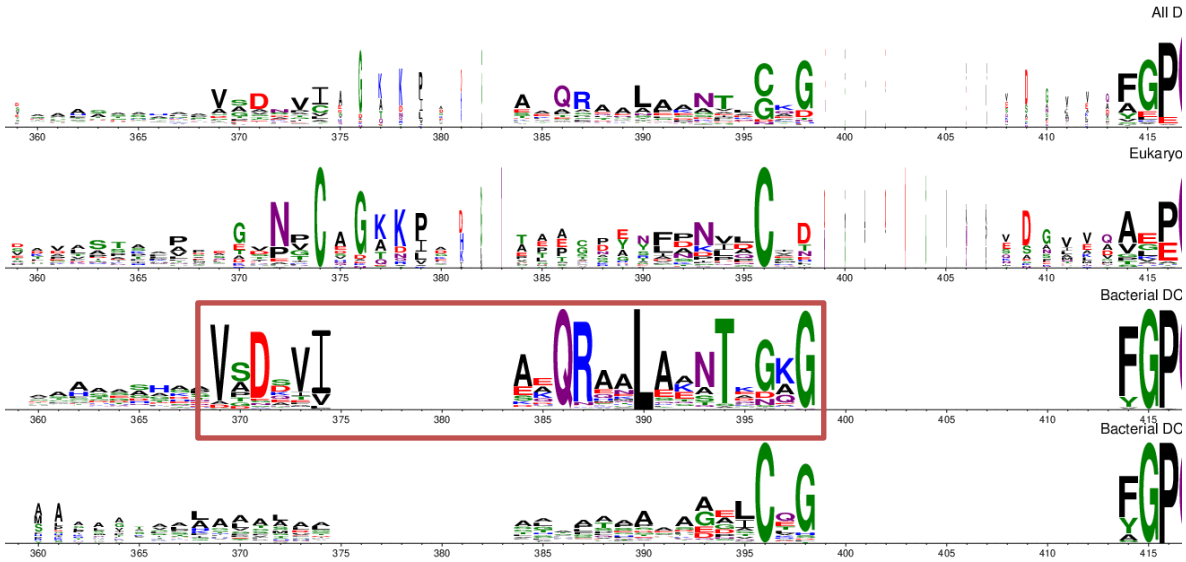


Figure 39. Bacteria DCA long form

The last G in the surrounded region is shared between long and short, and three residues before that are poorly conserved. The long-form-specific sequence would be between columns 369 "V" and 394 "T". In our prototype short DCA model, A0A0P1G5I0 this region is 28-44, as figure 40 (green alpha helix).

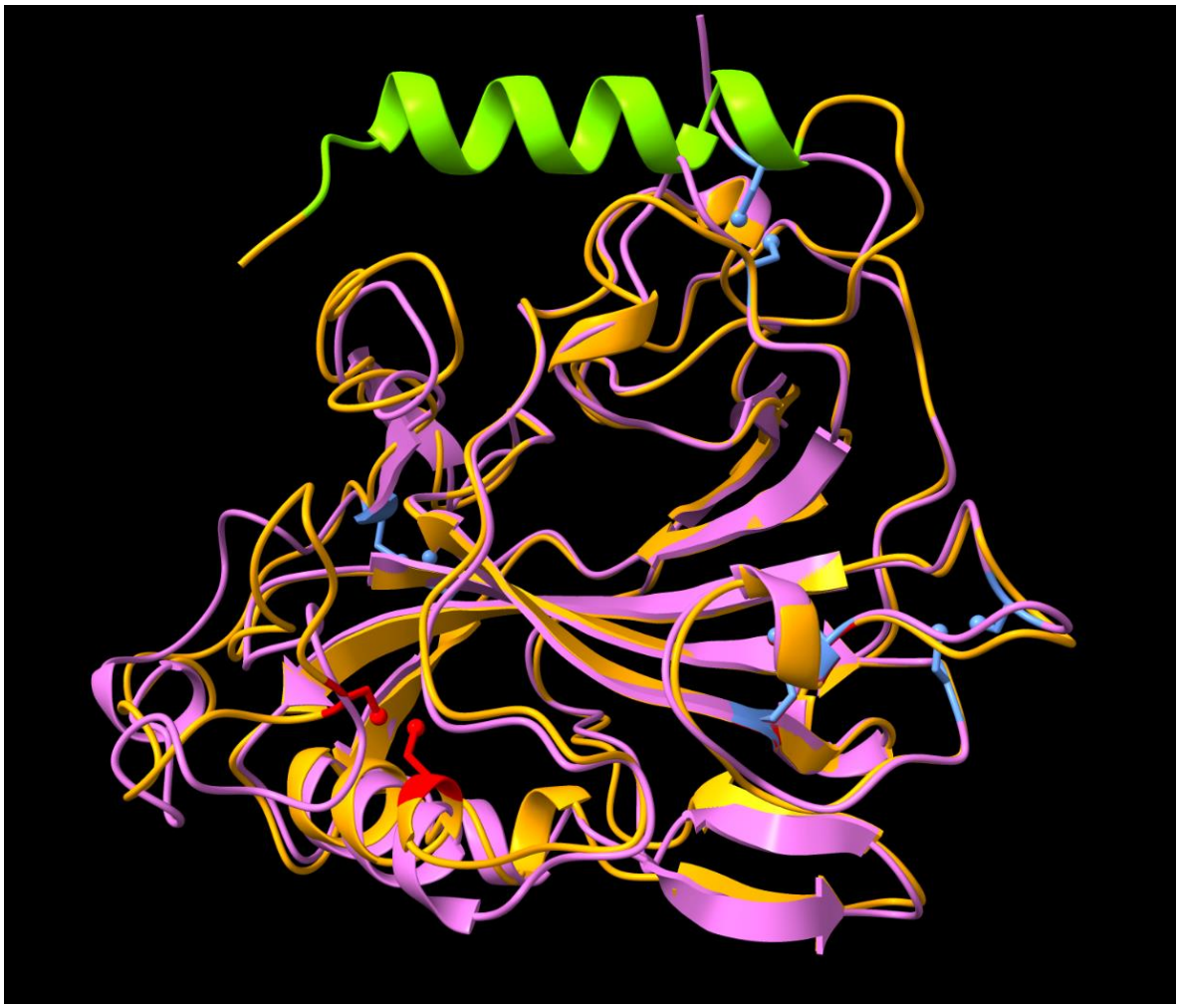


Figure 40. Long-form: orange, cysteines in red, short form: lilac, cysteines in blue. The conserved long-form-specific region in green

Showing the same image with coloring by conservation in 28-44 of the long DCA in figure 41.

Figure 41. long vs DCAb conservation, Red: high conservation, blue: low conservation

5 Discussion

5.1 Taxonomy analysis

Next, I extracted the data of taxonomy and linages parsing the JSON file (bacterial_ref_proteomes_9677_lineages). Technically, the science of taxonomy involves identifying, describing, and categorizing groupings of species according to their shared traits. Also, lineages are the successions of ancestors and offspring of a certain organism. They are frequently employed in the evaluation of the connections and evolutionary history of species. The resulting JSON data frame presents the following data: Id - description - taxonomy - modified - proteomeType - strain-components - citations - redundantProteomes - panproteome - annotationScore - superkingdom - proteomeCompletenessReport - genomeAssembly - geneCount - proteinCount - genomeAnnotation - taxonLineage. The results are as Figures 42, 43, 44, 45.

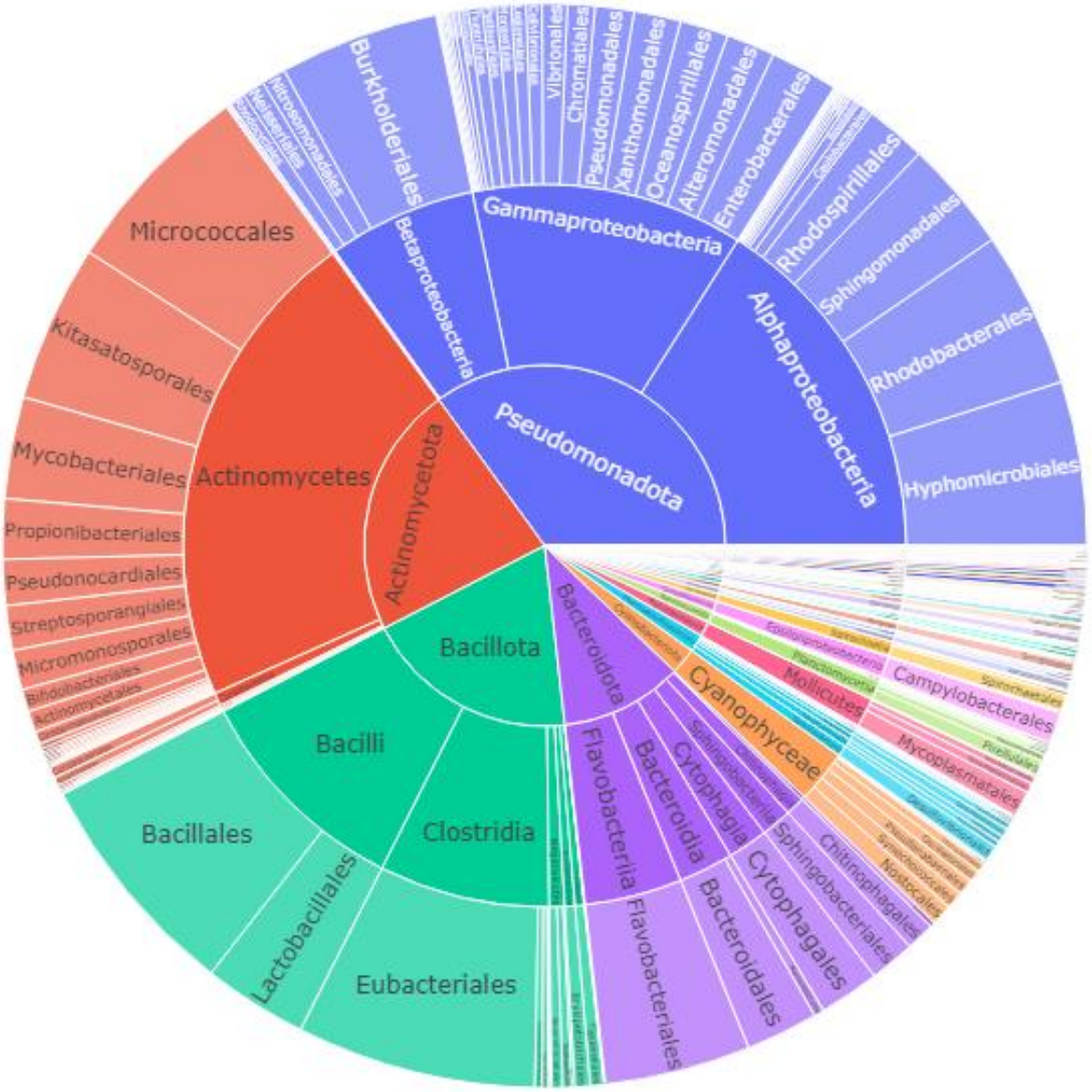


Figure 42. UniProt Reference Proteomes

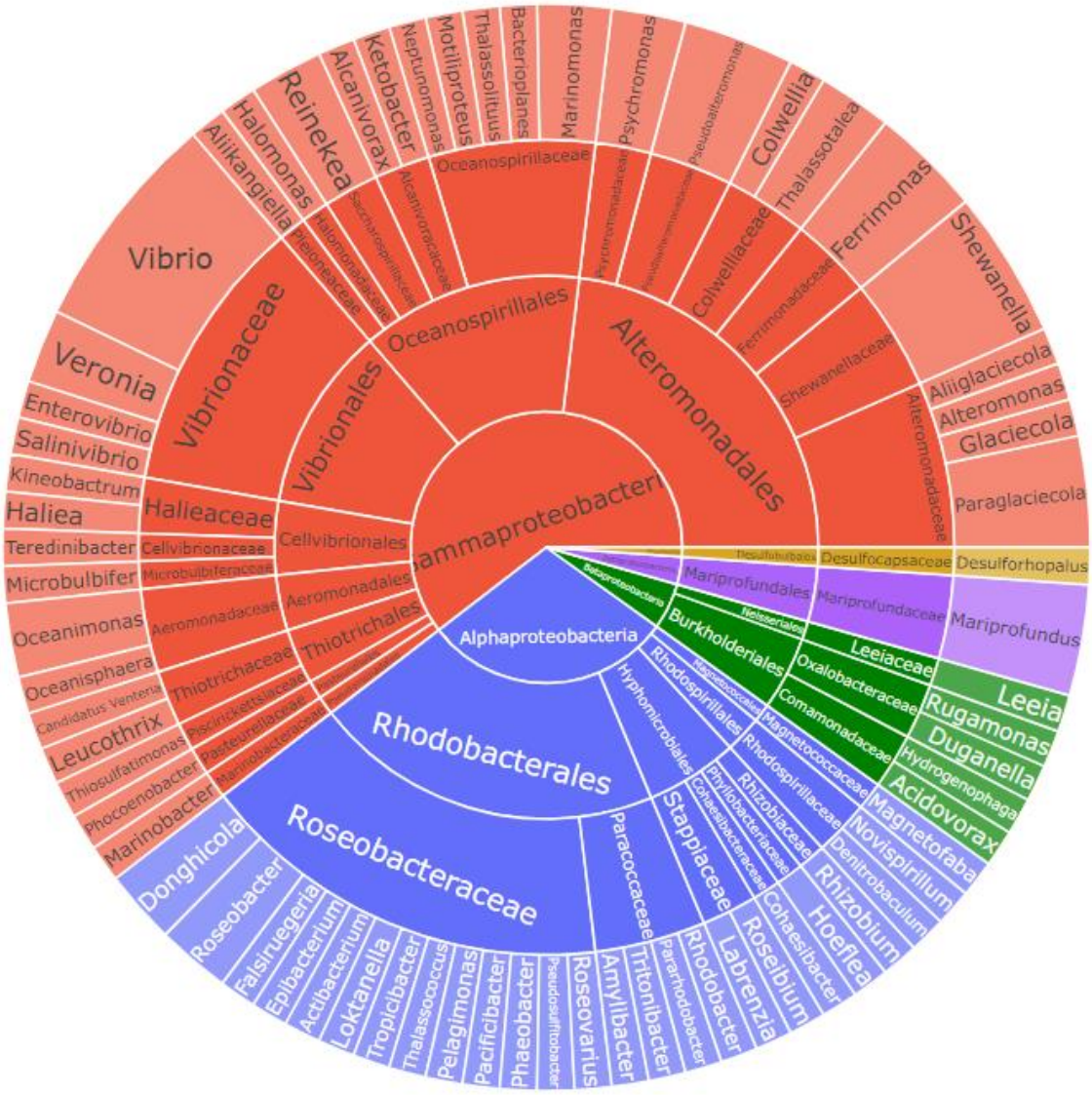


Figure 43. Long-form bacterial DCAs (96 seq)

