Johannes Kepler University in Linz
Faculty of Engineering and Natural Sciences
and
University of South Bohemia in České Budějovice
Faculty of Science

Amino Acid Distributions at Oligomeric Membrane Protein Interfaces

Bachelor Thesis

Author : Anna Drechslerová

Supervisor: Associate Professor Dr. Andreas Horner

Guarantor : Ing. Ph.D. Rudolf Vohnout

České Budějovice ,2021

Drechslerová, A., 2021: Amino Acid Distributions at Oligomeric Membrane Protein Interfaces ,Bc. Thesis, in English -34p., Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic and Faculty of Engineering and Natural Sciences, Johannes Kepler University, Linz, Austria.

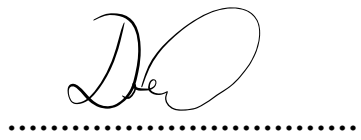
Annotation

A set of twelve exemplary oligomeric transmembrane proteins was selected and subjected to an analysis in order to discover more information about their amino acid distribution. These findings were compared to an aquaporin dataset and evaluated with regard to protein and lipid interfaces.

Declaration

I declare that I am the author of this qualification thesis and that in writing it I have used the sources and literature displayed in the list of used sources only.

In Linz, 12.04.2021



Abstract

Protein-protein interfaces determine the oligomerization of transmembrane proteins by providing an interface of complementary hydrophobic surfaces, which is stabilised by lipid-protein interactions and salt bridges as well as hydrogen bonds between single protomers. However, also the protein surfaces facing the lipid bilayer core are purely hydrophobic. This hydrophobic belt determines the location in the lipid bilayer. However, it is not established to which extend the respective protein-protein or protein-lipid interfaces differ in their overall amino acid distributions, thereby contributing to protein oligomerization besides specific amino acid interactions. Therefore, this study aims to analyse the amino acid distribution of twelve exemplary oligomeric membrane proteins to identify differences between protein interfaces facing the lipid bilayer and protomer-protomer interfaces. Furthermore, we analyse the corresponding internal amino acid distributions next to the respective interfaces. Additionally, this data set is compared to another dataset including a collection of seventeen aquaporins originating from the research group of Associate Professor Dr. Andreas Horner. Thereby, we want to evaluate, whether amino acid distribution trends found in aquaporins are specific or whether they prove to be more universal and correlate with the findings in our dataset.

Our data indicates a preference for leucine in the protein and lipid surfaces and for glycine and alanine in the internal regions. This correlates with the polarity of these sectors, as the surface favours large hydrophobic amino acids whereas internal parts prefer small hydrophobic amino acids. Overall the dataset composed in this work exhibits rather similar trends and propensities as compared to the aquaporin dataset.

These findings could be used in the field of protein engineering to tune the oligomeric state of membrane proteins, thereby optimising or manipulating their functionality.

Acknowledgements

First and foremost, I would like to thank my supervisor Associate Professor Dr. Andreas Horner for giving me the opportunity to work on this project with him and providing guidance and advice whenever it was needed.

My wholehearted gratitude goes to Dipl. Ing. Natasha Trajkovska, who allowed me to use her tools and findings as well as granted me invaluable insights to the problematics surrounding the topic of this paper. I have no doubt that without your help and great ideas many aspects of this paper wouldn't have come to life.

I would like to thank the Institute for Machine Learning in Linz and Institute of Applied Informatics in České Budějovice for making my studies possible by creating the cross-border Bioinformatics program, without which I most probably wouldn't get to the place where I currently find myself.

Last but definitely not least I want to thank BSc. Laura-Nadine Kroll. I'm infinitely grateful to you for standing by me every step of the way no matter how troublesome the road might have been. I'm not exaggerating when I say I couldn't have done this without you and your support.

1. Introduction

The cell membrane is a complex and fascinating structure, which still up to this day possesses many challenges for researchers. Its two main components are well known and shared across the kingdoms of life - proteins and lipids (Cooper, 2000). These compounds are crucial for the proper work of cells and membranes.

Lipids form the phospholipid bilayer (Cooper, 2000). The structure is built in a way, where the heads of lipids create the outer surface and the interior is made of lipid tails (Cooper, 2000). Heads are hydrophilic, whereas tails are hydrophobic, which makes this bilayer very beneficial for the cell, as it is secure and due to its intrinsic properties self assembles in the correct orientation (Cooper, 2000).

The second very important component of a cell membrane are proteins. Various kinds of proteins exist, however one is especially important for our work - integral membrane proteins (Cooper, 2000). Members of this group are inserted within the phospholipid bilayer, which most of these proteins cross via the formation of α-helices (Lodish et al., 2000a). These proteins can be useful for a cell in many ways, as they define the selectivity and permeability of the cell membrane for certain solutes. A major part of these membrane proteins is not monomeric but they exhibit an oligomeric state of two or more protomers (Forrester, 2015). The process which describes that proteins create bigger aggregates is called protein oligomerization (Forrester, 2015).We are particularly interested in a sub-population of these proteins which form oligomeric complexes within the membrane, where the functional unit resides within each of the single protomers. 12 of such membrane proteins form our exemplary dataset.

Questions which are still unresolved in this respect are : Why do some of the membrane proteins undergo oligomerization even though their functional units reside within each of the respective protomers ? What are the driving forces of this protein oligomerization process?

The answers to these questions are not fully established and therefore we can only hypothesise about them having a potential correlation with e.g. hydrophobic interactions at complementary interfaces, specific interactions like hydrogen bonding or salt bridges, hydrophobic mismatch of the protein and the membrane etc.

To advance the understanding of membrane protein oligomerization we aim to analyse, compare and interpret the principle differences of amino acid distributions between protein-protein and protein-lipid interfaces. Our desire to better understand the oligomerization process is to potentially use the acquired knowledge in the field of protein engineering, specifically when it comes to optimising the functionality and the improvement of the oligomeric state of the respective proteins.

There has been already various research done in connection to amino acid distributions within the transmembrane proteins, however information which we are aiming for is completely lacking.

Ulmschneider and Sansom studied amino acid distributions in integral membrane protein structures (Ulmschneider & Sansom, 2001). They focused mainly on the differences between α-helical and β-barrel proteins and assessed different sequence and structure-based methods to do so (Ulmschneider & Sansom, 2001). This study describes a preference of α-helices for certain hydrophobic residues such as alanine, isoleucine, leucine and valine (Ulmschneider & Sansom, 2001). They also mention that "leucine appears more than twice as often in α-helical proteins than in β-barrels" (Ulmschneider & Sansom, 2001). In this work we aim to acquire more in-depth analysis of α-helical transmembrane proteins and their amino acid distributions to understand these specific proteins better.

Duarte and Biyani performed an analysis of oligomerization interfaces in transmembrane proteins (Duarte, Biyani, Baskaran & Capitani, 2013). The work examines and compares soluble protein interfaces and transmembrane protein interfaces, evaluating both α-helices and β-barrels. Various methods and criteria to evaluate the datasets of interest were applied, one of them being the amino acid distribution. Their findings suggest that amino acid frequencies display rather comparable trends for their sets of proteins, however these results do not seem to hold for certain amino acids (Durate et al., 2013). In the respective case alanine and glycine in α-helical structures displays high values of distribution and the same applies to leucine in β-barrels (Durate et al., 2013). The amino acid composition was also examined based on different groups, which these acids belong to (e.g. based on charge, size, polarity etc.) (Durate et al., 2013).

Our approach differs from the above mentioned as it uses structurally dissected interfaces. We analysed and compared amino acid distributions across these interfaces, combining the data from all exemplary proteins to gain distinct knowledge about the specific regions.

In this paper we aim to analyse the amino acid distribution of exemplary oligomeric membrane proteins and consequently identify differences between protein interfaces facing the lipid bilayer and protomer-protomer interfaces. In addition we aim to understand if the amino acid differences at the interfaces determine the overall properties of the amino acid distributions facing inside the proteins.

1.1. Membrane transport proteins

Transport of particles across the cell membrane is a process highly dependent on the properties of the given molecule, which aims to pass in or out of the cell. The main characteristic, which plays a role in this action is polarity. When a substance is non-polar it can diffuse through the lipid bilayer and enter the cell (Alberts, 2002).

However, this is in general not the case for polar molecules. These particles are usually transported inside the cell by a group of proteins fittingly named transport proteins (Alberts, 2002). These compounds vastly vary in many aspects, however probably the most significant difference to point out is their specialisation. Different transport proteins are able to provide a passage through the membrane only to certain groups of compounds (Alberts, 2002). This leads to a division of the respective proteins into many classes and subcategories, dependent on the kinds of atoms and molecules, which pass through them. Nevertheless, despite all the differences there are also certain similarities that membrane transport proteins share. To illustrate, all of these substances are transmembrane polypeptides and as such they pass, usually on more than one occasion, through the membrane creating looping structures (Alberts, 2002). To provide a more concrete case, we can mention our samples, which are all oligomeric transmembrane proteins, meaning they posses several polypeptide chains(subunits) (Garratt, Valadares, Bachega, 2013).

There are many different groups into which transport proteins can be divided into, however, we would like to depict two of them, channels and carriers (Alberts, 2002).

These classes, together with ATP-powered pumps (Lodish et al., 2000b), serve as main subcategories of membrane transport proteins and carry a major importance for our work, as all our samples are representatives of either channels or carriers.

Before any further elaboration, we would like to note that all samples in our dataset were extracted from the Protein Data Bank (Berman et al., 2000) and therefore we choose to refer to them by their Protein Data Bank identification code (PDB ID).

1.1.1.Channel proteins

These structures provide a path shielding the passing compounds, which are in many cases ions or other polar particles, from the hydrophobic membrane (Berg, Tymoczko & Stryer, 2002a). Usage of channels is very beneficial for the cells, as these transport proteins do not require any energy to function. The transport of molecules is solely driven by the electrochemical gradient of the polar compounds passing through the channel (Berg et al., 2002a). Due to said properties functionality of these transport proteins is often referred to as passive transport or facilitated diffusion (Berg, Tymoczko & Stryer, 2002b).

1.1.1.1.Channel proteins present in our dataset

1.1.1.1.1. Chloride channel - ClC

Chloride channels are, as the name suggests, responsible for transport of chloride ions, and as such they are considered a member of a larger group of proteins called ion channels (Jentsch, 2002). The first protein of this group was cloned in 1990 (Jentsch, Steinmeyer & Schwarz, 2002) and since then different ClCs have been discovered in all established taxonomical kingdoms (Jentsch, 2002). Our dataset possesses two samples, which belong to this group and despite their source organisms being different, one (1KPK) originates from *Escherichia coli* and the second protein (1KPL) is derived from *Salmonella typhimurium*, their domain of origin - Bacteria - is the same.

1.1.1.1.2. Ammonia channel - AmtB

Similarly to ClC channels, there are many proteins, which specialise on the transport of ammonia molecules over the membrane (Gruswitz, O'Connell & Stroud, 2007). There are also differences, when it comes to the approach of organisms to ammonium in the first place. Many different members of e.g. Bacteria use it as a source of nitrogen, whereas in case of animals this compound is considered poisonous (Mirandela, Tamburrino, Hoskisson, Zachariae & Javelle, 2019). The organism of origin of our protein 1U7C is *Escherichia coli,* belonging to the domain of Bacteria, and it represents AmtB.

1.1.1.1.3. Formate channel - FocA

These proteins belong to a larger group of formate-nitrite anion channels (Lü et al., 2013). The structure of FocA seems to display similarities to aquaporin channels, however, despite this aspect, FocA channels are specialised solely on formate transportation and molecules of water are not able to pass through them (Wang et al., 2009). Our dataset contains one protein (3KCU), which was derived from *Escherichia coli* and is associated with this channel.

1.1.1.1.4. TRIC channel - TRIC-B1

Trimeric intracellular cation (TRIC) channels are responsible for transport of calcium (Ca2+) (Yang et al., 2016). As this particle is important in various processes of a cell, these proteins are vital for the proper functionality of many organisms. Protein 5EGI(sourcing from *Caenorhabditis elegans*) in our dataset, is an example of TRIC-B1, which can be found in endoplasmic reticulum of various cells (Zhou et al., 2014).

1.1.1.1.5. SatP

Succinate-Acetate Permease proteins are a group of channels responsible for transport of acetate (Qiu et al., 2018). Acetate is crucial for many processes within the cells of organisms across various kingdoms, for example in bacteria it is considered to be a

crucial source of carbon (Gao et al., 2016). Our dataset contains two different proteins, which represent this channel : 5YS3, originating from *Citrobacter koseri,* and 5ZUG, which was extracted from *Escherichia coli,* both of which belong to the domain of Bacteria.

1.1.1.1.6. Channelrhodopsin 2 (ChR2)

Channelrhodopsins are a group of transport proteins originating from *Chlamydomonas reinhardtii,* which is a single cell green algae (Nagel et al., 2003).They differ from the rest of transport proteins listed up till this point, as this ion channel is responsive to light particles (Nagel et al., 2003). The sample of our dataset, representing ChR2 is 6EID, which is simultaneously our only sample originating from the kingdom of *Viridiplantae*.

1.1.1.1.7. Urea channel

This protein is responsible for transportation of urea, making it an acid activated channel (Weeks & Sachs, 2001). Such transport proteins can be found for example in pathogens such as *Helicobacter pylori*. This bacterium, which is also the source organism of a 6NSK protein in our dataset, utilises urea and UreI channel in order to protect itself from the acidic environment of the gastric system, where it can be often found (Marshall, Barrett, Prakash, McCallum & Guerrant, 1990). This channel displays a similarity to aquaporins as it is able to transport water molecules as well (McNulty, Ulmschneider, Luecke & Ulmschneider, 2013).

1.1.1.2. Aquaporins (AQP)

Aquaporins are a group of channels transporting water molecules and in certain cases glycerol (Verkman, 2012). They can be found for example in many animals, plants and bacteria, mostly in tissues connected to the transportation of fluids (Verkman, 2013). The UniProtKB (UniProt, Consortium, 2008) currently comprises over 65 000 proteins, which are a part

of the Aquaporin Protein Family. However, it appears that rather small portion of them are curated and even less have resolved structures.

Our dataset does not contain aquaporins specifically, however, some of our proteins share certain similarities with these water channels. Due to this fact we decided to use aquaporin dataset originating from the research group of Associate Professor Dr. Andreas Horner. This dataset consists of more than 2300 sequences, with 17 different aquaporins, which have resolved structures and had undergone the same procedure for annotation, blast and alignment as our dataset. These aquaporins originate from six different kingdoms of life : Animals (AQP0, AQP1, AQP2, AQP4, AQP5, AQP7, AQP10), Plants (PIP24, PIP21, TIP21), Fungi (AQY1), Protista (AQGP), Bacteria (AQPZ_Ecoli, AQPZ2_AGRF, GLPF) and Archaea (AQPM_METTM, AQPM_ARCFU).

Comparing our dataset to these aquaporins is interesting also because we want to see, whether certain findings are aquaporin specific or universal for oligomeric transmembrane proteins.

1.1.2. Carrier proteins

Carriers are in many aspects similar to channels and also provide a suitable passage through the cell membrane for various particles. The crucial difference is that transporters, as carrier proteins are often called, are only able to transfer a small amount of a given compound over the membrane (in many cases this amount is limited to one molecule) (Lodish et al., 2000b). Once the transport capacity of a carrier is full, the structure carries its "passenger" to the other side of the membrane by performing a conformational change (Lodish et al., 2000b). This distinctive process is more time and energy consuming compared to the fairly quick and efficient channel transport, however, many carrier proteins posses a significant advantage as they are able to pass the molecules in but also against the direction of the electrochemical gradient (Lodish et al., 2000b).

