Studying the effect of protein structure characteristics on the evolutionary rate of proteins

by

Theodora Kyprianidi

Master of Science in
Protein Biotechnology

2023

Department of Biology
University of Crete

Abstract

The aim of this study is to investigate if the evolutionary rate of proteins is affected, firstly by structural characteristics of proteins and then by protein localization. Information about the evolutionary rate of proteins can provide valuable insights into the functional and structural importance of specific protein regions; furthermore it yields information about the evolutionary history, relationships and functional divergence and finally it can be useful in evolutionary biology, bio-informatics and drug discovery fields. Regarding structural characteristics of proteins, secondary structure characteristics were studied, particularly Alpha helices and Beta sheets in a data set that was extracted from the Protein Data Bank. The intention was to investigate how the quantity of protein secondary structural elements affects its evolutionary rate. We found that proteins containing more amino acids assigned as Beta sheets rather than Alpha helices, tend to evolve faster. We classified our proteins in KEGG pathways, in order to conduct pathway analysis, and compared the evolutionary rates of proteins containing bigger percentage of Alpha helices and Beta sheets within each pathway. From the Uniprot database we extracted the possible locations for each protein and examined their impact on evolutionary rate. Furthermore, we performed a Gene Ontology analysis on both Alpha helix and Beta sheet proteins to identify their corresponding biological processes, molecular functions, and cellular components. This analysis aimed to investigate whether these two types of proteins participate in different gene ontology terms, which could potentially explain the observed differences in their evolutionary rates. We found that B proteins are found more in the extracellular region while A proteins inside the cell; this could possibly explain their difference in the evolutionary rates. For the second part of the study concerning protein localization we used a data set from the G.Gkouridis lab (IMBB-FORTH), there are three protein locations, i.e., heavy membranes, light membranes and solution to investigate (i) if the protein localization affects the evolutionary rate of proteins, and (ii) if secondary structure elements in different locations affect the evolutionary rate.

Contents

Chapter 1

Introduction

1.1 Phylogenetics

Phylogenetics is the study of the evolutionary history and relationships among organisms, and it plays a fundamental role in the understanding of the natural world. By analyzing the similarities and differences in genetic and morphological traits among different species, researchers can reconstruct the evolutionary relationships among those species and infer patterns of evolutionary change over time *Wiley and Lieberman (2011).* Phylogenetics has a long history dating back to the work of Charles Darwin and Alfred Russel Wallace, who proposed the theory of evolution by natural selection in the mid-19th century *Nei et al. (2000).*

1.1.1 Phylogenetic trees

A phylogenetic tree is a visual representation of the evolutionary relationships between different organisms, species, or genes that share a common ancestor. These trees are useful for organizing information about the diversity of life, structuring taxonomic classifications, and providing insights into evolutionary events. Additionally phylogenetic trees demonstrate the shared ancestry of organisms, which is a strong piece of evidence for evolution, a thorough understanding of these trees is necessary to fully comprehend the abundance of evidence supporting the theory of evolution. *Robinson and Foulds (1981).*

A phylogenetic tree consists of external nodes, which are the tips representing the existing sequences, and internal nodes representing hypothetical ancestors. The nodes are connected to each other by branches, where the length of each branch represents an estimation of the amount of change that occurred between two nodes. Currently, molecular data, such as DNA or protein sequences, are primarily used to construct phylogenetic trees. The initial goal of these trees was to determine the relationships among the species represented by the sequences, but their purposes have expanded. Now, they are also used to investigate the relationships among sequences themselves, without taking into account the host species. Additionally, phylogenetic trees are used to infer the functions of genes that have not yet been experimentally studied, as well as to uncover the mechanisms that cause microbial outbreaks, among other applications. *Hall (2013)*

1.1.2 Homologous sequences

Homology is a general term used to describe a relationship of shared ancestry between entities, such as genes, without specifying the exact evolutionary scenario. When entities share homology, they are called homologs. The terms ”orthologs” and ”paralogs” are subcategories of homologs. Orthologs are genes related through speciation, or vertical descent, while paralogs are genes related through duplication.*Koonin (2005)*

Vertical descent is the transmission of genes across generations through normal reproduction and replication mechanisms within a particular species. This process typically involves recombination within sexually reproducing populations. *Zhaxybayeva and Doolittle (2011)*

Gene duplication is a process in which a single gene produces two identical genes, which cannot be distinguished from each other. These duplicated genes are referred to as paralogues if they are present within the same genome or as orthologues if they are present in different genomes. Gene duplication is considered to be significant in evolution as it provides raw material for the development of new gene functions. By producing a duplicate gene, natural selection has more opportunities to create new and innovative functions with less constraints. *Magadum et al. (2013)*

1.1.3 Bulding phylogenetic trees

To construct a phylogenetic tree, four steps are necessary: (i) recognizing and obtaining a group of homologous DNA or protein sequences, (ii) aligning those sequences, (iii) inferring a tree from the aligned sequences, and (iv) presenting the tree in a manner that is understandable to others and communicates the pertinent information. *Hall (2013)* The accuracy and reliability of the phylogenetic tree depend on the quality of the data and the chosen methods. Therefore, it is essential to carefully select the sequences and algorithms used in each step. Additionally, phylogenetic analysis is an iterative process that may involve refining the data and methods to obtain a more accurate tree. It is also important to consider that phylogenetic trees are hypotheses, not definitive facts.

The most tricky part of building a phylogenetic tree is the tree estimation from the aligned sequences. Various methods are commonly used to estimate phylogenetic trees including Maximum Parsimony, Maximum Likelihood, Bayesian Inference, and Neighbor Joining *Hall (2013)*. Maximum Parsimony seeks to construct the tree with the fewest evolutionary events *Mount (2008)*. Maximum Likelihood estimates the tree that best fits the observed data *Felsenstein (1981)*. Bayesian Inference uses probabilities to estimate the tree *Huelsenbeck et al. (2001)*. Neighbor Joining constructs the tree by finding pairwise distances between sequences *Saitou and Nei (1987)*. In this study we used an open software (RAxML *Stamatakis (2014)*) which uses the Maximum Likelihood method to infer the phylogenetic trees.

1.1.4 Maximum likelihood method

Maximum likelihood is a method used in phylogenetics to estimate the most likely evolutionary history or relationship between a group of organisms based on molecular data, such as DNA or protein sequences. The method assumes a particular model of molecular evolution and calculates the likelihood of the data given a particular tree topology and model parameters. The likelihood of the data given a particular tree and branch lengths is calculated by multiplying the probabilities of observing the nucleotides or amino acids at each site in the sequence data. The probability of observing each nucleotide or amino acid is determined by the model of molecular evolution used. The likelihood of the tree is then calculated as the product of the likelihoods of the data at each site (Equation 1.1).

$$L = Pr(S_1, S_2, ..., S_n | T, \theta) \tag{1.1}$$

Where L is the likelihood function, $S_1, S_1, ..., S_n$ are the nucleotide or protein sequences in the alignment, T is the tree topology, and θ represents the model parameters, such as the nucleotide or protein substitution rates. The likelihood of each possible tree topology is calculated, and the tree with the highest likelihood is considered the most likely tree. *Felsenstein (1981)*

1.1.5 Comparing phylogenetic trees

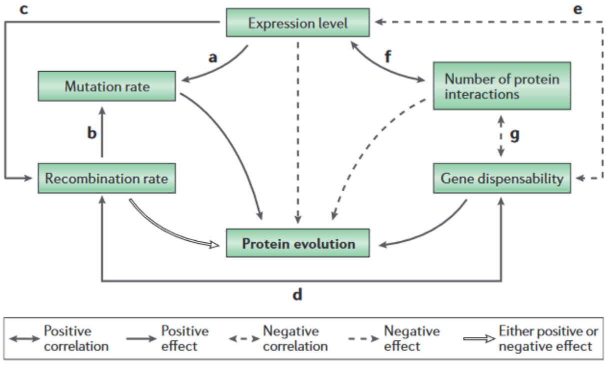
In this study we are interested in comparing the total branch lengths of the phylogenetic trees. This comparison can be challenging as there are certain restrictions that need to be considered to avoid leading to inaccurate results. Firtsly to compare two phylogenetic trees, it is necessary that they are inferred for the same set of organisms *Felsenstein and Felenstein (2004)*. The evolutionary relationships between different organisms are inferred based on the sequence similarity or other molecular markers, and if different organisms are included in the two trees being compared, the comparison would not be valid. And secondly the trees must be inferred with the same evolutionary model in order for the branch lengths to be comparable *Felsenstein and Felenstein (2004).* This ensures that the branch lengths represent the same evolutionary distances and can be directly compared.

1.2 Evolutionary rate of proteins

The evolutionary rate of proteins refers to the rate at which the amino acid sequences of proteins change over time as a result of genetic mutations and natural selection. The rate of evolution varies among different proteins, and can be influenced by factors such as the function of the protein, its expression level, and the selective pressures acting on it.*Pál et al. (2006).* The evolution of proteins can also be influenced by various other factors such as their structure, the location of the genes in the genome, their expression patterns, their position in biological networks, and their ability to tolerate errors during the process of translation. *Echave et al. (2016)*

According to the work of *Echave et al. (2016)* and previous researches Figure 1.1 shows the dependence relationship between some of the factors that affect protein evolution. Figure 1.1 : a) Transcription causes increased spontaneous mutation rates in Saccharomyces cerevisiae *Datta and Jinks-Robertson (1995)* and Escherichia coli *Wright et al. (1999)* , probably by exposing the non-transcribed ss-DNA to mutagenic chemicals, b) Recombinational repair of double-sheeted breaks in S. cerevisiae increases the frequency of nearby point mutations *Rattray and Strathern (2003)*, c) Genes that are close to recombination hotspots in S. cerevisiae are expressed at higher levels during vegetative growth than most other genes*Gerton et al. (2000)*, d) Essential genes are clustered in regions of low recombination in S. cerevisiae and Caenorhabditis elegans *Pál and Hurst (2003)*, e) Proteins that are more dispensable tend to be expressed at lower levels than less dispensable ones *Pál et al. (2003)*, f) More protein–protein interactions have been reported for highly expressed proteins than for low-abundance proteins in S. cerevisiae *Von Mering et al. (2002)*, however, this correlation is not supported by all interaction-detection methods *Von Mering et al. (2002)*, and might reflect a detection bias towards high-abundance proteins, g) It has been reported that essential genes have more protein–protein interactions than non-essential genes *Jeong et al. (2001)*, this correlation might be an artefact of biases in certain interaction data sets *Coulomb et al. (2005)*.

Figure 1.1: Interdependence between the factors that affect protein evolution.*Echave et al. (2016)*



Our knowledge regarding how functional and structural limitations combine to influence differences in evolutionary rates is currently limited. In order to obtain a comprehensive understanding of protein evolution, it is crucial to identify the specific structural and functional characteristics that ultimately determine protein evolutionary rates and develop mechanistic explanatory models. *Echave et al. (2016)* Proteins that are involved in essential cellular processes, such as DNA replication or protein synthesis, tend to evolve at a slower rate, as changes in their amino acid sequence can have a greater impact on protein function. In contrast, proteins that have more variable functions, such as those involved in immune recognition or sensory perception, may evolve more rapidly, as mutations may not have as strong an effect on protein function. *Echave et al. (2016)*

1.2.1 Importance of evolutionary rate of proteins

Understanding the reasons behind the variation in protein evolutionary rates is critical in several fields, such as molecular evolution, comparative genomics, and structural biology. Quantifying the rate of protein evolution is a powerful tool to determine the relative significance of genetic drift and selection, and to identify selective forces using genomic data. Protein evolution analyses also offer a unique approach to investigate speciation (*Webster et al. (2003)*), senescence (*Cutter and Ward (2005)*), and social lifestyle (*Bromham and Leys (2005)*). Additionally protein evolution analysis can identify functionally important sites that can be used in protein design, peptides associated with genetic diseases, drug targets, or protein interaction partners. The rate of protein evolution can be also utilized to predict the impact of various mutations on disease. Recognizing and accounting for confounding factors that influence protein evolution can greatly enhance the accuracy of these predictions. *Pál et al. (2006)*

1.2.2 Calculating evolutionary rate of proteins

The evolutionary rate of proteins can be quantified as the speed of genetic change between a taxonomic group over a certain period of time. The genetic change is measurable and can be calculated as the amino acid substitutions in a given protein alignment. For a group of organisms the amino acid substitutions can be calculated by the total branch length of their phylogenetic tree *Yang (1998).* In this study the period of time for which the evolutionary rate will be calculated can not be determined; therefore to overcome this obstacle a group of mammals was chosen for the structural characteristics analysis and a group of insects for the protein localization analysis. For every protein the phylogenetic gene tree of the chosen organisms is inferred. Under the assumption that the age of the common ancestor (root of the tree) is the same, the evolutionary rate for a specific protein can be calculated as the total branch lengths of each tree.

1.3 Protein structure characteristics

Proteins are large biomolecules composed of one or more chains of amino acids. They play a crucial role in many biological processes, including catalyzing biochemical reactions, replicating DNA, responding to stimuli, and transporting molecules from one location to another. They are composed of a linear sequence of amino acids, which are linked together by peptide bonds to form a polypeptide chain. All the information about the sequence of amino acids exists in the DNA of each organism. The sequence of amino acids determines the structure and function of the protein . In the majority of the organisms there are 20 different types of amino acids that can be arranged in a nearly infinite number of ways to form different proteins.

Protein structure refers to the three-dimensional arrangement of atoms in a protein molecule. The long chains of amino acids can fold into specific shapes, which are critical to their function. The structure of a protein can be described at 4 levels: primary, secondary, tertiary, and quaternary. The primary structure is the linear sequence of amino acids in the polypeptide chain. The secondary structure refers to local spatial arrangements of the polypeptide chain; the most common types are Alpha helices and Beta sheets. The tertiary structure describes the overall three-dimensional folding of the protein molecule, which is largely determined by non-covalent interactions such as hydrogen bonds, ionic bonds, van der Waals forces, and hydrophobic interactions between amino acid side chains. The quaternary structure refers to the association of two or more polypeptide chains into a larger functional unit.*Branden and Tooze (2012)*

Protein structure is essential for its function, as the three-dimensional shape of a protein determines its interactions with other molecules. The active site of an enzyme, for example, is a specific region of the protein with a unique structure that allows it to interact with substrates and catalyze chemical reactions. Similarly, the structure of a receptor protein determines its ability to recognize and bind to specific ligands. Protein structure can be determined experimentally using techniques such as X-ray crystallography, NMR spectroscopy, and electron microscopy. These methods provide detailed information about the three-dimensional arrangement of atoms in a protein. The Protein Data Bank (PDB *Berman et al. (2000)*), is one of the bigger databases that contains the 3D structural data of proteins and is a key resource for the scientific community for studying the structure and function of biological macromolecules and it is also used in this study.

In the recent past there was also a huge research effort on predicting the protein structure from the amino acid sequence. AlphaFold *Jumper et al. (2021)* was developed, which is a deep learning-based protein folding prediction algorithm. It uses a combination of deep neural networks and Monte Carlo sampling methods to predict the three-dimensional structure of a protein from its amino acid sequence. AlphaFold is also used in a part of this study.

1.3.1 Secondary structure characteristics

The secondary structure of a protein refers to the local spatial arrangement of its backbone atoms (mainly the Cα atoms) without regard to the conformations of its side chains.The two most common types of secondary structure are alpha helices and beta sheets. In this study we will focus on those two types of secondary structure. Alpha helices are tightly coiled structures that form a right-handed helix, with the amino acid side chains pointing outward from the helix axis. Beta sheets, on the other hand, are extended structures that are stabilized by hydrogen bonds between adjacent sheets, which can be either parallel or anti parallel.