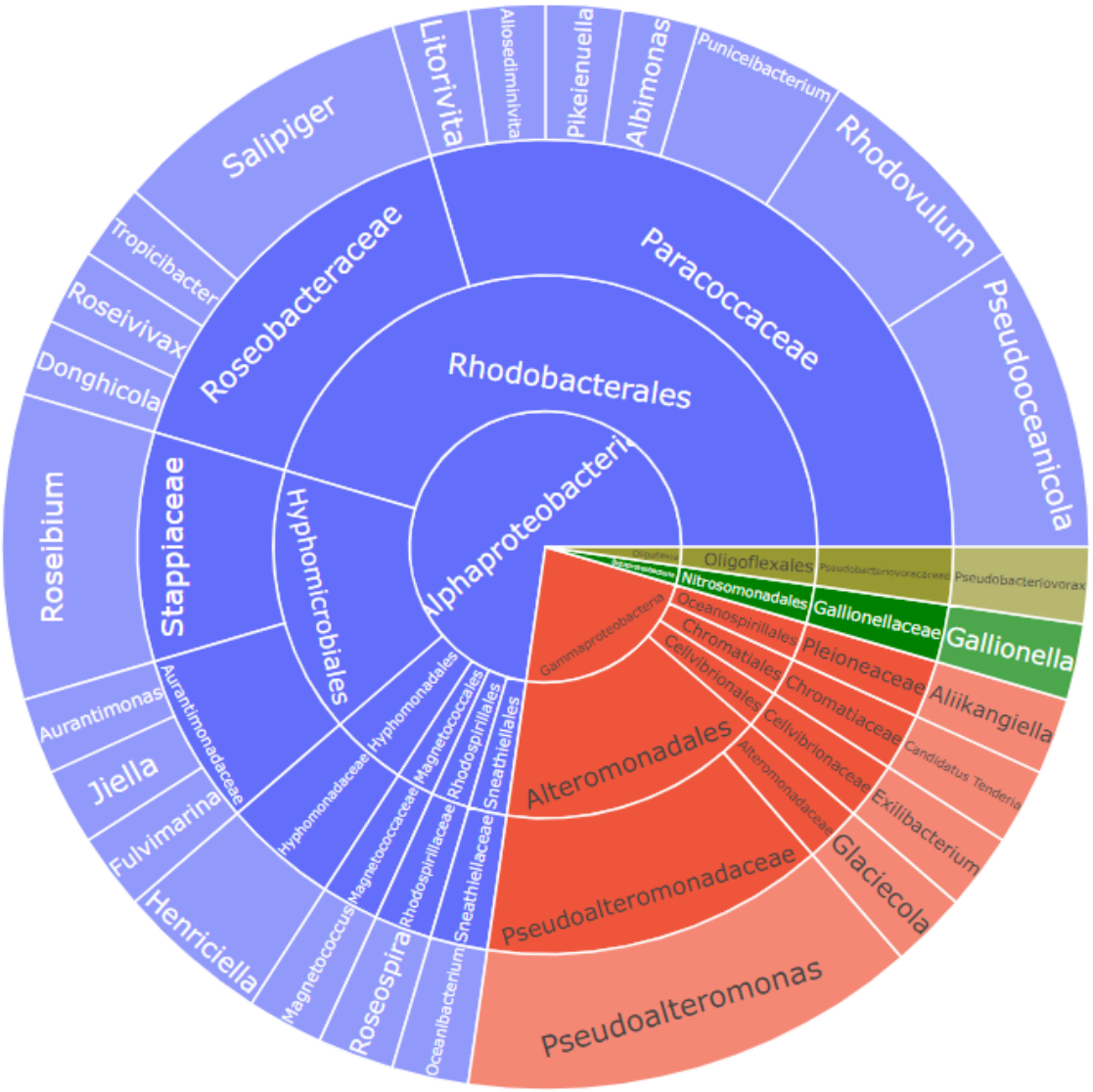


Figure 44. Short-form bacterial DCAs (50 seq)

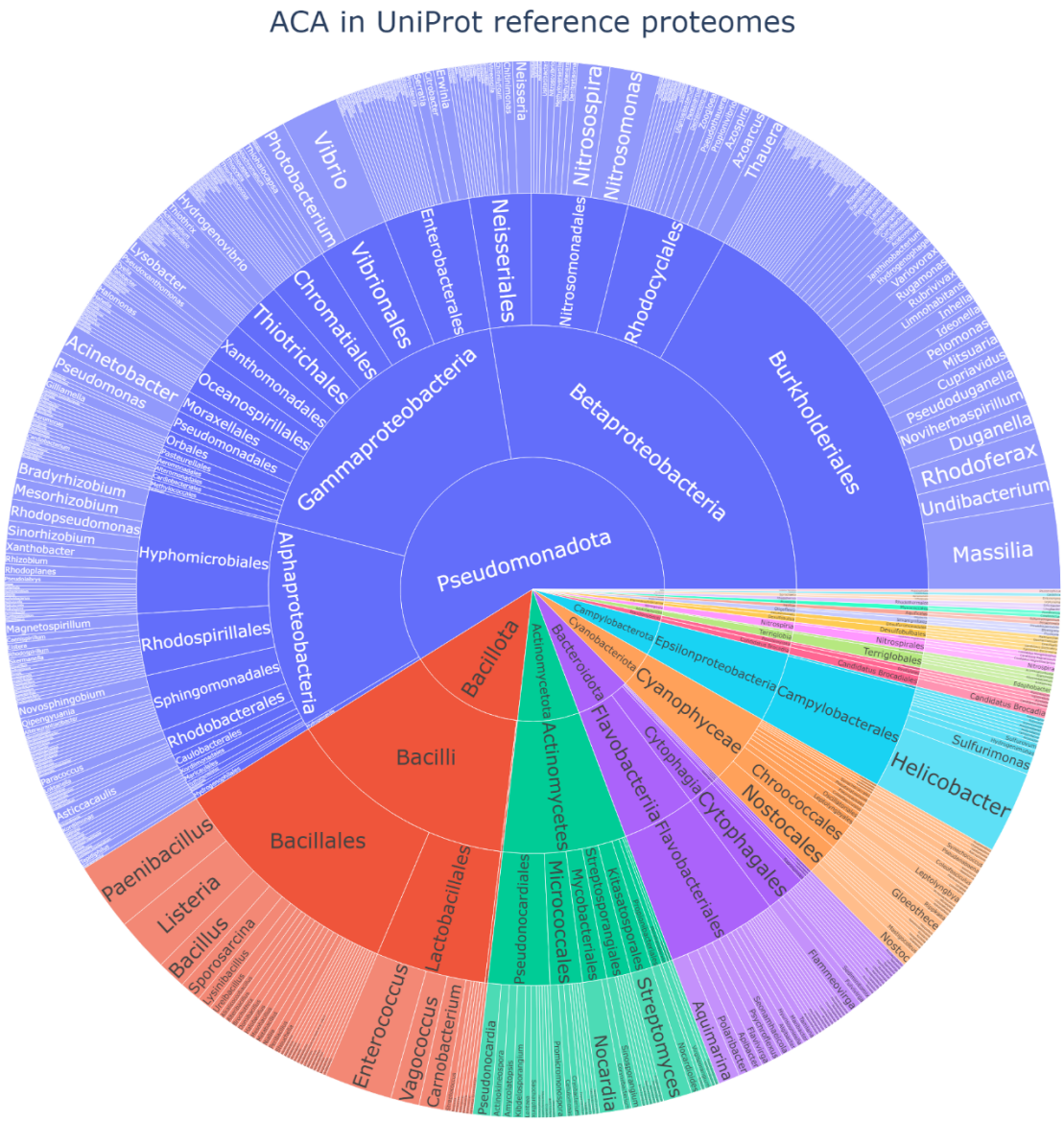


Figure 45. ACA taxa large based on 994 lineages

Figure 46 presents the bacterial taxa with pathogens, which is based on the data from Bartlett et al. (https://github.com/padpadpadpad/bartlett_et_al_2022_human_pathogens/tree/master/data).

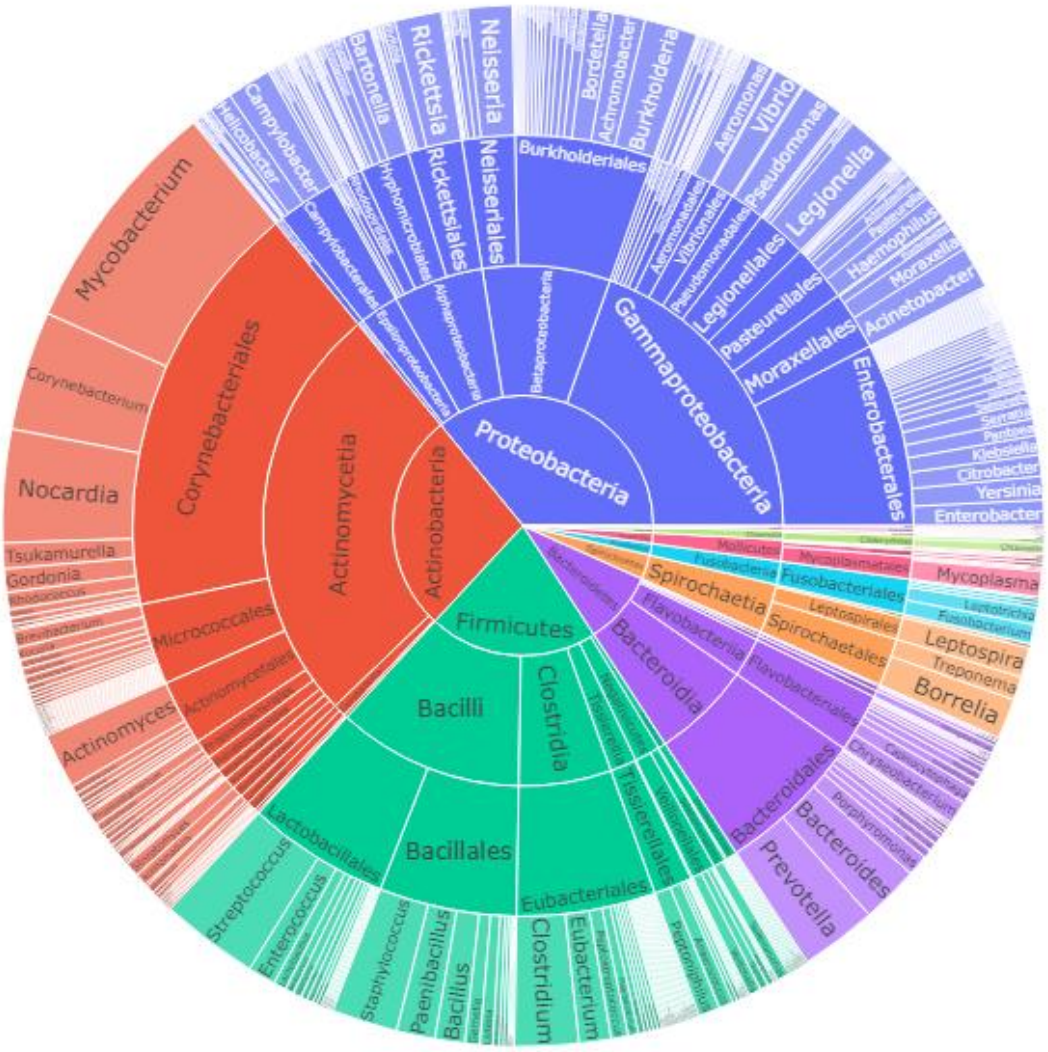


Figure 46. bacteria human pathogens

5.2 Species list comparisons

Initially exploring the shared organisms and genera between short and long DCAs two species were found, while five genera are shared between the short and long sets. Then we have explored the shared species of the DCA list and pathogens list yeiled in the five shared genera of 'Acidovorax', 'Halomonas', 'Vibrio', 'Rhizobium', 'Shewanella'. However, in the ACA list, 45 species and 60 genera were found in the pathogens. Consequently, 6 genomes, 9 species, and 17 genera shared with both ACA and DCA were found (Appendix 3).

6 Conclusion

Carbonic anhydrases are involved in a variety of physiological activities, and they may be found in many realms of life, such as respiration, acid-base balance, photosynthesis and ion transport. There are eight recognized families of CAs, each with a distinctive structure and function. The most typical form of CA is β-CA, which is required for diatom CO2 fixation and photosynthesis in C3 plants. A special variety of CA termed ζ-CA is found only in diatoms, and its structure is comparable to that of β-CA, but it is still unknown what it does. More study is required to completely understand the molecular processes of CAs and their role in various species.

This reach contributes to the ongoing advancement of AI. The innovative AI program of AlphaFold can predict 3D protein structures with high precision. By enabling the prediction of protein structures that are difficult or impossible to investigate experimentally, AlphaFold has the potential to transform drug development and other areas of biology. In the context of analyzing health data, bioinformatics methods are crucial since they employ ICT to manage and extract information from biological data. This is vital given the growing flow of data from DNA sequencing and the functionalities discovered as a result of the spread of Next-generation sequencing technology.

In addition, the cross-disciplinary aspect of bioinformatics is emphasised by contributions from social sciences, engineering, and decision science. Accordingly, the future area of research can be highlighted as: Blockchain may be used to provide safe and impenetrable medical data records, enhancing patient privacy, and boosting confidence in the healthcare system (Chukwu & Garg, 2020). Fair and impartial bioinformatics algorithms must be developed, and data privacy must be maintained (DeCamp & Lindvall, 2020). Using bioinformatics to better understand drug-food-interactions, so we can offer patients safer and more efficient therapies (Casalino et al., 2021). Improvement of the E-health infrastructure such as telemedicine, the creation of medical social networks, medical decision support systems, and electronic health records to improve accessibility and the lower cost of healthcare for all people (Seyhan & Carini, 2019). All in all, bioinformatics has the potential to improve the safety, equity, efficiency, and accessibility of healthcare for everybody.

References

alphafold. (2023). *Protein Structure Database*. Retrieved 2023, from https://www.alphafold.ebi.ac.uk/

Callaway, E. (2020). 'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures. *Nature, 588*(7837), 203-205.

Campestre, C., De Luca, V., Carradori, S., Grande, R., Carginale, V., Scaloni, A., Supuran, C.T. and Capasso, C. (2021). Carbonic anhydrases: new perspectives on protein functional role and inhibition in Helicobacter pylori. *Frontiers in Microbiology, 12*, 629163.

Casalino, G., Castellano, G., Consiglio, A., Nuzziello, N., & Vessio, G. (2021). MicroRNA expression classification for pediatric multiple sclerosis identification. *Journal of Ambient Intelligence and Humanized Computing*, 1-10.

Chichiarelli, S., Altieri, F., Paglia, G., Rubini, E., Minacori, M., & Eufemi, M. (2022). ERp57/PDIA3: new insight. *Cellular & Molecular Biology Letters, 27*(1), 12.

Chukwu, E., & Garg, L. (2020). A systematic review of blockchain in healthcare: frameworks, prototypes, and implementations. *Ieee Access, 8*, 21196-21214.

Crooks, G. E. (2017). On measures of entropy and information. *Tech. Note, 9*(4).

Crooks, G.E., Hon, G., Chandonia, J.M. and Brenner, S.E. (2004). WebLogo: a sequence logo generator. *Genome research, 14*(6), 1188-1190.

Cullen, T., & Garcia, J. E. (2021). Data Mining, Data Analytics, and Bioinformatics: Leveraging Big Data to Identify the Most Vulnerable and to Reduce Health Disparities . In *Innovations in Global Mental Health* (pp. 455-488). Cham: Springer International Publishing.

DeCamp, M., & Lindvall, C. (2020). Latent bias and the implementation of artificial intelligence in medicine. *Journal of the American Medical Informatics Association, 27*(12), 2020-2023.

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

García, S., Luengo, J., & Herrera, F. (2015). *Data preprocessing in data mining* (Vol. 72). Cham, Switzerland: Springer International Publishing.

Gouy, M., Guindon, S. and Gascuel, O. (2010). SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Molecular biology and evolution, 27*(2), .221-224. Retrieved from https://doua.prabi.fr/software/seaview

hang, J., Chapman, B., Friedberg, I., Hamelryck, T., de Hoon, M., Cock, P., Antao, T., Talevich, E. and Wilczynski, B. (2023). *Biopython tutorial and cookbook.* Retrieved 2023, from https://biopython.org/DIST/docs/tutorial/Tutorial.html

Jensen, E.L., Maberly, S.C. and Gontero, B. (2020). Insights on the functions and ecophysiological relevance of the diverse carbonic anhydrases in microalgae. *International Journal of Molecular Sciences, 21*(8), 2922.

Launay, H., Huang, W., Maberly, S. C., & Gontero, B. (2020). Regulation of carbon metabolism by environmental conditions: a perspective from diatoms and other chromalveolates. *Frontiers in Plant Science, 11*, 1033.