1.1.2.1. Carrier proteins present in our dataset

1.1.2.1.1. Carnitine transporter - CaiT

This carrier belongs to anti-porters, which are transport proteins able to move two particles at the same time in the opposing directions, and is a member of the betaine/choline/carnitine family of transport proteins (Bracher et al., 2019). In the case of CaiT the two molecules, which pass through are L-carnitine and γ-butyrobetaine (Tang, Bai, Wang & Jiang, 2010). This carrier is important for energy acquirement of the cell as it is processing fatty acids (Longo, Frigeni & Pasquali, 2016). A representative of this transporter in our dataset is 3HFX protein originating from *Escherichia coli.*

1.1.2.1.2. Glycine betaine transporter - BetP

This carrier is, similarly to CaiT, a member of betaine/choline/ carnitine family of transport proteins (Perez, Kosher, Yildiz & Ziegler, 2012). It is used in processes revolving around osmosis, such as sensing and regulation, and it is responsible for transport of betaine (Krämer & Morbach, 2004). In our dataset protein 4AIN derived from a bacteria *Corynebacterium glutamicum* represents this carrier.

1.1.2.1.3.Ammonia transporter - Amt1

This carrier, analogously to the ammonia channel, is responsible for transportation of ammonia over the membrane (Mayer & Ludewig, 2006). It can be often found in plants such as *Arabidopsis thaliana* where is helps with acquisition of nitrogen (Mayer & Ludewig, 2006). Our dataset contains protein 2B2F, which is connected to Amt1 and sources from *Archaeoglobus fulgidus* belonging to Archaea.

2. Materials and Methods

In this chapter we aim to describe the procedures and tools used to acquire the results and conclusions. The following steps are listed in chronological order in which they were applied, together with the tools, that were necessary for their completion. To achieve our result we adapted a pipeline created by Dipl. Ing. Natasha Trajkovska (this pipeline is a part of an ongoing master project of the Biophysics institute at Johannes Kepler University in Linz, Austria).

2.1. Step 1 - Our starting point : Acquiring input data and set-up

Before the work on this paper even started a set of twelve membrane transport proteins was preselected. As it is in its majority focused on processes connected to sequences, the very first step had to necessary be to acquire the required input data. This was done simply by manually accessing RCSB-PDB (Berman et al., 2000), searching each of the proteins by their Protein Data Bank identification code (PDB ID) and acquiring a FASTA file containing the corresponding sequence and a PDB file.

Once this initial part was finished we proceeded to protein Blast (BlastP) (Altschul, Gish, Miller, Myers & Lipman, 1990). The procedure here was to use the FASTA file acquired earlier from PDB and use it as a query sequence, which was then blasted against the non-redundant protein sequences (nr) database, with the maximum target sequences parameter set to 1000. As an outcome of this step we obtained a file containing all the sequences fitting our blast criteria as well as a Hit table and a Description table.

Finally, the last part of our input data necessary, was a set of Excel documents each of which contained annotation for one of our proteins. Once we acquired the annotation we compared the sequence, which it contained, with our query sequence gathered form PDB. This was done to prevent e.g. possible typos or misalignments and to be fully certain we are working with the right protein and its correct annotation. When it was confirmed that each member of our dataset possesses an annotation, i.e. the query sequence and the sequence in the annotation file were identical, we were able to continue to the next step.

2.1.1. Division of regions within our proteins

The proteins, which we worked with are oriented in their respective membranes. They can be divided into specific sections based on location and surrounding regions. This may be done in the following manner:

There are cytoplasmic (Cyto) and peripheral (Peri) regions, which represent the N and C-terminus as well as connections between e.g. α-helices traversing the membrane.

The protein surface (P-S) is able to interact with the neighbouring protomer in the oligomer. Similarly, the lipid surface (L-S) is oriented towards the lipid bilayer. Internal residues are ones, which are not directed towards either of the surfaces mentioned earlier. In case they are adjacent to the protein surface they described as protein internal (P-I). Correspondingly, lipid internal (L-I) regions are close to the lipid surface. The channel surface (C-S) are residues oriented towards the channel surface.

2.2. Step 2 - Multiple sequence alignment and work with annotation

The goal of this step was to acquire an Excel file for each of our proteins, which contained the annotation together with aligned set of sequences, divided so that each letter of a sequence and each label of the annotation were located in their own field.

First, the multiple sequence alignment was performed using ClustalOmega (Madeira et al., 2019). This resulted in a multiline fasta file, which was later transferred into a single line fasta and from there into a comma-separated values format (csv). These tables contained a column with information, such as accession number, description, organism of origin etc., extracted from the fasta file and a column containing aligned sequences.

Secondly, a new set of tables was created. The information column was preserved, however the sequences were divided so each of their characters (amino acids or gaps) were placed separately into columns. This step could be more-less considered a verification as it was crucial that the sequence alignment and division worked properly before introducing the annotation.

Finally, the annotation was extracted from the input file, divided by character and placed together with the corresponding sequences into a new Excel file, meeting the goal format we wanted to achieve.

2.3. Step 3 - Colour coding

The tables resulting from the previous step were rather overwhelming, due to their fairly large size, and because of that we decided to create an additional python script, which made the results easier to read. It resulted in a set of tables colour coded in three ways.

First, solely the annotation was colour coded, with each label having its own colour. This distinction was useful for the sake of quick orientation in the structure of the sequences, as it became rather easy and fast to pass through the file, focusing just on the colouring.

Second, the annotation and all the present amino acids were individually colour coded. As a result of this feature, it was easily possible to spot areas where the alignment was identical and where certain differences occurred.

Finally, we decided to divide the amino acids into the following groups : non-polar, polar, acidic and basic. These groups were then coloured together with the annotation. This allowed us to examine the distribution of these different groups over the labels and once more provided a less demanding way to uncover similarities and contrasting areas.

A set of tables demonstrating the colour coding on our data is possible to find in the attached supplement.

2.4. Step 4 - Removing duplicates

When we saw the contents of our full files we needed to remove sequences, which were not relevant for further analysis. The first choice to achieve this goal was to eliminate duplicate sequences in our queries. For this we had taken our original query fasta files and applied CD-HIT (Fu, Niu, Zhu, Wu & Li, 2012; Li & Godzik, 2006) on them. This tool removed all the duplicates within our files and allowed us to proceed further with only pertinent and unique sequences in our dataset.

2.5. Step 5 - Hit and description tables

Let us go back to the beginning for a moment, as that is when we together with the sequences also acquired hit and description tables. In this step, the data from these tables was finally utilised. We decided to create overview tables containing all the information about each of the proteins from out dataset. For this task the hit and description tables were very beneficial as they contain a lot of useful data.We merged these tables and added them together with other data (e.g. the raw sequences) about the queries. This resulted in a large overview table containing everything we knew about the proteins at the time and it served as an input, guidance and a verification in the future steps. These tables were adjusted so they would in their final version contain only information about the relevant sequences.

2.6. Step 6 - Removing different iso-forms of each organism

When examining our dataset closer we noticed that even though there were no more duplicates, there could still be found multiple sequences belonging to one organism. We have decided to keep only the longest protein isoform per organism to narrow down our queries even more. For this task our overview table from the previous step was useful, as it contains information about the organism of origin, and most importantly its NCBI taxonomy ID (Schoch et al., 2020; Sayers et al., 2019).We have used this information together with the sequence length as basis and over a python script we omitted all unwanted sequences. The resulting files at that point contained sequences, which were not duplicates, and each one of them originated from a different organism.

2.7. Step 7 - Gap removal

We have used the new shortened query files as an input and repeated step 2 and 3. We again acquired an alignment with annotation, at this point only containing relevant sequences. Our next step was to close gaps in our annotation. This was done by removing all the sequences which were opening such gaps, i.e. we had to omit all the proteins, which didn’t properly align with the query. We were trying to achieve a file, which would contain an annotation without any gaps, except for Cyto and Peri

regions. These two regions were not included in the gap removal procedure. This step was performed manually by deleting the unfitting sequences from the file in Excel.

2.8. Step 8 - Visualisation of our data and statistics

Once we acquired the final files for our protein dataset we visualised it to display our results more clearly. We again used python to obtain the graphical representation of our proteins. The main focus was to analyse the amino acid distributions - overall and also within the specific labels.

To acquire statistically more valuable data we also applied normalisation via two methods. First, was created using sklearn.preprocessing.normalize (Pedregosa et al., 2011) function. This works on the principle of normalisation to a unit norm, meaning the squared values of the normalised distribution sum to one. Second, was a manual normalisation, where we hard-coded the computations. These calculations were performed by simply dividing the number of occurrences of each amino acid by the total sum of amino acids. As a result we acquired a set of three graphs - one plot, which displayed the actual values and two graphs plating the two normalised functions - which we used for comparison.

2.9. Step 9 - Comparison with aquaporins

A set of aquaporins was used to display potential similarities or differences in the data. These proteins went through the same pipeline as our data so we could objectively analyse and compare these two datasets.

3. Results

3.1. Overview of our dataset

A brief summary of our dataset is displayed in the table below (Tab. I). All the information enclosed here originates from PDB (Berman et al., 2000).

TABLE I. OVERVIEW OF THE EXEMPLARY DATASET

PDB ID	Description	Organism of origin
4AIN	Crystal structure of BetP	Corynebacterium glutamicum
3KCU	Structure of formate channel	Escherichia coli
5EGI	Engineered human cystathionine gamma lyase	Caenorhabditis elegans
1KPL	Crystal Structure of the ClC Chloride Channel	Salmonella enterica
5YS3	Crystal structure of Succinate-Acetate Permease	Citrobacter koseri
1U7C	Crystal Structure of AmtB with Methyl Ammonium	Escherichia coli
5ZUG	Structure of the bacterial acetate channel SatP	Escherichia coli K-12
1KPK	Crystal Structure of the ClC Chloride Channel	Escherichia coli
6NSK	CryoEM structure of Helicobacter pylori urea channel in open state.	Helicobacter pylori
3HFX	Crystal structure of carnitine transporter	Escherichia coli
6EID	Crystal structure of wild-type Channelrhodopsin 2	Chlamydomonas reinhardtii
2B2F	Ammonium Transporter Amt-1	Archaeoglobus fulgidus

3.2. The resulting dataset

3.2.1. Evaluation of size

Here we present the twelve queries together with the amount of sequences within them, which we acquired as the final set for the analysis. These proteins had undergone all the steps described in the previous chapter and as it is visible in the table below (Tab. II) their sequence counts differ rather vastly from each other and from their original size, which was set to 1000 sequences. We have to mention here, that not all the present proteins were able to meet this set threshold - 6EID reached

only 406 results, when it was blasted. The size of the entire final dataset equals to 832 sequences.

TABLE II. SIZE PROGRESS OF OUR QUERIES

Query	Original size	Number of sequences	Number of reduced sequences	% remaining from the original
4AIN	1000	143	857	14,3
3KCU	1000	136	864	13,6
5EGI	1000	132	868	13,2
1KPL	1000	121	879	12,1
5YS3	1000	118	882	11,8
1U7C	1000	65	935	6,5
5ZUG	1000	40	960	4
1KPK	1000	30	970	3
6NSK	1000	25	975	2,5
3HFX	1000	19	981	1,9
6EID	406	2	404	0,49
2B2F	1000	1	999	0,1

3.2.2. Representation of different kingdoms

Our dataset contains mainly bacterial sequences, however there were other kingdoms of life present. We acquired results for Animals, Plants, Archaea and also one unspecified kingdom. The amount of bacterial sequences equaled to 697 and animal sequences only scored 132 entries. These two kingdoms are the most present in our dataset as the rest only contained one result each. It is also worth mentioning that we further investigated the origin of the sequence sourcing from an unknown kingdom (described simply as “other sequences”) in a worry of potential error. However, we have discovered that this protein classifies as a synthetic construct and as such does not have a usual kingdom of origin.

3.3. Analysis of sequence contents

We analysed our dataset with regards to three categories : sequence length, frequency of occurrence of each amino acid and frequency of occurrence of each label class. This was performed for the entire dataset together and then for each of the twelve queries.

3.3.1. The whole dataset

3.3.1.1.Sequence Length

The figure below (Fig. 1) displays the sequence length of all our sequences over density. As it is visible from the plot the sequences ranged from approximately 200 to 600 bp in length.

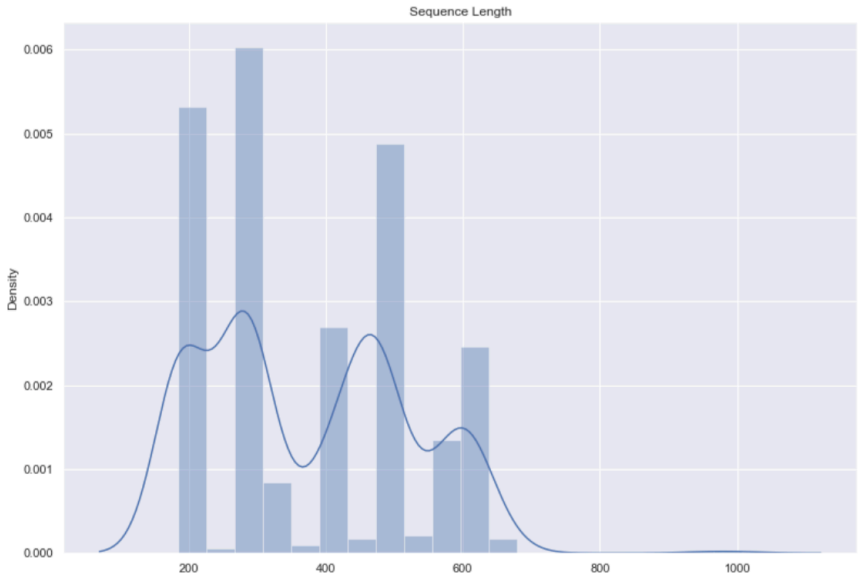


FIGURE 1. GRAPH OF SEQUENCE LENGTH OVER DENSITY FOR THE WHOLE DATASET

3.3.1.2. Amino Acid Frequency

The frequency of amino acids acquired in the whole dataset is displayed on the figure below (Fig. 2). As you can see the amino acid which was discovered in the most cases was leucine, closely followed by alanine and glycine. On the other hand the lowest scoring was cysteine, only slightly higher scored histidine and tryptophan.

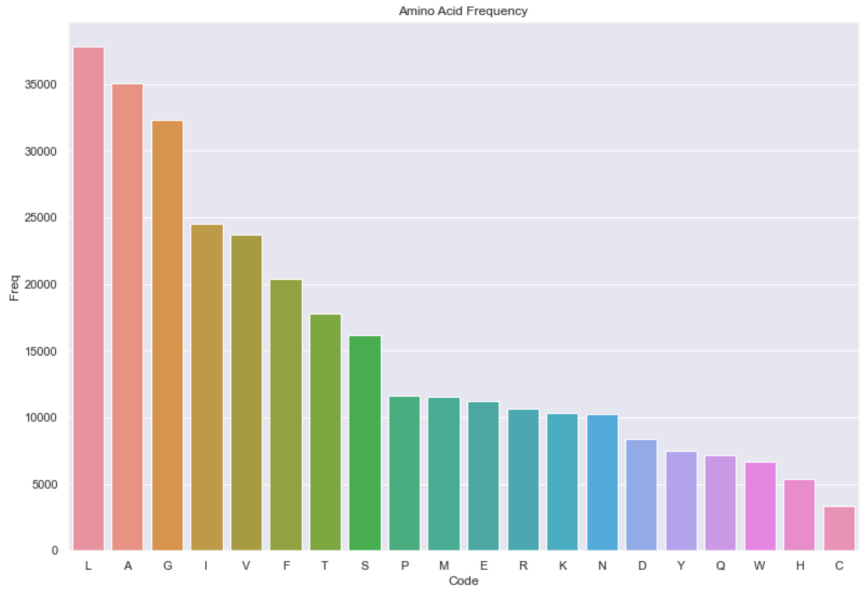


FIGURE 2. AMINO ACID FREQUENCY IN THE OVERALL DATASET

3.3.1.3. Label Frequency

We decided to display how frequently the different labels of the annotation occur in our dataset. The figure (Fig. 3) below shows that the most highly scoring label was L-S, i.e. the lipid-surface region, the second was P-S, i.e. the protein-surface region. The lowest were regions with gaps, which could be found within the Peri or Cyto regions.