The spatial orientation of the peptide backbone, defined by a set of dihedral angles (Φ, ϕ) and specific hydrogen bonds, determines the secondary structure of a protein. Regular secondary structures are formed when the backbone dihedral angles repeat specific values. The principal geometry for the alpha helix is $\Phi = 60°$ and $\phi = 45°$ with hydrogen bonds from the NH of the fifth residue in the chain to the C=O group on the first residue, or between residues i and (i + 4). To observe an alpha helix in a protein, one would see a right-handed helical structure when looking down its axis from the amino terminal end. In soluble proteins, the length of an alpha-helix is typically 11 amino acid residues, which corresponds to three turns of the helix. Since all the backbone amide groups participate in intra-chain hydrogen bonds, interactions of helices with other peptide domains or small molecules take place solely through side-chain interactions. The beta sheet is formed by dihedral angles of $\Phi = 130°$ and $\Phi = 120°$, creating an elongated structure with some right-handed twist. The stability of beta sheets in proteins and protein complexes is ensured by hydrogen bonds between protein chains, which can be arranged in either parallel or antiparallel orientations. An other crusial characteristic are turns which refer to a type of structure that connects two sheets of a beta sheet or two alpha helices. These turns are typically four residues long and stabilize the protein structure by reversing the direction of the polypeptide chain. There are several different types of turns, that are classified according to the number of residues involved in the hydrogen-bonded structure. The formation of turns is critical for the stability and function of many proteins. Unordered or random structure is generally defined as a conformation that is not helix, sheet, or turn. *Pelton and McLean (2000).*

According to SCOP classification (*Andreeva et al. (2014),Andreeva et al. (2020)*) alpha helices and beta sheets are divided in 4 groups all-α, all-β, α/β and $\alpha + \beta$. Where in α/β proteins, the secondary structure is composed of alternating Alpha helices and Beta sheets, while in the $\alpha + \beta$ proteins, the secondary structure is mainly composed of Alpha helices with some Beta sheets interspersed.

In the following analysis we will focus on Alpha helices and Beta sheets and the rest of the secondary structure will be considered as unordered and mentioned as coil, more information and details will be found in the section 2.2.4. The intention is to investigate how the quantity of secondary structural elements of a protein, in terms of Alpha helices and Beta sheets, affects its evolutionary rate.

1.3.2 Relative solvent accessible area (RSA)

Relative solvent accessible area (RSA) is the measure of the proportion of the solvent-accessible surface area of an amino acid residue that is actually accessible to the solvent, compared to the solvent-accessible surface area of that amino acid in a fully extended state. The RSA is useful for analyzing protein structures because it provides information about the accessibility of individual residues in a protein, which can be related to their biological function. Residues that are buried in the protein core are likely to be involved in stabilizing the protein structure, while residues on the protein surface are more likely to be involved in interactions with other molecules. RSA values range from 0% to 100%, with 0% indicating a completely buried residue and 100% indicating a completely exposed residue. *Petersen et al. (2009)*

We will use this measure to investigate if there is a difference in the mean RSA of proteins containing Alpha helices and Beta sheets.

1.4 Protein localization

Protein localization refers to the process by which proteins are targeted to specific locations within the cell, where they perform their biological functions. Proteins can be localized to different compartments within the cell, including the nucleus, cytoplasm, plasma membrane, mitochondria, and endoplasmic reticulum. In multicellular organisms, proteins can be found in specific tissues or organs, such as muscle tissue or the liver. The specific localization of a protein is determined by signals present in the protein itself, known as localization signals, which direct the protein to its correct destination *Janmey (1998).* There are several types of localization signals, including nuclear localization signals (NLSs), which direct proteins to the nucleus, and signal sequences, which direct proteins to the endoplasmic reticulum and the mitochondria *Lange et al. (2007)* . Other localization signals include transmembrane domains, which anchor proteins to the plasma membrane, and peroxisomal targeting signals, which target proteins to peroxisomes. Protein localization is a complex and dynamic process that can be regulated by various mechanisms. For example, post-translational modifications, such as phosphorylation and ubiquitination, can regulate protein localization by altering localization signals or by promoting protein degradation *Choudhary and Mann (2010).* Studying protein localization is important for understanding protein function and cellular processes, as well as for developing new therapeutics that target specific proteins in specific locations.

In a part of this study we will focus on proteins that are located in membranes and in solution. We will investigate whether proteins localization, in terms of membrane and soluble proteins, affect the evolutionary rate of proteins.

1.4.1 Membrane proteins

Biological membranes are critical barriers that separate living cells from their surroundings, and in eukaryotes, they also compartmentalize organelles within the cell. Organelle membranes include those that surround the nucleus, mitochondria, endoplasmic reticulum, Golgi apparatus, lysosomes, and secretory vesicles. These membranes have highly specialized functions that vary depending on their location within the organism and within each cell. Membrane proteins are essential for the proper functioning of the membrane because they perform important functions within it. The fundamental structure of biological membranes is established by the arrangement of lipids, which form a phospholipid bilayer.

Many of the functions of the membranes are carried out by proteins. Each biological membrane has its own specific functions, which are enabled by a unique set of proteins embedded within it. The amount and types of proteins present within a membrane vary depending on the specific needs of the cell or organelle. On average, biological membranes are made up of 50% proteins by mass, meaning that there are 50 lipid molecules for every protein molecule. This is due to the fact that lipids are relatively small molecules, while proteins are much larger in comparison. Lipids in the membrane bilayer function mainly as solvents for membrane proteins which tend to be hydro-phobic in nature. Depending on their mode of association with the membrane, membrane proteins can be broadly classified into integral and peripheral membrane proteins. Integral membrane proteins are tightly bound to the membrane and can only be removed by using detergents or other agents that disrupt the membrane. They are typically transmembrane proteins that span the entire lipid bilayer or have one or more hydrophobic domains that anchor them to the membrane. Peripheral membrane proteins, on the other hand, are loosely associated with the membrane and can be removed by changing the ionic strength or pH of the solution. They are typically found on the surface of the membrane or attached to integral membrane proteins. Peripheral membrane proteins often play a regulatory role in signaling pathways or membrane transport processes.*Tan et al. (2008)*

1.4.2 Soluble proteins

Soluble proteins are proteins that are not associated with biological membranes and are present in the aqueous environment of cells or extracellular fluids. These proteins have a wide range of functions, including catalyzing biochemical reactions, serving as structural components, regulating gene expression, and acting as signaling molecules. Soluble proteins can be further classified based on their physical and chemical properties, such as size, shape, charge, hydrophobicity, and solubility. Some examples of soluble proteins include enzymes, antibodies, hormones, and cytokines.

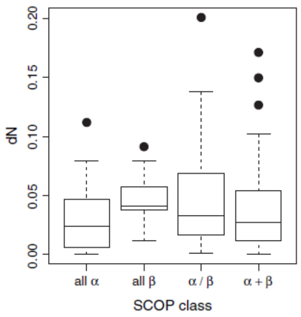
1.5 Results from previous research

Findings of previous research that analyzed the evolutionary rate of proteins based on their secondary structure characteristics, as well as the evolutionary rate of proteins based on their specific localization within cells, will be presented.

1.5.1 Secondary structure characteristics

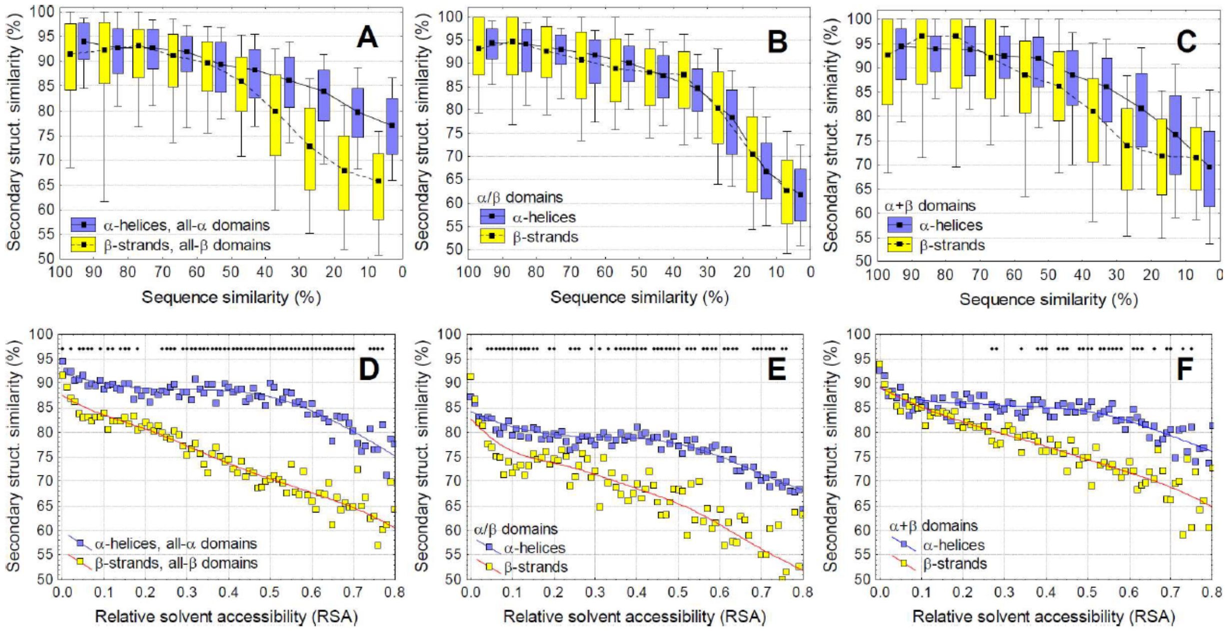
In the work of *Bloom et al. (2006)* structural determinants of the rate of protein evolution in yeast were studied. They quantified the evolutionary rate of proteins as the number of nonsynonymous substitutions per site, dN. All open reading frames (ORFs) in S. cerevisiae were Blasted against those in Saccharomyces bayanus and vice versa. Pairwise hits with an E value of $< 10^{-20}$ were retained and aligned with ClustalW, using the aligned protein sequences to align the nucleotide sequences. Evolutionary rate, the numbers of nonsynonymous substitutions per nonsynonymous site (dN) was computed for these hits using the Phylogenetic Analysis by Maximum Likelihood. They correlated dN with the fraction of helix sites, fraction of sheet sites, fraction of turn sites, and fraction of coil sites, and they found that none of these quantities correlated significantly with dN and neither did they correlate with the expression level. They investgated the relationship between protein structure classification and evolutionary rate for 137 proteins. The result can be seen in Figure 1.2; the median evolutionary rate of all-β proteins was higher than other types, but the highest individual evolutionary rates were found in α/β and $\alpha + \beta$ proteins. However, no class of proteins, including all-β proteins, showed a significantly increased evolutionary rate after correcting for multiple tests. They concluded that secondary structure composition and protein-fold classification had almost no effect on evolutionary rate, although they were skeptical that their analysis might underestimate the contribution of protein structure to evolutionary rate.

Figure 1.2: Box plots of the distributions of evolutionary rate for different SCOP classes.*Bloom et al. (2006)*



In an other study of *Abrusán and Marsh (2016)* they investigated whether alpha helices of beta sheets are more robust to mutation. Mutational robustness is the ability to accept mutations without change. The authors of this study illustrate that there is a variation in the mutation tolerance of different secondary structure elements of proteins, specifically helices and sheets. They explain that this variation is due to the dissimilarity in the count of non-covalent residue interactions within these secondary structure units. The researchers conducted a comprehensive study using SCOP domains and the Protein Data Bank (PDB) and found that alpha helices are more robust than beta sheets, meaning they can handle more mutations in the sequence without compromising their secondary structure. The study suggests that this is mainly due to the greater number of residue interactions in helices. Moreover, both helices and sheets are more robust than regions without any secondary structure (coils). They used a comparative method first, then made all possible pairwise structural alignments between all domains within all SCOP families. Next, they determined the secondary structure of the domains in the alignments, and examined how secondary structure similarity (the percent of aligned helix residues that remained helices in both proteins) changes with sequence similarity, a result that can be seen in Figure 1.3. Alpha helices change significantly less with sequence change than beta sheets in case of all-α, all-β and $\alpha + \beta$ domains. They also indicate that the higher robustness of helices is caused by their higher number of residue-residue interactions. They calculated the relative solvent accesible area (RSA) for each residue and an average RSA for each protein. Alpha helices tend to have bigger RSA than beta sheets. Figure 1.3 also indicates that residues of helices are significantly more robust for mutations than sheets in all SCOP classes, except for the most buried residues with RSA < 0.1.

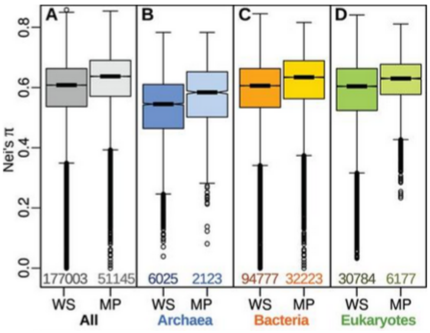
Figure 1.3: Secondary structure similarity of pairwise alignments as a function of sequence similarity and RSA. *Abrusán and Marsh (2016)*



1.5.2 Protein localization

Sojo et al. (2016) investigated whether membrane proteins are dramatically less conserved than water-soluble proteins across the tree of life. The study findings indicate that membrane proteins have fewer orthologs compared to water-soluble proteins throughout the evolutionary history. This could either be due to higher evolutionary rates leading to sequence divergence beyond a detectable threshold by sequence-searching algorithms, or it could be actual gene loss. The study provides evidence for both possibilities. The results show that evolutionary rates for membrane proteins are generally faster than those of water-soluble proteins across all domains of life, including archaea, bacteria, and eukaryotes Figure 1.4. They used Nei's sequence-diversity measure for a measure of evolutionary rate, which was calculated by averaging the number of differences per alignment position per pair of sequences, and then averaging these over the number of pairs, for each group of orthologs. Notably, the evolutionary rates of membrane proteins are faster in the aqueous regions that face outside of the membrane than in the inside-facing regions. They also demonstrated that aqueous sections evolve faster overall than membrane-spanning sections in membrane proteins. Splitting the aqueous sequences into outside- and inside-facing sections confirms that regions exposed to the environment evolve faster than those facing the cytosol.

Figure 1.4: Nei's sequence diversity measure for membrane and soluble proteins in archaea, bacteria and eukaryotes. *Sojo et al. (2016)*



In other previous studies, it was found that the evolutionary rate of a protein is strongly correlated with its subcellular localization. Proteins that are secreted from the cell or are located on the external parts of membranes evolve faster than intracellular proteins in both mammals and yeast *Tourasse and Li (2000),Julenius and Pedersen (2006)Liao et al. (2010)*, although the reasons for this are not fully understood. It is believed that structural and packing constraints, such as the exposure of amino acid residues to the solvent *Oberai et al. (2009),Franzosa et al. (2013)*, as well as the subcellular localization of the proteins and their portions *Julenius and Pedersen (2006),Liao et al. (2010)*, are the strongest predictors of evolutionary rate. In parasites, membrane proteins diverge faster than intracellular water-soluble proteins, likely due to the pressure to avoid detection by the host *Volkman et al. (2002),Plotkin et al. (2004)*. This pattern may be specific to the "red-queen" (the co-evolutionary arms race between parasites and their hosts) dynamics of parasitic interactions, which require constant adaptation just to maintain fitness.