PDB, R. (2023, 9). *rcsb*. Retrieved 2023, from RCSB Protein Data Bank (RCSB PDB): https://www.rcsb.org/stats/growth/growth-released-structures

Roberts, S. B., Lane, T. W., & Morel, F. M. (1997). Carbonic anhydrase in the marine diatom Thalassiosira weissflogii (Bacillariophyceae). *Journal of Phycology, 33*(5), 845-850.

Sawaya, M.R., Cannon, G.C., Heinhorst, S., Tanaka, S., Williams, E.B., Yeates, T.O. and Kerfeld, C.A. (2006). The structure of β-carbonic anhydrase from the carboxysomal shell reveals a distinct subclass with one active site for the price of two. *Journal of Biological Chemistry, 281*(11), 7546-7555.

Schneider, T. D., & Stephens, R. M. (1990). Sequence logos: a new way to display consensus sequences. *Nucleic acids research, 18*(20), 6097-6100.

Scozzafava, A., Mastrolorenzo, A. and Supuran, C.T. (2006). Carbonic anhydrase inhibitors and activators and their use in therapy. *Expert Opinion on Therapeutic Patents, 16*(12), 1627-1664.

Seyhan, A. A., & Carini, C. (2019). Are innovation and new technologies in precision medicine paving a new era in patients centric care? *Journal of translational medicine, 17*, 1-28.

Shalileh, F., Gheibzadeh, M.S., Lloyd, J.R., Fietz, S., Shahbani Zahiri, H. and Zolfaghari Emameh, R. (2023). Evolutionary analysis and quality assessment of ζ-carbonic anhydrase sequences from environmental microbiome. *Journal of Basic Microbiology*, 1-14.

Supuran, C. (2008). Carbonic anhydrases: novel therapeutic applications for inhibitors and activators. *Nature reviews Drug discovery, 7*(2), 168-181.

Supuran, C. (2016). Structure and function of carbonic anhydrases. *Biochemical Journal, 473*(14), 2023-2032.

www.forbes.com. (2021). Retrieved 2023, from https://www.forbes.com/sites/robtoews/2021/10/03/alphafold-is-the-most-important-achievement-in-ai-ever/?sh=132c44ae6e0a

Zimmerman, S.A. and Ferry, J.G. (2008). The β and γ classes of carbonic anhydrase. *Current pharmaceutical design, 14*(7), 716-721.

Appendices

Appendix 1. Distances between all residues in five groups of sequences

60seq

	ACC	Q-H1	H2-E	E-H3	H3-T	Total
0	A0A812XTI8	34	4	69	78	187
1	A0A812RQ93	34	4	69	79	188
2	A0A812ZQG1	34	4	69	79	188
3	A0A1Q9DSV6	34	4	69	79	188
4	A0A812H5T8	34	4	68	86	194
5	A0A812XEL1	34	4	67	84	191
6	A0A812NUI5	34	4	67	84	191
7	A0A812UAJ1	34	4	65	84	189
8	A0A812HUV3	34	4	67	84	191
9	A0A812HTU8	34	4	67	84	191
10	A0A090N4Y7	36	4	51	73	166
11	A4S200	35	4	61	68	170
12	C1E037	35	4	77	74	192
13	A0A7S1FBW0	34	4	64	70	174
14	A0A7S1FC43	34	4	64	70	174
15	A0A7S1WJC0	34	4	65	79	184
16	A0A1Q9D410	34	4	65	77	182
17	A0A812U275	34	4	65	78	183
18	A0A812XZF5	34	4	65	78	183
19	A0A812JW14	34	4	65	78	183
20	A0A812RYT3	34	4	65	78	183
21	A0A812ULI5	34	4	65	78	183
22	A0A812TUP3	34	4	65	78	183
23	F0XWW1	34	4	86	92	218
24	A0A7S4GBX2	36	4	58	82	182
25	A0A7S4GBX6	36	4	58	82	182
26	A0A7S2U9R3	36	4	60	79	181
27	A0A7S0IBT1	41	4	78	106	231
28	C1EG17	41	4	78	105	230
29	A0A1E7EZ89	41	4	88	95	230
30	A0A7S2XUT0	35	4	65	82	188
31	A0A7S2V5F9	35	4	67	81	189
32	A0A448ZPF4	35	4	98	88	227
33	A0A7S2M259	36	4	72	90	204
34	K0RA12	36	4	72	86	200
35	K0R419	36	4	72	86	200
36	W8VUA7	36	4	73	86	201
37	W8VYH0	36	4	72	88	202
38	A0A7S4MK07	36	4	77	85	204
39	A0A7S3AZA3	36	4	73	85	200
40	K0SZQ0	35	4	78	83	202
41	X5I1I9	35	4	64	86	191
42	K8EQL3	36	4	64	87	193
43	A0A1E7EUG8	35	4	86	86	213
44	A0A7S2LSL9	36	4	67	84	193
45	A0A7S0KBZ5	36	4	60	83	185
46	A0A7S3L7P6	35	4	86	79	206

47	A0A7S0UKW0	35	4	78	88	207
48	A0A7S1UU45	35	4	62	86	189
49	A0A7S0CDJ9	35	4	64	82	187
50	A0A7S4JEJ7	35	4	65	83	189
51	A0A6V2KCI5	35	4	67	86	194
52	A0A7S4S7Z7	35	4	67	86	194
53	A0A830HI75	36	4	69	85	196
54	A0A7S2HVS4	35	4	68	82	191
55	A0A830HHC5	36	4	71	81	194
56	A0A7S2BCN7	36	4	69	81	192
57	A0A7S0U0B0	35	4	61	79	181
58	A0A7S0R5U4	36	4	64	90	196
59	A7UC72	36	4	71	85	198

DCAb short 50

	ACC	Q-H1	H2-E	E-H3	H3-T	Total
0	A0A0F4PPN0	31	4	52	78	167
1	W4LFV2	31	4	55	78	170
2	A0A1Y6BK61	31	4	58	69	164
3	D9SD75	31	4	57	79	173
4	A0A177W0K2	31	4	54	79	170
5	A0A0S2TAA2	31	4	59	79	175
6	A0A7I8MS99	31	4	54	79	170
7	A0LDG0	31	4	54	79	170
8	A0A3R8U0F6	31	4	51	78	166
9	A0A545UCK1	31	4	54	79	170
10	A0A545U8G7	31	4	51	79	167
11	A0A2Z6UFR9	31	4	51	79	167
12	A0A0A7EH23	31	4	51	79	167
13	U1JGJ1	31	4	51	79	167
14	A0A0F4QK96	31	4	51	79	167
15	A0A166YEI2	31	4	51	79	167
16	A0A1Y6CF06	31	4	51	73	161
17	A0A399R9T2	31	4	50	78	165
18	A0A399R781	31	4	50	78	165
19	A0A1V0PQ74	31	4	53	80	170
20	A0A1I1LXM6	31	4	56	79	172
21	A0A6L7G0J1	31	4	53	79	169
22	A0A418SDM0	31	4	53	79	169
23	A0A2V4NVW8	31	4	56	79	172
24	A0A1I2D4U9	31	4	51	79	167
25	A0A1I3GC59	31	4	53	82	172
26	A0A177H5X1	31	4	53	79	169
27	A0A1V0PQ83	31	4	53	79	169
28	A0A1W1Z9Z7	31	4	53	80	170
29	A0A1M4MX71	31	4	53	79	169
30	A0A5M6IEN9	31	4	53	79	169
31	A0A1Y5S412	31	4	53	79	169
32	A0A238YK89	31	4	53	79	169
33	A0A2G8RB38	31	4	53	79	169
34	A0A1X0SY01	31	4	53	79	169
35	A0A4Y8RV32	31	4	53	79	169
36	A0A2Z4UND1	31	4	53	79	169
37	A0A0P1G5I0	31	4	53	78	168
38	A0A366C3U2	31	4	53	79	169

39	A0A2T6B2K7	31	4	53	78	168
40	A0A0M6Y0M0	31	4	51	79	167
41	A0A0M7AB92	31	4	51	79	167
42	A0A1M7GS86	31	4	55	79	171
43	A0A562T9J6	31	4	51	79	167
44	A0A7L5BWJ0	31	4	47	77	161
45	A0A844WAP3	31	4	53	79	169
46	A0A327YRR2	31	4	53	79	169
47	A0A1G8T7Z6	31	4	53	79	169
48	A0A1G7GQA1	31	4	53	79	169
49	A0A1P8UZE2	31	4	53	79	169

DCAb long 96

	ACC	Q-H1	H2-E	E-H3	H3-T	Total
0	A0A1Y0CV22	31	4	53	78	168
1	E1SMY4	31	4	57	78	172
2	A0A1M5Z5J9	31	4	57	78	172
3	A0A6G8D3H9	31	4	51	78	166
4	U3A2Q2	31	4	50	78	165
5	A0A366WDQ9	31	4	57	78	172
6	A0A0M3B8C0	31	4	57	78	172
7	A0A0F4R6A1	31	4	57	78	172
8	A0A0N8TXR8	31	4	57	78	172
9	A0A1C3ER30	31	4	57	78	172
10	A0A4Q0YV57	31	4	57	78	172
11	A0A545T6W0	31	4	53	78	168
12	A0A1R4B4D1	31	4	57	78	172
13	A0A5P9EVL0	31	4	57	78	172
14	Q083D6	31	4	57	78	172
15	A0A7Y4C3A8	31	4	52	78	167
16	A0A4P7JL61	31	4	57	78	172
17	A0A099KGP0	31	4	57	78	172
18	A0A0S2JBM8	31	4	57	78	172
19	A0A1I1KL32	31	4	57	78	172
20	A0A2N0X064	31	4	57	78	172
21	B1KH45	31	4	57	78	172
22	D4ZG09	31	4	57	78	172
23	A0A6G8AD04	31	4	57	78	172
24	A0A0F5ASE0	31	4	57	78	172
25	A0A3N5XXH8	31	4	57	78	172
26	A0A553L480	31	4	57	78	172
27	G4QE78	31	4	57	78	172
28	A0A2N1CUU8	31	4	57	78	172
29	A0A2N5Y243	31	4	57	78	172
30	A0A4Z0LUQ5	31	4	57	78	172
31	A0A559QTE1	31	4	57	78	172
32	K6Y571	31	4	57	78	172
33	K6Z6Q1	31	4	57	78	172
34	A0A222FHA2	31	4	56	78	171
35	E3BNS4	31	4	51	78	166
36	A0A120DHN5	31	4	57	78	172
37	A0A379C9V9	31	4	57	78	172
38	A0A483AP18	31	4	57	78	172
39	Q0VNZ5	31	4	57	78	172
40	A0A2K8KPA8	31	4	57	78	172

41	A0A1G7P8L1	31	4	57	78	172
42	A0A233RDF1	31	4	57	78	172
43	H2G072	31	4	57	78	172
44	F2JVM1	31	4	56	79	172
45	S6GD57	31	4	57	78	172
46	S6GJ28	31	4	57	78	172
47	A0A7W2IBK2	31	4	49	78	164
48	A0A845HVN5	31	4	49	78	164
49	A0A1V3RVU2	31	4	57	78	172
50	A0A2K9LHM4	31	4	57	78	172
51	A0A4V5NVC8	31	4	57	78	172
52	A0A1Y2K408	31	4	57	78	172
53	A0A1E7PVU8	31	4	57	78	172
54	A0A1H6F9G6	31	4	57	78	172
55	A0A4U0ZVH1	31	4	57	78	172
56	A0A1N7JI91	31	4	57	78	172
57	A4BHH6	31	4	57	78	172
58	A0A847SI98	31	4	57	78	172
59	C5BPF1	31	4	57	78	172
60	A0A498DI92	31	4	58	80	175
61	A0A1T4W5S3	31	4	57	78	172
62	A0A5P1RBS2	31	4	57	78	172
63	A0A0S2K106	31	4	57	78	172
64	A0A317CE19	31	4	57	78	172
65	A0A5R9GKY7	31	4	60	79	176
66	A0A6F8PUZ7	31	4	57	78	172
67	A0A2N0WVI4	31	4	57	78	172
68	A0A269PJB8	31	4	57	78	172
69	A0A1R1MGM7	31	4	57	78	172
70	A0A2K8L2J7	31	4	57	78	172
71	A0A7X0DNL4	31	4	57	78	172
72	A0A1Y5TU44	31	4	57	78	172
73	A0A238JJR4	31	4	57	78	172
74	A0A2T7HJE9	31	4	57	78	172
75	A0A843YJC0	31	4	57	78	172
76	A0A640VQP1	31	4	57	78	172
77	A0A1G5PN50	31	4	57	77	171
78	A0A238KEQ0	31	4	57	77	171
79	A0A285PDA5	31	4	57	77	171
80	A0A6N7ZYF5	31	4	57	78	172
81	A0A1N7MAR0	31	4	57	78	172
82	A0A1B0ZW06	31	4	57	78	172
83	Q169G0	31	4	57	78	172
84	A0A6I1EDA4	31	4	57	78	172
85	A0A545TRT5	31	4	57	78	172
86	A0A1H5SWY5	31	4	57	78	172
87	A0A1Y5TJ40	31	4	57	78	172
88	A0A2U2C7N3	31	4	57	78	172
89	A0A1M4MVA5	31	4	57	78	172
90	A0A1M7DPW7	31	4	57	78	172
91	A0A2T0WID3	31	4	57	78	172
92	A9CYX8	31	4	57	78	172
93	A0A0N7LZK6	31	4	57	78	172
94	A0A073J1D9	31	4	57	78	172
95	A0A0F4RJL2	31	4	57	78	172

Bacterial ACA

	ACC	Q-H1	H2-E	E-H3	H3-T	Total	
0	0A254TJH6		57	4	13	67	143
1	0A0Q5HDE4		57	4	13	67	143
2	0A7G5ZKD7		57	4	13	67	143
3	0A0M2WIR2		57	4	13	67	143
4	0V0V8		57	4	13	67	143
5	0A086W8Y0		57	4	13	67	143
6	9DFX6		57	4	13	67	143
7	0A1S9AB93		57	4	13	67	143
8	0A2U2I7P6		57	4	13	67	143
9	0A418XSR9		57	4	13	67	143
10	0A848HJ88		57	4	13	67	143
11	0A4Y9SQB3		57	4	13	67	143
12	0A0Q8QD40		57	4	13	67	143
13	0A0Q6SYX4		57	4	13	67	143
14	0A7Z2ZSM0		57	4	13	67	143
15	0A0J1DHW8		57	4	13	67	143
16	0A4Y9SQ24		57	4	13	67	143
17	0A1M5JBU9		57	4	13	67	143
18	0A2D2DK21		57	4	13	67	143
19	0A2G8T3B4		57	4	13	67	143
20	0A2G8TJ65		57	4	13	67	143
21	0A1E7X678		57	4	13	67	143
22	0A1I4HZY9		57	4	13	67	143
23	0A1H8MXM1		57	4	13	67	143
24	0A7Y9KZQ9		57	4	13	67	143
25	0A7W2ID34		57	4	13	67	143
26	0A845HW60		57	4	13	67	143
27	0A843SLV6		57	4	13	67	143
28	0A2R4CFY2		57	4	13	67	143
29	0A562R591		57	4	13	67	143
30	0A6L6PTM6		57	4	13	67	143
31	0A0K1K3M2		57	4	13	67	143
32	0A1I7FVV6		57	4	13	67	143
33	0A0Q6WXW9		57	4	13	67	143
34	0A944DPI2		57	4	13	67	143
35	0A6C1B8A4		57	4	13	68	144
36	0A4R5W4P9		57	4	13	67	143
37	0A2S9H170		57	4	13	67	143
38	0A923HQN1		57	4	13	67	143
39	0A941E8P7		57	4	13	67	143
40	0A916XMG8		57	4	13	67	143
41	0A850QK10		57	4	13	67	143
42	0A3Q9BPU4		57	4	13	67	143
43	0A6N4TBN9		57	4	13	67	143
44	0A923KII4		57	4	13	67	143
45	0A941DK72		57	4	13	67	143
46	0A923HZL4		57	4	13	67	143
47	0A3A3FGD8		57	4	13	67	143
48	0A0C2BP44		57	4	13	67	143
49	0A410UQW4		57	4	13	67	143
50	0A3A3GAD3		57	4	13	67	143
51	0A4V6NXY6		57	4	13	67	143
52	0A0N0JCX9		57	4	13	67	143
53	0A916A5Q2		57	4	13	67	143
54	0A4R6R6R1		57	4	13	67	143
55	0A4Q9GZD3		57	4	13	67	143
56	0A3R8S8Y7		57	4	13	67	143