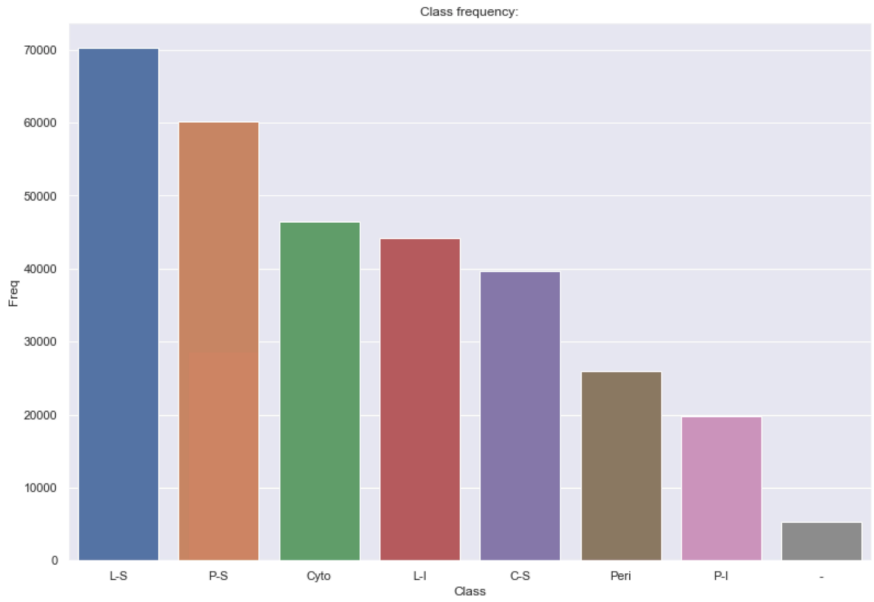


FIGURE 3. CLASS FREQUENCY IN THE WHOLE DATASET

3.3.2. Individual sequences

All the sequences were also individually visualised in the same manner as the overall dataset. All the results of these analyses can be found in Appendix A. Not all the sequences match the exact proportions of distribution depicted on the overall graphs, for example not all our proteins possess leucine as the most frequently present amino acid.

3.4. Amino acid distribution in specific regions (whole dataset)

In this chapter we describe the composition of selected regions in the proteins based on their amino acid distribution. As mentioned earlier, we used two different normalisation methods and we display them here for comparison side by side next to their absolute value equivalent. We also chose to add a comparison of all the protein and lipid regions to provide an addition point of view on our dataset.

3.4.1. Protein-Surface vs Protein-Internal

When comparing these two protein regions one can not miss the differentiation between them (Fig. 4). The Protein-Surface region is rather rich in leucine (L) and isoleucine (I), as these two amino acids are the most occurring. On the other hand the Protein-Internal section is abundant in alanine (A) and glycine (G). All of the mentioned amino acids classify as non-polar.

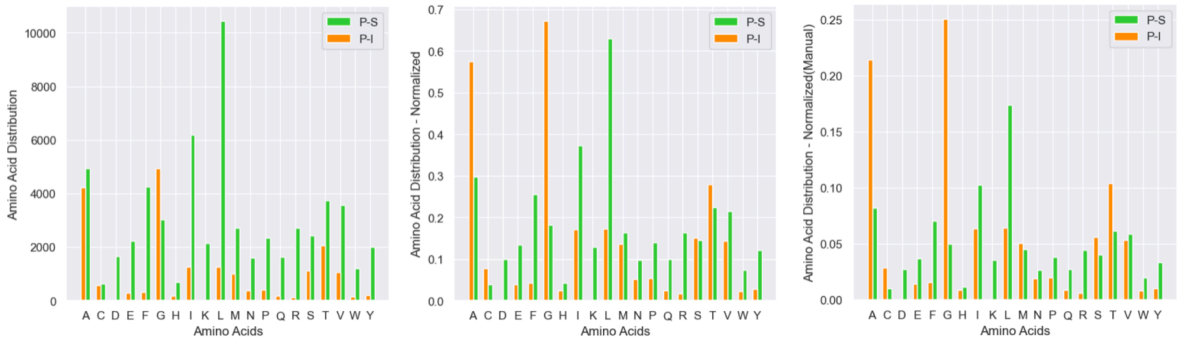


FIGURE 4. AMINO ACID DISTRIBUTION IN P-S AND P-S REGIONS OF THE WHOLE DATASET, CONTAINING THE ABSOLUTE VALUE RESULTS (LEFT), NORMALISED RESULTS (MIDDLE) AND "MANUALLY NORMALISED" RESULTS (RIGHT)

3.4.2. Lipid-Surface vs Lipid-Internal

Comparison of the lipid region seems a bit more varying, however the preferences of certain amino acids appear to have lasted (Fig. 5). The surface region L-S is favouring leucine (L), isoleucine (I) and valine (V). This is indeed very similar to the P-S layout we have mentioned earlier, where even valine was represented more in the surface region rather than in the internal, although its results in that case did not appear as that notable. For Lipid-Internal region the similarity remains, as it again displays a preference for alanine (A) and glycine (G). All of these amino acids are once more non-polar and so there was no difference in this area either, when it comes to protein and lipid interfaces.

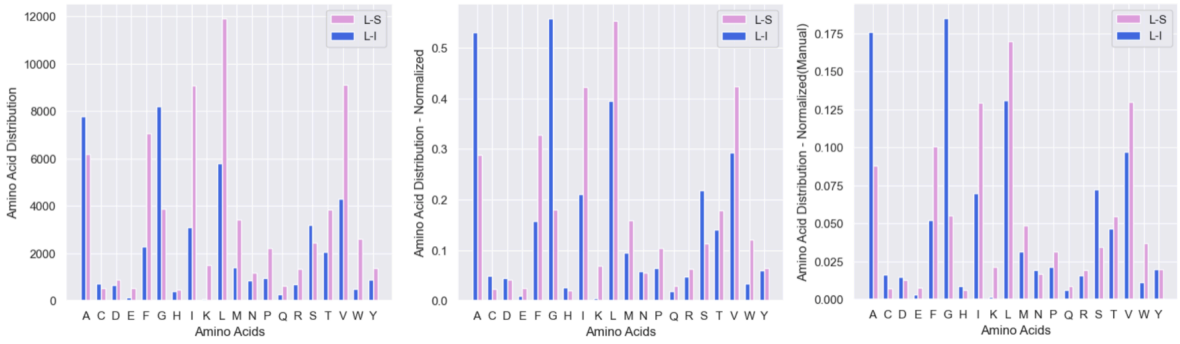


FIGURE 5. AMINO ACID DISTRIBUTION IN L-S AND L-I REGIONS OF THE WHOLE DATASET, CONTAINING THE ABSOLUTE VALUE RESULTS (LEFT), NORMALISED RESULTS (MIDDLE) AND "MANUALLY NORMALISED" RESULTS (RIGHT)

3.4.3. Distribution over all Protein and Lipid regions

This depiction of all the lipid and protein sectors together can potentially provide a more compact demonstration of the amino acid distributions (Fig. 6). The distributions are again normalised using our "manual normalisation" procedure.

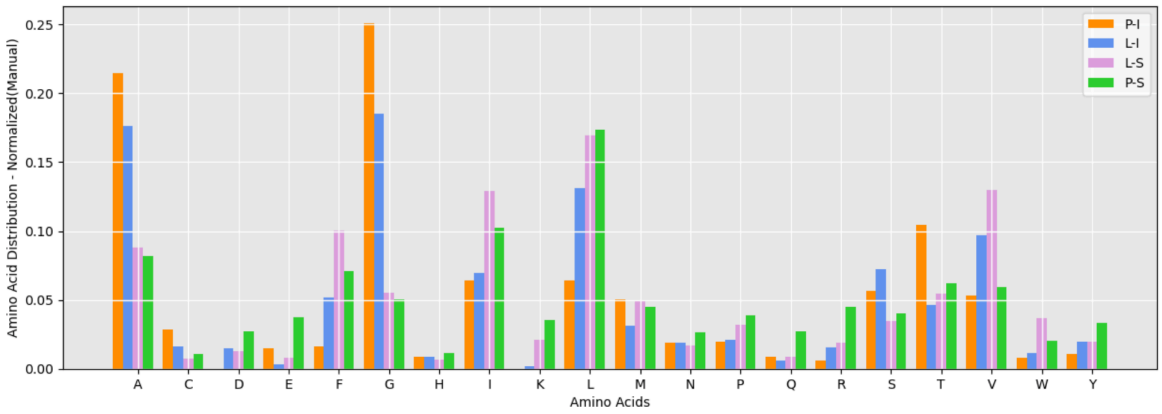


FIGURE 6. AMINO ACID DISTRIBUTION IN PROTEIN AND LIPID INTERFACES (SURFACE AND INTERNAL) IN THE WHOLE DATASET

3.4.4. Distribution in other regions

As the comparison of lipid and protein regions was our main focus we do not include here the additional results which we acquired for the Cyto and Peri regions. However, we have still taken them into account and they can be found in the Appendix B.

3.5. Amino acid distribution in specific regions (individual proteins)

3.5.1. Protein-Surface vs Protein-Internal

The protein-surface region displays in nine out of twelve cases in our dataset leucine clearly as the most frequent amino acid (Fig. 7). Alanine was the most present in two proteins and for sample 2B2F the distribution of A and L was nearly identical. In general this seems quite similar to the amino acid distribution of the whole dataset.

The distributions of the protein-internal region, however, show more dissimilarities. All our proteins together showed four different amino acids that scored the highest frequencies in the P-I sector. Half of our queries obtained G as the most occurring amino acid (Fig. 7- 3KCU,6NSK), three proteins displayed A (Fig. 7-1KPK), two proteins acquired L and one (Fig. 7 - 5EGI) scored methionine (M) the highest.

The amino acid distribution of all proteins with absolute values can be seen in the appendix A.

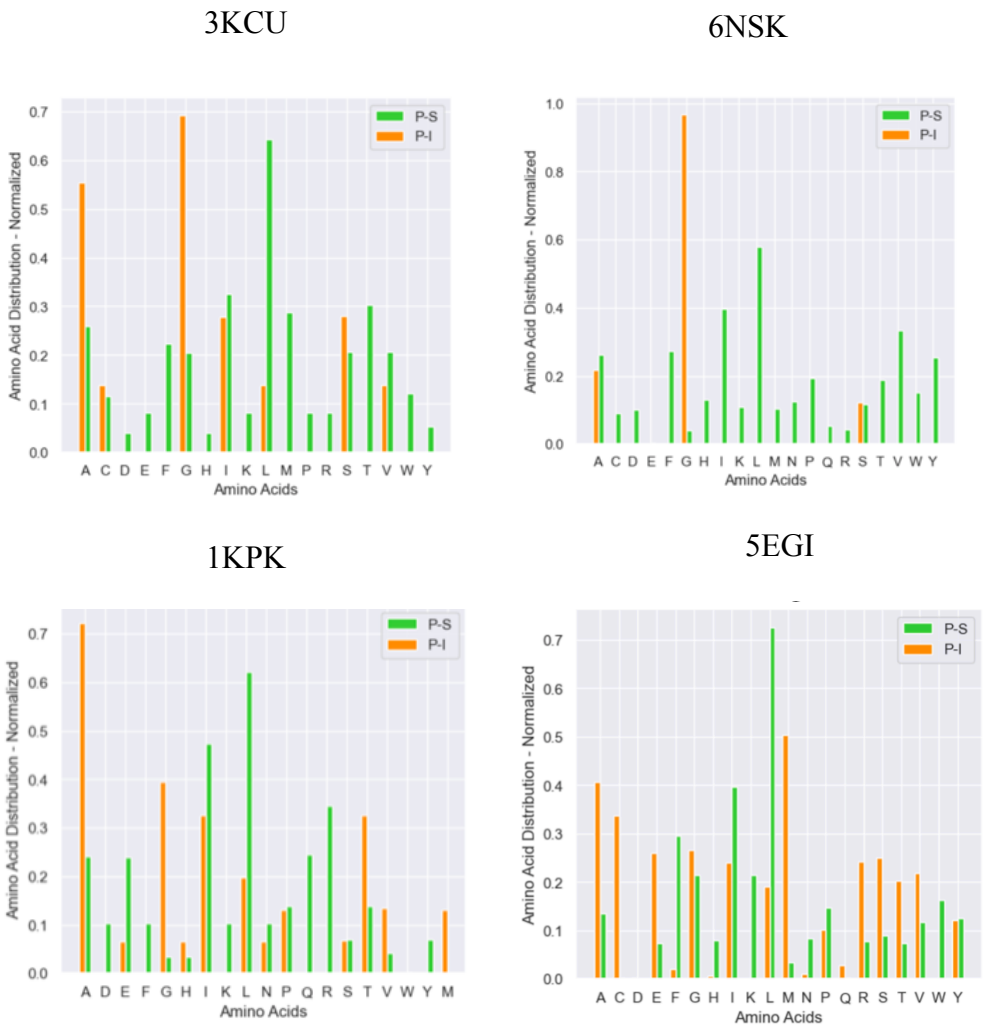


FIGURE 7. EXEMPLARY AMINO ACID DISTRIBUTIONS IN P-S AND P-I REGIONS OF PROTEINS 3KCU (UPPER LEFT), 6NSK (UPPER RIGHT), 1KPK(LOWER LEFT) AND 5EGI (LOWER RIGHT) - NORMALISED TO UNIT NORM

3.5.2. Lipid-Surface vs Lipid-Internal

The lipid-surface region even exceeds the P-S in the amount of proteins, which obtained leucine as the most frequent amino acid, acquiring eight out of twelve (Fig. 8 - 4AIN). Isoleucine was on the second place, having the highest distribution for one protein (6NSK). The rest of the proteins didn't show so apparent results and some amino acids gained similar scores (such as L and I in Fig. 7 - 3HFX).

The distributions once more became more diverse, when examining the L-I region. Similarity to P-I, we obtained several different amino acids. Half of our dataset favoured G as the leading amino acid (Fig. 8 - 5YS3). Alanine was the most frequent for three proteins (Fig. 8 - 4AIN). Leucine and valine (V) each acquired the highest

distribution for one query. The rest of the proteins again displayed similarly high distributions for more than one acid (Fig. 8 - 6NSK).

This comparison of all our proteins individually provides an interesting insight to their amino acid distributions. It also shows how similar can the protein and lipid interfaces be. However, based on the acquired results it appears that P-S and L-S of the individual proteins display more resemblance to the overall dataset compared to their internal counterparts.