Chapter 2

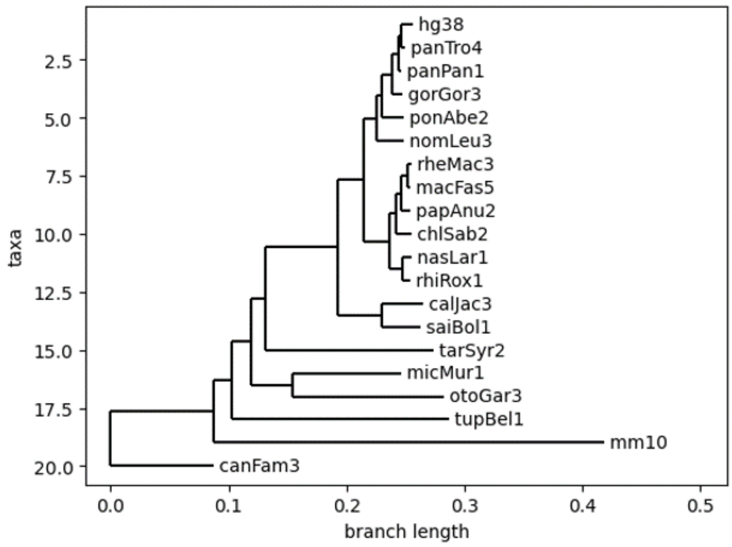
Materials & Methods

2.1 Simulation with INDELible

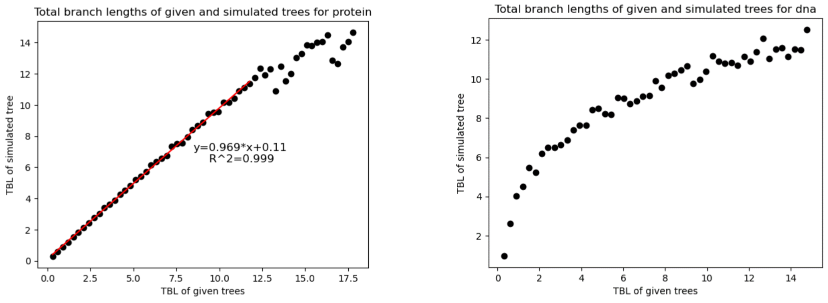
Simulated data are usually used to investigate the accuracy and efficiency of a method. In this case the method is phylogenetics; the phylogenetic tree is obtained through the maximum likelihood approach and true phylogenetic relationships are rarely known with certainty. Therefore simulations help to investigate the accuracy and efficiency of phylogenetic relationships. INDELible which is a tool for biological sequences simulation is used; it can simulate nucleotides, amino acids and codons. INDELible takes as input a given phylogenetics tree with branch lengths assigned and gives as an output the simulated biological sequence. The input tree was taken from The Human genome Browser at UCSC (*Kent et al. (2002)*); it is the following tree containing 20 mammals. (Figure 2.1)

The input to INDELible is a tree and the output is a simulated sequence that can be explained by the given tree; both nucleotide and amino acids sequences were simulated. Then the simulated sequence was used as an input to calculate again the phylogenetic tree with a program for phylogenetic analysis (RAxML subsection 2.2.6), and the final step is the calculation of the total branch length of the simulated tree. The aim here was to investigate how the total branch length

Figure 2.1: Phylogenetic tree of 20 mammals from USCS



of the simulated tree responds to changes in the total branch lengths of the given tree. Thus the UCSC tree was used changing it's branch lengths multiplying by numbers from 0.2 to 18 for amino acid sequences and from 0.2 to 14 for nucleotide sequences with a step of 0.2. The following plots (Figure 2.2) show the relation between total branch lengths of the given and the simulated tree *Fletcher and Yang (2009)*).



(a) Given vs simulated total branch length in proteins

(b) Given vs simulated total branch length in dna

Figure 2.2: Relation of total branch length between given and simulated trees for protein and dna.

According to the plots, in protein simulations the total branch length (tbl) of given versus simulated trees is almost 1 : 1 up to total branch length of 12 in the given tree; this means that both increase with the same rate. Above 12 as the given total branch length increases the simulated one increases with a smaller rate. On the other hand for dna sequences up to total branch length of 6 the simulated tbl increases faster than the given tbl; after that number the simulated one starts to increase slower than the given one, and it seems that it reaches a plateau for big total branch lengths. This is maybe due to number of possible nucleotides which is 4 (A, C, G, T). Having increased total branch lengths means that there is a higher rate of mutation; this can cause the multiple mutation of a site which can end to the same nucleotide as the one before the mutation and this position will not be detectable in the simulated sequence. This explanation may account for the difference in rates between the given and simulated branch lengths. This is different for proteins because the possible amino acids are 20 and it's much more difficult to have multiple mutations that with lead to the same amino acid as the one before the mutation; it is thus more difficult to reach to saturation. Since protein sequences up to a total branch length of 12 are comparable in given vs simulated trees this threshold will be used in further analysis.

2.2 Data set from PDB

One of the data sets analysed was retrieved from the Protein Data Bank (PDB) (*Berman et al. (2000)*) with applied filters for SCOP (Structural Classification of Proteins) classification. SCOP is another database for all proteins with known structure that aims to provide detailed and comprehensive description between those proteins (*Andreeva et al. (2014),Andreeva et al. (2020)*). The aim in this part of the study was to find the effect of proteins that contained mostly Alpha helices and proteins that contained mostly Beta sheet on evolutionary rate. Therefore according to SCOP classification 19,126 Alpha and 29,546 Beta proteins exist and retrieved from the PDB. This data set can be provided in a .txt file containing the PDB id of the proteins due to the size of the file. The PDB files downloading was done with Python (version 3.9.13) and the Biopython package (version 1.81) (*Cock et al. (2009)*).

2.2.1 Organisms used for Phylogenetic analysis

Phylogenetic analysis and particularly building of phylogenetic trees is essential in this study for the calculation of the evolutionary rate or proteins, through the total branch length of each evolutionary tree. As mentioned in subsection 1.1.5, in order to have comparable evolutionary trees they much be built with the exact same organisms, and to have comparable branch lengths the trees much be inferred with the same evolutionary model. For this data set 16 mammal organism were chosen (Table 2.1) covering a wide range of the mammalian species tree.

Table 2.1: Table with the mammals selected for the analysis

scintific name	common name
h.sapiens	human
c.l.familiaris	dog
m.musculus	house mouse
b.taurus	bull
b.b.bison	buffalo
e.a.asinus	donkey
d.leucas	beluga whale
e.caballus	horse
f.catus	cat
l.africana	African bush elephant
u.americanus	American black bear
p.troglofytes	chimpanzee
p.paniscus	Bonobo
p.leo	African lion
r.norvegicus	Norwegian rat
s.scrofa	Eurasian wild pig

2.2.2 g:Profiler

g:Profiler is a collection of tools that are commonly used in standard pipelines of genes or proteins for computational analysis. g:GOSt performs the functional enrichment analysis of individual or multiple gene lists, and g:Orth allows to map orthologous genes across species (*Raudvere et al. (2019),Reimand et al. (2007)*). The data set from PDB consist of protein PDB ids which are used in order to find the gene code for the protein in h.sapiens and other 15 organisms that will be used in order to build the phylogenetic tree. The gene codes for each protein for the 16 organisms are collected and used in the next steps of the analysis. g:GOSt can measure the enrichment of the gene set provided in a given Gene Ontology category i.e. "GO:PB→ Biological Process", "GO:MF→ Molecular Function", "GO:CC→ Cellular Component" and more. It can be used as a statistical measure to evaluate the significance of the functional annotations obtained for a set of genes. The analysis was done with Python (version 3.9.13) and gProfiler module for python.

2.2.3 Ensemble and Biomart tool

Ensemble (*Smedley et al. (2009),Cunningham et al. (2022),Van Rossum and Drake (2009)*) is a flexible infrastructure for access to genomic data and annotation; it can deliver reference data for genome interpretation for any species. It consists of several different tools; one of them is Biomart which was used in this study. Biomart can perform advanced querying of biological data sources through a web interface and an interface with Python (pyBiomart). The gene codes obtained from g:Profiler for each protein and each of the 16 organism are used in Biomart in order to retrieve their amino acid sequences (fasta files). For this procedure, first the organism is specified, then the gene code is provided and finally the required output is selected, which is the amino acid sequence in our case. This procedure is done in Python environment, not in web interface, and it is automated in order to be able to run it for a big number of different queries . The analysis performed using Python (version 3.10.4) and pyBiomart which is a simple pythonic interface for Biomart.

2.2.4 DSSP

The DSSP program (*Kabsch and Sander (1983),Touw et al. (2015)*) was designed to standardize secondary structure assignment, DSSP is a database of secondary structure assignments for all protein entries in the PDB. It is also the program that can calculate the secondary structure from a PDB file. The secondary structure of proteins in this data set is essential for this study, in order to investigate how it impacts the evolutionary rate. This program was used taking as input the PDB files of the proteins and receiving as an output the secondary structure of proteins which contain the 8-states of secondary structure [H,B,E,G,I,T,S,-]. The 8-states Table 2.2 was then converted to 3-state Alpha-helix, Beta-sheet and Coil; for this conversion the following dictionary was used ($H : H, B : E, E : E, G : H, I : C, T : C, S : C, - : C$). The 3-state secondary structure was used for further analysis. The percentage of Alpha helices, Beta sheets and coil was determined for each protein, and the proteins were classified as A proteins if A% was bigger than B% and also classified as B proteins if the B% was bigger than A%.

Table 2.2: Table 8-state secondary structure description of code letters

Code	Description
H	Alpha helix
B	Beta bridge
E	Sheet
G	Helix-3
I	Helix-5
T	Turn
S	Bend

2.2.5 ClustalW

ClustalW is a widely used system for aligning any number of homologous nucleotide or protein sequences. For multi-sequence alignments, ClustalW uses progressive alignment methods. In these, the most similar sequences, i.e. those with the best alignment score are aligned first. Then progressively more distant groups of sequences are aligned until a global alignment is obtained. This heuristic approach is necessary because finding the global optimal solution is prohibitive concerning both memory and time requirements. ClustalW performs very well in practice. The algorithm starts by computing a rough distance matrix between each pair of sequences based on pairwise sequence alignment scores. These scores are computed with the pairwise alignment parameters for DNA and protein sequences. Next, the algorithm uses the neighbor-joining method with midpoint rooting to create a guide tree, which is used to generate a global alignment. The guide tree serves as a rough template for clades that tend to share insertion and deletion features. This generally provides a close-to-optimal result, especially when the data set contains sequences with varied degrees of divergence, so the guide tree is less sensitive to noise (*Thompson et al. (2003).* The alignment of the protein sequences for the 16 organisms and for every protein in the data set is done with ClustalW. The fasta files gathered before are turned into multiple sequence alignments of 16 organisms and are ready for use in the next step. The analysis was done using Python (version 3.9.13) and the Biopython package (version 1.81) (*Cock et al. (2009)*).

2.2.6 RAxML

RAxML is an open source software for sequential and parallel Maximum Likelihood based inference of large phylogenetic trees (*Stamatakis (2014)*).

It takes as an input the multiple sequence alignment, acceptable formats are relaxed interleaved or sequential PHYLIP or FASTA, and it has many options as an input depending on the users preferences. One of the main options is -m which is used to specify the substitution model to be used for phylogenetic inference. The substitution model describes the probabilities of different types of nucleotide or amino acid substitutions that occur over evolutionary time. In this study we used the WAG model, a substitution model for amino acid sequences in phylogenetic analysis. The name ”WAG” stands for Whelan and Goldman (*Whelan and Goldman (2001)*), the authors who developed the model. The WAG model assumes that different amino acids have different propensities for substitution and that these propensities vary according to the evolutionary distance between sequences. The model uses a maximum likelihood approach to estimate the rates of amino acid substitutions from a multiple sequence alignment. The WAG model is a popular choice for analyzing protein-coding genes because it accounts for the compositional heterogeneity and variable evolutionary rates among different amino acid residues.

RAxML produces several different outputs during the course of a phylogenetic analysis. In this study the output used is RAxML_bestTree; it contains the best scoring maximum likelihood tree found by the program. The tree is used to calculate the total branch length which will be used further in the analysis.The analysis was done with Python (version 3.9.13) and the Biopython package (version 1.81), (*Cock et al. (2009)*).

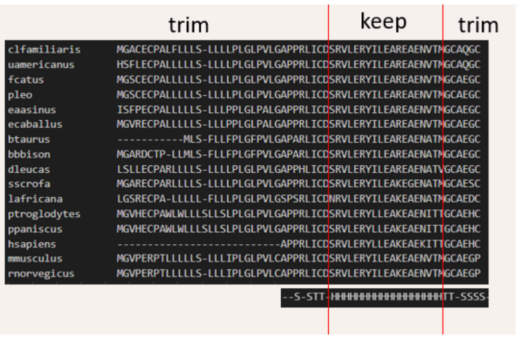
2.2.7 Data set and protein chains

A protein can be formed from one or more identical or different chains. Different proteins can contain one or more identical chains. For each protein in the data set all different chains were used for this study. The data set with SCOP classification from PDB contained 19,126 Alpha and 29,546 Beta proteins. These numbers were reduced to 4,148 Alpha and 5,466 Beta, since only for those chains we found fasta files for the 16 organisms chosen for this study. Also each chain used was classified as Alpha or Beta according to the percentage of Alpha helices and Beta sheets in their secondary structure. These numbers were further changed according to the classification to Alpha and Beta proteins, and the data set then consisted of 4,270 Alpha and 3,867 Beta proteins. Each chain is encoded by a gene; when a chain is found in multiple proteins then the chain was kept only once in the data set; this check was done with the gene name of each chain. This step is important in order to avoid multiple identical chains in the data set that could change the results due to their multiple occurrence; this is not the actual information we wanted to obtain. The final step was to keep only once the chains that appeared more than once; thus the final data set consisted of 662 Alpha and 579 Beta chains with unique genes.

2.2.8 Alignment trimming

In this step of the analysis the aim is to trim the multiple sequence alignments of the 16 organisms obtained for each protein, according to their secondary structure. The desired result is the alignment containing only regions that are Alpha helices for A proteins and Beta sheets for B proteins, in order to use only those regions to investigate if the evolutionary rate of these regions is affected by the secondary structure of the region. We kept for Alpha helices the regions with code letter ”H” and for Beta sheets the regions with code letter ”E”. For each multiple sequence alignment the trimmed regions were again connected in order to obtain one trimmed multiple sequence alignment for each protein. An example of alignment trimming can be seen in Figure 2.3. After the alignment trimming was applied, the trimmed alignments are used as an input to RAxML to infer the trees for each alignment and calculate the rates through the total branch lengths. The data set consisted of 662 Alpha and 579 Beta chains was further reduced to 497 Alpha and 499 Beta chains due to the removal of proteins with rates over than 12. Those proteins were removed for two reasons: firstly because of the simulation results with INDELible and secondly because some proteins had very high evolutionary rates (the trimmed alignment was done based on the h.sapiens sequence and had big gaps with the rest of the organisms) This analysis step was done with Python (version 3.9.13) and the Biopython package (version 1.81), (*Cock et al. (2009)*).