57	0A924PQ21	57	4	13	67	143
58	0A519ZJJ4	57	4	13	67	143
59	0A6S6Y5A3	57	4	13	67	143
60	0A518UA11	58	4	13	67	144
61	1K3Q4	57	4	13	68	144
62	0SI93	57	4	13	67	143
63	0A4Y4CVZ7	57	4	13	68	144
64	0A6C2D5T9	57	4	13	68	144
65	0A848GEE3	57	4	13	68	144
66	0A840BS55	57	4	13	67	143
67	0A931NDL0	57	4	13	67	143
68	0A931J617	57	4	13	67	143
69	0A0U3CAA1	57	4	13	67	143
70	0A246JMI3	57	4	13	67	143
71	0A1I6K7P8	57	4	13	67	143
72	0A318HCC1	57	4	13	67	143
73	0A7Y9R0X2	57	4	13	67	143
74	0A5C1PYZ6	57	4	13	67	143
75	0A4V2EYK5	57	4	13	67	143
76	1XY18	57	4	13	67	143
77	0A3N7CCP3	57	4	13	67	143
78	0A4R6NCE6	57	4	13	67	143
79	0A2R7RZJ8	57	4	13	70	146
80	0A3S2XTQ9	57	4	13	67	143
81	0A3S3SEF7	57	4	13	67	143
82	5WIY7	57	4	13	67	143
83	0A480AR27	57	4	13	68	144
84	0A7C9PHQ3	57	4	13	67	143
85	0A0K8P3E9	57	4	13	67	143
86	0A643FJZ2	57	4	13	67	143
87	0A3N7JJC6	57	4	13	67	143
88	0A924TCE8	57	4	13	67	143
89	0A937A1G7	57	4	13	67	143
90	0A0Q6XHL1	57	4	13	67	143
91	0A2Z5G2A4	58	4	13	67	144
92	0A1R1I2V7	56	4	13	68	143
93	0A974SSX3	56	4	13	68	143
94	0A3B7AA07	56	4	13	68	143
95	0A5C1EDS6	57	4	13	68	144
96	0A1A8XL21	56	4	13	68	143
97	0A840G335	56	4	13	68	143
98	0A011PUS7	56	4	13	68	143
99	0A011Q4S7	56	4	13	68	143
100	0A011NYQ4	56	4	13	68	143
101	0A1A8XZ98	56	4	13	68	143
102	0A080M863	56	4	13	68	143
103	0A2R5EQ31	56	4	13	69	144
104	0A974SQH8	56	4	13	68	143
105	0A4Q0T0W3	57	4	13	67	143
106	0A1B3LP64	57	4	13	67	143
107	5RFA0	57	4	13	67	143
108	0A0Q6WY30	57	4	13	68	144
109	0A2N4XV26	57	4	13	67	143
110	0A0K1K5P8	57	4	13	68	144
111	0A3R9NW77	57	4	13	68	144
112	0A7G8BDR1	57	4	13	68	144
113	0A848H4Q0	59	4	13	67	145
114	0A1H8L0H8	59	4	13	68	146

115	0A0Q7SDU0	59	4	13	68	146
116	0A940M364	59	4	13	68	146
117	0A177QY45	57	4	13	68	144
118	0A0Q8QKY6	57	4	13	68	144
119	0A6M4GTU1	57	4	13	68	144
120	0A1V3RYW4	59	4	13	68	146
121	0A1W6LHK0	57	4	13	67	143
122	0A0Q8PA20	57	4	13	67	143
123	0A254MXV3	57	4	13	70	146
124	0A4R3VEN8	57	4	13	70	146
125	0A0Q7T2L7	57	4	13	70	146
126	0A1A9HSI9	57	4	13	67	143
127	0A257FTZ6	57	4	13	67	143
128	0A839LQU2	57	4	13	67	143
129	0A7S9CD38	57	4	13	67	143
130	0HMP7	57	4	13	67	143
131	0A940YRQ3	57	4	13	67	143
132	0A257CD88	57	4	13	67	143
133	0A924MWE6	57	4	13	67	143
134	0A257KHI5	57	4	13	67	143
135	0A257D9Q9	57	4	13	67	143
136	0A840S2X7	57	4	13	67	143
137	0A7Y6NP14	57	4	13	67	143
138	0A5C6U6S5	57	4	13	67	143
139	0A2U8FVC8	57	4	13	67	143
140	0A4R6QLD8	57	4	13	67	143
141	0A0L6TA93	57	4	13	67	143
142	0A0Q6MNT8	57	4	13	67	143
143	0A975E5N5	57	4	13	67	143
144	0A0Q7AKZ4	57	4	13	67	143
145	0A0C1E2M9	58	4	13	66	143
146	5F930	58	4	13	66	143
147	7IHQ3	58	4	13	66	143
148	2B9H8	58	4	13	66	143
149	0EZH5	58	4	13	66	143
150	0A0K6IWZ5	57	4	13	66	142
151	0A2Z6E0L1	58	4	14	64	142
152	0A7W8EFK2	72	4	14	71	163
153	0A8J3PT60	74	4	14	69	163
154	0A5M3XWM1	73	4	14	69	162
155	0A4U3MHZ3	73	4	14	68	161
156	0A239GRS9	74	4	14	69	163
157	0A1G7W0L5	74	4	13	69	162
158	0A919V5V2	73	4	13	74	166
159	0A1G7TH32	73	4	13	74	166
160	0A429ZUY1	60	4	13	67	146
161	0A6G8AN51	61	4	13	67	147
162	0A430ACP5	61	4	13	67	147
163	0A518G9J2	69	4	13	85	173
164	0A142X0S6	81	4	13	92	192
165	0A1S9CRC0	58	4	13	63	140
166	0A5D0CJP3	57	4	13	68	144
167	0A090Z8L4	57	4	13	68	144
168	2QTJ1	58	4	13	62	139
169	0A242KD62	59	4	13	62	140
170	0A2A5RXZ4	59	4	13	62	140
171	0A1E5KSQ2	59	4	13	62	140
172	0A430AL21	59	4	13	62	140

173	0A143YSC0	58	4	13	67	144
174	0A1B3XJJ2	58	4	13	65	142
175	0A099W1R4	58	4	13	65	142
176	8E726	58	4	13	65	142
177	0A099W002	58	4	13	65	142
178	0A2X3GVQ2	58	4	13	65	142
179	7CHB3	58	4	13	65	142
180	0A3D8TKW5	58	4	13	65	142
181	0A1H7RHC4	63	4	13	65	147
182	0A843YV92	63	4	13	65	147
183	0A1T5LX12	61	4	13	64	144
184	0A3D8VG43	61	4	13	63	143
185	0A0S2FCQ8	62	4	13	67	148
186	0A4R6YQ46	62	4	13	64	145
187	0A3N4VG17	62	4	13	65	146
188	0A0S2DG19	62	4	13	65	146
189	0A4P6YRB8	61	4	13	60	140
190	0A3R9YDQ7	63	4	13	67	149
191	0A430AJF3	61	4	13	64	144
192	0A1L8WSB1	60	4	13	62	141
193	0A179ETH3	60	4	13	64	143
194	0A4Y3JKX1	60	4	13	64	143
195	0A377KP34	60	4	13	64	143
196	0A2C9XQV8	60	4	13	64	143
197	0A242KD51	62	4	13	64	145
198	2TEU2	61	4	13	64	144
199	834E4	59	4	13	64	142
200	0A1E5KWG9	61	4	13	64	144
201	2RYG5	64	4	13	64	147
202	1NH67	61	4	13	64	144
203	1RNT9	61	4	13	64	144
204	1NUQ5	61	4	13	64	144
205	6LFX9	62	4	13	64	145
206	3TYG7	60	4	13	64	143
207	0NXA6	60	4	13	64	143
208	0A1E8GPX5	61	4	13	64	144
209	0A1I5WNS1	63	4	13	69	151
210	8EGZ2	63	4	13	69	151
211	0A0R2HP78	63	4	13	69	151
212	4BRB6	63	4	13	69	151
213	0A0U3E5K9	63	4	13	69	151
214	0A1X7NF36	63	4	13	69	151
215	0A4D7CVY4	60	4	13	61	140
216	0A430B5S7	60	4	13	66	145
217	6QBQ1	61	4	13	66	146
218	0A7H1N1S4	58	4	13	67	144
219	0A1N7PJI5	57	4	13	67	143
220	9GVU7	57	4	13	67	143
221	9GWP2	57	4	13	67	143
222	0A5J6NA72	57	4	13	67	143
223	5SFX3	57	4	13	67	143
224	0A973X8H5	57	4	13	67	143
225	0A1B3NRY5	57	4	13	66	142
226	0A257HA78	57	4	13	66	142
227	0A212Q545	57	4	13	66	142
228	0A3L7AK23	58	4	13	66	143
229	0A848K4U2	58	4	13	66	143
230	7IHH9	58	4	13	66	143

231	0A974SJV2	58	4	13	66	143
232	6IV30	57	4	13	67	143
233	0A5M6INU1	57	4	13	67	143
234	6UHL3	57	4	13	66	142
235	0A327KP99	57	4	13	77	153
236	0A974AQQ2	58	4	13	67	144
237	0A327KZQ3	57	4	13	67	143
238	8ETC8	57	4	13	66	142
239	6NBN4	57	4	13	67	143
240	0A0D7EA29	58	4	13	67	144
241	0A974AHK7	57	4	13	67	143
242	0A7W7Z8Z5	57	4	13	66	142
243	0A318TIK1	57	4	13	67	143
244	4Z2G9	58	4	13	67	144
245	0U0K6	58	4	13	67	144
246	0A4Q8R6B1	57	4	13	67	143
247	89VB0	57	4	13	67	143
248	0A975RXP4	57	4	13	67	143
249	0A1M7UC28	57	4	13	68	144
250	0A345ZXP1	57	4	13	66	142
251	0A0Q7TDH4	57	4	13	67	143
252	0A1S1TY55	57	4	13	66	142
253	0KTC4	57	4	13	67	143
254	6YIH4	57	4	13	67	143
255	5F6A6	57	4	13	67	143
256	0A7R7CP25	57	4	13	67	143
257	0A509EKU3	57	4	13	68	144
258	0HXF3	57	4	13	67	143
259	0A7W9FQS6	56	4	13	67	142
260	0A178Y345	56	4	13	67	142
261	0A0B4XAB3	56	4	13	67	142
262	0A1E3V5J1	55	4	13	67	141
263	931C3	55	4	13	67	141
264	0A7W8X983	56	4	13	68	143
265	3KM93	55	4	13	67	141
266	0A6N7LD56	55	4	13	67	141
267	0A126T7G3	59	4	13	68	146
268	0A1Y6CZB2	63	4	13	68	150
269	0A1I4A971	60	4	13	69	148
270	0A1I4R245	60	4	13	69	148
271	0A1I2D4H3	63	4	13	71	153
272	0A1H9QPK2	61	4	13	72	152
273	0A1N6FRG5	63	4	13	72	154
274	0A1I4HE11	63	4	13	72	154
275	82WQ7	62	4	13	73	154
276	0A4Y1YKF6	62	4	13	73	154
277	0A1H3HL44	61	4	13	71	151
278	0A0F7KE57	62	4	13	71	152
279	0A1I4RI16	63	4	13	71	153
280	0A1H8REZ4	62	4	13	73	154
281	0A2V3WB43	62	4	13	73	154
282	9ZIU2	62	4	13	72	153
283	0A7W5D4Q2	61	4	13	65	145
284	0A4V3WET5	60	4	13	68	147
285	0A1H3Q5H3	60	4	13	68	147
286	0A2W1L2J3	62	4	13	68	149
287	9KFW1	59	4	13	68	146
288	0A371PH43	61	4	13	68	148

289	6NKL8	62	4	13	68	149
290	0A368W6Q3	61	4	13	68	148
291	0A926KYK8	61	4	13	68	148
292	0A366XWZ8	61	4	13	68	148
293	0A398BN21	61	4	13	68	148
294	0A852UMH6	61	4	13	68	148
295	0A1V2A5R6	61	4	13	68	148
296	0A135WRK7	61	4	13	68	148
297	0A0K9GSL4	61	4	13	68	148
298	0A135WP91	61	4	13	68	148
299	3ECH2	61	4	13	67	147
300	0A917B6H6	61	4	13	68	148
301	0A5J5HQ48	63	4	13	67	149
302	0A1G8G137	62	4	13	68	149
303	0A1G8EF64	62	4	13	68	149
304	0A562QMQ9	61	4	13	68	148
305	3FRE5	61	4	13	68	148
306	0A0F5HYB5	61	4	13	68	148
307	0A4R2B0D4	61	4	13	68	148
308	0A1E7DUP2	61	4	13	67	147
309	0A135W4U0	61	4	13	68	148
310	0A380CBW1	61	4	13	68	148
311	0A975KJV6	61	4	13	68	148
312	0A2X4W6N8	61	4	13	68	148
313	0A6L3V550	61	4	13	68	148
314	0A385NUH0	61	4	13	68	148
315	0A024P8D4	61	4	13	68	148
316	2F9J9	61	4	13	68	148
317	0A2S1GXK1	61	4	13	68	148
318	0A417YB18	61	4	13	68	148
319	0A5J5GVH4	61	4	13	68	148
320	0A0M2PAV4	61	4	13	68	148
321	0A179SLX4	61	4	13	68	148
322	0A0Q4Y2E4	59	4	13	68	146
323	0A5M8SLZ6	58	4	13	69	146
324	0A1G7VM16	56	4	13	68	143
325	0A972FGB1	57	4	13	68	144
326	0A2T4IBN8	57	4	13	68	144
327	0A4S4ATP1	57	4	13	68	144
328	0A4S4A8C6	57	4	13	68	144
329	0A2K9LD00	57	4	13	68	144
330	0PT46	57	4	13	69	145
331	0A1N6ZAU7	57	4	13	68	144
332	0A290ZIN8	57	4	13	68	144
333	0A2U8GSA8	57	4	13	68	144
334	0A4R6E3C5	57	4	13	68	144
335	9ZSW3	57	4	13	68	144
336	0A323UV77	57	4	13	68	144
337	6YE17	57	4	13	68	144
338	0A235EYI7	57	4	13	68	144
339	6YKC3	57	4	13	68	144
340	4KCV8	57	4	13	68	144
341	0A504J9T2	60	4	13	68	147
342	0A563D9N3	59	4	13	67	145
343	0A2S8AAM8	61	4	13	67	147
344	0A0X3ARM4	59	4	13	67	145
345	0A937K169	60	4	13	68	147
346	0A2H1E9S4	62	4	13	67	148