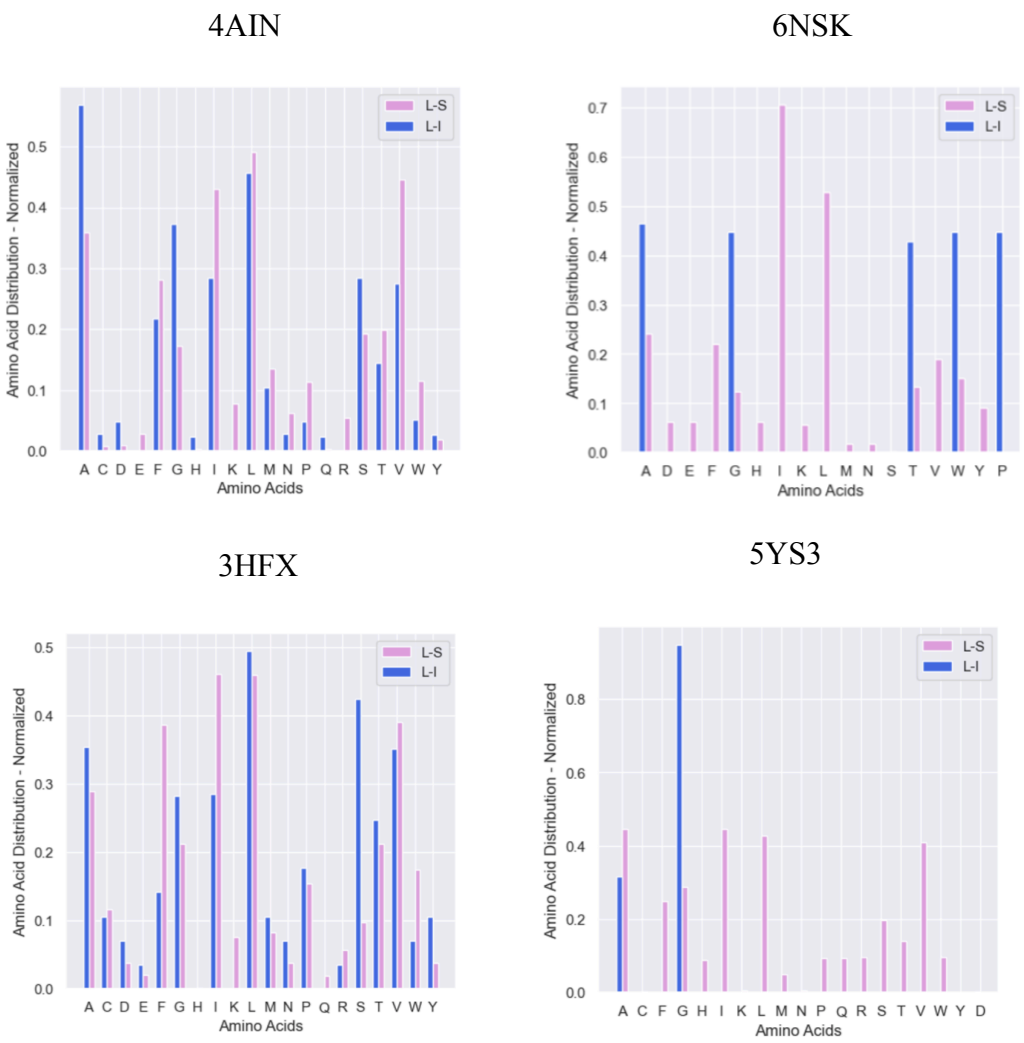


FIGURE 8. EXEMPLARY AMINO ACID DISTRIBUTIONS IN L-S AND L-I REGIONS OF PROTEINS 4AIN (UPPER LEFT), 6NSK (UPPER RIGHT), 3HFX (LOWER LEFT) AND 5YS3(LOWER RIGHT) - NORMALISED TO UNIT NORM

3.6. Comparison with the Aquaporin dataset

3.6.1. Amino acid distribution

This first set of graphs (Fig. 9) displays normalised amino acid distributions. The purpose of this comparison was to see how similar these two datasets were in their amino acid content. We can conclude that in this category there are no major differences and only a few dissimilarities, and as such the two datasets have very similar trends.

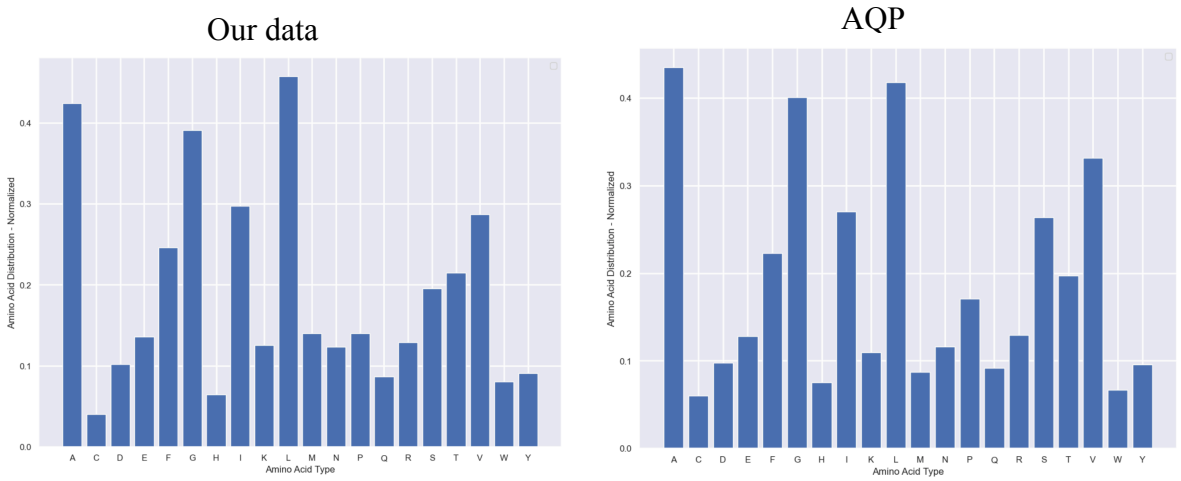


FIGURE 9. NORMALISED AMINO ACID DISTRIBUTIONS IN OUR DATASET (LEFT) AND THE AQUAPORIN DATASET (RIGHT)

3.6.2. Annotation label distribution

Similar analysis was performed in the case of labels and their normalised distributions were compared between the two datasets. In this category the aquaporins and our data seems to have rather different tendencies (Fig. 10) for example our most frequent label is L-S and in the case of aquaporins it is the Cyto label. On the other hand there can be certain similarities seen in Peri region.

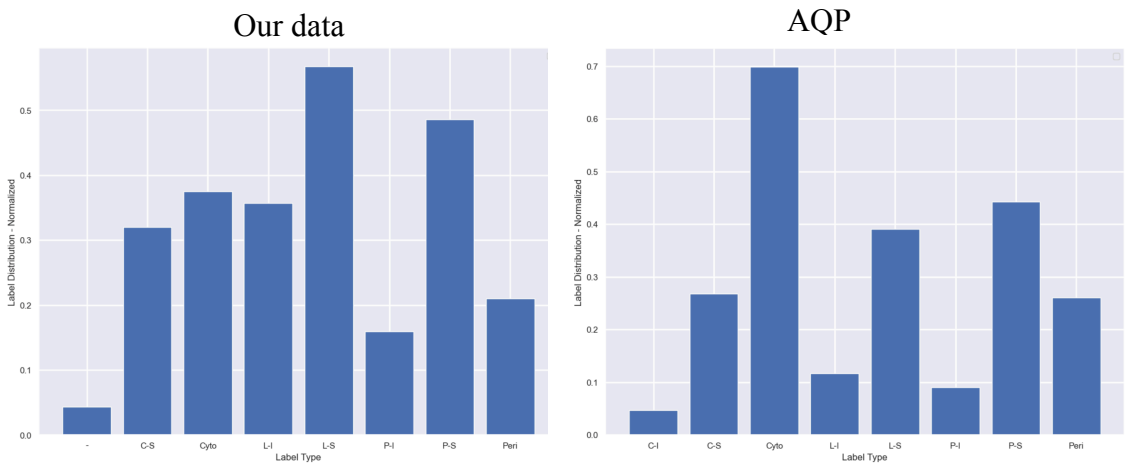


FIGURE 10. NORMALISED INTERFACE DISTRIBUTIONS IN OUR DATASET (LEFT) AND THE AQUAPORIN DATASET (RIGHT)

3.6.3. Amino acid distribution in specific regions

In this part of our paper we present the comparison of aquaporin and our dataset. The results show several differences and similarities, when it comes to the amino acid distributions, which we will address presently.

3.6.4. Protein-Surface vs Protein-Internal

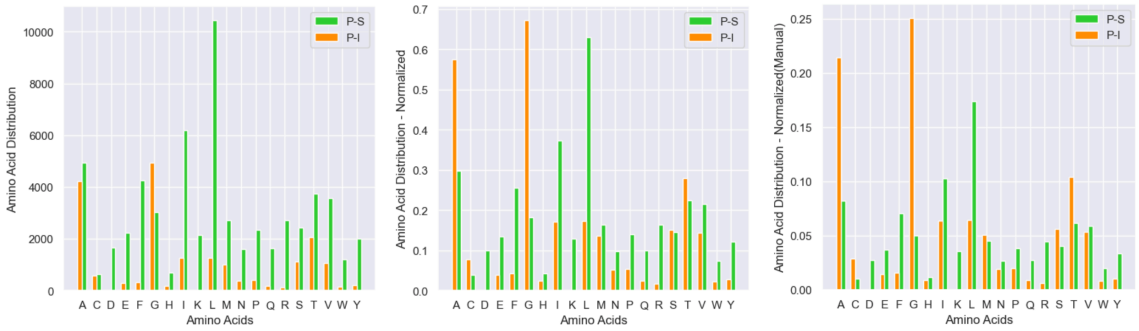
The graphs below (Fig. 11) show certain trends supporting the ones described earlier.

When talking about the internal region, aquaporins seem to exhibit a preference for G and A, with amounts of A similar to threonine (T), which we do not necessarily observe in our data. In our case G and A occurrences appear to outnumber the rest by far greater amount. However, in both cases T still holds the third place when it comes to amino acid distribution.

Moving on to the surface sector, where the results actually differ more than in the previous case. Let us start with the main similarity, which is the clear preference for leucine, however the remaining order slightly varies. While our datasets favours I and later A that is not true for aquaporins. There A is the second most occurring amino acid and I displays similar amounts to phenylalanine (F), placing it to a third or potentially fourth place.

Despite these variations we can still see that for both these datasets there is a preference for certain amino acids, which seems to be region specific.

Our data



AQP

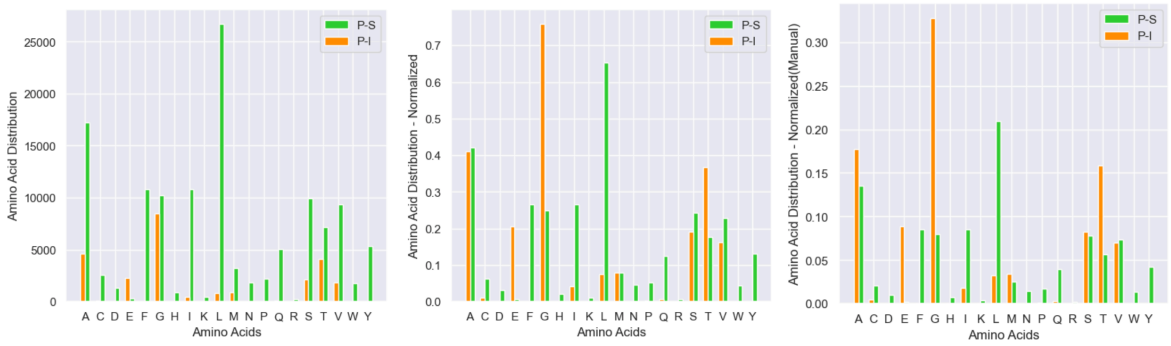


FIGURE 11. AMINO ACID DISTRIBUTION IN P-S AND P-I REGIONS FOR OUR DATASET (UPPER ROW) AND THE AQUAPORIN DATASET (LOWER ROW)

3.6.5. Lipid-Surface vs Lipid-Internal

When we look at the graphs below (Fig. 12) there are several spikes in the distributions.

The lipid-surface for our data displays the following amino acids as the most occurring: L>I>V>F>A, with leucine having the highest recorded distribution. In case of aquaporins the order is slightly different: L>V>I>F>A. As we can see the first place for both datasets is still taken by leucine and the fourth and fifth place are again identical. The slight difference seems to be V and I, which in our dataset have approximately the same distribution, whereas in case of aquaporins occurrence of V clearly surpasses I. The precise order is hence not fully set, however both datasets seem to agree on the type of amino acids which are the most frequently present within their surface regions.

The lipid-internal region across all the data displays a certain misbalance, as there is in both datasets a clear preference for A and G, however when we look at the remaining amino acids they vary. For aquaporins, there is basically no notable result for several amino acids, such as histidine (H), I, lysine (K), L, M, asparagine (N) etc. This is not the case for our dataset, where all the present amino acids acquired distribution amounts clearly visible on the plots.

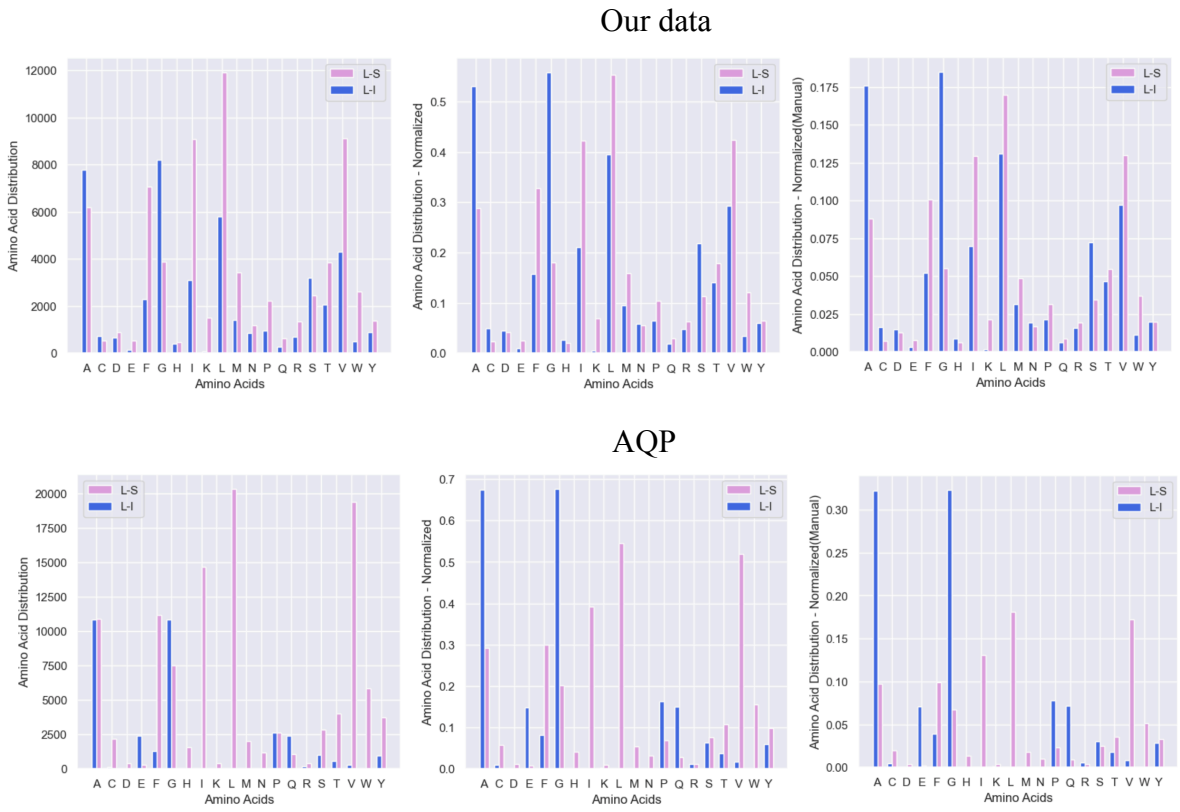


FIGURE 12. AMINO ACID DISTRIBUTION IN L-S AND L-I REGIONS FOR OUR DATASET (UPPER ROW) AND THE AQUAPORIN DATASET (LOWER ROW)

3.6.6. Protein-Surface vs Lipid-Surface

The following figure (Fig. 13) represents the side by side comparison between the two surface interfaces - protein and lipid. The main preference for leucine is valid for both the interfaces and the datasets. In our case the difference exhibited by the leucine occurrences compared to the other amino acids seems to be slightly higher, whereas in the aquaporin dataset the L-S leucine distribution is fairly close to valine and as such does not have so clear head start.

The leading amino acids in both datasets in the P-S region are L, I, A, F. The distribution of these amino acids in our case is the following: L>I>A>F, and for aquaporins it is slightly altered: L>A>F>I.