Figure 2.3: Example of alignment trimming



```
clfamiliaris     MGACECPALFLLLS-LLLLPLGLPVLGAPPRLICDSRVLERYILEAREAENVTMGCAQGC
uamericanus      HSFLECPALLLLLS-LLLLPLGLPVLGAPPRLICDSRVLERYILEAREAENVTMGCAQGC
fcatus           MGSCECPALLLLLS-LLLLPLGLPVLGAPPRLICDSRVLERYILEAREAENVTMGCAEGC
pleo             MGSCECPALLLLLS-LLLLPLGLPVLGAPPRLICDSRVLERYILEAREAENVTMGCAEGC
eaasinus         ISFPECPALLLLLS-LLLPPLGLPALGAPPRLICDSRVLERYILEAREAENVTMGCAEGC
ecaballus        MGVRECPALLLLLS-LLLPPLGLPALGAPPRLICDSRVLERYILEAREAENVTMGCAEGC
btaurus          -----------MLS-FLLFPLGFPVLGAPARLICDSRVLERYILEAREAENATMGCAEGC
bbbison          MGARDCTP-LLMLS-FLLFPLGFPVLGAPARLICDSRVLERYILEAREAENATMGCAEGC
dleucas          LSLLECPARLLLLS-LLLLPLGLPVLGAPPHLICDSRVLERYILEAREAENATVGCAEGC
sscrofa          MGARECPARLLLLS-LLLLPLGLPVLGAPPRLICDSRVLERYILEAKEGENATMGCAESC
lafricana        LGSRECPA-LLLLL-FLLLPLGLPVLGSPSRLICDNRVLERYILEAKEAENATMGCAEDC
ptroglodytes     MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRVLERYLLEAKEAENITTGCAEHC
ppaniscus        MGVHECPAWLWLLLSLLSLPLGLPVLGAPPRLICDSRVLERYLLEAKEAENITTGCAEHC
hsapiens         ---------------------------APPRLICDSRVLERYLLEAKEAEKITTGCAEHC
mmusculus        MGVPERPTLLLLLS-LLLIPLGLPVLCAPPRLICDSRVLERYILEAKEAENVTMGCAEGP
rnorvegicus      MGVPERPTLLLLLS-LLLIPLGLPVLCAPPRLICDSRVLERYILEAKEAENVTMGCAEGP
                                           --S-STT-HHHHHHHHHHHHHHHHHHTT-SSSS-
```

2.2.9 Relative solvent accessible area

RSA is a measure to quantify the accessible area of a residue. In the file obtained from DSSP, also used for secondary structure calculation, there is a column called ACC which stands for the accessible surface area of a residue. ACC is the number of water molecules in contact with this residue *10, or residue water exposed surface in Å^2 (*Kabsch and Sander (1983),Touw et al. (2015)*). To calculate the RSA we used the following formula

$$RSA = \frac{ACC}{MaxASA}$$

. MaxASA Maximal Accessible Surface Area (MaxASA) refers to the total area of a given amino acid residue that is exposed to the solvent when the protein is in its native state. It is typically measured in square angstroms (Å^2). MaxASA values for different amino acid residues have been experimentally determined and can be used to calculate RSA values for each residue in a given protein. The MaxASA used to calculate RSA can be found in Table 2.3. After having calculated RSA for each residue of each protein, we calculate a mean RSA for each protein and investigate if the mean RSA is different for A and B proteins. Also we calculate another mean RSA for the residues that are assigned to Alpha helices or Beta sheets, and then investigate if the mean RSA in residues of Alpha helices and Beta sheets is different.

Table 2.3: Maximum solvent accessible surface area (MaxASA) values for common amino acids.

Amino acid	**MaxASA (Å^2)**
Ala (A)	129.0
Arg (R)	274.0
Asn (N)	195.0
Asp (D)	193.0
Cys (C)	167.0
Gln (Q)	225.0
Glu (E)	223.0
Gly (G)	104.0
His (H)	224.0
Ile (I)	197.0
Leu (L)	201.0
Lys (K)	279.0
Met (M)	224.0
Phe (F)	240.0
Pro (P)	159.0
Ser (S)	155.0
Thr (T)	172.0
Trp (W)	285.0
Tyr (Y)	263.0
Val (V)	174.0

2.2.10 Kegg pathways analysis

For the pathway analysis, we used the KEGG (Kyoto Encyclopedia of Genes and Genomes) database *Kanehisa and Goto (2000),Kanehisa (2019),Kanehisa et al. (2023)*. Proteins were searched in the KEGG database to find pathways related to each protein. In order to search proteins in KEGG database the gene name of each protein was collected from PDB since the PDB name of proteins was used in the analysis so far. Proteins can take part in more than one pathway; thus all pathways related with the gene name were collected. We wanted to gather all the pathways that contained A and B proteins, with the intention of investigating if in each pathway there is a difference on the evolutionary rate of A and B proteins. All pathway analysis was done with Python(version 3.9.13) and the Biopython package (version 1.81), (*Cock et al. (2009)*).

2.2.11 Proteins localization through Uniprot

UniProt (Universal Protein Resource) is a comprehensive protein database that provides a central repository of protein sequence and function information. Uniprot contains high-quality protein sequences, curated functional annotations, and other related information such as protein-protein interactions, gene ontology (GO) terms, and cross-references to other biological databases (*uni (2023)*). In this study Uniprot is used to collect protein localization information from the section subcellular location. Protein localization in this part of the study is divided in three teams, membrane proteins which are found only in membranes, soluble proteins which are found anywhere else except membranes and both proteins that can be found both in membranes and solution. The analysis was done with Python (version 3.9.13) (*Cock et al. (2009)*). 662 Alpha and 579 Beta chains were searched in Uniprot for their sub-cellular location. 610 Alpha and 547 Beta chains a subcellular location existed in Uniprot.

2.2.12 Gene Ontology analysis

We used g:GOSt and performed gene set enrichment analysis for the two gene protein sets, for Alpha helices proteins and Beta sheet proteins. For each Gene Ontology category (Biological process, Molecular function and Cellular component) we kept the 11 GO terms with the smallest p-values for each protein category (A and B proteins). Then we used the python library GOatools *Klopfenstein et al. (2018)* and Python (version 3.11.3) to plot the GO terms for each Gene ontology category and protein set (we used different colors) in the gene ontology hierarchy tree and compared A and B proteins according to gene ontology terms.

2.3 Data set from G.Gouridis lab

To investigate the protein localization effect on the evolutionary rate, a data set from G.Gouridis group at IMBB in FORTH was used, and can be provided in an excel file. This data set consisted of 2.789 proteins from organism *Spodoptera frugiperda* and their protein localization which consisted of three possible options: light membrane (LM), heavy membrane (HM) and Soluble (SUP). A similar process as for the structural characteristics was applied in this part. We chose different organisms for the phylogenetic analysis and a different source for the orthologs of the 16 organisms. The remaining procedure for inferring the phylogenetic tree, calculating the rate of evolution and calculating the secondary structure was the same as for the structural characteristics. ClustalW was used for the multiple sequence alignment, RAxML for the inference of the phylogenetic trees, and DSSP for the secondary structure calculation.

2.3.1 Organisms used for Phylogenetic analysis

This time 16 insects were chosen(Table 2.4) covering a wide range of the insects species tree.

2.3.2 OrthoDB

OrthoDB is a comprehensive database of orthologs. OrthoDB provides an interface for exploring the evolutionary relationships between genes from different

Table 2.4: Table with the insects selected for the analysis

scientific name	common name
s.frugipedra	Fall armyworm
d.elegans	flower-feeding fruit fly
d.melanogaster	fruit fly
c.capitata	Mediterranean fruit fly
b.coprophila	darkwinged fungus gnat
a.gifuensis	aphidius wasp
d.similis	white pine sawfly
d.plexippus	monarch butterfly
v.cardui	painted lady
p.glaucus	eastern Tiger Swallowtail
b.mori	silkworm moth
o.brumata	Winter moth
c.carnea	green lacewing
i.elegans	blue-tailed damselfly
p.h.corporis	Body and head lice
f.candida	springtail

species. It contains orthologous groups that are computationally predicted using a variety of methods, including reciprocal best hits, clustering algorithms, and phylogenetic analysis. The database includes data from a wide range of organisms, including animals, plants, fungi, bacteria, and archaea. OrthoDB provides a standardized nomenclature for orthologs and paralogs across all species, which facilitates cross-species comparisons and functional annotation. *Waterhouse et al. (2013),Kriventseva et al. (2019)*

The orthologs and the fasta files in this part of the study were retrieved from OrthoDB, not from the web interface, but using Python (version 3.9.13) and URL extraction in order to run it automatically for big sets of proteins.

2.3.3 AlphaFold

AlphaFold *Jumper et al. (2021)* is an artificial intelligence system developed by the research group at DeepMind that predicts the 3D structure of a protein based on its amino acid sequence. It uses a deep neural network trained on a large database of protein structures to generate predictions with high accuracy. The system is based on a novel machine learning method called attention-based neural networks, which allows it to effectively capture the complex relationships between different parts of the protein sequence and predict its 3D structure with unprecedented accuracy. We used AlphFold to get the predicted 3D structure for the proteins of the G.Gouridis data set. The 3D structure was used to classify proteins according to their secondary structure characteristics with DSSP (*Kabsch and Sander (1983),Touw et al. (2015)*), to Alpha and Beta proteins depending on their percentage of Alpha helices and Beta sheets. We wanted to investigate if the secondary structure characteristics affected evolutionary rate.

2.4 Statistical methods

2.4.1 t-test

T-test is a type of statistical test that is used to compare the means of two groups, and to determine if there is a significant difference between these means. In our case the independent t-test was used, which means that the two groups under comparison are independent from each other. Mathematically, the t-test takes a sample from each of the two sets and establishes the problem statement by assuming a null hypothesis, i.e. that the two means are equal. Based on the applicable formulas, certain values are calculated and compared against the standard values, and the assumed null hypothesis is accepted or rejected accordingly. If the null hypothesis qualifies to be rejected, it indicates that data readings are strong and probably not due to chance. The data values required for the T-test are the mean difference, the difference between the mean values of each data set, the standard deviation of each group, and the number of data values of each group. The outcomes of a t-test is the t statistic.The t-statistic is calculated as the difference between the means of two groups

$$t = \frac{x_1 - x2}{\sqrt{\frac{{s_1}^2}{n1} + \frac{{s_2}^2}{n_2}}}$$

divided by an estimate of the standard error of the difference ($x_{1,2}$ =sample means of each group, $s_{1,2}$ =sample standard deviation of each group, $n_{1,2}$ =sample size of each group). This estimate takes into account the sample sizes and variances of each group. Large t statistic indicates that the difference between the means of the two groups is relatively large compared to the variability within each group, suggesting that the difference is unlikely to have occurred by chance.

The p-value in a t-test is the probability of obtaining a test statistic (such as the t-statistic) as extreme or more extreme than the one observed in the sample, assuming the null hypothesis is true. In other words, it represents the probability of observing a difference between the means of two groups as large or larger than the one observed in the sample, if there is actually no difference between the population means. The p-value is calculated as the area under the t-distribution curve, that is greater than the absolute value of the calculated t-value. The p-value is typically compared to a predetermined significance level, such as 0.05, to determine whether the result is statistically significant. If the p-value ≤ 0.05, the result is considered statistically significant, and the null hypothesis is rejected in favor of the alternative hypothesis that there is a significant difference between the population means. If the p-value > 0.05, the result is not statistically significant, and the null hypothesis is not rejected. (*Kim (2015)*)

The t-test was performed with Python (version 3.9.13), numpy package *Harris et al. (2020)*, pandas package *McKinney et al. (2010)* and scipy package *Virtanen et al. (2020)*

2.4.2 Linear Regression

Linear regression is a statistical method used to model the relationship between two variables by fitting a linear equation to the observed data. In simple linear regression, we are interested in predicting a dependent variable (y) based on a single independent variable (x). The linear equation can be represented as $y = a + b * x$, where a is the intercept of the line and b is the slope of the line. The goal of linear regression is to find the best-fitting line through the data points by the least squares method. This is done by minimizing the sum of the squared residuals, the differences between the predicted values and the actual values. Once the best-fitting line is found, it can be used to make predictions about the dependent variable based on values of the independent variable. The most common way to evaluate the resulting equation from linear regression is the R-squared R^2 value. This metric measures the proportion of variance in the dependent variable that is explained by the independent variables in the model. R^2 values range from 0 to 1, with higher values indicating a better fit. R^2 alone is not always sufficient to evaluate the quality of a regression model, although in this study it will be used to evaluate the linear regression models. *Seber and Lee (2003)*

The t-test was performed with Python (version 3.9.13), numpy package *Harris et al. (2020)*, pandas package *McKinney et al. (2010)*, sklearn package *Pedregosa et al. (2011)* and matplotlib package *Hunter (2007)*.

2.4.3 Compositional data analysis

Compositional data analysis is a statistical framework for the analysis of data that measures the relative proportions of different parts that make up a whole, where the sum of the parts is constant. These types of data are common in many fields, including geology, biology, chemistry, economics, and more. Compositional data analysis methods are designed to deal with the special nature of compositional data, where traditional statistical methods may not be appropriate due to the constraints imposed by the sum of the parts being constant. This constant sum constraint can create spurious correlations and lead to invalid statistical inferences if not accounted for properly. Techniques for compositional data analysis aim to extract meaningful information from compositional data by transforming them into unconstrained data sets that can be analyzed using standard statistical methods. This involves various methods for data preprocessing, transformation, and analysis, such as log ratio and additive log ratio (alr), centered log-ratio transformation, principal component analysis, regression models, and more. *Aitchison (1982)*

In this analysis the percentages of Alpha helices, Beta sheets and Coil are considered as compositional data; this means that they contain relative information and that they are parts of some whole. The total branch length is the dependent variable and the percentages the compositional variables; the aim was to investigate the relationship between the structural characteristics (A%, B%, C%) and the evolutionary rate of proteins (total branch length). We use log ratio and additive log ratio transformations for the analysis *Pawlowsky-Glahn and Egozcue (2006)*. For the log regression the formula used was

$$y = b_1 * ln(x_1) + b_2 * ln(x_2) + b_3 * ln(x_3) + a$$

and the additive log ratio formula is

$$y = b_1 * ln(\frac{x_1}{x_3}) + b_2 * ln(\frac{x_2}{x_3}) + a$$

.

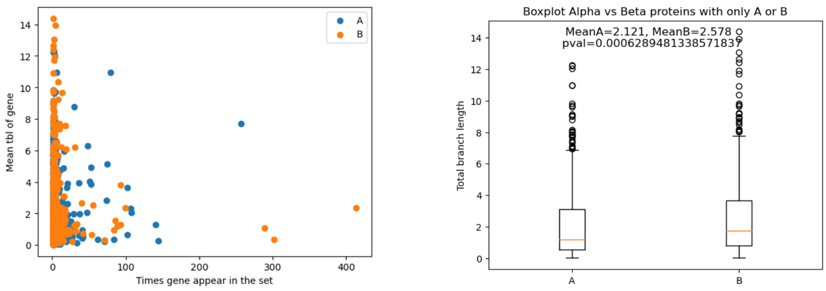
Chapter 3

Results and Discussion

3.1 Results form PDB data set

3.1.1 Analysis including all proteins from the data set

The data set with all Alpha helices (A) and Beta sheets (B) according to SCOP classification, consist of 19,126 A and 29,546 B proteins. After running the process to obtain the trees for each protein, many proteins were excluded from the set because fasta files for some organisms were missing, hence fasta files for all 16 organisms existed for 4,148 A and 5,466 B protein chains. Proteins may consist of one or more identical amino acid chains or of two or more different amino acid chains. When proteins consisted of two or more different chains, every chain was scanned to check if the amount of A-helices or B-sheets was higher in their secondary structure, in order to place the chain in the correct group. Following this step, the set consisted of 4,270 A and 3,867 B proteins. Each chain is encoded by a gene. In order to check if a chain is found in many proteins, the gene name of the chain was checked for each protein. In Figure 3.1(a) the total branch length (tbl) is plotted against the times a gene name appears in the data set. Some genes are over-represented in the data set appearing many times, in some cases over a 100 times. It seems that the genes that are over-represented in Alpha helices tend to evolve faster than those in Beta sheets which tend to evolve slower. This is only an optical observation and it is important to keep the chains with the same gene names only once in the data set to avoid misleading results. After removing multiple chains, the final data set consisted of 662 Alpha and 579 Beta chains with unique genes.