347	0A1Y0M6C7	63	4	13	67	149
348	0A099XYU5	63	4	13	67	149
349	0A2U2J7Y2	63	4	13	67	149
350	0A918V8V6	62	4	13	67	148
351	0A6N6MDB5	62	4	13	67	148
352	0A8J6PS27	61	4	13	67	147
353	0A5C7AT19	62	4	13	67	148
354	0A3D9HFG8	62	4	13	67	148
355	0A5C4SEY7	62	4	13	67	148
356	0A1U7DN16	62	4	13	67	148
357	0A1H9K6E5	62	4	13	67	148
358	0A923KH27	62	4	13	67	148
359	0A0P0D8R2	62	4	13	67	148
360	0A1E5SHA1	62	4	13	67	148
361	0A1M6C6H3	62	4	13	67	148
362	0A2K9PX19	62	4	13	67	148
363	0A4S1DUE1	62	4	13	67	148
364	0A848IVL8	60	4	13	67	146
365	0A1M4XAH6	60	4	13	67	146
366	0A967AI21	60	4	13	67	146
367	0A916ZP52	60	4	13	67	146
368	0A7H8PNT1	60	4	13	67	146
369	0A6I2MRK8	60	4	13	67	146
370	0A2G5GCQ4	62	4	13	67	148
371	0A975CRE9	62	4	13	67	148
372	6L4V9	60	4	13	67	146
373	0A0P7C7G2	62	4	13	67	148
374	0A4Q1IQ21	62	4	13	67	148
375	0A1G9JK24	62	4	13	67	148
376	8ZT21	60	4	13	67	146
377	0A1Z9Z2M8	59	4	13	67	145
378	0A1B2M3P4	59	4	13	67	145
379	0A3M6QMM3	58	4	13	67	144
380	0A0A7S2V9	58	4	13	66	143
381	0A495RBZ3	58	4	13	66	143
382	0A484GEJ2	58	4	13	66	143
383	0A1P8KQZ6	57	4	13	67	143
384	6Q1X3	57	4	13	67	143
385	0A4P9VUS1	57	4	13	66	142
386	0A418YAG8	57	4	13	68	144
387	0A090SVC1	59	4	13	67	145
388	0Z0I1	57	4	13	67	143
389	0A1Y6L192	57	4	13	67	143
390	0A0N0DK79	57	4	13	67	143
391	0A2T3MSE1	57	4	13	67	143
392	0A0J1HD96	57	4	13	67	143
393	0A0C5WP56	57	4	13	67	143
394	0A0J1JFL9	57	4	13	67	143
395	6LM17	57	4	13	67	143
396	0A2Z3ID50	57	4	13	67	143
397	8W5L1	57	4	13	67	143
398	0A0N0Z8B7	57	4	13	67	143
399	8W6R9	57	4	13	67	143
400	0A379AF48	57	4	13	67	143
401	0A2I0FZ21	57	4	13	67	143
402	0A6M8UAA6	57	4	13	67	143
403	0A443IC50	57	4	13	67	143
404	0A427KCE6	57	4	13	67	143

405	0A0M2KJ78	57	4	13	67	143
406	0A2L0PLW5	57	4	13	69	145
407	8YC59	57	4	13	67	143
408	3PP89	57	4	13	67	143
409	0A519JF11	57	4	13	67	143
410	2UL02	57	4	13	67	143
411	2SNS4	57	4	13	67	143
412	0A380N1K0	58	4	13	67	144
413	9CJT6	58	4	13	67	144
414	0A1I5NIT9	58	4	13	68	145
415	0A1M6LPW6	57	4	13	68	144
416	0A6F8PKC7	57	4	13	68	144
417	0A7M1AVQ6	60	4	13	68	147
418	0A5P8P3S4	60	4	13	68	147
419	6BNC3	60	4	13	68	147
420	30R67	60	4	13	68	147
421	0A7S7M363	61	4	13	68	148
422	0A367RZB7	57	4	13	68	144
423	6Q9G6	58	4	13	67	144
424	0A317N1G0	57	4	13	69	145
425	1DTU5	60	4	13	66	145
426	1SUY1	57	4	13	67	143
427	0A4R3I844	57	4	13	67	143
428	0A1H4CK56	58	4	13	67	144
429	0A370QGQ2	57	4	13	67	143
430	2IUB4	57	4	13	68	144
431	0A2N4XS96	57	4	13	67	143
432	0A225N3I0	57	4	13	67	143
433	0A366HKR2	58	4	13	67	144
434	0A1B9N785	57	4	13	67	143
435	0A080KB29	57	4	13	67	143
436	0A6I6EG77	57	4	13	67	143
437	0A5J5FZ86	57	4	13	69	145
438	0A0G4JSX4	57	4	13	69	145
439	6DAJ6	57	4	13	67	143
440	0A250B2D0	57	4	13	67	143
441	0A084A3X0	57	4	13	67	143
442	0A1S8CJR1	57	4	13	67	143
443	0A381E9F0	58	4	13	66	143
444	0A1X3D2G0	58	4	13	66	143
445	0A6N1YAP7	57	4	13	67	143
446	0A4P8YDX2	60	4	13	67	146
447	0A8K0V490	60	4	13	67	146
448	0A807LHS8	58	4	13	67	144
449	5V5N0	59	4	13	67	145
450	0A085G216	59	4	13	67	145
451	0A0L0AIH8	59	4	13	67	145
452	0A4R6EI69	59	4	13	67	145
453	0A9J9L0T8	59	4	13	67	145
454	0A7W3E4B1	59	4	13	67	145
455	0A0W1RRG3	57	4	13	67	143
456	0A0D8D4A3	57	4	13	67	143
457	0A1G8A4M8	57	4	13	66	142
458	0A1M5XZ74	57	4	13	67	143
459	87JM7	57	4	13	66	142
460	0A3N9TL64	57	4	13	67	143
461	0A1R4LFS7	57	4	13	64	140
462	0A975YLA2	57	4	13	66	142

463	0A0Q2Y6Y0	57	4	13	66	142
464	9KMP6	57	4	13	66	142
465	0A099LTI5	57	4	13	66	142
466	0A2J8GU72	57	4	13	66	142
467	87G15	57	4	13	66	142
468	0A0A5I1F9	57	4	13	66	142
469	0A2S5KH84	57	4	13	67	143
470	0A1I3AMJ6	57	4	13	67	143
471	0A6M4H9G5	57	4	13	69	145
472	0A5B8SRX5	57	4	13	67	143
473	3U2V7	57	4	13	53	129
474	0A2N5ESA4	57	4	13	67	143
475	0A080KLN6	57	4	13	67	143
476	0A0F7H7T3	57	4	13	67	143
477	3VIZ7	57	4	13	67	143
478	0A2S2FI03	57	4	13	67	143
479	0PVZ8	57	4	13	67	143
480	0A0C5VPY8	57	4	13	67	143
481	0A261GTK0	57	4	13	67	143
482	0A0F5VCQ7	57	4	13	67	143
483	0A0W7U042	57	4	13	67	143
484	0A081NEE4	57	4	13	67	143
485	4L9B2	57	4	13	67	143
486	0A975H3G3	57	4	13	67	143
487	0A6L6Q4E6	57	4	13	67	143
488	0A1I7L4Z8	57	4	13	67	143
489	0A0Q6WCF2	57	4	13	67	143
490	0KNP2	57	4	13	67	143
491	0A1H0WIT1	57	4	13	67	143
492	0A941E8M9	57	4	13	67	143
493	0A2T4ZA86	58	4	13	68	145
494	0A3N9TYP1	57	4	13	67	143
495	0A127JVM7	57	4	13	67	143
496	0A1X0Y371	57	4	13	67	143
497	0A345ZK96	57	4	13	67	143
498	0A514BSZ7	57	4	13	67	143
499	0A1Y0I587	57	4	13	67	143
500	0A8J7KB07	57	4	13	67	143
501	0A369C117	57	4	13	67	143
502	0A7W4ZBJ6	57	4	13	67	143
503	2LH63	57	4	13	66	142
504	0A656HHX1	58	4	13	67	144
505	0A975F8D6	58	4	13	67	144
506	5E4A0	57	4	13	65	141
507	0A3G9HIR8	59	4	13	68	146
508	0A7U4RRN5	58	4	13	67	144
509	0A0C3HQS3	57	4	13	66	142
510	0QRB5	57	4	13	67	143
511	0A1A8T5K2	57	4	13	67	143
512	0A0M1J962	58	4	13	67	144
513	0A0M1JCL3	58	4	13	69	146
514	0A1M7Y327	57	4	13	67	143
515	0A444IR55	57	4	13	67	143
516	0A1X1QTX6	57	4	13	67	143
517	0A368N552	57	4	13	67	143
518	0A5C2HFI6	57	4	13	67	143
519	0A656HBV7	59	4	13	68	146
520	0A839HCU7	57	4	13	67	143

521	0A5M8FSJ4	57	4	13	67	143
522	0A4R3MVQ5	57	4	13	67	143
523	9VT92	57	4	13	67	143
524	0A6G7VGH4	57	4	13	67	143
525	0A1H3DM17	57	4	13	67	143
526	3RV11	57	4	13	67	143
527	9UFC4	57	4	13	67	143
528	0A6M0JVI1	57	4	13	67	143
529	0A2S7XMX2	57	4	13	67	143
530	3Y9R0	57	4	13	67	143
531	0A105T3E7	57	4	13	66	142
532	0A4R2N4U8	59	4	13	68	146
533	0A1H9MW73	59	4	14	68	147
534	0A368L7V7	64	4	13	68	151
535	0A377PTI4	60	4	13	66	145
536	0A347VPX9	59	4	13	66	144
537	3XEB6	59	4	13	66	144
538	0A3D8IL90	59	4	13	66	144
539	0A315ZCV1	61	4	13	77	157
540	0A316A3R7	59	4	13	80	158
541	0A150ABK1	57	4	13	78	154
542	0A1S1Z5I7	57	4	13	78	154
543	0A3Q9FM84	57	4	13	78	154
544	0A1J0LJ07	58	4	13	80	157
545	0A2W2AGI1	59	4	13	78	156
546	0A553FZ54	58	4	13	77	154
547	0A1Z5H8C4	59	4	13	78	156
548	0A6P0UJ14	59	4	13	79	157
549	7Y3S1	59	4	13	79	157
550	0A968GIA3	58	4	13	63	140
551	7JZV8	61	4	13	73	153
552	7KGJ9	61	4	13	74	154
553	0UL07	61	4	13	74	154
554	0A1I2XSY6	66	4	13	68	153
555	0A1G5B2P5	66	4	13	66	151
556	0A1I5E6V6	66	4	13	65	150
557	0A1W6SMQ7	69	4	13	66	154
558	0A1G5F831	69	4	13	66	154
559	0A1H7N366	66	4	13	66	151
560	0A1K1ME06	67	4	13	67	153
561	2YBV2	66	4	13	67	152
562	2YBU8	67	4	13	66	152
563	0A1I5CT63	66	4	13	66	151
564	0A1I3B2T9	66	4	13	65	150
565	0A1T4KUA6	76	4	13	70	165
566	0A4Q7KCU3	76	4	13	68	163
567	0A6B2U1R0	77	4	13	67	163
568	0A7Y4KWA3	77	4	13	67	163
569	0A8J3ZD69	73	4	13	62	154
570	0A8J3ZZ97	73	4	13	62	154
571	0A6M6JID0	77	4	13	67	163
572	0A4Y3WPN1	77	4	13	67	163
573	0A1Q9TC53	77	4	13	66	162
574	0A543G9U7	77	4	13	66	162
575	0A1Q9S8R4	77	4	13	66	162
576	0A3N1HRD0	77	4	13	65	161
577	0A0H5CCD5	72	4	13	66	157
578	0A1Y5WVD9	78	4	13	66	163

579	0A1Q9LJE7	72	4	13	65	156
580	7J5K9	72	4	13	65	156
581	0A0N9IA08	76	4	13	65	160
582	0A1H0QUG7	76	4	13	65	160
583	0A1V2QKU8	76	4	13	65	160
584	0A7W7CE49	76	4	13	65	160
585	0A563F328	76	4	13	65	160
586	0A1B2HW54	76	4	13	65	160
587	0A1W2B0Y0	76	4	13	65	160
588	0A5Q0H0Z7	77	4	13	65	161
589	0A1Q8CY06	76	4	13	65	160
590	0A1H3SSE1	76	4	13	65	160
591	4T9E2	76	4	13	65	160
592	0A5P9PX90	76	4	13	65	160
593	0A2T0T4S0	76	4	13	65	160
594	2V0Z9	59	4	13	70	148
595	0A1Y6BWZ0	56	4	13	68	143
596	0A2C0ZNK4	57	4	13	66	142
597	0A428J370	60	4	13	61	140
598	0A2I6S3Y3	57	4	13	68	144
599	0KJF2	57	4	13	67	143
600	0A5B8CQT3	57	4	13	67	143
601	0A0X3TBS5	60	4	13	68	147
602	0A1Y2KAJ3	57	4	13	67	143
603	0A370DD53	57	4	13	69	145
604	0A5M8FJI6	59	4	13	69	147
605	4LFG5	59	4	13	69	147
606	0A0R2HVY7	62	4	13	68	149
607	0A099WE44	65	4	13	68	152
608	0A099VYP7	64	4	13	68	151
609	0A099WFW4	65	4	13	68	152
610	0A099WCR9	65	4	13	68	152
611	0A1D2KNE6	62	4	13	69	150
612	7CKB6	62	4	13	69	150
613	8Y8T3	60	4	13	69	148
614	0A3D8TW03	60	4	13	69	148
615	0A2X3GUN1	60	4	13	69	148
616	0A928Z8X0	57	4	13	67	143
617	0A7W6J4X6	57	4	13	68	144
618	0A2U2DU38	57	4	13	68	144
619	0A328RTQ7	56	4	13	68	143
620	9RLW0	57	4	13	68	144
621	0A238ZUJ0	57	4	13	67	143
622	0A1H6HEB9	57	4	13	69	145
623	0A4Q1KFH5	57	4	13	68	144
624	0A1H9PQX7	57	4	13	69	145
625	0A1I4T636	57	4	13	67	143
626	0A1I6IV77	57	4	13	67	143
627	0A1A9HPS8	57	4	13	67	143
628	0A839HI72	58	4	13	67	144
629	0A519MFX2	57	4	13	67	143
630	0A0Q4RUW2	57	4	13	67	143
631	0A0G3BQK6	57	4	13	68	144
632	0A1S8FCW8	57	4	13	68	144
633	0A7X6EBJ2	57	4	13	66	142
634	0A4V2FHR6	57	4	13	66	142
635	0A2D2DJ22	59	4	13	68	146
636	0A316EYD1	57	4	13	67	143

637	3SGX3	58	4	13	68	145
638	0A1E4US35	57	4	13	67	143
639	0A286D739	57	4	13	66	142
640	0A1G9I543	57	4	13	67	143
641	0A0C4WPJ3	59	4	13	67	145
642	0A2T5HBY4	59	4	13	65	143
643	0A923HHC5	59	4	13	65	143
644	0A8J7FPL7	59	4	13	66	144
645	0A1W6YVI4	57	4	13	67	143
646	0A1E4QWB2	59	4	13	67	145
647	0A916IQS8	59	4	13	67	145
648	0A1I4R7R6	57	4	13	67	143
649	0A257CVK6	57	4	13	67	143
650	0A6B3SRF7	57	4	13	67	143
651	0A1S9AFH5	57	4	13	67	143
652	0A4Y9SU23	59	4	13	66	144
653	0A429WBA0	60	4	13	67	146
654	0A1G6PT42	60	4	13	67	146
655	0A934W7U0	59	4	13	67	145
656	0JYI9	59	4	13	67	145
657	0A516SLZ9	57	4	13	68	144
658	0A7H9BGJ7	57	4	13	69	145
659	0A6M8SRC4	57	4	13	69	145
660	0A4E0QUT2	55	4	13	68	142
661	0A3E1RA38	59	4	13	68	146
662	0A1V4CDK1	59	4	13	68	146
663	0A1T1AYT6	59	4	13	68	146
664	0A975I269	59	4	13	68	146
665	0A257CQY8	59	4	13	68	146
666	0A257GVH0	59	4	13	68	146
667	0A925JLT1	59	4	13	68	146
668	0A1P8K6V2	59	4	13	68	146
669	0A255ZJR8	59	4	13	68	146
670	0A1Y0NAK6	59	4	13	69	147
671	0A0N1B541	59	4	13	68	146
672	0A1H0RA26	59	4	13	68	146
673	5NEN9	59	4	13	68	146
674	0A1Q8YGD0	59	4	13	68	146
675	0A7Y8GZ87	59	4	13	68	146
676	0A162Z1X5	59	4	13	68	146
677	0A1I0F8A3	58	4	13	67	144
678	0A1K2HRE0	57	4	13	68	144
679	0A2G3K9R9	57	4	13	68	144
680	0A1W1XMP5	57	4	13	68	144
681	0A1I4X9G6	57	4	13	68	144
682	0A8J7KEF9	57	4	13	69	145
683	0A840RC34	60	4	13	69	148
684	4SVW8	58	4	13	66	143
685	0A975SKD1	56	4	13	68	143
686	0UHC8	57	4	13	67	143
687	0A0M1JWG8	68	4	13	67	154
688	0C685	58	4	13	66	143
689	9UJB1	57	4	13	67	143
690	0A0M2PZ81	57	4	13	67	143
691	0A930SLU4	57	4	13	67	143
692	9ZQ32	57	4	13	67	143
693	0A1Z4UTG3	57	4	13	67	143
694	0A916PVM7	57	4	13	67	143