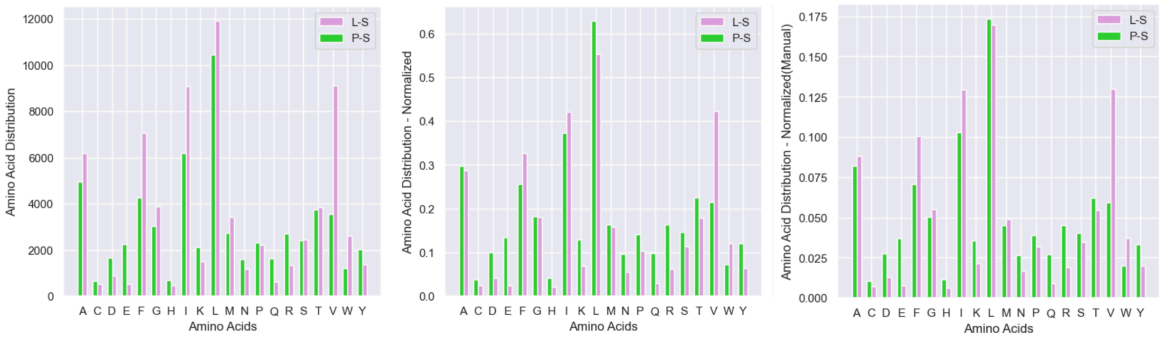
The L-S region shows L, V, I, F distributions as the highest in both the datasets and in the same order (L>V>I>F).

From the data it is visible that lipid and protein surface regions share several similarities, for example the fact that three out of four leading amino acids in

the distribution of both interfaces are matching. In other words, L, I and F can be found among the top four for P-S and L-S regions in both datasets.

When it comes to the differences between P-S and L-S in both datasets, then the lipid interface displays higher amounts of mainly valine, isoleucine, phenylalanine and tryptophan. The protein surface scores more in leucine and most other amino acids. In comparison with aquaporins, these trends are mostly valid as well.

Our data



AQP

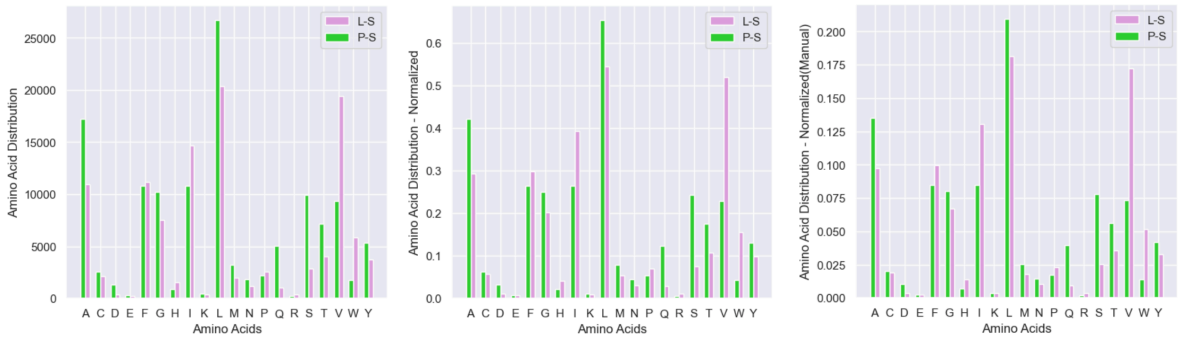


FIGURE 13. AMINO ACID DISTRIBUTION IN P-S AND L-S REGIONS FOR OUR DATASET (UPPER ROW) AND THE AQUAPORIN DATASET (LOWER ROW)

3.6.7. Protein-Internal vs Lipid-Internal

Analogously to the previous evaluation we compared also the internal sections of protein and lipid regions (Fig. 14).

Both the internal regions in both of the datasets display a preference for G and A.

When we look at the P-I sector in both of the datasets there is a match in the order of distributions of the leading three amino acids making it: G>A>T. Our dataset follows with L and I (having nearly identical distributions), the aquaporins however score higher for glutamic acid (E) and serine (S).

The L-I region does not share the three most frequently occurring amino acids but only two G>A. The third most common amino acid for our dataset is L(>V>S), and for the aquaporins it is proline (P)(> glutamine (Q) > E).

Therefore, the internal regions seem to display more differences in the distributions, as they only share the leading two amino acids (G and A) and vary in the rest, comparing to the surface regions we have examined earlier, which shared three out of four leading amino acids.

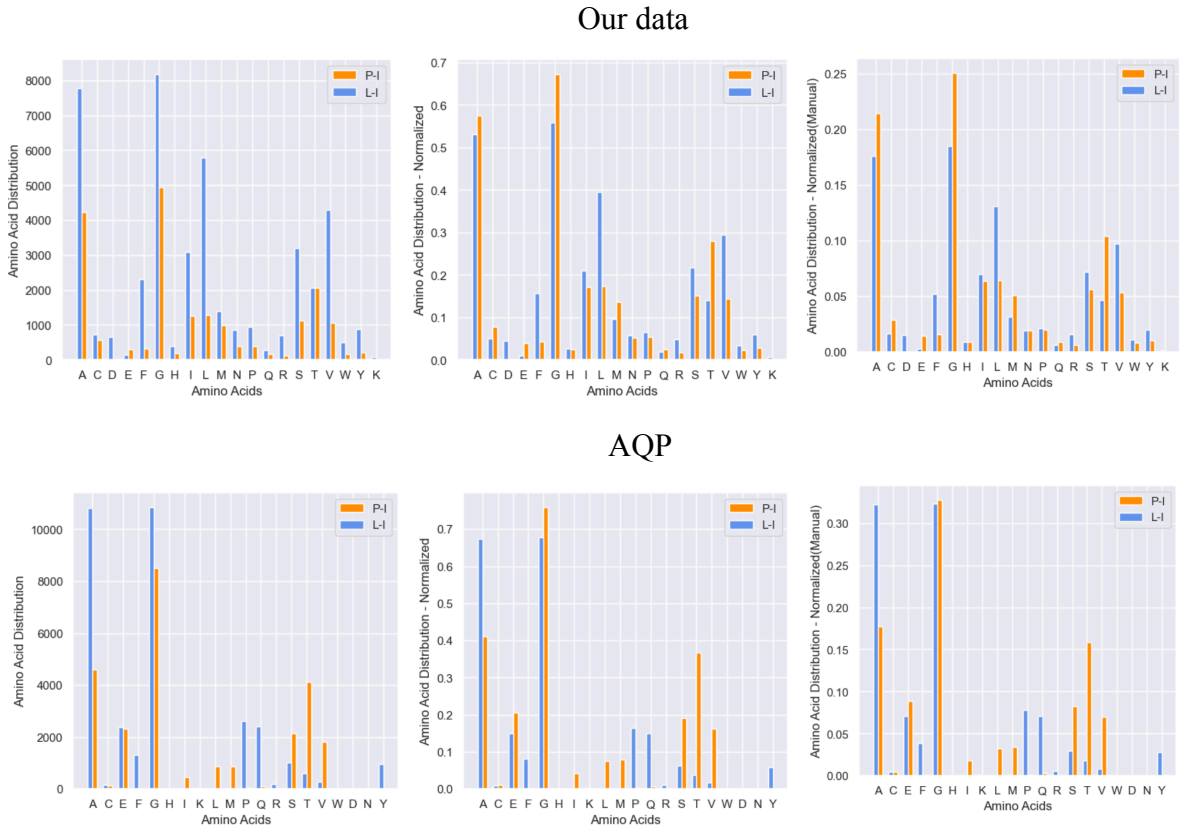


FIGURE 14. AMINO ACID DISTRIBUTION IN P-I AND L-I REGIONS FOR OUR DATASET (UPPER ROW) AND THE AQUAPORIN DATASET (LOWER ROW)

3.6.8. Cyto vs Peri

The Cyto and Peri regions of our two datasets are not in the primary focus of our work, however we were still interested to see how they might compare to one another (Fig. 15).

The Peri region of our data has the following leading amino acids: G>A>S>P, and the aquaporin dataset: G>A>S>L. As such they seem to share G,A,S as the three most frequent amino acids.

The Cyto region is more diverse with A> arginine (R) >E>G leading distribution in our data and S>R>K>G in the aquaporins, having the second (R) and fourth (G) amino acid in common.

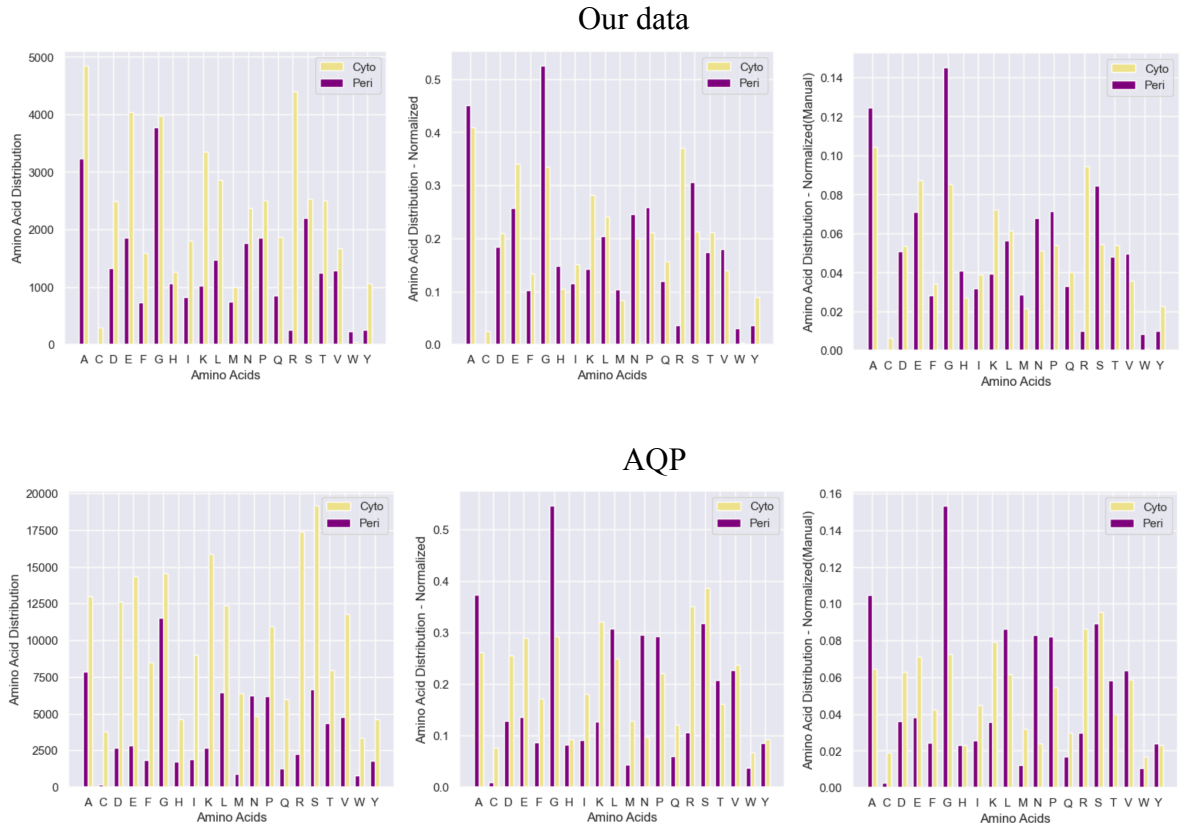


FIGURE 15. AMINO ACID DISTRIBUTION IN CYTO AND PERI REGIONS FOR OUR DATASET (UPPER ROW) AND THE AQUAPORIN DATASET (LOWER ROW)

3.6.9. Amino acid group distribution

When comparing the two datasets, we also decided to categorise the amino acids into four main groups : Polar, Non-polar, Basic and Acidic, and display the distributions of the respective groups in the protein and lipid interfaces (Fig. 16).

Across all the four regions the most frequent amino acid group is Non-polar. The second most occurring group in the data is Polar, where slightly bigger differences are showing between the interfaces. Our dataset displays alike distribution for P-S and P-I in the Polar region whereas the aquaporins favour P-I over P-S. It appears that for the lipid interface the two datasets have more in common and share similar trends in Polar, Non-polar and Basic groups,

however, this is not true for the Acidic category, where our data remains rather similar, but the aquaporins show higher distribution in the L-I region.. As the main likeness in the protein interfaces we can potentially mention the higher share of Basic amino acids in the P-S region comparing to the P-I, which is valid for both datasets.

Comparing the two surface regions it is visible for both datasets that L-S has more Non-polar amino acids and P-S has more Polar ones. For our dataset P-S also contains more Basic and Acidic amino acids compared to L-S.

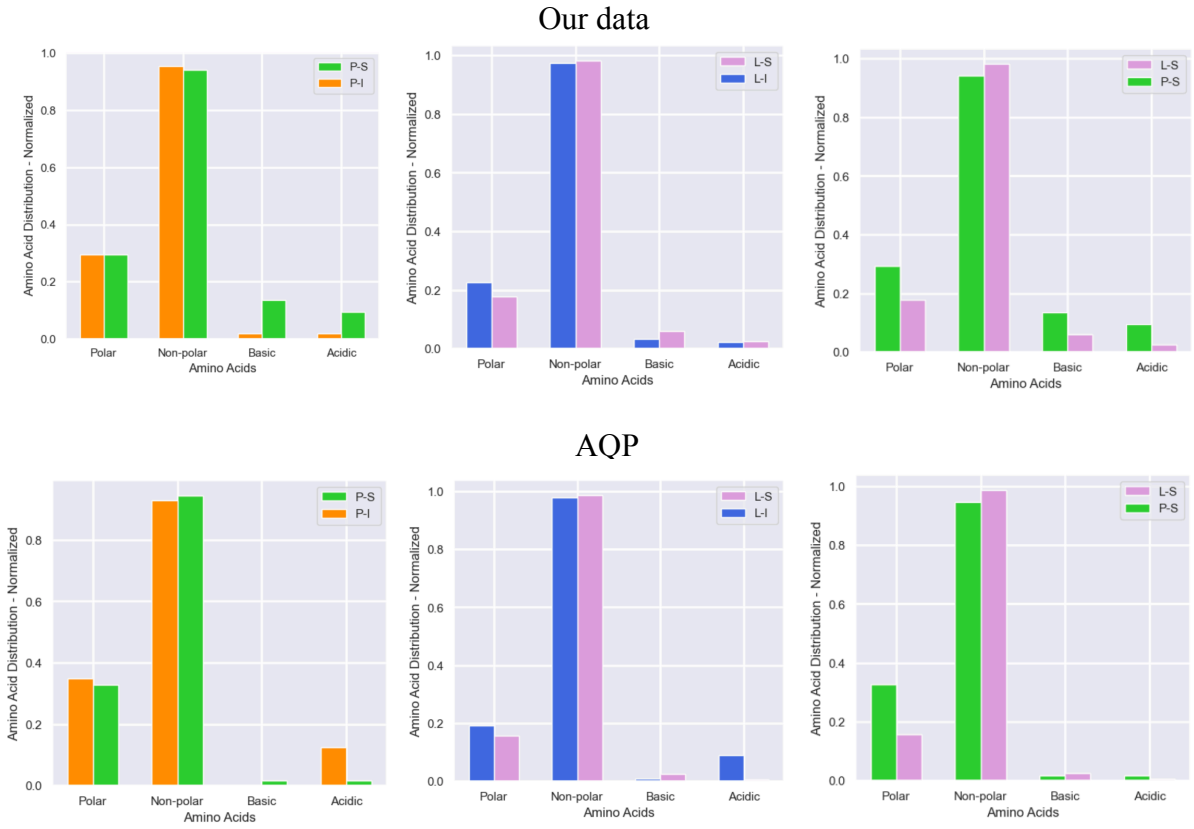


FIGURE 16. AMINO ACID GROUP DISTRIBUTIONS IN PROTEIN AND LIPID INTERFACES OF OUR DATASET (UPPER ROW) AND THE AQUAPORIN DATASET (LOWER ROW) - NORMALISED TO UNIT NORM

4. Discussion

First, when comparing the distributions, there seems to be a preference for large hydrophobic amino acids (such as leucine) in the surface regions (P-S and L-S). On the other hand, the internal regions (P-I and L-I) display a preference for glycine and alanine, hence for small hydrophobic amino acids. These findings correspond to the conclusions depicted in the work of Ulmschneider and Sansom as they also state leucine is more likely to be located in the surface regions (Ulmschneider & Sansom, 2001). Among other things they also pointed out the preference of aromatic amino acids to occur in the internal regions (Ulmschneider & Sansom, 2001). Nevertheless, this trend is not observable for our dataset.