(a) Times of appearance of gene name in the data set and their tbl

(b) t-test for all A and B proteins

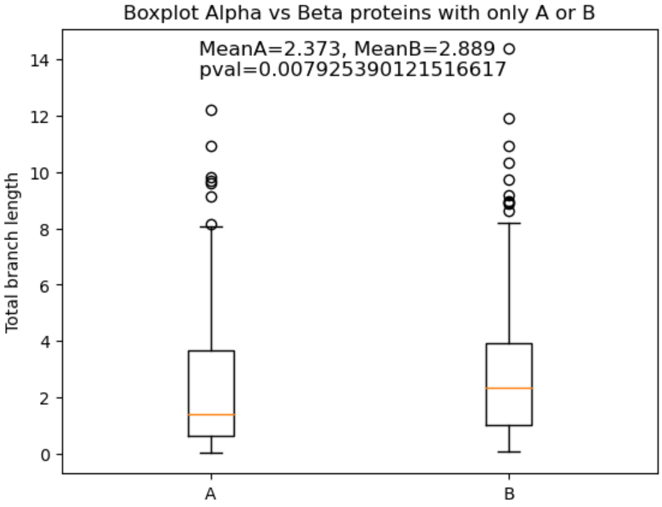
Figure 3.1: Analysis of all A and B proteins from PDB

In these 662 Alpha and 579 Beta chains with unique genes, a t-test was applied to find out if the means of the total branch lengths (evolutionary rate) of the two groups have a significant difference between them. The results of the t-test can be visualised in the following box plot Figure 3.1 (b). The p-value of the t-test is $6.28 * 10^{-4}$,much smaller than 0.05. The mean tbl of A is 2.12, the mean tbl of B is 2.58, which indicates that the mean of total branch length of B proteins is significant different from the mean of A proteins. Hence B proteins tend to have higher evolutionary rates than A proteins for the data set used for this analysis.

3.1.2 Analysis for proteins that contain only A or B structure

Following the previous analysis, proteins with only A or B structure and coil in their chains were chosen from the original data set (662 A, 579 B proteins) and the new set, consisted of 359 A and 253 B proteins. Again a t-test was applied and the results are shown in Figure 3.2: the p-value of the t-test is $7.92 * 10^{-3}$, the mean of A proteins is 2.37, and mean of B is 2.89. The significant difference between the two means imply that and B proteins tend to evolve faster. The means in both groups have slightly higher values than in the previous analysis, and the difference between the two means is also slightly higher for this set of proteins.

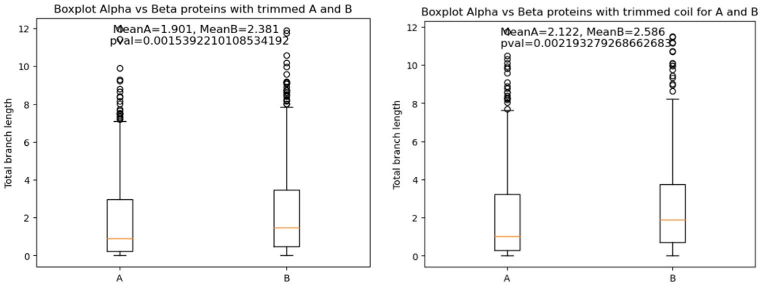
Figure 3.2: t-test for proteins containing only A or only B secondary structure



3.1.3 Analysis for trimmed A and B proteins

The next step was to trim the alignments obtained for each protein according to their secondary structure, whereby the aim was to have the alignment containing only regions that are Alpha helices for A proteins and Beta sheets for B proteins; the evolutionary rate has been checked only in those regions. On the other hand, we also trimmed the alignments and kept for A proteins the part that does not contain Alpha helices and for Beta proteins the part with no Beta sheets. We name that parts "coil". In the trimmed alignments a t-test was applied and the results are in Figure 3.3. For only secondary structure characteristics, the p-value of the t-test is equal to $1.5 * 10^{-3}$, the mean of A is 1.901, and mean of B is 2.381. There is a significant difference between the means of trimmed alignments, and proteins with trimmed alignments containing Beta sheets tend to evolve faster.

For the remaining trimmed part that did not contain Alpha helices for A proteins and Beta sheets for B proteins, the p-value is equal to $2.2 * 10^{-3}$, the mean of A is 2.122, and mean of B is 2.586. We also observe that for both A and B protein coil regions, have higher mean evolutionary rates than the secondary structure regions.



(a) Box plot for trimmed alignments with Alpha helices and Beta sheets

(b) Box plot for trimmed alignments with coils

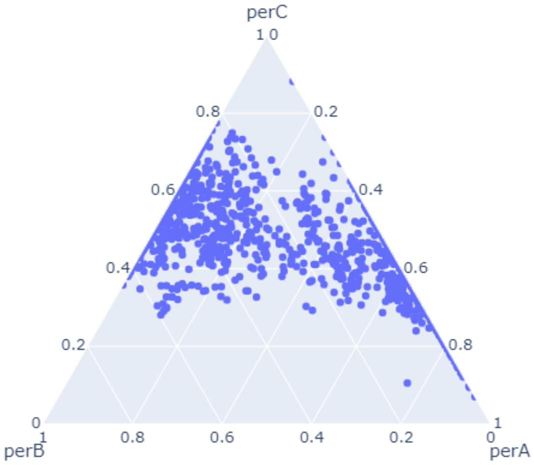
Figure 3.3: Box plot for trimmed alignments

3.1.4 A and B percentage analysis

We then considered the percentage of Alpha helices, Beta-stands and Coils in order to investigate if there is a relation between the evolutionary rate and the percentages of secondary structure. The percentage for each group was calculated for each protein using the secondary structure and the length of the protein sequence. The following image shows a ternary plot with the three groups and the proteins, whereby the contribution of each group is explained for each protein in Figure 3.4.

Linear regression was applied in 2-D using the percentage of Alpha helices from both A and B proteins and the total branch length,in order to investigate the effect of A% on evolutionary rate. The same method was then applied for B% effect. The plots are shown in Figure 3.5, whereby the slope for B% is positive and for A% is negative. This indicates that B% has a positive effect on the evolutionary

Figure 3.4: Ternary plot for A, B and Coil percentages



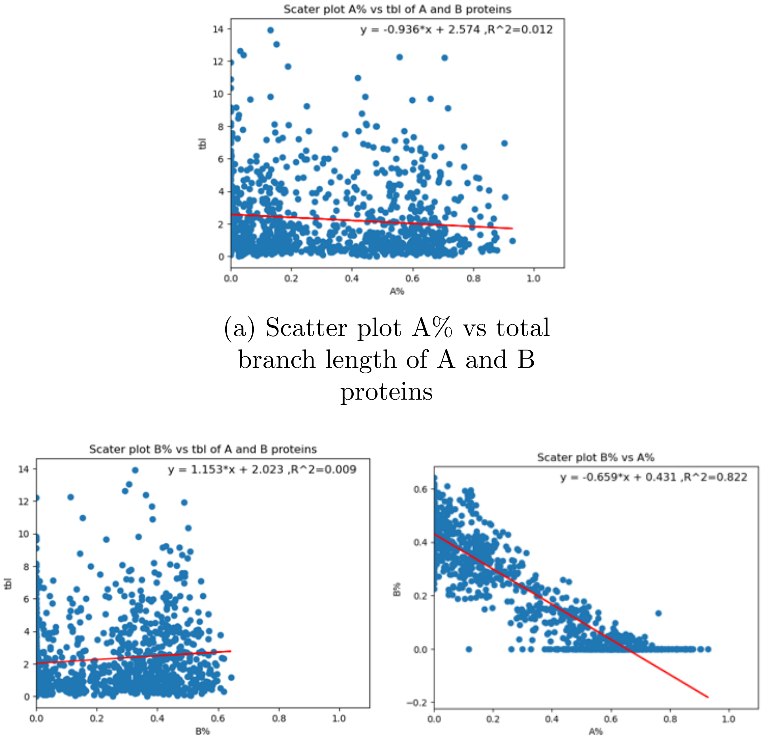
rate and the A% has a negative effect on evolutionary rate. This inspite the fact the coefficient of determination R^2 has a very low value close to zero, which means that the model does not explain any aspect of the variation of the data and that this results of linear regression are unreliable. Plot (c) in Figure 3.5 shows the relationship of A% and B% in the proteins of the data set. Those two values are inversely proportional, which means that while one value increases the other will decrease, which is expected since we are talking about percentages adding to 1.

3.1.5 Compositional data analysis

Linear regression was also applied in 4-D to percentages of A, B, Coil and total branch length. The percentages are considered as compositional data which means that they contain relative information and are parts of the some whole. As mentioned in 2.4.3 we used two transformation methods, the log ratio and additive log ratio (alr) transformation. The total branch length was the dependent variable and the percentages the compositional variables.

For the log regression the equation obtained after linear regression was

$$y = -0.10 * ln(perA) - 0.03 * ln(perB) + 0.31 * ln(perC) + 2.13$$



(a) Scatter plot A% vs total branch length of A and B proteins

(b) Scatter plot B% vs total branch length of A and B proteins

(c) Scatter plot A% vs B%

Figure 3.5: Linear Regression for A% and B% vs total branch length

. The slopes of each percentage show that A and B have a slightly negative effect on total branch length, whereby A has a more negative a effect than B, and Coil has a positive effect on total branch length. The coefficient of determination $R^2 = 0.019$ is very low, almost zero, and therefore the results of linear regression are unreliable because the variation of the data can not be explained from the linear model.

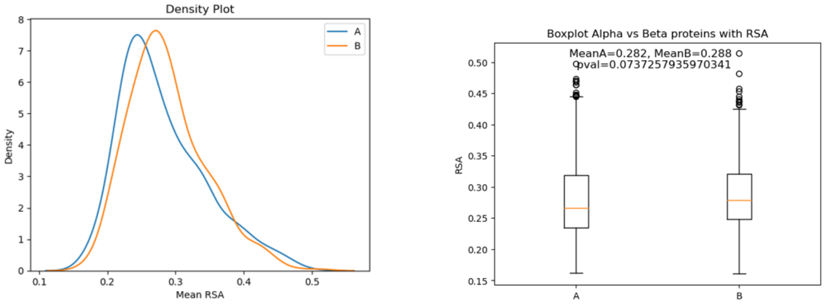
For the additive log ratio (alr) the equation obtained after linear regression was

$$y = -0.11 * ln(\frac{perA}{perC}) - 0.03 * ln(\frac{perB}{perC}) + 1.98$$

. Both slopes are negative indicating a negative effect on total branch length, whereby the slope for B is less negative than the one for A. The coefficient of determination $R^2 = 0.018$ is still very low, close to zero, therefore the results of the linear regression are unreliable because the variation of the data can not be explained from the linear model.

3.1.6 Relative Solvent Accessible area analysis

RSA is calculated for each amino acid in the proteins of the data set. Then for each protein a mean RSA is calculated in order to have one single value assigned to each protein. Proteins with small length < 30 amino acids are excluded from the data set. The data set contains 627 A and 565 B chains. Thus each protein is described by its mean RSA, and the t-test is applied to the values of mean RSA of A and B proteins. The results of the t-test are shown in Figure 3.6. The density plot shows the distribution of mean RSA values for each group of proteins; the two distributions are very close, with B proteins having slightly more proteins with mean RSA values in the region $0.3 - 0.4$. Regarding the t-test, the means of the two groups are almost identical, with the mean for A proteins equal to 0.282 and for B proteins equal to 0.288. The p-value > 0.05 which indicates that the means of the two groups do not have a significant difference.

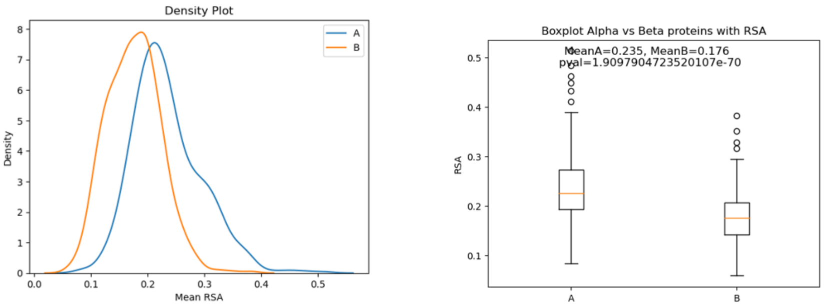


(a) Density plot for mean RSA values of A and B proteins

(b) Box plot for mean RSA of A and B proteins

Figure 3.6: Result of the RSA analysis of A and B proteins

Furthermore, did the same analysis this time containing only residues that are assigned to Alpha helices for A proteins and Beta sheets for B proteins. RSA is calculated for each residue; A mean RSA is then calculated for the residues that belong to the desired secondary structure for each protein. Proteins with small length < 30 amino acids are excluded from the data set, and the data set now contains 647 A and 565 B chains. We have a measure for each protein, this time containing information for RSA for Alpha helices and Beta sheets residues only; and the t-test is again applied for the mean RSA of A and B proteins. The results can be seen in Figure 3.7, the density plot for A proteins is slightly displaced to the right, which means that A proteins may tend to have an increased mean RSA for their amino acids that belong to Alpha helices. The t-test can also confirm this result, whereby the mean for A proteins equals to 0.235 and for B proteins to 0.179; the p-value is equal to $1.9 * 10^{-70}$, thus showing that the means of the two groups are significantly different.

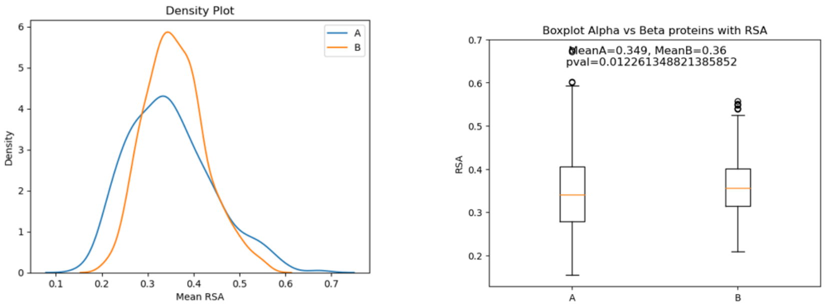


(a) Density plot for mean RSA values of A and B proteins

(b) Box plot for mean RSA of A and B proteins

Figure 3.7: Result of RSA analysis of A and B proteins regarding residues that belong to their secondary structures

We also applied the same analysis including only the amino acid that correspond to the coil region for A and B proteins. The results can be seen in Figure 3.8, the mean RSA of coils of A proteins is 0.34 and of B proteins 0.36, and the p-value is 0.012, which indicates that there is a significant difference between the mean RSA values of A and B proteins. Coil regions of B proteins have slightly larger mean RSA values.



(a) Density plot for mean RSA values of A and B proteins

(b) Box plot for mean RSA of A and B proteins

Figure 3.8: Result of RSA analysis of A and B proteins regarding residues that belong to coil

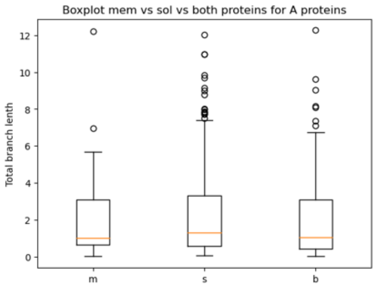
3.1.7 Protein localization analysis

To find the sub-cellular location for each protein we used Uniprot. 662 Alpha and 579 Beta chains were searched in the Uniprot database for their sub-cellular location; the sub-cellular location existed for 610 Alpha and 547 Beta chains. Then those sub-cellular locations where categorized in three possible groups: Membrane proteins for proteins that where found only in membranes, Soluble proteins for proteins that where not found in membranes and Both for proteins that could be found both in solution and membranes. Following this categorization of the data set, the results can be seen in Table 3.1.