695	0A256BID6	57	4	13	67	143
696	0A7W9Z367	61	4	13	68	148
697	0A7V8SWC8	62	4	13	68	149
698	0A833JDL7	60	4	13	67	146
699	0A1L4CZ52	60	4	13	68	147
700	0A1G8FNA4	57	4	13	69	145
701	0A0M2UXZ8	57	4	13	69	145
702	3IND1	57	4	13	69	145
703	0A0M2UYR7	57	4	13	69	145
704	0A916FTC1	57	4	13	69	145
705	0A0C9PTG1	57	4	13	69	145
706	0A0F2J7T5	61	4	13	69	149
707	0A0F3GWA5	57	4	13	68	144
708	0A7X6DLN3	57	4	13	68	144
709	4W0E2	57	4	13	67	143
710	0YNQ5	57	4	13	67	143
711	0A0F5YHS2	56	4	13	67	142
712	0A401ILZ0	57	4	13	68	144
713	7JY43	57	4	13	68	144
714	94170	62	4	13	68	149
715	9QFY1	62	4	13	68	149
716	9QTD3	57	4	13	68	144
717	0A218QSN5	57	4	13	68	144
718	0A8J6XET3	57	4	13	68	144
719	0A0V7ZBP0	58	4	13	68	145
720	0A0V8A038	57	4	13	68	144
721	0A0N1BIE2	57	4	13	66	142
722	0A0F3IQ30	57	4	13	66	142
723	0A109BC18	57	4	13	66	142
724	0A4R3JAC6	57	4	13	66	142
725	0A7T5E091	57	4	13	68	144
726	0A8G2BFV5	58	4	13	65	142
727	0A2N3PWA4	57	4	13	67	143
728	0A6J4F4J9	57	4	13	67	143
729	2W9U0	57	4	13	67	143
730	0A178MSE1	57	4	13	67	143
731	2CFR5	59	4	13	67	145
732	0A1I4D1B5	57	4	13	67	143
733	0A1M5BV72	57	4	13	67	143
734	0A7S8C5G6	57	4	13	68	144
735	0A1W6P227	59	4	13	68	146
736	0A8J7U0Y0	59	4	13	68	146
737	0A3R7MB70	57	4	13	68	144
738	0A562NX85	57	4	13	67	143
739	0A844ZKI5	57	4	13	68	144
740	0A6H2DRC4	57	4	13	68	144
741	0A257J3Y8	57	4	13	68	144
742	0A2N4X4A2	57	4	13	68	144
743	0A2U1SV93	56	4	13	66	141
744	2RTM0	57	4	13	67	143
745	0A918WKU1	58	4	13	65	142
746	0A192IE56	60	4	13	65	144
747	0A0U3NEU3	58	4	13	66	143
748	0A5D0VJ35	58	4	13	66	143
749	0A559IPU8	61	4	13	68	148
750	0A931N956	61	4	13	68	148
751	0A089ITG6	61	4	13	68	148
752	0A5J5FWP7	61	4	13	68	148

753	0A089KQV5	61	4	13	68	148
754	0A917C8L4	61	4	13	68	148
755	0A098MAZ0	61	4	13	68	148
756	0A2Z2KD18	61	4	13	68	148
757	0A089M7S8	61	4	13	68	148
758	0A172ZFD3	61	4	13	68	148
759	0A2S0U6Q0	61	4	13	68	148
760	0A7S8IFC0	57	4	13	69	145
761	0A930ULN4	62	4	13	71	152
762	0A437QYD3	60	4	13	66	145
763	3UG46	57	4	13	65	141
764	0A1G9TJX6	58	4	13	67	144
765	1XJX0	59	4	13	66	144
766	0A2I2A601	69	4	13	69	157
767	9H2S2	59	4	13	70	148
768	0A4R6WUE4	60	4	13	65	144
769	0A8H9GFR8	59	4	13	64	142
770	0A4Q7DFD4	58	4	13	64	141
771	0A4R8XU78	58	4	13	64	141
772	0A3M0BYZ5	55	4	13	68	142
773	0A1G7E7L6	57	4	13	68	144
774	0A0H3K2I0	55	4	13	66	140
775	31NF1	55	4	13	66	140
776	0A918Q9P3	57	4	13	68	144
777	2K7G1	57	4	13	68	144
778	0GWR7	57	4	13	68	144
779	0A2K8UDA5	57	4	13	69	145
780	0DWJ7	57	4	13	69	145
781	0A6M0K3V1	57	4	13	69	145
782	3YH20	57	4	13	69	145
783	9UBP8	57	4	13	69	145
784	7RTP1	59	4	13	66	144
785	0A1E2VEE2	59	4	13	67	145
786	0A7W3E476	61	4	13	67	147
787	6WWH4	57	4	13	67	143
788	0K7K0	59	4	13	69	147
789	0A223P7L4	63	4	13	65	147
790	0LQY9	59	4	13	69	147
791	0A0K6IWV4	60	4	13	63	142
792	0A369Q5W3	57	4	13	63	139
793	0A6I4UYR7	57	4	13	63	139
794	0A1C7D934	57	4	13	63	139
795	0A844YBI6	59	4	13	63	141
796	0A7S8F4M8	59	4	13	63	141
797	0A1Y6ESN9	59	4	13	63	141
798	0A345YEW0	57	4	13	63	139
799	0A246K0K4	59	4	13	63	141
800	0A419QYN6	57	4	13	62	138
801	0A6I4TZG1	57	4	13	62	138
802	7G500	60	4	13	62	141
803	8KUA4	60	4	13	62	141
804	5VAH2	60	4	13	62	141
805	0A428MLN5	58	4	13	68	145
806	0A150AEP4	64	4	13	73	156
807	0A223V436	64	4	13	70	153
808	0A1L7I766	64	4	13	70	153
809	0A099W5Q2	61	4	13	68	148
810	7UXI2	60	4	13	69	148

811	7C7B0	60	4	13	69	148
812	0A553ZTQ4	63	4	13	66	148
813	0A1N6NYA0	61	4	13	67	147
814	0A974SNN9	57	4	13	68	144
815	0A0P0MD32	57	4	13	68	144
816	0A6G8NNH1	60	4	13	69	148
817	0A0N0GR34	60	4	13	69	148
818	0A016XEC4	60	4	13	68	147
819	0A5S3P8E3	57	4	13	63	139
820	0A7Y9XSE5	59	4	13	63	141
821	0A255XLT7	57	4	13	66	142
822	0A2A8CWV9	57	4	13	64	140
823	0A919ANZ4	57	4	13	68	144
824	4NF13	57	4	13	68	144
825	4Q810	57	4	13	68	144
826	4P8W5	57	4	13	68	144
827	0A918Z897	60	4	13	71	150
828	1L790	60	4	13	77	156
829	0A7X6R157	61	4	13	67	147
830	5EM87	61	4	13	66	146
831	0A2M9CZD7	59	4	13	65	143
832	0A512PHT7	57	4	13	66	142
833	0A1H0CFP1	57	4	13	66	142
834	0A1B1AG81	59	4	13	67	145
835	0A7G6WPD9	57	4	13	67	143
836	0A3N2IIW9	58	4	13	63	140
837	0A653XLC9	62	4	13	70	151
838	0A2T6L7C1	59	4	13	64	142
839	0A7W3JCK4	59	4	13	63	141
840	3UGG4	58	4	13	66	143
841	3JC10	62	4	14	69	151
842	0A3D8L8C3	59	4	13	69	147
843	0A367GLC1	61	4	13	68	148
844	0A1T5MM01	59	4	13	68	146
845	0A368JGH7	59	4	13	68	146
846	0A848LES9	59	4	13	68	146
847	1FAZ5	66	4	13	68	153
848	0A975GDN9	61	4	13	61	141
849	0A518E3J9	60	4	13	67	146
850	0A0S2HX11	58	4	13	66	143
851	1XWM8	60	4	13	66	145
852	0A162ZH30	60	4	13	68	147
853	0A937D6U5	60	4	13	68	147
854	0A554VJC6	60	4	13	68	147
855	0A162ZH48	60	4	13	68	147
856	0A1J0LNU7	58	4	13	68	145
857	0A5Q0JSE4	59	4	13	69	147
858	0A6L9EDZ9	58	4	13	68	145
859	0A2I2A583	59	4	13	74	152
860	0A7W3FQ27	61	4	13	70	150
861	0A3L7AXU9	61	4	13	70	150
862	0A543I4H5	61	4	13	70	150
863	0A679HTE1	58	4	13	66	143
864	0A4R1FUU6	58	4	13	66	143
865	0A1Y0KNA3	61	4	13	70	150
866	0A4U8UC68	58	4	13	66	143
867	0A366XZH3	61	4	13	68	148
868	0A0A3JS54	61	4	13	68	148

869	0A0A3I491	61	4	13	68	148
870	0A3N9UJ80	61	4	13	67	147
871	0A0A3IRI4	65	4	13	68	152
872	0A3S0JSG6	61	4	13	68	148
873	0A494Z7D4	61	4	13	68	148
874	0A917THM3	63	4	13	67	149
875	0A6A8DLG7	61	4	13	67	147
876	0A1H7WMP5	60	4	13	67	146
877	0A521FZ84	57	4	13	67	143
878	8CVC3	58	4	13	67	144
879	5M2Y3	58	4	13	67	144
880	7DKN1	57	4	13	67	143
881	0A2Z3I273	58	4	13	67	144
882	0A974RXH8	58	4	13	67	144
883	0A7X4LPR7	57	4	13	66	142
884	0A5P9CSF8	57	4	13	67	143
885	0A3S8ZRN4	57	4	13	69	145
886	5V5J1	60	4	13	68	147
887	0A3D8J8E9	58	4	13	66	143
888	0A268U0E8	58	4	13	66	143
889	0A3D8ILV7	58	4	13	66	143
890	0A377PWA0	60	4	13	66	145
891	7VHP3	60	4	13	66	145
892	0A3D8IFV9	58	4	13	66	143
893	3XFH2	58	4	13	66	143
894	0A2A2GTF9	58	4	13	66	143
895	0A347VNU5	58	4	13	66	143
896	0A348HBC8	59	4	13	66	144
897	0A377IWL3	59	4	13	66	144
898	0A0A3ASR9	58	4	13	66	143
899	7RTP2	60	4	13	66	145
900	0A220S4X2	59	4	13	66	144
901	1NKA1	58	4	13	66	143
902	8NA28	58	4	13	66	143
903	0A221MC69	60	4	13	67	146
904	0A2N1J0Z5	58	4	13	68	145
905	8CD88	60	4	13	66	145
906	1CPH1	61	4	13	66	146
907	0A1X9SQA1	57	4	13	66	142
908	0A1B8J6Q6	58	4	13	66	143
909	0A3D8HLH8	58	4	13	66	143
910	0A978S3T6	59	4	13	67	145
911	0UHC7	57	4	13	67	143
912	4F2N2	61	4	13	68	148
913	0A6G6Y8T0	57	4	13	68	144
914	0A2A8HV68	57	4	13	68	144
915	0A0B1ZNK5	57	4	13	68	144
916	0A091AMR7	61	4	13	61	141
917	8AQF8	59	4	13	67	145
918	0A2D0NA22	59	4	13	67	145
919	0A850L000	57	4	13	67	143
920	0A1N6EUJ4	61	4	13	68	148
921	0A4P9K646	60	4	13	68	147
922	0A6F8PWL7	60	4	13	68	147
923	0A066ZZ32	60	4	13	68	147
924	0A6F8PQ29	60	4	13	68	147
925	0A4P7P0H1	60	4	13	68	147
926	0A2I2A807	60	4	13	68	147

927	0A7C9L081	59	4	13	68	146
928	0A0R0MET7	59	4	13	68	146
929	0A1P8K395	59	4	13	68	146
930	3U4N2	58	4	13	66	143
931	0A4U1HNN8	60	4	13	70	149
932	0A0Q5PD36	60	4	13	69	148
933	0A847RYQ9	57	4	13	68	144
934	0A7X0HNU3	61	4	13	61	141
935	1SDL8	57	4	13	67	143
936	0A1H2QJQ5	59	4	13	67	145
937	0A554A0P7	61	4	13	66	146
938	5WD44	60	4	13	66	145
939	0A060LX93	60	4	13	66	145
940	0A286G913	59	4	13	69	147
941	0A2P6BWE1	56	4	13	67	142
942	6XQK9	58	4	13	66	143
943	0A4V3RYL1	58	4	13	66	143
944	0A929FAU4	56	4	13	68	143
945	0A090AP97	57	4	13	68	144
946	0A2M6V0L1	57	4	13	68	144
947	0A315CB47	57	4	13	68	144
948	0A2M6VN05	57	4	13	68	144
949	0A964FG15	57	4	13	67	143
950	9VK84	57	4	13	70	146
951	0A6J4YMQ8	66	4	13	68	153
952	0A5J4F9M3	57	4	13	68	144
953	1X0B4	57	4	13	68	144
954	8MBG6	57	4	13	68	144
955	0A2Z5T0Y6	57	4	13	68	144
956	0A369KLY7	60	4	13	68	147
957	0A929ARE4	65	4	13	66	150
958	1WTG1	58	4	13	67	144
959	4NK12	57	4	13	68	144
960	9QJV3	57	4	13	64	140
961	8P3S8	57	4	13	70	146
962	0A1G6XP73	57	4	13	70	146
963	0A257JJ77	57	4	13	65	141
964	0A2K9EP71	60	4	13	65	144
965	0A6B8KDG2	58	4	13	66	143
966	0A5B8FTC6	57	4	13	66	142
967	0A5C4N3D9	61	4	13	65	145
968	0A329ZJZ3	58	4	13	66	143
969	0A1Z4BU49	61	4	13	68	148
970	0A2P0B555	57	4	13	63	139
971	0A4T3F2N2	57	4	13	68	144
972	0A1B6ZAY8	57	4	13	68	144
973	0A4R6YTA6	61	4	13	65	145
974	0A1G8A0F2	57	4	13	68	144
975	0A4R0YQV1	60	4	13	70	149
976	0A4V2NMF2	60	4	13	70	149
977	0A1B6ADS4	58	4	13	67	144
978	3C8M9	58	4	13	67	144
979	0A4R4T9P2	60	4	13	67	146
980	0A505DMS3	60	4	13	67	146
981	0A5N8W227	60	4	13	68	147
982	0A7D6Z090	61	4	13	67	147
983	0A4V3CPP5	61	4	13	66	146
984	0A285KVG5	61	4	13	67	147

985	0EWN3	61	4	13	65	145
986	0A379JMS1	61	4	13	65	145
987	0A931ICS3	61	4	13	64	144
988	9S943	61	4	13	62	142
989	8MF79	59	4	13	69	147
990	0A2A9DPZ8	57	4	13	67	143
991	0A1F1URG0	57	4	13	67	143
992	0A542UR49	58	4	13	66	143
993	0A0Q5VC30	57	4	13	66	142