Second, the surface areas tend to be basic rather than acidic. On the contrary, the L-I region is more basic. In the P-I region the acidity is more balanced.

Even though both surface regions are quite similar, they both still display distinct trends.

P-S region contains more polar and charged amino acids while L-S is a little more hydrophobic. In more detail, the amino acid distributions of the surface regions show that the lipid interface displays higher amounts of mainly valine, isoleucine, phenylalanine and tryptophan. The protein surface scores more in leucine and most other amino acids.

Third, in comparison the aquaporin dataset is quite similar to ours. The main observed trends in amino acid distribution and charge distribution matched relatively well. Despite some variations in the surface and the internal protein region we can still see that for both these datasets there is a preference for certain amino acids, which seems to be region specific. Based on this case we can hypothesise that a region strongly favouring leucine is more likely going to belong to the surface sector rather than to the internal. In like manner it appears that a section with a preference for A and G should plausibly be in the internal region.

To summarise, the results support the hypothesis stated earlier, as the surface region for both datasets again favours L and the internal region prefers A and G. For all regions we found specific trends, that are for the most part consistent between our dataset and the aquaporins. It is noteworthy that there are still differences between our dataset and the latter, however these results still suggest that an unknown amino acid distribution of a transmembrane protein interface could be recognised.

It is worth mentioning, that the amount of sequences for the proteins in our dataset varies. This could potentially create a slight discrepancy in our results. This should not cause issues for the normalised data, nevertheless the error in datasets with lower amount of amino acids may be higher.

These trends, however, are distributed over the dataset and therefore should not have substantial influence on our results.

Some differences between proteins are plausibly due to their specific functions. 1KPK and 1KPL, which both transport chlorine, display alike trends. Our two other proteins 2B2F and 1U7C are both proteins transporting ammonium and they also exhibit quite similar distributions, however they are less similar to each other than 1KPK and 1KPL.

When observing the amino acid distributions of the individual proteins it can be said that the surface regions appear to be more stable and constant over the dataset, whereas the internal regions show higher variability. This once more could potentially be due to the different functionalities.

5. Conclusion

After analysing the amino acid distribution of exemplary oligomeric membrane proteins we successfully identified differences between protein interfaces facing the lipid bilayer and protomer-protomer interfaces. For each region we found specific trends in the amino acid distribution.

These were comparable to the aquaporin dataset and its specific patterns as there were only slight dissimilarities. The variability between the two sets of data seems logical regarding their different origin.

In future research we are considering to handle parts of the dataset analysis differently. For example we would like to apply a different gap removal method, which allows gaps in certain loop regions, where we expect the sequences to be more variable. This approach could hinder large sequence drops as was observed for e.g. 2B2F. Another potential expansion is the incorporation of z-coordinates, which could be aligned according to peaks of different amino acids in various interfaces.

This work will hopefully aid to a better understanding of the effects of the oligomerization process, which could later be applied to the engineering of new proteins.

6. References

1. Alberts B., Johnson A., Lewis J., Raff M., Roberts K., & Walter P. (2002). Principles of Membrane Transport. In Molecular Biology of the Cell. 4th edition. New York: Garland Science.
2. Altschul, S., Gish, W., Miller, W., Myers, E., & Lipman, D. (1990). Basic local alignment search tool.. J Mol Biol, 215(3), 403–410.
3. a) Berg J., Tymoczko J, & Stryer L. (2002). The Transport of Molecules Across a Membrane May Be Active or Passive. In Biochemistry. 5th edition. New York: W H Freeman
4. b) Berg J., Tymoczko J., & Stryer L. (2002). Membrane Channels and Pumps. In Biochemistry. 5th edition. New York: W H Freeman
5. Berman, H., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T., Weissig, H., Shindyalov, I., & Bourne, P. (2000). The Protein Data Bank.. Nucleic Acids Res, 28(1), 235–242.
6. Bracher, S., Hilger, D., Guérin, K., Polyhach, Y., Jeschke, G., Krafczyk, R., Giacomelli, G., & Jung, H. (2019). Comparison of the functional properties of trimeric and monomeric CaiT of Escherichia coli. Scientific Reports, 9(1), 3787.
7. Cooper G. (2000). Cell Membranes. In he Cell: A Molecular Approach. 2nd edition.Sunderland (MA): Sinauer Associates.
8. Duarte, J., Biyani, N., Baskaran, K., & Capitani, G. (2013). An analysis of oligomerization interfaces in transmembrane proteins. BMC Structural Biology, 13(1), 21.
9. Forrest, L. (2015). Structural Symmetry in Membrane Proteins. Annual Review of Biophysics, 44(1), 311–337.
10. Fu, L., Niu, B., Zhu, Z., Wu, S., & Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data.. Bioinformatics, 28, 3150–3152.
11. Gao, X., Lin, S.H., Ren, F., Li, J.T., Chen, J.J., Yao, C.B., Yang, H.B., Jiang, S.X., Yan, G.Q., Wang, D., Wang, Y., Liu, Y., Cai, Z., Xu, Y.Y., Chen, J., Yu, W., Yang, P.Y., & Lei, Q.Y. (2016). Acetate functions as an epigenetic metabolite to promote lipid synthesis under hypoxia.. Nat Commun, 7, 11960.

12. Garratt, R., Valadares, N., & Bachega, J. (2013). Oligomeric Proteins. In Encyclopedia of Biophysics (pp. 1781–1789). Springer Berlin Heidelberg.
13. Gruswitz, F., O'Connell, J., & Stroud, R. (2007). Inhibitory complex of the transmembrane ammonia channel, AmtB, and the cytosolic regulatory protein, GlnK, at 1.96 Å. Proceedings of the National Academy of Sciences, 104(1), 42–47.
14. Jentsch, T. (2002). Chloride channels are different. Nature, 415(6869), 276–277.
15. Jentsch, T., Steinmeyer, K., & Schwarz, G. (1990). Primary structure of Torpedo marmorata chloride channel isolated by expression cloning in Xenopus oocytes. Nature, 348(6301), 510–514.
16. Krämer, R., & Morbach S. (2004). BetP of Corynebacterium glutamicum, a transporter with three different functions: betaine transport, osmosensing, and osmoregulation. Biochimica et Biophysica Acta (BBA) - Bioenergetics, 1658(1), 31-36.
17. Li, W., & Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.. Bioinformatics, 22(13), 1658–1659.
18. a) Lodish H., Berk A., Zipursky S., Matsudaira P., Baltimore D., & Darnell J. (2000). Membrane Proteins. In Molecular Cell Biology. 4th edition. New York: W. H. Freeman.
19. b) Lodish H., Berk A., Zipursky S., Matsudaira P., Baltimore D., & Darnell J. (2000). Overview of Membrane Transport Proteins. In Molecular Cell Biology. 4th edition. New York: W. H. Freeman.
20. Longo, N., Frigeni, M., & Pasquali, M. (2016). Carnitine transport and fatty acid oxidation.. Biochim Biophys Acta, 1863(10), 2422–2435.
21. Lü, W., Du, J., Schwarzer, N., Wacker, T., Andrade, S., & Einsle, O. (2013). The formate/nitrite transporter family of anion channels.. Biol Chem, 394(6), 715–727.
22. Madeira, F., Park, Y., Lee, J., Buso, N., Gur, T., Madhusoodanan, N., Basutkar, P., Tivey, A., Potter, S., Finn, R., & Lopez, R. (2019). The EMBL-EBI search and sequence analysis tools APIs in 2019. Nucleic acids research, 47(W1), W636–-W641
23. Marshall, B., Barrett, L., Prakash, C., McCallum, R., & Guerrant, R. (1990). Urea protects Helicobacter (Campylobacter) pylori from the bactericidal effect of acid.. Gastroenterology, 99(3), 697–702.

24. Mayer, M., & Ludewig, U. (2006). Role of AMT1;1 in NH4+ acquisition in Arabidopsis thaliana. Plant biology (Stuttgart, Germany), 8(4), 522–-528.
25. McNulty, R., Ulmschneider, J., Luecke, H., & Ulmschneider, M. (2013). Mechanisms of molecular transport through the urea channel of Helicobacter pylori.. Nat Commun, 4, 2900.
26. Mirandela, G., Tamburrino, G., Hoskisson, P., Zachariae, U., & Javelle, A. (2019). The lipid environment determines the activity of the Escherichia coli ammonium transporter AmtB.. FASEB J, 33(2), 1989–1999.
27. Nagel, G., Szellas, T., Huhn, W., Kateriya, S., Adeishvili, N., Berthold, P., Ollig, D., Hegemann, P., & Bamberg, E. (2003). Channelrhodopsin-2, a directly light-gated cation-selective membrane channel.. Proc Natl Acad Sci U S A, 100(24), 13940–13945.
28. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research, 12, 2825-2830.
29. Perez, C., Koshy, C., Yildiz, O., & Ziegler, C. (2012). Alternating-access mechanism in conformationally asymmetric trimers of the betaine transporter BetP.. Nature, 490(7418), 126–130.
30. Qiu, B., Xia, B., Zhou, Q., Lu, Y., He, M., Hasegawa, K., Ma, Z., Zhang, F., Gu, L., Mao, Q., Wang, F., Zhao, S., Gao, Z., & Liao, J. (2018). Succinate-acetate permease from Citrobacter koseri is an anion channel that unidirectionally translocates acetate. Cell Research, 28(6), 644–654.
31. Sayers, E., Cavanaugh, M., Clark, K., Ostell, J., Pruitt, K., & Karsch-Mizrachi, I. (2019). GenBank.. Nucleic Acids Res, 47(D1), D94-D99.
32. Schoch, C., Ciufo, S., Domrachev, M., Hotton, C., Kannan, S., Khovanskaya, R., Leipe, D., Mcveigh, R., O'Neill, K., Robbertse, B., Sharma, S., Soussov, V., Sullivan, J., Sun, L., Turner, S., & Karsch-Mizrachi, I. (2020). NCBI Taxonomy: a comprehensive update on curation, resources and tools.. Database (Oxford).

33. Tang, L., Bai, L., Wang, W.h., & Jiang, T. (2010). Crystal structure of the carnitine transporter and insights into the antiport mechanism. Nature Structural & Molecular Biology, 17(4), 492–496.
34. Ulmschneider, M., & Sansom, M. (2001). Amino acid distributions in integral membrane protein structures. Biochimica et Biophysica Acta (BBA) - Biomembranes, 1512(1), 1-14.
35. UniProt, Consortium. (2008). The universal protein resource (UniProt).. Nucleic Acids Res, 36(Database issue), D190-5.
36. Verkman, A. (2012). Aquaporins in clinical medicine.. Annu Rev Med, 63, 303–316.
37. Verkman, A. (2013). Aquaporins. Curr Biol, 23(2), R52-5.
38. Wang, Y., Huang, Y., Wang, J., Cheng, C., Huang, W., Lu, P., Xu, Y.N., Wang, P., Yan, N., & Shi, Y. (2009). Structure of the formate transporter FocA reveals a pentameric aquaporin-like channel.. Nature, 462(7272), 467–472.
39. Weeks, D., & Sachs, G. (2001). Sites of pH regulation of the urea channel of Helicobacter pylori. Molecular microbiology, 40(6), 1249–1259.
40. Yang, H., Hu, M., Guo, J., Ou, X., Cai, T., & Liu, Z. (2016). Pore architecture of TRIC channels and insights into their gating mechanism.. Nature, 538(7626), 537–541.
41. Zhou, X., Lin, P., Yamazaki, D., Park, K., Komazaki, S., Chen, S., Takeshima, H., & Ma, J. (2014). Trimeric intracellular cation channels and sarcoplasmic/endoplasmic reticulum calcium homeostasis.. Circ Res, 114(4), 706–716.

7. List of tables

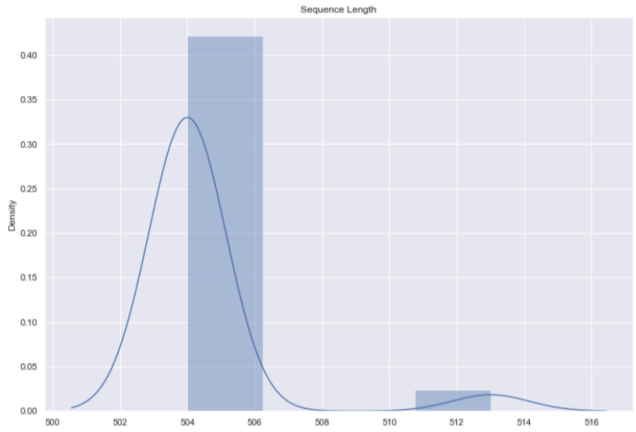
8. List of figures

9. Appendixes

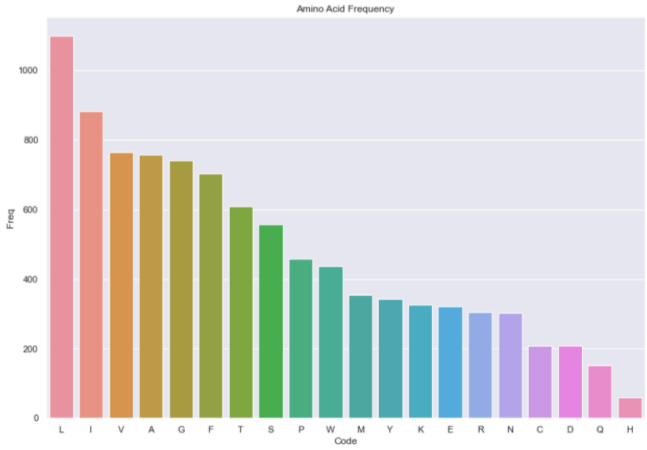
A. Analysis of sequence contents for each query

A.1. Query 3HFX

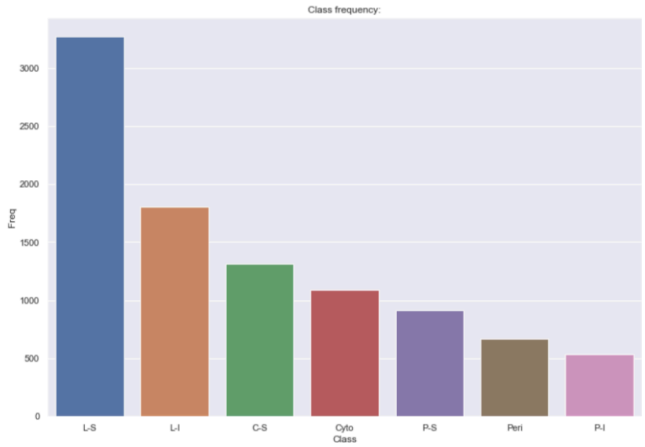
A.1.1.Query 3HFX - plot of sequence length over density