Table 3.1: Table with groups and number of proteins for A and B proteins

	A proteins	B proteins
Membrane	69	129
Soluble	391	276
Both	150	142

Firstly we wanted to check whether there is a significant difference in the mean of each localization group in A and B proteins. We applied a pairwise t-test for each possible pair (membrane-soluble, membrane-both and soluble-both) for A and B proteins separately. The box plots,the means and p-values for each pairwise t-test can be found in Figure 3.9 for A proteins and in Figure 3.10 for B proteins. The t-test results show no significant difference in the means of the groups, neither for A nor for B proteins. This indicates that within A proteins and within B proteins the protein localization does not affect the evolutionary rate of the proteins.

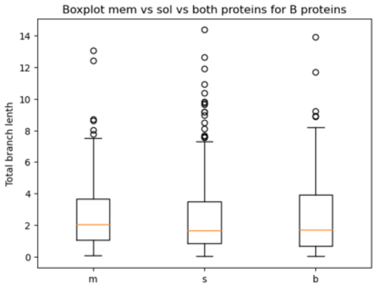


(a) Box plot for Membrane, Soluble and Both groups against tbl

	Mean	p-value
Membrane / (m-s)	1.98	0.43
Soluble / (m-b)	2.19	0.67
Both / (b-s)	2.11	0.71

(b) t-test results for each possible pairwise groups

Figure 3.9: Protein localization analysis for A proteins



(a) Box plot for Membrane, Soluble and Both groups against tbl

	Mean	p-value
Membrane / (m-s)	2.71	0.48
Soluble / (m-b)	2.53	0.65
Both / (b-s)	2.58	0.85

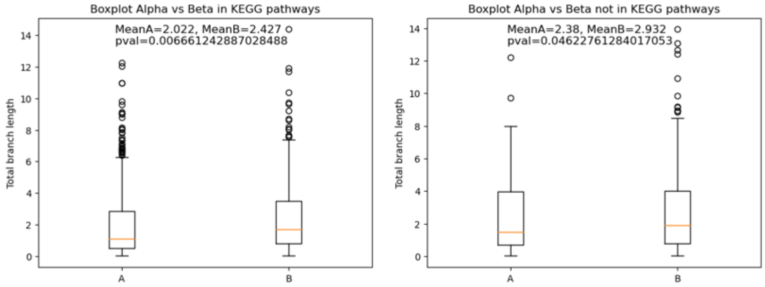
(b) t-test results for each possible pairwise groups

Figure 3.10: Protein localization analysis for B proteins

3.1.8 Pathway analysis

We used the 662 unique A protein chains and the 579 unique B protein chains to determine in which pathways they take part. The KEGG database was used for the determination of pathways. The search found 479 A and 406 B proteins in KEGG pathways, the remaining 183 A and 173 B were not assinged to any

KEGG pathways. Then a t-test has been applied to the proteins that were found in KEGG pathways; the box plot and the results can be found in Figure 3.11. B proteins tend to evolve faster since the p-value is $6.7 * 10^{-3}$ and the means of the two groups have a significant difference; while the mean of A proteins is 2.02 and the mean of B proteins is 2.43.



(a) Box plot and t-test result for proteins that exist in KEGG pathways

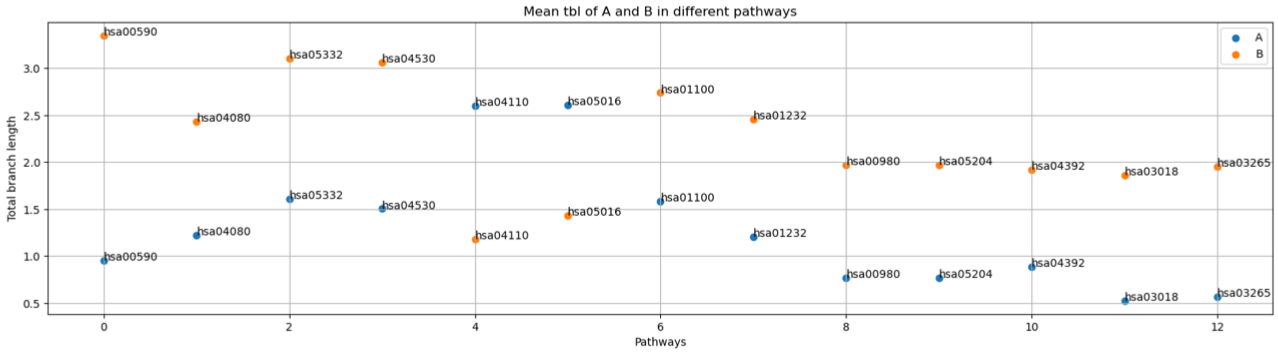
(b) Box plot and t-test results for proteins that do not exist in KEGG pathways

Figure 3.11: Box plots for proteins that exist or not in KEGG pathways

A t-test was also applied to the proteins that did not exist in KEGG pathways (Figure 3.11). Again B proteins tend to evolve faster since the p-value is $4.6 * 10^{-2}$ and the mean of the two groups have a significant difference; while the mean of A proteins is 2.38, and the mean of B proteins is 2.93.

In each pathway that has both A and B proteins, a t-test was again applied to check if there is a significant difference of means of the total branch length between A and B proteins. Since in 297 pathways both A and B proteins exist, the t-test was applied 297 times. We found that in 297 pathways that contained both A and B proteins, only 13 pathways had p-values¡0.05. In those 13 pathways, in 11 pathways B proteins had a bigger mean, and in the remaining 2 pathways A proteins had a bigger mean. The results of the mean total branch lengths for the two groups and the 13 pathways can be found in Figure 3.12. In most of the pathways A and B proteins do not have a significant difference in their evolutionary rate. In pathways with significant difference, B proteins tend to evolve faster in most of the cases.

Figure 3.12: Mean total branch length of A and B proteins in pathways with significant difference between the groups



We then performed a Gene Ontology analysis for the proteins that exist the 13 pathways with significant tbl difference between A and B proteins. We searched with g:GOSt from g:Profiler 4 different protein categories: A proteins that evolve faster than B proteins, A proteins that evolve slower than B proteins, B proteins that evolve faster than A, and B proteins that evolve slower than A. Then for the three gene ontology categories (biological process, molecular function and cellular component) we kept the 10 terms for each 4 protein categories that had the larger p-values when the hierarchical test g:GOSt applied. We then applied another filter and kept in each protein category only the terms that appears only in one specific protein category and removed the terms that appear in more than one protein category. Three Gene Ontology hierarchical trees were plotted for the three Gene Ontology categories and for the terms remaining in each protein category.

The tree for biological processes Figure 3.13 indicates that: A proteins that evolve faster than B proteins are involved in regulation of cell communication and response to oxygen-containing compound; A proteins that evolve slower than B are involved in response to endogenous stimulus and positive regulation of metabolic process; B proteins that evolve faster than A are involved in signal transduction and regulation of response to stimulus; while B proteins that evolve slower than A are involved in regulation of metabolic process.

Figure 3.13: Gene ontology hierarchical tree for Biological process

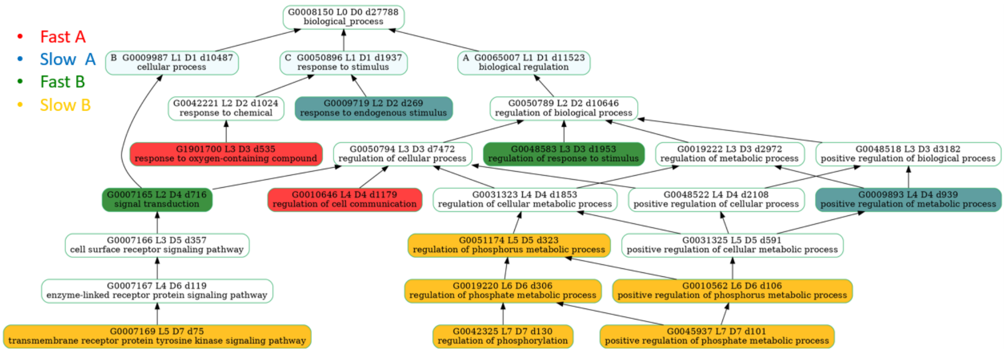
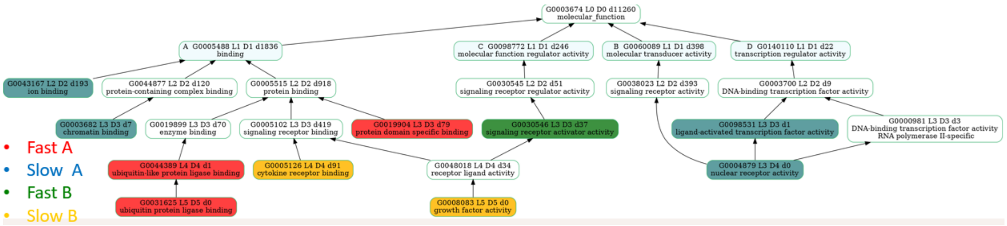
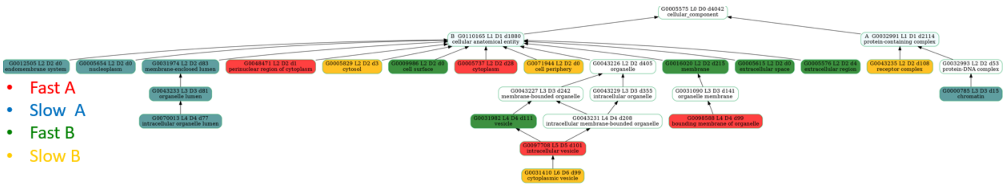

Fast A
Slow A
Fast B
Slow B
G0008150 L0 D0 d27788 biological_process
B G0009987 L1 D1 d10487 cellular process
C G0050896 L1 D1 d1937 response to stimulus
A G0065007 L1 D1 d11523 biological regulation
G0042221 L2 D2 d1024 response to chemical
G0009719 L2 D2 d269 response to endogenous stimulus
G0050789 L2 D2 d10646 regulation of biological process
G1901700 L3 D3 d535 response to oxygen-containing compound
G0050794 L3 D3 d7472 regulation of cellular process
G0048583 L3 D3 d1953 regulation of response to stimulus
G0019222 L3 D3 d2972 regulation of metabolic process
G0048518 L3 D3 d3182 positive regulation of biological process
G0007165 L2 D4 d716 signal transduction
G0010646 L4 D4 d1179 regulation of cell communication
G0031323 L4 D4 d1853 regulation of cellular metabolic process
G0048522 L4 D4 d2108 positive regulation of cellular process
G0009893 L4 D4 d939 positive regulation of metabolic process
G0007166 L3 D5 d357 cell surface receptor signaling pathway
G0051174 L5 D5 d323 regulation of phosphorus metabolic process
G0031325 L5 D5 d591 positive regulation of cellular metabolic process
G0007167 L4 D6 d119 enzyme-linked receptor protein signaling pathway
G0019220 L6 D6 d306 regulation of phosphate metabolic process
G0010562 L6 D6 d106 positive regulation of phosphorus metabolic process
G0007169 L5 D7 d75 transmembrane receptor protein tyrosine kinase signaling pathway
G0042325 L7 D7 d130 regulation of phosphorylation
G0045937 L7 D7 d101 positive regulation of phosphate metabolic process

Figure 3.14: Gene ontology hierarchical tree for Molecular function


G0003674 L0 D0 d11260 molecular_function
A G0005488 L1 D1 d1836 binding
C G0098772 L1 D1 d246 molecular function regulator activity
B G0060089 L1 D1 d398 molecular transducer activity
D G0140110 L1 D1 d22 transcription regulator activity
G0043167 L2 D2 d193 ion binding
G0044877 L2 D2 d120 protein-containing complex binding
G0005515 L2 D2 d918 protein binding
G0030545 L2 D2 d51 signaling receptor regulator activity
G0038023 L2 D2 d393 signaling receptor activity
G0003700 L2 D2 d9 DNA-binding transcription factor activity
G0003682 L3 D3 d7 chromatin binding
G0019899 L3 D3 d70 enzyme binding
G0005102 L3 D3 d419 signaling receptor binding
G0019904 L3 D3 d79 protein domain specific binding
G0030546 L3 D3 d37 signaling receptor activator activity
G0098531 L3 D3 d1 ligand-activated transcription factor activity
G0000981 L3 D3 d3 DNA-binding transcription factor activity RNA polymerase II-specific
Fast A
Slow A
Fast B
Slow B
G0044389 L4 D4 d1 ubiquitin-like protein ligase binding
G0005126 L4 D4 d91 cytokine receptor binding
G0048018 L4 D4 d34 receptor ligand activity
G0004879 L3 D4 d0 nuclear receptor activity
G0031625 L5 D5 d0 ubiquitin protein ligase binding
G0008083 L5 D5 d0 growth factor activity

The tree for molecular function Figure 3.14 indicates that: A proteins that evolve faster than B are involved in protein binding; A proteins that evolve slower than B are involved in ion and chromatin binding and nuclear receptor activity; B proteins that evolve faster than A are involved in signaling receptor activator activity; while B proteins that evolve slower than A are involved in cytokine receptor binding and growth factor activity.

Figure 3.15: Gene ontology hierarchical tree for Cellular component

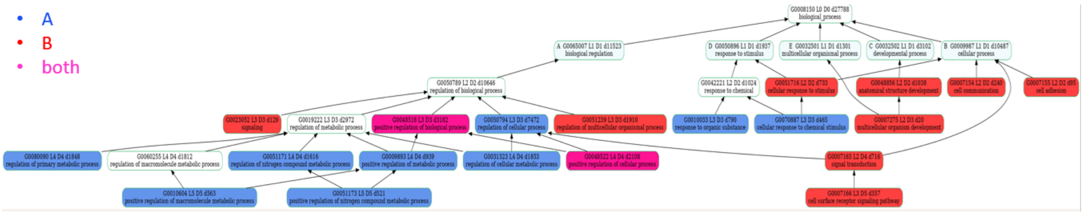

G0005575 L0 D0 d4042 cellular_component
B G0110165 L1 D1 d1880 cellular anatomical entity
A G0032991 L1 D1 d2114 protein-containing complex
G0012505 L2 D2 d0 endomembrane system
G0005654 L2 D2 d0 nucleoplasm
G0031974 L2 D2 d83 membrane-enclosed lumen
G0048471 L2 D2 d1 perinuclear region of cytoplasm
G0005829 L2 D2 d3 cytosol
G0009986 L2 D2 d0 cell surface
G0005737 L2 D2 d28 cytoplasm
G0071944 L2 D2 d0 cell periphery
G0043226 L2 D2 d405 organelle
G0016020 L2 D2 d215 membrane
G0005615 L2 D2 d0 extracellular space
G0005576 L2 D2 d4 extracellular region
G0043235 L2 D2 d108 receptor complex
G0032993 L2 D2 d53 protein-DNA complex
Fast A
Slow A
Fast B
Slow B
G0043233 L3 D3 d81 organelle lumen
G0043227 L3 D3 d242 membrane-bounded organelle
G0043229 L3 D3 d355 intracellular organelle
G0031090 L3 D3 d141 organelle membrane
G0000785 L3 D3 d15 chromatin
G0070013 L4 D4 d77 intracellular organelle lumen
G0031982 L4 D4 d111 vesicle
G0043231 L4 D4 d208 intracellular membrane-bounded organelle
G0098588 L4 D4 d99 bounding membrane of organelle
G0097708 L5 D5 d101 intracellular vesicle
G0031410 L6 D6 d99 cytoplasmic vesicle

The tree for Cellular Component Figure 3.15 indicates that: A proteins that evolve faster than B are found in cytoplasm and its perinuclear region and in bounding membranes of organelle; A proteins that evolve slower than B are found in nucleus and inside organelles; B proteins that evolve faster than A are found in membranes,cell surface and extracellular regions; while B proteins that evolve slower than A are found in cell periphery, cytosol and cytoplasmic vesicles.