Metazoan ACA

ACC	Q-H1	H2-E	E-H3	H3-T	Total	
0	A0A0B4LHX0	70	10	13	86	181
1	A0A8M2BBM0	66	10	13	83	174
2	A0A8M2BGJ2	65	10	13	82	172
3	A0A8M3AWN7	64	11	13	81	171
4	A0A8M3B325	65	10	13	81	171
5	A8KB74	66	10	13	81	172
6	A9JSU9	64	10	13	83	172
7	E9QB97	64	12	13	82	173
8	F1Q816	66	11	13	89	181
9	F1QK51	64	10	13	82	171
10	O43570	64	11	13	81	171
11	O70354	66	10	13	83	174
12	O75493	66	10	13	83	174
13	P00915	66	10	13	80	171
14	P00918	66	10	13	79	170
15	P00920	66	10	13	79	170
16	P07451	66	10	13	79	170
17	P13634	66	10	13	80	171
18	P16015	66	10	13	79	170
19	P18761	64	12	13	82	173
20	P22748	64	10	13	85	174
21	P23280	64	12	13	82	173
22	P23589	66	10	13	80	171
23	P28651	67	10	13	82	174
24	P35218	66	10	13	80	171
25	P35219	67	10	13	82	174
26	P43166	66	10	13	80	171
27	P61215	66	10	13	83	174
28	Q16790	65	10	13	81	171
29	Q3B739	66	10	13	83	174
30	Q4V4S9	70	10	13	87	182
31	Q568S6	67	10	13	82	174
32	Q64444	64	10	13	79	168
33	Q6DBS1	64	10	13	82	171
34	Q6PBI7	66	10	13	80	171
35	Q8CI85	65	11	13	81	172
36	Q8IMV4	71	10	13	86	182
37	Q8N1Q1	66	10	13	80	171
38	Q8VHB5	65	10	13	81	171
39	Q92051	66	10	13	79	170
40	Q99N23	71	10	13	84	180
41	Q9D6N1	66	10	13	80	171
42	Q9ERQ8	66	10	13	80	171

43	Q9NS85	66	10	13	83	174
44	Q9QZA0	66	10	13	80	171
45	Q9ULX7	64	11	13	82	172
46	Q9V396	67	10	13	81	173
47	Q9V9Y6	70	10	13	86	181
48	Q9V9Y8	70	10	13	86	181
49	Q9VB76	75	10	13	83	183
50	Q9VH26	65	4	13	84	168
51	Q9VTU8	72	10	13	82	179
52	Q9W316	73	10	13	81	179
53	Q9W3C8	69	10	13	83	177
54	Q9W3P7	66	10	13	81	172
55	Q9WVT6	64	11	13	82	172
56	Q9Y2D0	66	10	13	80	171
57	R4GDY8	66	11	13	77	169

Appendix 2. Developed codes for analysis of DCA

```
#!/usr/bin/env python
# coding: utf-8

# In[1]:


import numpy as np
import ruptures as rpt
import pandas as pd
import matplotlib.pyplot as plt
from Bio import AlignIO
from Bio.Align import MultipleSeqAlignment
import requests
import time
from Bio.PDB import PDBParser
import os
import seaborn as sns
import plotly.express as px


# In[2]:


logofile = "./data/logo60seq.txt"
msafile = "./data/60seq.aln"
alignment = AlignIO.read(msafile, "clustal")
print(f"Length of alignment file {len(alignment)} records")
lencheck=[]
for record in alignment:
    lencheck.append(len(record.seq))
print(f"Unique seq lengths in MSA: {set(lencheck)}")
if len(set(lencheck)) > 1:
    raise Exception(f"Sequences of inequal length in {msafile}")

dfweights0 = pd.read_csv(logofile, sep="\t", header=7)
dfweights0.rename(columns={"#": "position"}, inplace=True)
# dfweights0.to_excel("./data/logo466.xlsx")
logo_width = len(dfweights0)
print(f"MSA columns in logofile {logo_width}")
if lencheck[0] != logo_width:
    raise Exception(f"Mismatch between {logofile} and {msafile}")


# In[3]:


SMALL_SIZE = 16
MEDIUM_SIZE = 20
BIGGER_SIZE = 24
plt.rc('font', size=SMALL_SIZE)          # controls default text sizes
plt.rc('axes', titlesize=SMALL_SIZE)     # fontsize of the axes title
plt.rc('axes', labelsize=BIGGER_SIZE)    # fontsize of the x and y labels
plt.rc('xtick', labelsize=MEDIUM_SIZE)    # fontsize of the tick labels
plt.rc('ytick', labelsize=SMALL_SIZE)    # fontsize of the tick labels
```

```
plt.rc('legend', fontsize=SMALL_SIZE)    # legend fontsize
plt.rc('figure', titlesize=BIGGER_SIZE)  # fontsize of the figure title


# In[4]:


dfweights0 = pd.read_csv(logofile, sep="\t", header=7)  # with 60 records
(62-2=60)
dfweights0.rename(columns={"#": "position"}, inplace=True)
fig, ax1 = plt.subplots(figsize=(16, 5))
ax2 = ax1.twinx()
dfweights0.plot(x='position', y='Entropy', ax=ax1, legend=False)
dfweights0.plot(x='position', y='Weight', ax=ax2, legend=False, color='r')
ax1.set_xlabel('Position')
ax1.set_ylabel('Entropy', color='b')
ax2.set_ylabel('Weight', color='r')
#plt.xlim(left=350, right=450) # to check start of the domain
plt.show()


# In[5]:


# running average of n residues
fig, ax1 = plt.subplots(figsize=(16, 5))
ax2 = ax1.twinx()
dfweights0['Entropy_mean'] = dfweights0['Entropy'].rolling(window=10,
center=True).mean() # n=10 residues
dfweights0.plot(x='position', y='Entropy_mean', ax=ax1, legend=False)
dfweights0.plot(x='position', y='Weight', ax=ax2, legend=False, color='r')
ax1.set_xlabel('Position')
ax1.set_ylabel('Entropy', color='b')
ax2.set_ylabel('Weight', color='r')
ax1.set_title('with running average of 10 residues for Entropy')
plt.show()


# In[6]:


# Normalizing the weight and entropy to the range 0 to 1
weight_norm = (dfweights0['Weight'] - dfweights0['Weight'].min()) /
(dfweights0['Weight'].max() - dfweights0['Weight'].min())
entropy_norm = (dfweights0['Entropy_mean'] - dfweights0['Entropy_mean'].min()) /
(dfweights0['Entropy_mean'].max() - dfweights0['Entropy_mean'].min())
# the average of the normalized values
joint_quantity = (weight_norm + entropy_norm) / 2
# Adding the joint_quantity column to the df
dfweights0['Average'] = joint_quantity
# Plotting the joint_quantity as a line
fig, ax = plt.subplots(figsize=(16, 5))
dfweights0.plot(x='position', y='Average', ax=ax, legend=False, color='g')
ax.set_xlabel('Position')
```

```
ax.set_ylabel('Average', color='g')
ax.set_title('Normalized Average of weight and entropy of residues')
ax.grid()
plt.show()


# In[7]:


algorithm =
rpt.Pelt(model="l1",min_size=1).fit(np.asanyarray(dfweights0["Weight"]))
result = algorithm.predict(pen=2)
rpt.display(np.asanyarray(dfweights0["Weight"]), result, figsize=(16, 5))
#plt.xlim(left=300, right=400)
plt.xlabel('Position')
plt.ylabel('Weight')
plt.show()


# In[8]:


dfweights= dfweights0.loc[:, ["position", 'Weight']]


# In[9]:


#downloading of alphafold models for each sequence in an MSA
#if they are not previously in model_dir rewrites the alignment if models were
missing and could not be downloaded
def download_file(url, save_path):
    response = requests.get(url)
    response.raise_for_status()  # Raise an exception if the request was not
successful
    with open(save_path, "wb") as file:
        file.write(response.content)

model_dir = "./data/Models/"
data1= []
uniprot_ids = [record.id.split("|")[1] for record in alignment]
found = 0
not_found = 0
dl_ok = 0
dl_fail = 0

if os.path.isdir(model_dir):
    #checking both that the pathname exist and that it's a directory
    print(f"Using folder {model_dir} to save models")
else:
    raise FileNotFoundError(f"The folder {model_dir} does not exist!")
    # crashes with this error if model_dir is not present

for uniprot_id in uniprot_ids:
```

```
#    file_name = os.path.join(model_dir, f"AF-{uniprot_id}-F1-model_v4.pdb")
    file_name = f"{model_dir}AF-{uniprot_id}-F1-model_v4.pdb"
    if os.path.isfile(file_name) == False:
        not_found +=1
        url =
f"https://alphafold.ebi.ac.uk/files/AF-{uniprot_id}-F1-model_v4.pdb"
        save_path = file_name
        try:
            download_file(url, save_path)
            dl_ok +=1
        except:
            dl_fail +=1
            print(f"{file_name} download failed. Manual download link:")
            print (url)
    else:
        found +=1

print(f"Found {found} models, {not_found} were missing; Downloaded {dl_ok},
failed {dl_fail}.")

if (dl_fail > 0):
    #creating a new MSA if not all models could be downloaded
    print("Rewriting new alignment file of the found records.")
    new_msa = MultipleSeqAlignment([]) #empty MSA object
    seqs = 0

    for record in alignment:
        uniprot_id = record.id.split("|")[1]
        file_name = f"{model_dir}AF-{uniprot_id}-F1-model_v4.pdb"
        if os.path.isfile(file_name): #if the model is there we write it into
the new MSA
            new_msa.append(record)
            seqs += 1
            #would be better if the previous had written a list of record
indices
            #as the files were retrieved

    new_msafile = f"{msafile}.new{seqs}.fasta"
    AlignIO.write(new_msa, new_msafile, "fasta")
    print(seqs, "sequences as", new_msafile)
    print("Now realign the MSA, make a new logo and start with new msafile and
logofile")


# In[10]:


alignment = AlignIO.read(msafile, "clustal")  #alignment with 62 records
data1= []
for record in alignment:
    for i, residue in enumerate(record.seq):
        if residue != "-":
            sequence_id = record.id.split("|")[1]
            data1.append({"Uniprot ID": sequence_id, "Residue": residue,
```

```
"position": i+1})
df1 = pd.DataFrame(data1)
aa_dict = {
    "A": "ALA",
    "R": "ARG",
    "N": "ASN",
    "D": "ASP",
    "C": "CYS",
    "E": "GLU",
    "Q": "GLN",
    "G": "GLY",
    "H": "HIS",
    "I": "ILE",
    "L": "LEU",
    "K": "LYS",
    "M": "MET",
    "F": "PHE",
    "P": "PRO",
    "S": "SER",
    "T": "THR",
    "W": "TRP",
    "Y": "TYR",
    "V": "VAL"
}
df1["Residue"] = df1["Residue"].apply(lambda x: aa_dict.get(x))
#df1 # data frame for Uniprot ID & Residue & column number


# In[11]:


new_df = pd.DataFrame(columns=['weights_collection'])
positions = df1['position']
for position in positions:
    value = dfweights.loc[dfweights['position'] == position].drop('position',
axis=1).values[0][0]
    new_df = pd.concat([new_df, pd.DataFrame({'weights_collection': [value]})],
ignore_index=True)
#new_df # mapping the weights


# In[12]:


df2 = pd.merge(df1, new_df, left_index=True, right_index=True, suffixes=('_df1',
'_dfweights'))
#df2 # data frame for Uniprot ID & Residue & column number & weights


# In[13]:


uniprot_ids = [record.id.split("|")[1] for record in alignment]
model_dir="./data/Models/"
```

```
parser = PDBParser()
data = []
for uniprot_id in uniprot_ids:
#    file_name = os.path.join(model_dir, f"AF-{uniprot_id}-F1-model_v4.pdb")
    file_name = f"{model_dir}AF-{uniprot_id}-F1-model_v4.pdb"
    try:
        structure = parser.get_structure(f"AF-{uniprot_id}-F1-model_v4",
file_name)

        for model in structure:
            for chain in model:
                for residue in chain:
                    b_factor = residue["CA"].get_bfactor()
                    data.append({"Uniprot ID": uniprot_id, "Residue":
residue.get_resname(),
                                 "Residue ID": residue.get_id()[1], "B Factor":
b_factor})
    except FileNotFoundError:
        print(f"File {file_name} does not exist. Continuing to the next
sequence.")
        continue
df3 = pd.DataFrame(data)
#df3  # data frame for Uniprot ID & Residue & B Factor


# In[14]:


ready_df = pd.concat([df3, df2.iloc[:, [2,3]]], axis=1)
ready_df


# In[15]:


x = ready_df['B Factor']
y = ready_df['weights_collection']
sns.regplot(x=x, y=y,line_kws={'color':'orange'}, scatter_kws={'s': 5})
plt.xlabel('pLDDT')
plt.ylabel('MSA weights')
plt.title('Whole MSA positions')
A = [20, 50, 50, 20]
B = [0.6, 0.6, 1, 1]
plt.fill(A, B, 'r',alpha=0.2)
plt.gcf().set_size_inches(16, 5)
plt.show()


# In[16]:


sns.displot(ready_df['B Factor'], kde=True, height=5, aspect=4, bins=39)
plt.title('Whole MSA positions')
plt.show()
```

```
# In[17]:


print("correlation of whole MSA:",ready_df['B
Factor'].corr(ready_df['weights_collection']))
# data is skewed, so total correlation can not be used


# In[18]:


filtered_RED_box = ready_df[(ready_df['weights_collection'] > 0.6) &
(ready_df['B Factor'] < 50)]
# plt.scatter(filtered_RED_box['position'], filtered_RED_box['B Factor'])
# #plt.xlabel('MSA position')
# plt.ylabel('pLDDT')
# # plt.title('Scatter Plot of the RED box', size=24)
# plt.gca().set_facecolor('mistyrose')
# plt.show()
fig, ax = plt.subplots(figsize=(16,5))
ax.scatter(filtered_RED_box['position'], filtered_RED_box['B Factor'])
ax.set_ylabel('pLDDT')
ax.set_facecolor('mistyrose')
plt.show()

# plt.hist(filtered_RED_box['position'], bins=220)
# plt.xlabel('MSA position')
# plt.ylabel('Frequency')
# #plt.title('Histogram Plot of the RED box', size=24)
# plt.gca().set_facecolor('mistyrose')
# plt.show()
fig, ax = plt.subplots(figsize=(16,5))
ax.hist(filtered_RED_box['position'], bins=220)
ax.set_xlabel('MSA position')
ax.set_ylabel('Frequency')
ax.set_facecolor('mistyrose')
plt.show()


# In[19]:


dca_start = 360
dca_end = 800
df_dcadom = ready_df[(ready_df['position'] >= dca_start) & (ready_df['position']
<= dca_end)]
x = df_dcadom['B Factor']
y = df_dcadom['weights_collection']
# sns.set(rc={'figure.figsize':(20,5)})
fig, ax = plt.subplots(figsize=(16,5))
sns.regplot(x=x, y=y,line_kws={'color':'orange'}, scatter_kws={'s': 5})
plt.xlabel('pLDDT')
```

```
plt.ylabel('MSA weight')
plt.title(f'DCA domain {dca_start} to {dca_end}')
# plt.show()


# In[20]:


df_dcadom=ready_df[(ready_df['position'] >= dca_start) & (ready_df['position']
<= dca_end)]
sns.displot(df_dcadom['B Factor'], kde=True, bins=39, height=5, aspect=4)
plt.title(f'Histogram of DCA domain {dca_start} to {dca_end}')


# In[21]:


ready_df_copy = ready_df.copy()
df_360_800=ready_df_copy.loc[(ready_df_copy['position'] >= 360) &
(ready_df_copy['position'] <= 800)]
print("correlation of 360-800:", df_360_800['B
Factor'].corr(df_360_800['weights_collection']))


# In[22]:


ready_df_copy = ready_df.copy()


# In[23]:


ready_df_copy.groupby('position')['B Factor'].mean().plot(figsize=(16,5))
#average pLDDT across the MSA
plt.ylabel('Mean pLDDT')
# the aggregated “bfactor” column by mean based on the “position” column id in
the dataframe


# In[24]:


fig, ax = plt.subplots(figsize=(16, 5))
ax2 = ax.twinx()
dfweights.plot(x="position", y=["Weight"], kind="line", ax=ax, color='red',
ylim=[0,1.25])
ready_df_copy.groupby('position')['B Factor'].mean().plot(ax=ax2, color='b')
ax1.set_ylabel('pLDDT', color='b')
ax2.set_ylabel('MSA weight', color='r')
plt.xlim(330,710)
# plt.show()
```

```
# In[25]:


normalized_b_factors = (ready_df_copy.groupby('position')['B Factor'].
                        mean()-min(ready_df_copy.groupby('position')['B
Factor'].

mean()))/(max(ready_df_copy.groupby('position')['B Factor'].

mean())-min(ready_df_copy.groupby('position')['B Factor'].mean()))
fig, ax = plt.subplots(figsize=(16, 5))
dfweights['pLDDT times MSA weight'] = normalized_b_factors * dfweights['Weight']
dfweights.plot(x="position", y=["pLDDT times MSA weight"], kind="line", ax=ax,
color='g')
ax.set_ylabel('mean pLDDT * MSA weight', color='g')
plt.xlabel("position")
plt.xlim(330,710)
plt.show()
plt.plot(normalized_b_factors) #normalized Mean pLDDT plot similar to above
non-normalized; for double_check


# ## Statistics of good-quality residues in AF models

# In[26]:


limit0 = pd.DataFrame(ready_df_copy.groupby('position')['B Factor'].mean())
limits = limit0.reset_index()
limits.columns = ['position', 'B Factor']
column_values1=limits.loc[(limits['position'] >= dca_start)
                          &(limits['position'] <= dca_end)
                          & (limits['B Factor'] >= 70) ]['position'].values
len(column_values1) #for catalytic domain limits &&& 70 < B Factor


# In[27]:


column_values2= limits.loc[(limits['position'] >= dca_start)
                           &(limits['position'] <= dca_end)
                           &(limits['B Factor'] >= 90) ]['position'].values
len(column_values2) #for 360 <'position'<800 &&& 90 < B Factor


# In[28]:


column_values3 = limits.loc[(limits['position'] >= dca_start)
                           &(limits['position'] <= dca_end)
                            & (limits['B Factor'] < 70) ]['position'].values
len(column_values3) #for 360 <'position'<800 &&&  B Factor < 70
```

```
# In[29]:


column_values4= limits.loc[(limits['position'] >= dca_start)
                           &(limits['position'] <= dca_end)
                           &(limits['B Factor'] >= 95) ]['position'].values
len(column_values4)  #supergood residues 95 < B Factor


# ### Statistics of good-quality residues  for each Uniprot ID

# In[30]:


filtered_df1 = pd.DataFrame(columns=['Uniprot ID', 'Column Values'])
for uniprot_id in ready_df['Uniprot ID'].unique():
    column_values = ready_df.loc[(ready_df['position'] >= dca_start)
                                 & (ready_df['position'] <= dca_end)
                                 & (ready_df['B Factor'] >= 70)
                                 & (ready_df['Uniprot ID'] ==
uniprot_id)]['position'].values
    filtered_df1 = pd.concat([filtered_df1, pd.DataFrame({'Uniprot ID':
[uniprot_id],
                                                          'Column Values':
[column_values]})])
    residue_count = ready_df[(ready_df['position'] >= dca_start)
                             & (ready_df['position'] <= dca_end) &
                             (ready_df['Uniprot ID'] ==
uniprot_id)].groupby('Uniprot ID')['Residue'].count()
    filtered_df1.loc[filtered_df1['Uniprot ID'] == uniprot_id,
                     f"{dca_start} < Residue Count < {dca_end}"] =
residue_count[0]
filtered_df1 = filtered_df1.reset_index(drop=True)
filtered_df1['Len of Column Values no. of > 70'] = filtered_df1['Column
Values'].apply(len)


# In[31]:


# no. of > 90
filtered_df2 = pd.DataFrame(columns=['Uniprot ID', 'Column Values'])
for uniprot_id in ready_df['Uniprot ID'].unique():
    column_values = ready_df.loc[(ready_df['position'] >= dca_start)
                                 & (ready_df['position'] <= dca_end)
                                 & (ready_df['B Factor'] >= 90)
                                 & (ready_df['Uniprot ID'] ==
uniprot_id)]['position'].values
    filtered_df2 = pd.concat([filtered_df2, pd.DataFrame({'Uniprot ID':
[uniprot_id], 'Column Values': [column_values]})])
filtered_df2 = filtered_df2.reset_index(drop=True)
filtered_df2['Len of Column Values no. of > 90'] = filtered_df2['Column
Values'].apply(len)
```

```
# In[32]:


# no. of < 70
filtered_df3 = pd.DataFrame(columns=['Uniprot ID', 'Column Values', 'Residue'])
for uniprot_id in ready_df['Uniprot ID'].unique():
    column_values = ready_df.loc[(ready_df['position'] >= dca_start) &
                                 (ready_df['position'] <= dca_end) &
                                 (ready_df['B Factor'] < 70) &
                                 (ready_df['Uniprot ID'] ==
uniprot_id)]['position'].values
    residues = ready_df.loc[(ready_df['position'] >= dca_start) &
                            (ready_df['position'] <= dca_end) &
                            (ready_df['B Factor'] < 70) &
                            (ready_df['Uniprot ID'] ==
uniprot_id)]['Residue'].values
    filtered_df3 = pd.concat([filtered_df3, pd.DataFrame({'Uniprot ID':
[uniprot_id],
                                                          'Column Values':
[column_values],
                                                          'Residue':
[residues]})])

filtered_df3 = filtered_df3.reset_index(drop=True)
filtered_df3['Len of Column Values no. of < 70'] = filtered_df3['Column
Values'].apply(len)


# In[33]:


#70…80
filtered_df4 = pd.DataFrame(columns=['Uniprot ID', 'Column Values'])
for uniprot_id in ready_df['Uniprot ID'].unique():
    column_values = ready_df.loc[(ready_df['position'] >= dca_start)
                                 & (ready_df['position'] <= dca_end) &
                                 (ready_df['B Factor'] >= 70) & (ready_df['B
Factor'] <  80) &
                                 (ready_df['Uniprot ID'] ==
uniprot_id)]['position'].values
    filtered_df4 = pd.concat([filtered_df4, pd.DataFrame({'Uniprot ID':
[uniprot_id], 'Column Values': [column_values]})])
filtered_df4 = filtered_df4.reset_index(drop=True)
filtered_df4['80>Len of Column Values no. of > 70'] = filtered_df4['Column
Values'].apply(len)


# In[34]:


#80…90
filtered_df5 = pd.DataFrame(columns=['Uniprot ID', 'Column Values'])
for uniprot_id in ready_df['Uniprot ID'].unique():
```

```
    column_values = ready_df.loc[(ready_df['position'] >= dca_start)
                                 & (ready_df['position'] <= dca_end) &
                                 (ready_df['B Factor'] >= 80) & (ready_df['B
Factor'] <  90) &
                                 (ready_df['Uniprot ID'] ==
uniprot_id)]['position'].values
    filtered_df5 = pd.concat([filtered_df5, pd.DataFrame({'Uniprot ID':
[uniprot_id], 'Column Values': [column_values]})])
filtered_df5 = filtered_df5.reset_index(drop=True)
filtered_df5['90>Len of Column Values no. of > 80'] = filtered_df5['Column
Values'].apply(len)


# In[35]:


filtered_df = pd.concat([filtered_df1.drop('Column Values', axis=1),
                         filtered_df4.drop(['Column Values', "Uniprot
ID"],axis=1),
                         filtered_df5.drop(['Column Values', "Uniprot
ID"],axis=1),
                         filtered_df2.drop(['Column Values', "Uniprot ID"],
axis=1),
                         filtered_df3.drop(["Uniprot ID"], axis=1)], axis=1)
# filtered_df
filtered_df.style.set_table_styles([{'selector': 'th:nth-child(8),
th:nth-child(9), th:nth-child(10)',
                                     'props': [('background-color', 'yellow')]},
                                    {'selector': 'th:nth-child(7)', 'props':
[('background-color', 'blue')]},
                                    {'selector': 'th:nth-child(5),
th:nth-child(6)',
                                     'props': [('background-color',
'#87CEFA')]}])



# In[36]:


fig = px.line(ready_df, x='position', y='B Factor', color='Uniprot ID')
fig.update_layout(legend=dict(orientation="h",yanchor="bottom", y=1.02,
xanchor="right", x=1), width=950, height=600)
fig.show()


# In[41]:


fig, axs = plt.subplots(16, 4, figsize=(20, 80))
for i, uniprot_id in enumerate(ready_df['Uniprot ID'].unique()):
    row = i // 4
    col = i % 4
    axs[row][col].plot(ready_df[ready_df['Uniprot ID'] ==
```

```
uniprot_id]['position'], ready_df[ready_df['Uniprot ID'] == uniprot_id]['B
Factor'])
    axs[row][col].set_title(uniprot_id)
    axs[row][col].set_xlim([360, 800]) #suggestion by Martti
plt.show()
```