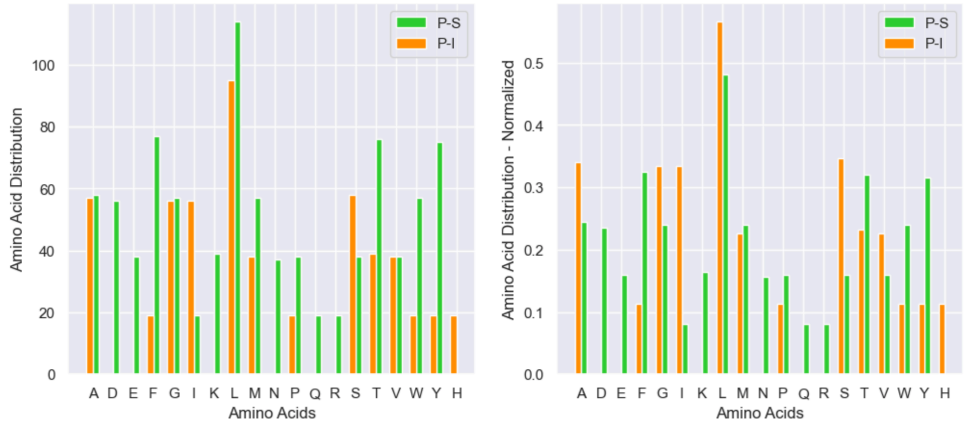
A.1.2. Query 3HFX - plot of frequency of occurrence of amino acid



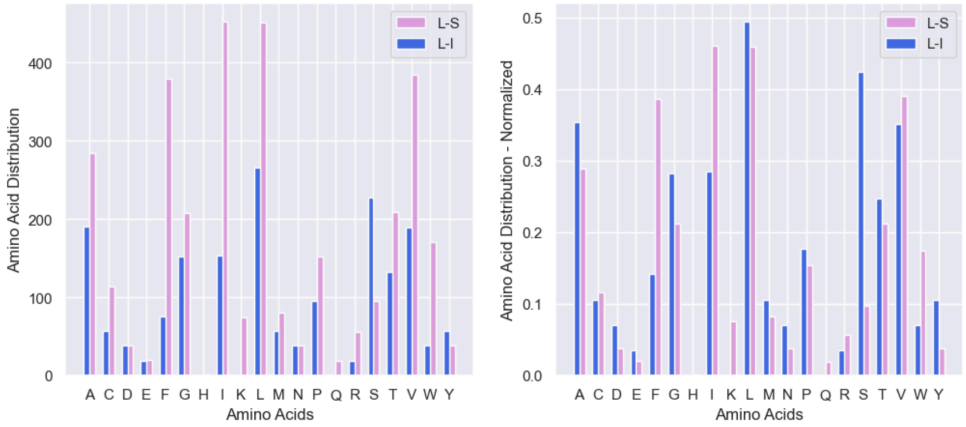
A.1.3. Query 3HFX - plot of frequency of occurrence of annotation classes



A.1.4. Query 3HFX- Amino acid distribution in the P-S and P-I region

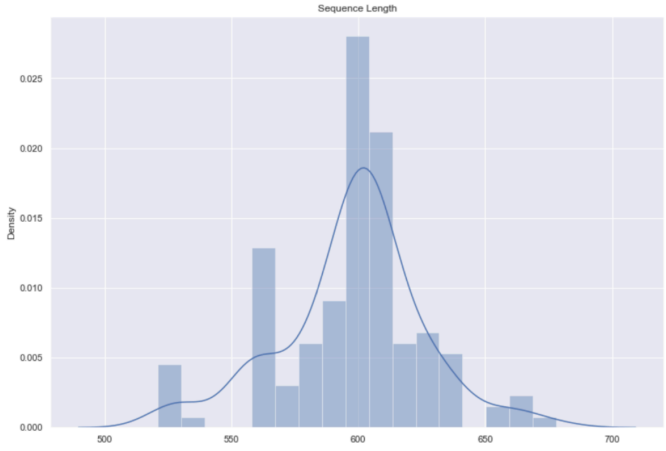


A.1.5. Query 3HFX- Amino acid distribution in the L-S and L-I region

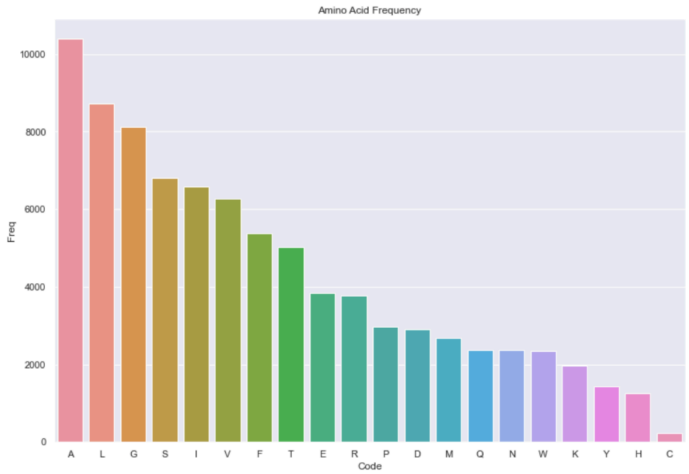


A.2. Query : 4AIN

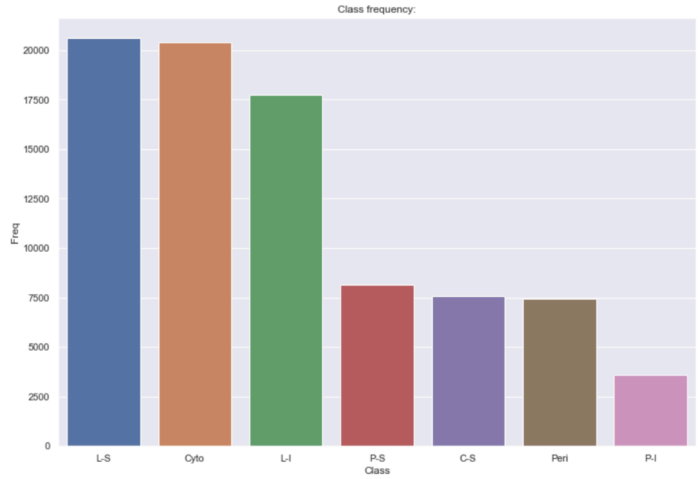
A.2.1.Query 4AIN - plot of sequence length over density



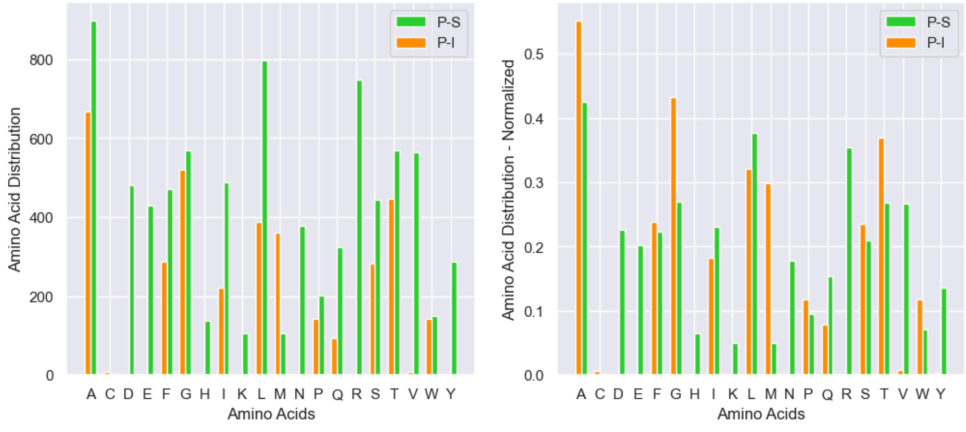
A.2.2. Query 4AIN - plot of frequency of occurrence of amino acid



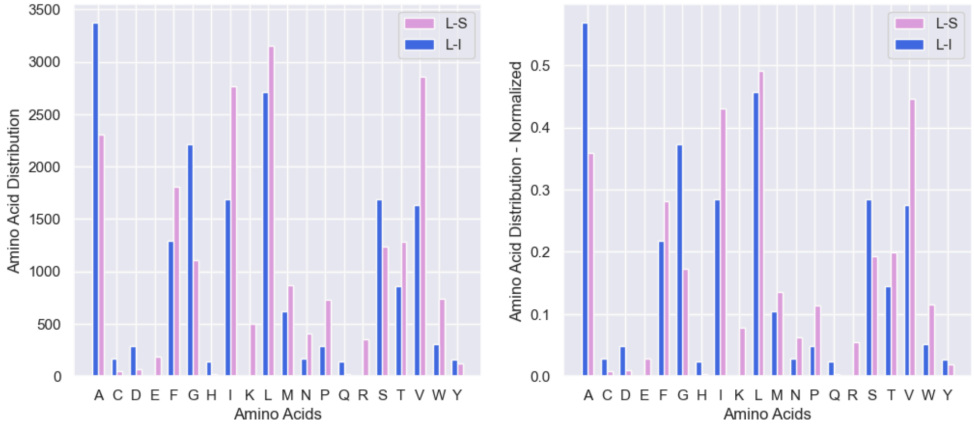
A.2.3. Query 4AIN - plot of frequency of occurrence of annotation classes



A.2.4. Query 4AIN- Amino acid distribution in the P-S and P-I region

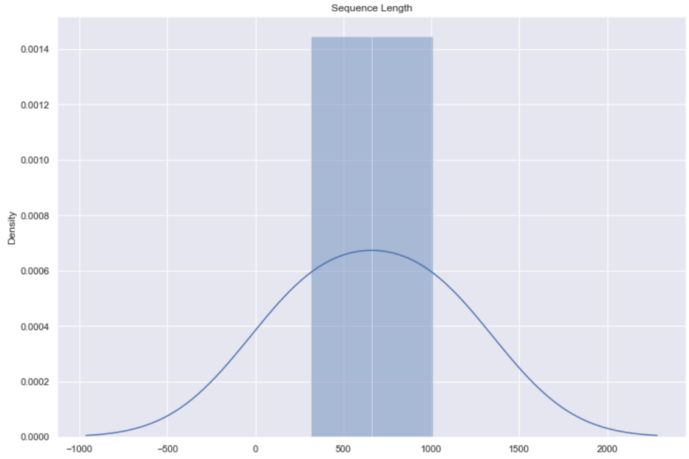


A.2.5. Query 4AIN- Amino acid distribution in the L-S and L-I region

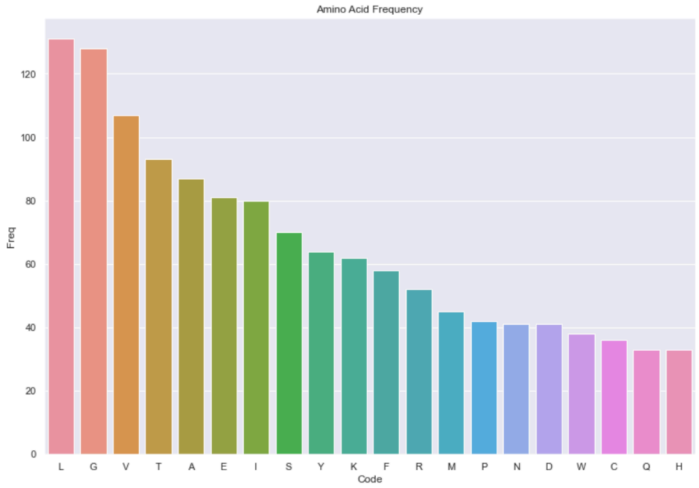


A.3.Query : 6EID

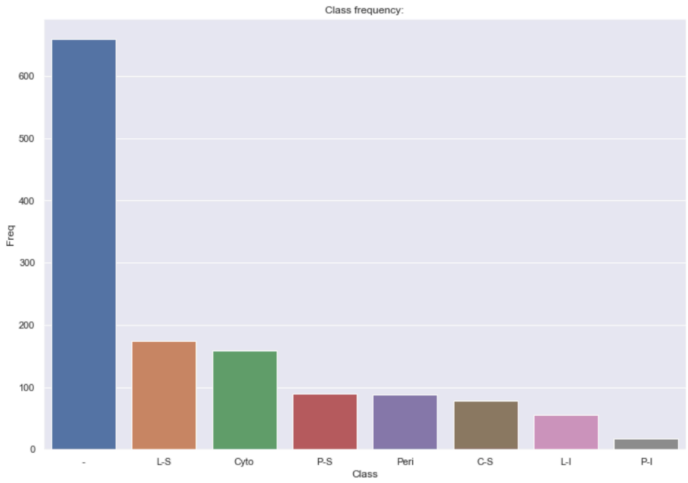
A.3.1.Query 6EID - plot of sequence length over density



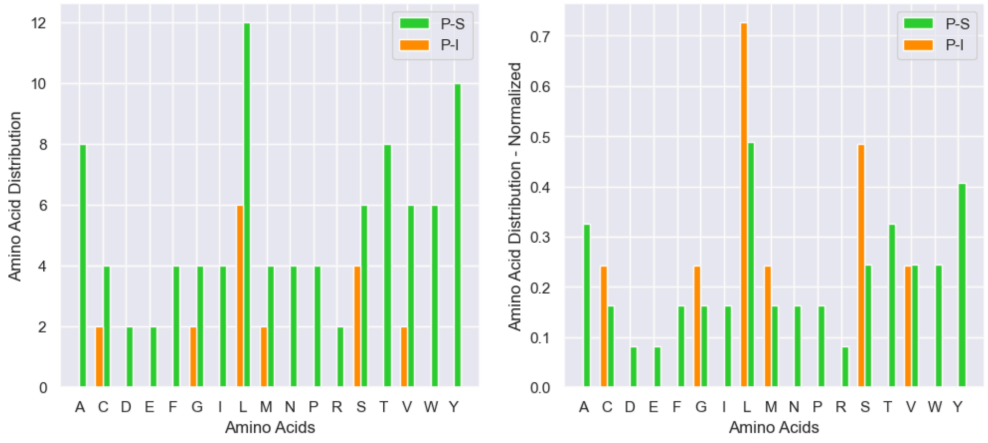
A.3.2. Query 6EID - plot of frequency of occurrence of amino acid



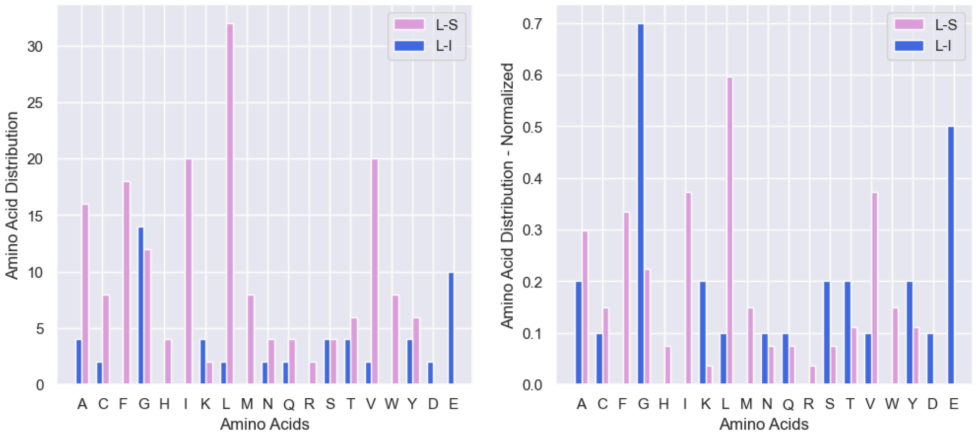
A.3.3. Query 6EID - plot of frequency of occurrence of annotation classes