3.1.9 Gene Ontology analysis

The purpose of this analysis was to investigate whether A and B proteins tend to take part in different Gene ontology categories (biological process, molecular function and cellular component) that could also explain their difference in their in evolutionary rate we observed from our analysis. The gene ontology hierarchical tree for biological process can be found in Figure 3.16. A proteins tend to get involved in the regulation of biological processes and the response in chemical and organic stimulus, while B proteins take part in signaling, cell communication and adhesion, and anatomical structure development

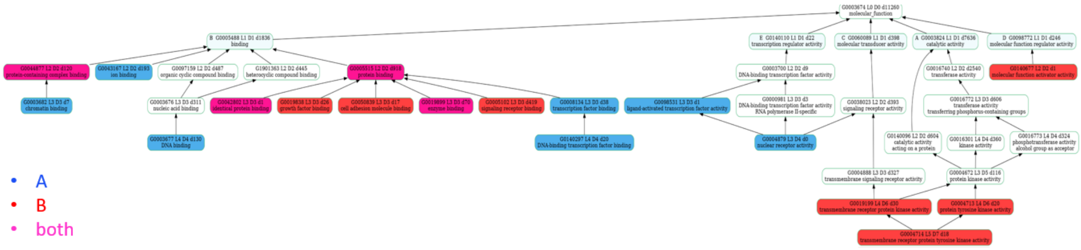
Figure 3.16: Gene ontology hierarchical tree for Biological process


A
B
both
G0008150 L0 D0 d27788 biological_process
A G0065007 L1 D1 d11523 biological regulation
D G0050896 L1 D1 d1937 response to stimulus
E G0032501 L1 D1 d1301 multicellular organismal process
C G0032502 L1 D1 d3102 developmental process
B G0009987 L1 D1 d10487 cellular process
G0050789 L2 D2 d10646 regulation of biological process
G0042221 L2 D2 d1024 response to chemical
G0051716 L2 D2 d733 cellular response to stimulus
G0048856 L2 D2 d1038 anatomical structure development
G0007154 L2 D2 d240 cell communication
G0007155 L2 D2 d95 cell adhesion
G0023052 L3 D3 d129 signaling
G0019222 L3 D3 d2972 regulation of metabolic process
G0048518 L3 D3 d3182 positive regulation of biological process
G0050794 L3 D3 d7472 regulation of cellular process
G0051239 L3 D3 d1910 regulation of multicellular organismal process
G0010033 L3 D3 d790 response to organic substance
G0070887 L3 D3 d465 cellular response to chemical stimulus
G0007275 L2 D3 d20 multicellular organism development
G0080090 L4 D4 d1848 regulation of primary metabolic process
G0060255 L4 D4 d1812 regulation of macromolecule metabolic process
G0051171 L4 D4 d1616 regulation of nitrogen compound metabolic process
G0009893 L4 D4 d939 positive regulation of metabolic process
G0031323 L4 D4 d1853 regulation of cellular metabolic process
G0048522 L4 D4 d2108 positive regulation of cellular process
G0007165 L2 D4 d716 signal transduction
G0010604 L5 D5 d563 positive regulation of macromolecule metabolic process
G0051173 L5 D5 d521 positive regulation of nitrogen compound metabolic process
G0007166 L3 D5 d357 cell surface receptor signaling pathway

The Gene ontology hierarchical tree for molecular function (Figure 3.17) indicates that both A and B proteins take part in binding, A proteins take also part in DNA-binding transcription factor activity while B proteins take part in transmembrane receptor protein kinase activity and proteins kinase activity.

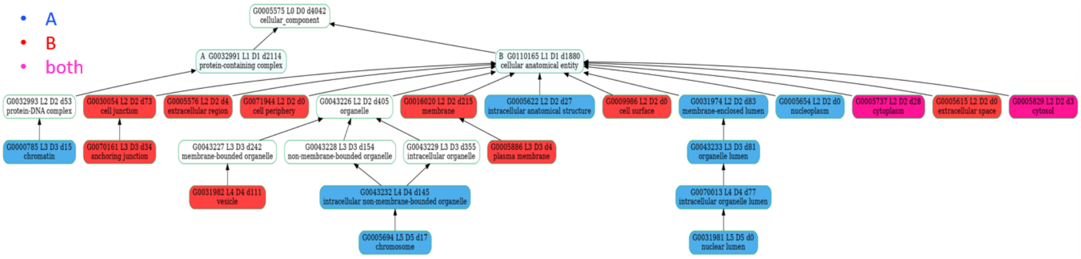
Concerning the Gene ontology hierarchical tree for cellular component (Figure 3.18), we can observe that both A and B proteins can be found in cytoplasm and cytosol. A proteins can be found in chromatin, intra-cellular anatomical structure, membrane enclosed lumen and nucleoplasm. B proteins can be found in membranes, cell surface, extracellular regions, cell junction and cell periphery. This result shows that A proteins tend to be inside the cell and close to the nucleus, while B proteins are found more in cell surface and extracellular region. The existence of B proteins in the extracellular region and cell surface could explain

Figure 3.17: Gene ontology hierarchical tree for Molecular function


G0003674 L0 D0 d11260 molecular_function
B G0005488 L1 D1 d1836 binding
E G0140110 L1 D1 d22 transcription regulator activity
C G0060089 L1 D1 d398 molecular transducer activity
A G0003824 L1 D1 d7636 catalytic activity
D G0098772 L1 D1 d246 molecular function regulator activity
G0044877 L2 D2 d120 protein-containing complex binding
G0043167 L2 D2 d193 ion binding
G0097159 L2 D2 d487 organic cyclic compound binding
G1901363 L2 D2 d445 heterocyclic compound binding
G0005515 L2 D2 d918 protein binding
G0003700 L2 D2 d9 DNA-binding transcription factor activity
G0016740 L2 D2 d2540 transferase activity
G0140677 L2 D2 d1 molecular function activator activity
G0003682 L3 D3 d7 chromatin binding
G0003676 L3 D3 d311 nucleic acid binding
G0042802 L3 D3 d1 identical protein binding
G0019838 L3 D3 d26 growth factor binding
G0050839 L3 D3 d17 cell adhesion molecule binding
G0019899 L3 D3 d70 enzyme binding
G0005102 L3 D3 d419 signaling receptor binding
G0008134 L3 D3 d38 transcription factor binding
G0098531 L3 D3 d1 ligand-activated transcription factor activity
G0000981 L3 D3 d3 DNA-binding transcription factor activity RNA polymerase II-specific
G0038023 L2 D2 d393 signaling receptor activity
G0016772 L3 D3 d606 transferase activity transferring phosphorus-containing groups
G0003677 L4 D4 d130 DNA binding
G0140297 L4 D4 d20 DNA-binding transcription factor binding
G0004879 L3 D4 d0 nuclear receptor activity
G0140096 L2 D2 d604 catalytic activity acting on a protein
G0016301 L4 D4 d360 kinase activity
G0016773 L4 D4 d324 phosphotransferase activity alcohol group as acceptor
G0004888 L3 D3 d327 transmembrane signaling receptor activity
G0004672 L3 D5 d116 protein kinase activity
G0019199 L4 D6 d30 transmembrane receptor protein kinase activity
G0004713 L4 D6 d20 protein tyrosine kinase activity
G0004714 L5 D7 d18 transmembrane receptor protein tyrosine kinase activity
A
B
both

their tendency to evolve faster than A proteins, as reported in previous researches mentioned in section 1.4, indicating that proteins outside the cell tend to evolve faster. This statement can have many possible explanations: proteins may have greater selective pressure due to the interaction with the environment, and also proteins inside the cell are more crucial for the function of the cell and may be more conserved.

Figure 3.18: Gene ontology hierarchical tree for Cellular component

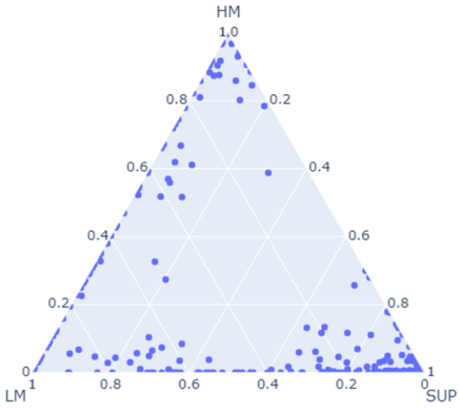

A
B
both
G0005575 L0 D0 d4042 cellular_component
A G0032991 L1 D1 d2114 protein-containing complex
B G0110165 L1 D1 d1880 cellular anatomical entity
G0032993 L2 D2 d53 protein-DNA complex
G0030054 L2 D2 d73 cell junction
G0005576 L2 D2 d4 extracellular region
G0071944 L2 D2 d0 cell periphery
G0043226 L2 D2 d405 organelle
G0016020 L2 D2 d215 membrane
G0005622 L2 D2 d27 intracellular anatomical structure
G0009986 L2 D2 d0 cell surface
G0031974 L2 D2 d83 membrane-enclosed lumen
G0005654 L2 D2 d0 nucleoplasm
G0005737 L2 D2 d28 cytoplasm
G0005615 L2 D2 d0 extracellular space
G0005829 L2 D2 d3 cytosol
G0000785 L3 D3 d15 chromatin
G0070161 L3 D3 d34 anchoring junction
G0043227 L3 D3 d242 membrane-bounded organelle
G0043228 L3 D3 d154 non-membrane-bounded organelle
G0043229 L3 D3 d355 intracellular organelle
G0005886 L3 D3 d4 plasma membrane
G0043233 L3 D3 d81 organelle lumen
G0031982 L4 D4 d111 vesicle
G0043232 L4 D4 d145 intracellular non-membrane-bounded organelle
G0070013 L4 D4 d77 intracellular organelle lumen
G0005694 L5 D5 d17 chromosome
G0031981 L5 D5 d0 nuclear lumen

3.2 Data set from G.Gouridis lab

This data set consisted of 2,789 proteins from the organism *Spodoptera frugiperda* and their protein localization; consisted of three possible options: light membrane

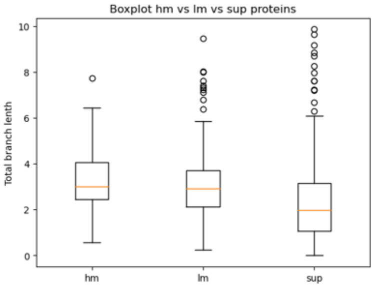
(LM), heavy membrane (HM) and Soluble (SUP). Those proteins were reduced to 651 proteins that exist in OrthoDB and have fasta files for the 16 organism selected. Then the 651 proteins were assigned to the three groups, light membrane (LM), heavy membrane (HM) and soluble (SUP), whereby a protein was assigned to a group if it was found in that location more than 50% of the times. Thus the numbers of proteins in each group was 60 HM, 118 LM and 466 SUP. The proteins and the percentage of their occurrence in the three possible locations can be seen in Figure 3.19 ternary plot.

Figure 3.19: Ternary plot of the percentage of protein occurrence in the three possible locations



3.2.1 Analysis of evolutionary rate and protein localization

First a pairwise t-test was applied on the total branch lengths of the proteins, on all possible pairs in the 3 locations, to check if the protein location affects the evolutionary rate. The results of the t-test and the box plots for each location can be found in Figure 3.20. The means of the two membrane locations LM and HM do not have a significant difference, but for both t-tests for LM-SUP and HM-SUP the p-values < 0.05, which means that membrane proteins tend to evolve faster than soluble proteins in this data set.



(a) Box plot for LM, HM and SUP groups against tbl

	Mean	p-value
LM / (LM-HM)	2.71	0.63
HM / (LM-SUP)	2.53	$1.14 * 10^{-5}$
SUP / (HM-SUP)	2.58	$7.39 * 10^{-6}$

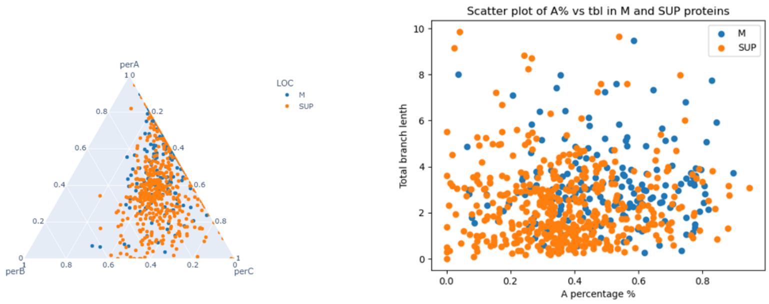
(b) t-test results for each possible pairwise groups

Figure 3.20: t-test results for all possible location pairs

3.2.2 Analysis for evolutionary rate and structural characteristics

We then proceeded to investigate if the percentage of Alpha helices, Beta sheets and coils is different in the different protein locations. We used AlphaFold to get the structure of the proteins and DSSP to calculate the secondary structure. Proteins with a percentage of Alpha helices bigger than the Beta sheet percentage ($A\% > B\%$) were assigned as A proteins and the opposite ($B\% > A\%$) as B proteins. After this classification the 648 protein were split to 528 A proteins and 120 B proteins. The two membrane locations LM and HM are merged to one location as membrane proteins (M). We established visual representations plotted to check if a difference can be visually spotted (Figure 3.21). The two groups, membrane and soluble proteins, can not be visually distinguished as shown by the plots, and a statistical method should be applied for further investigation.

Thereafter t-test is applied on A% for membrane and soluble proteins and on B% for membrane and soluble proteins. The results are shown on Figure 3.22 and they indicate that the mean percentage of Alpha helices is significantly different in membrane and soluble proteins, which means that membrane proteins tend to have a larger percentage of Alpha helices. Regarding beta-sheet percentage, the mean of the beta-sheet percentage is significantly different in membrane and

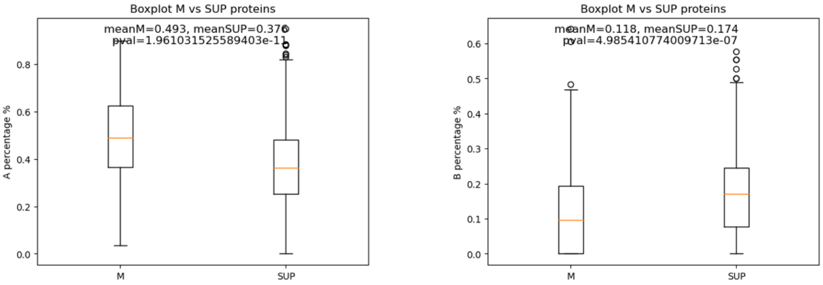


(a) Ternary plot for A%, B% and coil%

(b) 2D plot of A% vs total branch length

Figure 3.21: Visual representation of structural characteristics

soluble proteins; thus soluble proteins tend to have a larger percentage of Beta sheet in their structure in this data set.