Appendix 3. Species list comparisons

#Shared organisms and genera between short and long DCAs

```
['Donghicola eburneus', 'Tropicibacter naphthalenivorans']
['Roseibium', 'Donghicola', 'Glaciecola', 'Aliikangiella', 'Tropicibact
er', 'Pseudoalteromonas']
28
['Pseudoalteromonas' 'Candidatus Entotheonella' nan 'Gallionella'
 'Candidatus Tenderia' 'Magnetococcus' 'Aliikangiella' 'Exilibacterium'
 'Glaciecola' 'Pseudobacteriovorax' 'Henriciella' 'Rhodovulum'
 'Pseudooceanicola' 'Litorivita' 'Roseivivax' 'Albimonas' 'Fulvimarina'
 'Donghicola' 'Roseospira' 'Oceanibacterium' 'Puniceibacterium'
 'Aurantimonas' 'Jiella' 'Tropicibacter' 'Allosediminivita' 'Roseibium'
 'Pikeienuella' 'Salipiger']
68
['Oceanisphaera' 'Ferrimonas' 'Vibrio' 'Psychromonas' 'Rhizobium'
 'Marinomonas' 'Shewanella' 'Veronia' 'Aliikangiella' 'Microbulbifer'
 'Thalassotalea' 'Colwellia' 'Pseudoalteromonas' 'Acidovorax'
 'Salinivibrio' 'Alteromonas' 'Aliiglaciecola' 'Glaciecola'
 'Paraglaciecola' 'Kineobactrum' 'Haliea' nan 'Bacterioplanes'
 'Phocoenobacter' 'Alcanivorax' 'Reinekea' 'Halomonas' 'Oceanimonas'
 'Rugamonas' 'Duganella' 'Hydrogenophaga' 'Ketobacter' 'Magnetofaba'
 'Marinobacter' 'Candidatus Venteria' 'Desulforhopalus' 'Thalassolituus
'
 'Leeia' 'Teredinibacter' 'Mariprofundus' 'Enterovibrio' 'Neptunomonas'
 'Leucothrix' 'Thiosulfatimonas' 'Motiliproteus' 'Novispirillum'
 'Falsiruegeria' 'Actibacterium' 'Labrenzia' 'Tritonibacter' 'Roseobact
er'
 'Epibacterium' 'Pelagimonas' 'Cohaesibacter' 'Roseibium' 'Rhodobacter'
 'Phaeobacter' 'Amylibacter' 'Denitrobaculum' 'Thalassococcus'
 'Pacificibacter' 'Pararhodobacter' 'Donghicola' 'Roseovarius' 'Hoeflea
'
 'Tropicibacter' 'Pseudosulfitobacter' 'Loktanella']
```

#Shared species of the DCA list and pathogens list

```
Shared species []
Shared genera ['Acidovorax', 'Halomonas', 'Vibrio', 'Rhizobium', 'Shewa
nella']
```

Shared species of the ACA list and pathogens

```
45 species with ACA found in the pathogens list
Acinetobacter guillouiae
Aeromonas hydrophila
Aeromonas schubertii
Atlantibacter hermannii
Campylobacter lanienae
Cardiobacterium hominis
Cardiobacterium valvarum
Citrobacter freundii
Citrobacter koseri
```

```
Desmospora activa
Eikenella corrodens
Enterococcus cecorum
Enterococcus durans
Enterococcus faecalis
Haemophilus pittmaniae
Helicobacter bilis
Helicobacter fennelliae
Helicobacter hepaticus
Helicobacter suis
Kingella denitrificans
Kosakonia cowanii
Listeria grayi
Listeria monocytogenes
Massilia timonae
Moellerella wisconsensis
Neisseria bacilliformis
Neisseria gonorrhoeae
Nocardia amikacinitolerans
Nocardia asteroides
Nocardia brasiliensis
Nocardia ignorata
Paenibacillus macerans
Paenibacillus polymyxa
Pantoea agglomerans
Pasteurella multocida
Photobacterium damselae
Rahnella aquatilis
Scandinavium goeteborgense
Serratia fonticola
Streptococcus mutans
Suttonella indologenes
Vagococcus lutrae
Vibrio cholerae
Vibrio furnissii
Vibrio parahaemolyticus

60 genera with ACA found in the pathogens list
Shared genera:
 ['Acidovorax', 'Acinetobacter', 'Aeromonas', 'Amycolatopsis', 'Arcobac
ter', 'Atlantibacter', 'Bacillus', 'Bordetella', 'Brevundimonas', 'Burk
holderia', 'Buttiauxella', 'Campylobacter', 'Cardiobacterium', 'Cellulo
monas', 'Chryseobacterium', 'Citrobacter', 'Corynebacterium', 'Cupriavi
dus', 'Desmospora', 'Dyella', 'Eikenella', 'Enterobacter', 'Enterococcu
s', 'Erwinia', 'Haemophilus', 'Halomonas', 'Helicobacter', 'Herbaspiril
lum', 'Kingella', 'Klebsiella', 'Kosakonia', 'Lactococcus', 'Listeria',
'Lysinibacillus', 'Malaciobacter', 'Massilia', 'Methylobacterium', 'Moe
llerella', 'Neisseria', 'Nocardia', 'Paenibacillus', 'Pantoea', 'Paraco
ccus', 'Pasteurella', 'Photobacterium', 'Pontibacter', 'Providencia', '
Pseudomonas', 'Pseudonocardia', 'Pseudoxanthomonas', 'Rahnella', 'Rhizo
bium', 'Scandinavium', 'Serratia', 'Stenotrophomonas', 'Streptococcus',
'Streptomyces', 'Suttonella', 'Vagococcus', 'Vibrio']
```

Shared species of the ACA list and DCA lists

```
6 exact genomes with both ACA and DCA found
```

```
Duganella fentianensis
Leeia aquatica
Magnetofaba australis IT-1
Pseudobacteriovorax antillogorgiicola
Rugamonas brunnea
Thiosulfatimonas sediminis
9 species with both ACA and DCA found
Acidovorax sp.
Duganella fentianensis
Hydrogenophaga sp.
Leeia aquatica
Magnetofaba australis
Pseudobacteriovorax antillogorgiicola
Rhodovulum sp.
Rugamonas brunnea
Thiosulfatimonas sediminis
17 genera with both ACA and DCA found
Shared genera:
 ['Acidovorax', 'Duganella', 'Halomonas', 'Hydrogenophaga', 'Leeia', 'L
oktanella', 'Magnetofaba', 'Marinomonas', 'Microbulbifer', 'Pseudobacte
riovorax', 'Psychromonas', 'Reinekea', 'Rhizobium', 'Rhodovulum', 'Ruga
monas', 'Thiosulfatimonas', 'Vibrio']
```