(a) Box plot for A% in membrane and soluble protein

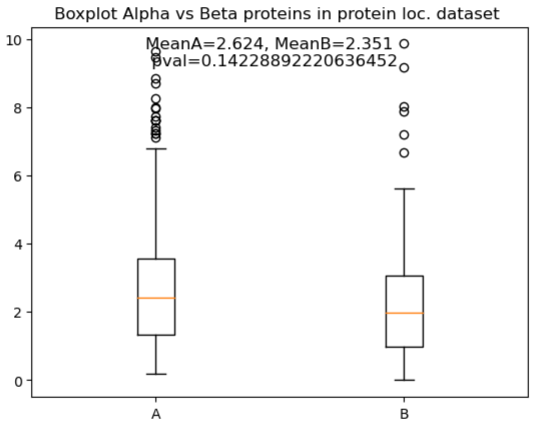
(b) Box plot for B% in membrane and soluble proteins

Figure 3.22: t-test results for A% and B% in membrane and soluble proteins

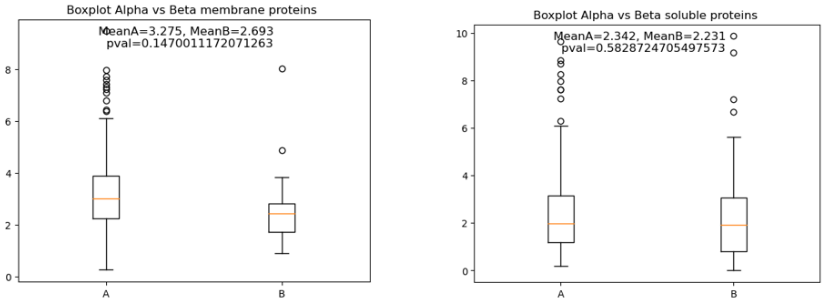
In order to investigate whether proteins with Alpha helices or Beta sheets tend to evolve faster, a t-test was applied in the total branch lengths of A and B proteins and the result (Figure 3.23) indicates that the means of the two groups do not have a significant difference: the secondary structure characteristics do not affect evolutionary rate in this data set.

Finally, we investigated whether there is a difference in the evolutionary rate of A and B proteins within the protein location groups (Membrane, Soluble). The Membrane protein group consists of 178 proteins, 159 A and 19 B proteins, and the Soluble group consists of 463 protein, 363 A and 100 B proteins. A t-test was again applied on the evolutionary rate of proteins within each group for A and B

Figure 3.23: Box plot of total branch length in A and B proteins



proteins. The results (Figure 3.24) indicate that there is no significant difference in A and B proteins neither in Membrane proteins nor in Soluble ones. This result leads again to the conclusion that secondary structure characteristics do not affect the evolutionary rate of this data set.



(a) Box plot for total branch length in A and B Membrane protein

(b) Box plot for total branch length in A and B Soluble protein

Figure 3.24: t-test results for total branch length in A and B, membrane and soluble proteins

Chapter 4

Conclusions

This study examined the effect of structural characteristics and protein location on evolutionary rate of proteins. For the data set from the PDB containing all Alpha helices and Beta sheets according to the SCOP classification, a pipeline has been developed to evaluate the evolutionary rate of the proteins in the data set. Due to constraints in the pipeline, the number of proteins was reduced and the analysis has been conducted with the proteins that fulfill the requirements of the method we used. We observed that Beta sheet proteins tend to evolve faster than Alpha helix proteins in all analyses, including whole proteins, proteins containing only Alpha helices or Beta sheets in their structure, and trimmed proteins containing only the desired secondary structure (H for helices or E for sheets) as well as trimmed proteins containing only coil. In the PDB data set, we calculated the mean Relative Solvent Accessible Area (RSA). The results of the statistical test showed that there was no significant difference in the mean RSA values of the A and B proteins; however, concerning only parts of the proteins that are helices or sheets, the mean RSA value was larger in helices, and for parts that are neither helices nor sheets, i.e. coil parts, the mean RSA values are slightly higher in proteins with bigger percentage of Beta sheets. Concerning the pathway analysis performed in both proteins that existed and did not exist in Kegg pathways, Beta sheet proteins tend to evolve faster. We found that in most pathways no significant difference has been observed in A and B proteins, whereby only in 13 pathways there was a difference in the evolutionary rates; in the 11 of them the B proteins were faster while in the remaining 2 A evolved faster. Additionally, protein localization (Membrane and soluble) within A and B proteins did not affect the evolutionary rate for the PDB data set. Gene ontology analysis provided some interesting results, concerning biological processes: A proteins seem to take part in the regulation of biological processes while B proteins participate in cell communication and adhesion as well as developmental process. According to molecular function term results both A and B take part in binding processes and B proteins also take part in trans-membrane signaling receptor activity. As for Cellular component term it seems that A proteins are found mostly inside the cell and near the nucleus while B proteins are found in extracellular regions, cell periphery and membranes; this could possibly yield an explanation for their faster evolutionary rates.

For the second part of the study examining the G.Gouridis data set, we found that membrane proteins tend to evolve faster than soluble proteins, and that membrane proteins tend to have higher percentages of A helices and soluble proteins higher percentages of Beta sheets. However, secondary structure analysis has shown that A and B proteins of this data set did not have a significant difference in their evolutionary rates.

Chapter 5

Bibliography

Bibliography

(2023). Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531.

Abrusán, G. and Marsh, J. A. (2016). Alpha helices are more robust to mutations than beta strands. *PLoS computational biology*, 12(12):e1005242.

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):139–160.

Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., and Murzin, A. G. (2014). Scop2 prototype: a new approach to protein structure mining. *Nucleic acids research*, 42(D1):D310–D314.

Andreeva, A., Kulesha, E., Gough, J., and Murzin, A. G. (2020). The scop database in 2020: expanded classification of representative family and superfamily domains of known protein structures. *Nucleic acids research*, 48(D1):D376–D382.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The protein data bank. *Nucleic acids research*, 28(1):235–242.

Bloom, J. D., Drummond, D. A., Arnold, F. H., and Wilke, C. O. (2006). Structural determinants of the rate of protein evolution in yeast. *Molecular biology and evolution*, 23(9):1751–1761.

Branden, C. I. and Tooze, J. (2012). *Introduction to protein structure.* Garland Science.

Bromham, L. and Leys, R. (2005). Sociality and the rate of molecular evolution. *Molecular biology and evolution*, 22(6):1393–1402.

Choudhary, C. and Mann, M. (2010). Decoding signalling networks by mass spectrometry-based proteomics. *Nature reviews Molecular cell biology*, 11(6):427–439.

Cock, P. J., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics*, 25(11):1422–1423.

Coulomb, S., Bauer, M., Bernard, D., and Marsolier-Kergoat, M.-C. (2005). Gene essentiality and the topology of protein interaction networks. *Proceedings of the Royal Society B: Biological Sciences*, 272(1573):1721–1725.

Cunningham, F., Allen, J. E., Allen, J., Alvarez-Jarreta, J., Amode, M. R., Armean, I. M., Austine-Orimoloye, O., Azov, A. G., Barnes, I., Bennett, R., et al. (2022). Ensembl 2022. *Nucleic acids research*, 50(D1):D988–D995.

Cutter, A. D. and Ward, S. (2005). Sexual and temporal dynamics of molecular evolution in c. elegans development. *Molecular Biology and Evolution*, 22(1):178–188.

Datta, A. and Jinks-Robertson, S. (1995). Association of increased spontaneous mutation rates with high levels of transcription in yeast. *Science*, 268(5217):1616–1619.

Echave, J., Spielman, S. J., and Wilke, C. O. (2016). Causes of evolutionary rate variation among protein sites. *Nature Reviews Genetics*, 17(2):109–121.

Felsenstein, J. (1981). Evolutionary trees from dna sequences: a maximum likelihood approach. *Journal of molecular evolution*, 17:368–376.

Felsenstein, J. and Felenstein, J. (2004). *Inferring phylogenies*, volume 2. Sinauer associates Sunderland, MA.

Fletcher, W. and Yang, Z. (2009). Indelible: a flexible simulator of biological sequence evolution. *Molecular biology and evolution*, 26(8):1879–1888.

Franzosa, E. A., Xue, R., and Xia, Y. (2013). Quantitative residue-level structure–evolution relationships in the yeast membrane proteome. *Genome Biology and Evolution*, 5(4):734–744.

Gerton, J. L., DeRisi, J., Shroff, R., Lichten, M., Brown, P. O., and Petes, T. D. (2000). Global mapping of meiotic recombination hotspots and coldspots in the yeast saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences*, 97(21):11383–11390.

Hall, B. G. (2013). Building phylogenetic trees from molecular data with mega. *Molecular biology and evolution*, 30(5):1229–1235.

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., et al. (2020). Array programming with numpy. *Nature*, 585(7825):357–362.

Huelsenbeck, J. P., Ronquist, F., Nielsen, R., and Bollback, J. P. (2001). Bayesian inference of phylogeny and its impact on evolutionary biology. *science*, 294(5550):2310–2314.

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in science & engineering*, 9(03):90–95.

Janmey, P. A. (1998). The cytoskeleton and cell signaling: component localization and mechanical coupling. *Physiological reviews*, 78(3):763–781.

Jeong, H., Mason, S. P., Barabási, A.-L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833):41–42.

Julenius, K. and Pedersen, A. G. (2006). Protein evolution is faster outside the cell. *Molecular biology and evolution*, 23(11):2039–2048.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589.

Kabsch, W. and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers: Original Research on Biomolecules*, 22(12):2577–2637.

Kanehisa, M. (2019). Toward understanding the origin and evolution of cellular organisms. *Protein Science*, 28(11):1947–1951.

Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M., and Ishiguro-Watanabe, M. (2023). Kegg for taxonomy-based analysis of pathways and genomes. *Nucleic Acids Research*, 51(D1):D587–D592.

Kanehisa, M. and Goto, S. (2000). Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30.

Kent, W. J., Sugnet, C. W., Furey, T. S., Roskin, K. M., Pringle, T. H., Zahler, A. M., and Haussler, D. (2002). The human genome browser at ucsc. *Genome research*, 12(6):996–1006.

Kim, T. K. (2015). T test as a parametric statistic. *Korean journal of anesthesiology*, 68(6):540.

Klopfenstein, D., Zhang, L., Pedersen, B. S., Ramírez, F., Warwick Vesztrocy, A., Naldi, A., Mungall, C. J., Yunes, J. M., Botvinnik, O., Weigel, M., et al.

(2018). Goatools: A python library for gene ontology analyses. *Scientific reports*, 8(1):1–17.

Koonin, E. V. (2005). Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet.*, 39:309–338.

Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simão, F. A., and Zdobnov, E. M. (2019). Orthodb v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic acids research*, 47(D1):D807–D811.

Lange, A., Mills, R. E., Lange, C. J., Stewart, M., Devine, S. E., and Corbett, A. H. (2007). Classical nuclear localization signals: definition, function, and interaction with importin α. *Journal of Biological Chemistry*, 282(8):5101–5105.

Liao, B.-Y., Weng, M.-P., and Zhang, J. (2010). Impact of extracellularity on the evolutionary rate of mammalian proteins. *Genome Biology and Evolution*, 2:39–43.

Magadum, S., Banerjee, U., Murugan, P., Gangapur, D., and Ravikesavan, R. (2013). Gene duplication as a major force in evolution. *Journal of genetics*, 92(1):155–161.

McKinney, W. et al. (2010). Data structures for statistical computing in python. In *Proceedings of the 9th Python in Science Conference*, volume 445, pages 51–56. Austin, TX.

Mount, D. W. (2008). Maximum parsimony method for phylogenetic prediction. *Cold Spring Harbor Protocols*, 2008(4):pdb–top32.

Nei, M., Kumar, S., et al. (2000). *Molecular evolution and phylogenetics*. Oxford University Press, USA.

Oberai, A., Joh, N. H., Pettit, F. K., and Bowie, J. U. (2009). Structural imperatives impose diverse evolutionary constraints on helical membrane proteins. *Proceedings of the National Academy of Sciences*, 106(42):17747–17750.

Pál, C. and Hurst, L. D. (2003). Evidence for co-evolution of gene order and recombination rate. *Nature genetics*, 33(3):392–395.

Pál, C., Papp, B., and Hurst, L. D. (2003). Rate of evolution and gene dispensability. *Nature*, 421(6922):496–497.

Pál, C., Papp, B., and Lercher, M. J. (2006). An integrated view of protein evolution. *Nature reviews genetics*, 7(5):337–348.

Pawlowsky-Glahn, V. and Egozcue, J. J. (2006). Compositional data and their analysis: an introduction. *Geological Society, London, Special Publications*, 264(1):1–10.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al. (2011). Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830.

Pelton, J. T. and McLean, L. R. (2000). Spectroscopic methods for analysis of protein secondary structure. *Analytical biochemistry*, 277(2):167–176.

Petersen, B., Petersen, T. N., Andersen, P., Nielsen, M., and Lundegaard, C. (2009). A generic method for assignment of reliability scores applied to solvent accessibility predictions. *BMC structural biology*, 9:1–10.

Plotkin, J. B., Dushoff, J., and Fraser, H. B. (2004). Detecting selection using a single genome sequence of m. tuberculosis and p. falciparum. *Nature*, 428(6986):942–945.

Rattray, A. J. and Strathern, J. N. (2003). Error-prone dna polymerases: when making a mistake is the only way to get ahead. *Annual review of genetics*, 37(1):31–66.

Raudvere, U., Kolberg, L., Kuzmin, I., Arak, T., Adler, P., Peterson, H., and Vilo, J. (2019). g: Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic acids research*, 47(W1):W191–W198.

Reimand, J., Kull, M., Peterson, H., Hansen, J., and Vilo, J. (2007). g: Profiler—a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic acids research*, 35(suppl_2):W193–W200.

Robinson, D. F. and Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical biosciences*, 53(1-2):131–147.

Saitou, N. and Nei, M. (1987). The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, 4(4):406–425.

Seber, G. A. and Lee, A. J. (2003). *Linear regression analysis*, volume 330. John Wiley & Sons.

Smedley, D., Haider, S., Ballester, B., Holland, R., London, D., Thorisson, G., and Kasprzyk, A. (2009). Biomart–biological queries made easy. *BMC genomics*, 10(1):1–12.

Sojo, V., Dessimoz, C., Pomiankowski, A., and Lane, N. (2016). Membrane proteins are dramatically less conserved than water-soluble proteins across the tree of life. *Molecular biology and evolution*, 33(11):2874–2884.

Stamatakis, A. (2014). Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.

Tan, S., Tan, H. T., and Chung, M. C. (2008). Membrane proteins and membrane proteomics. *Proteomics*, 8(19):3924–3932.

Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2003). Multiple sequence alignment with clustalw and clustalx. *Current protocols in bioinformatics*, (1):2–3.

Tourasse, N. J. and Li, W.-H. (2000). Selective constraints, amino acid composition, and the rate of protein evolution. *Molecular biology and evolution*, 17(4):656–664.

Touw, W. G., Baakman, C., Black, J., Te Beek, T. A., Krieger, E., Joosten, R. P., and Vriend, G. (2015). A series of pdb-related databanks for everyday needs. *Nucleic acids research*, 43(D1):D364–D368.

Van Rossum, G. and Drake, F. L. (2009). *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272.

Volkman, S. K., Hartl, D. L., Wirth, D. F., Nielsen, K. M., Choi, M., Batalov, S., Zhou, Y., Plouffe, D., Le Roch, K. G., Abagyan, R., et al. (2002). Excess polymorphisms in genes for membrane proteins in plasmodium falciparum. *Science*, 298(5591):216–218.

Von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, 417(6887):399–403.

Waterhouse, R. M., Tegenfeldt, F., Li, J., Zdobnov, E. M., and Kriventseva, E. V. (2013). Orthodb: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic acids research*, 41(D1):D358–D365.

Webster, A. J., Payne, R. J., and Pagel, M. (2003). Molecular phylogenies link rates of evolution and speciation. *Science*, 301(5632):478–478.

Whelan, S. and Goldman, N. (2001). A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Molecular biology and evolution*, 18(5):691–699.

Wiley, E. O. and Lieberman, B. S. (2011). *Phylogenetics: theory and practice of phylogenetic systematics*. John Wiley & Sons.

Wright, B. E., Longacre, A., and Reimers, J. M. (1999). Hypermutation in derepressed operons of escherichia coli k12. *Proceedings of the National Academy of Sciences*, 96(9):5089–5094.

Yang, Z. (1998). On the best evolutionary rate for phylogenetic analysis. *Systematic Biology*, 47(1):125–133.

Zhaxybayeva, O. and Doolittle, W. F. (2011). Lateral gene transfer. *Current Biology*, 21(7):R242–R246.