A.3.4. Query 6EID - Amino acid distribution in the P-S and P-I region

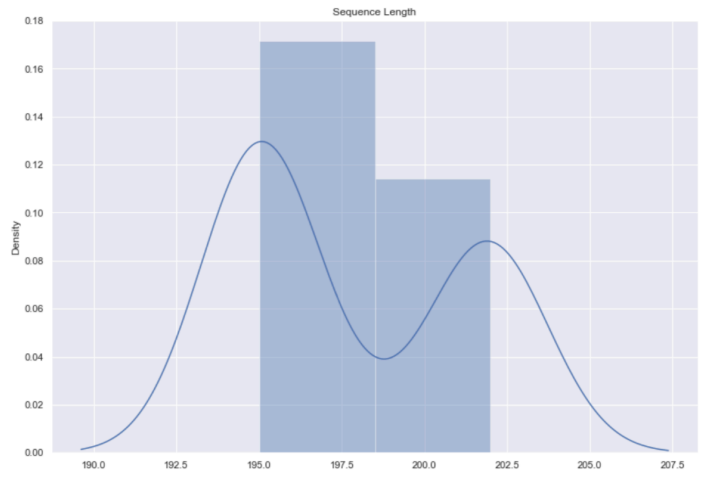


A.3.5. Query 6EID - Amino acid distribution in the L-S and L-I region

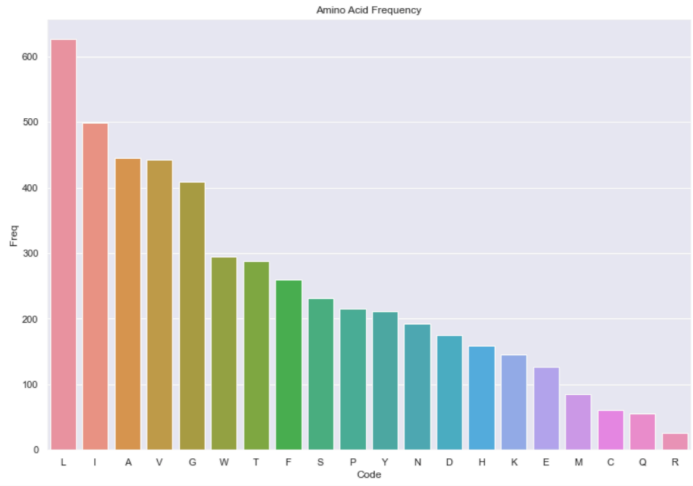


A.4.Query : 6NSK

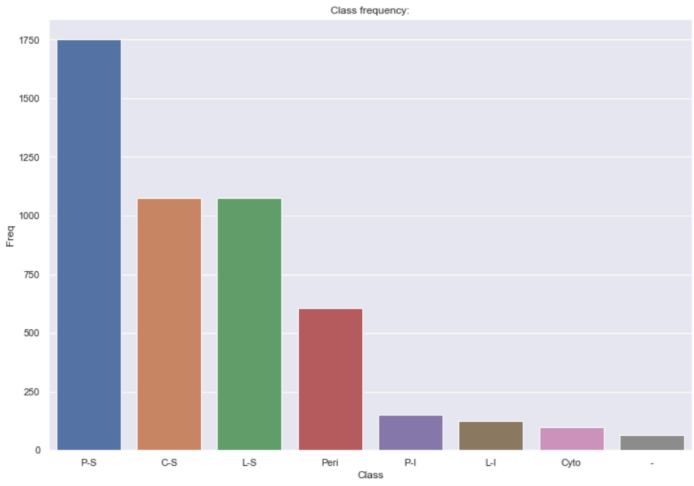
A.4.1.Query 6NSK - plot of sequence length over density



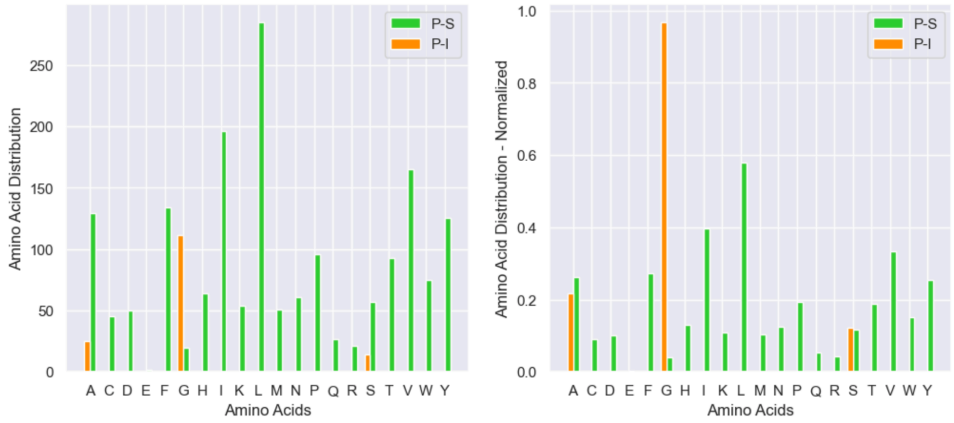
A.4.2. Query 6NSK - plot of frequency of occurrence of amino acid



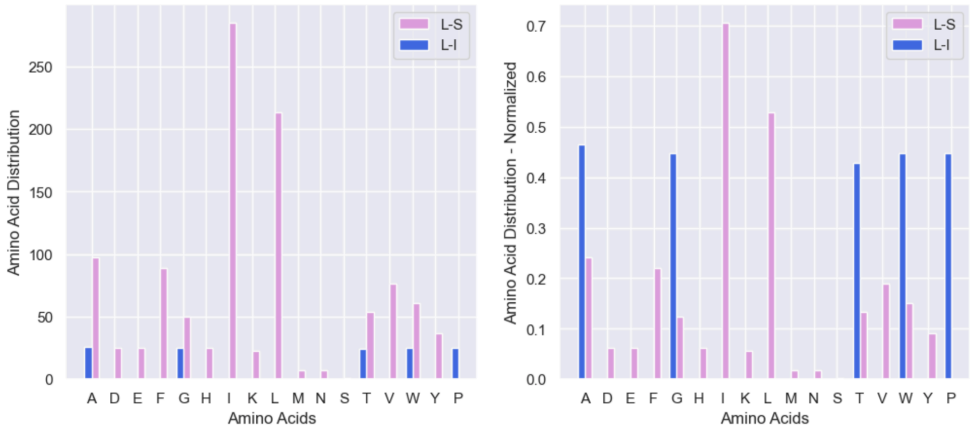
A.4.3. Query 6NSK - plot of frequency of occurrence of annotation classes



A.4.4. Query 6NSK- Amino acid distribution in the P-S and P-I region

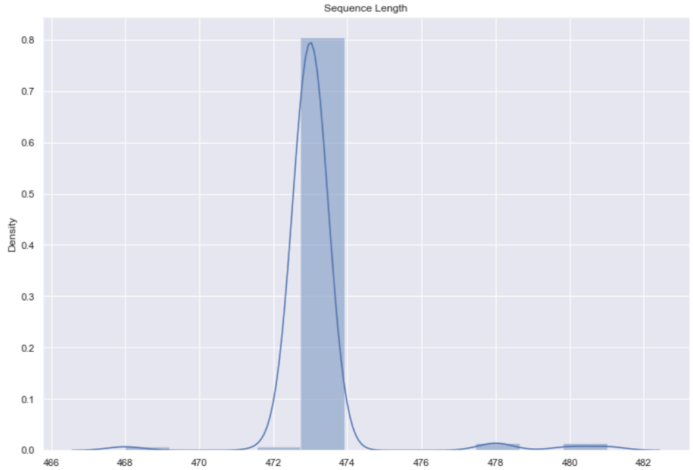


A.4.5. Query 6NSK- Amino acid distribution in the L-S and L-I region

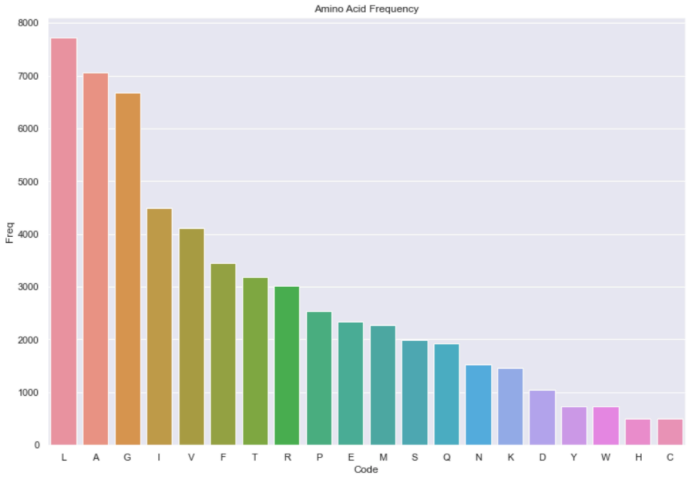


A.5. Query : 1KPL

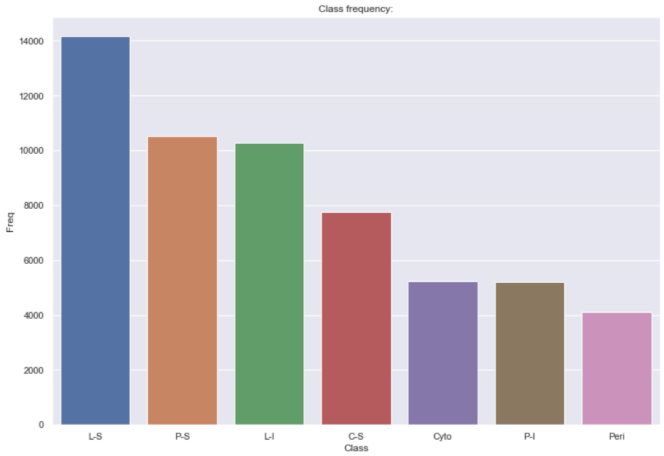
A.5.1.Query 1KPL - plot of sequence length over density



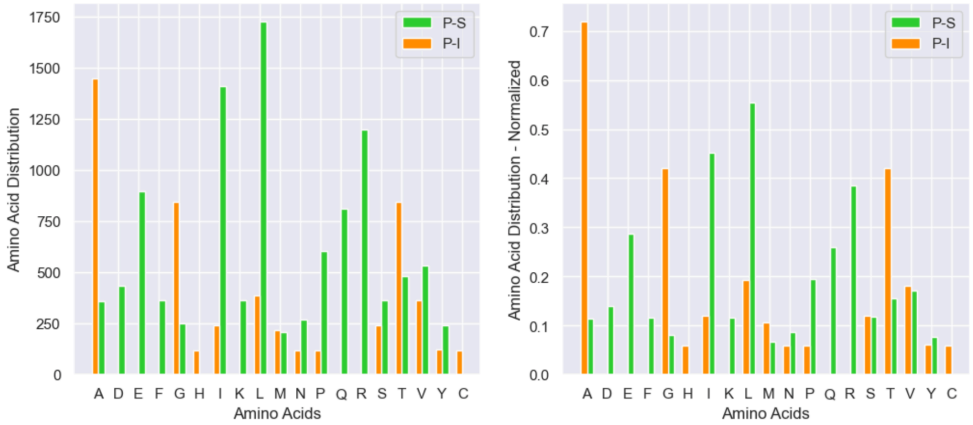
A.5.2. Query 1KPL - plot of frequency of occurrence of amino acid



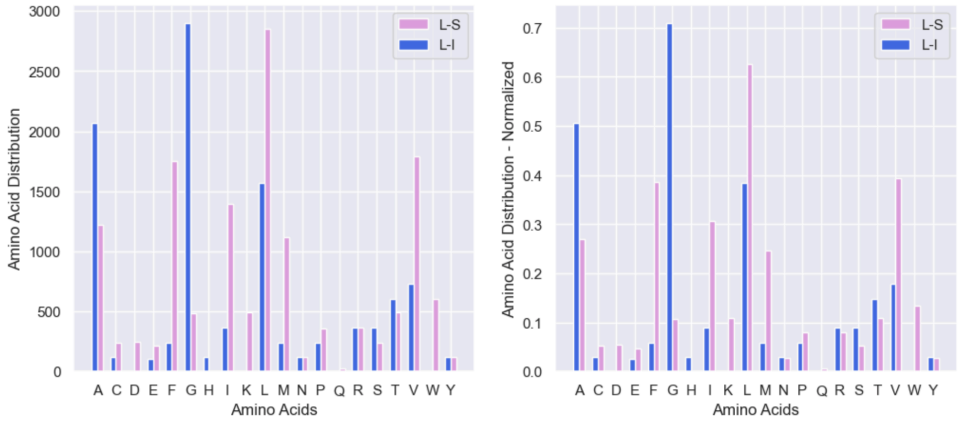
A.5.3. Query 1KPL - plot of frequency of occurrence of annotation classes



A.5.4. Query 1KPL - Amino acid distribution in the P-S and P-I region

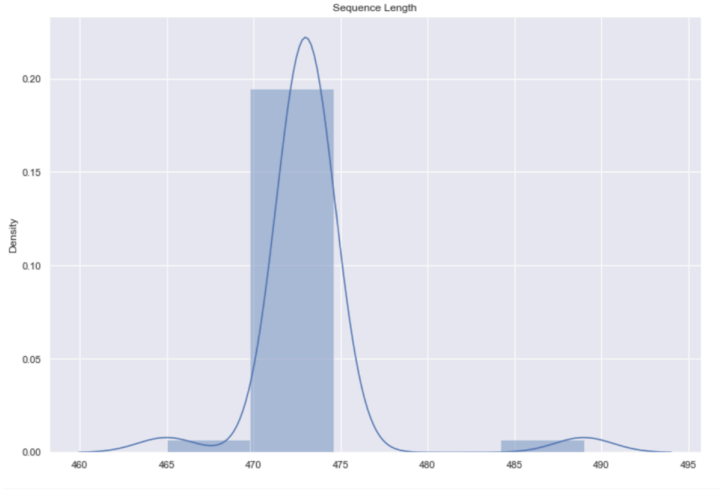


A.5.5. Query 1KPL - Amino acid distribution in the L-S and L-I region

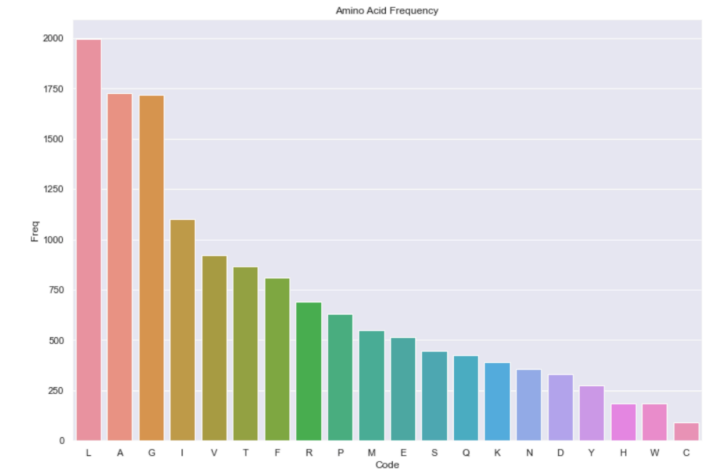


A.6.Query : 1KPK

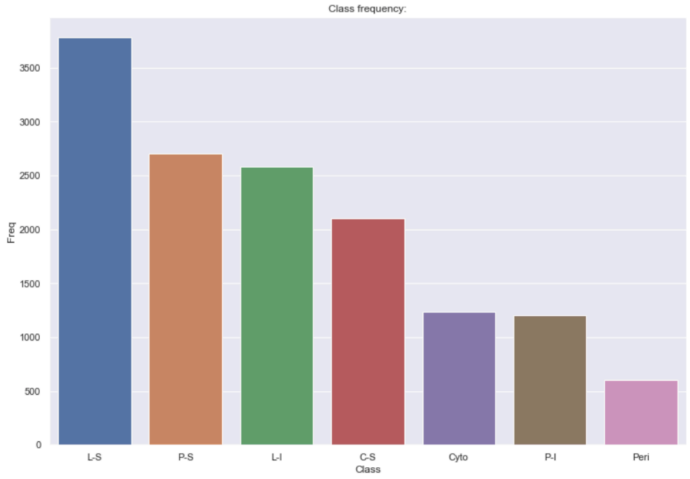
A.6.1.Query 1KPK - plot of sequence length over density



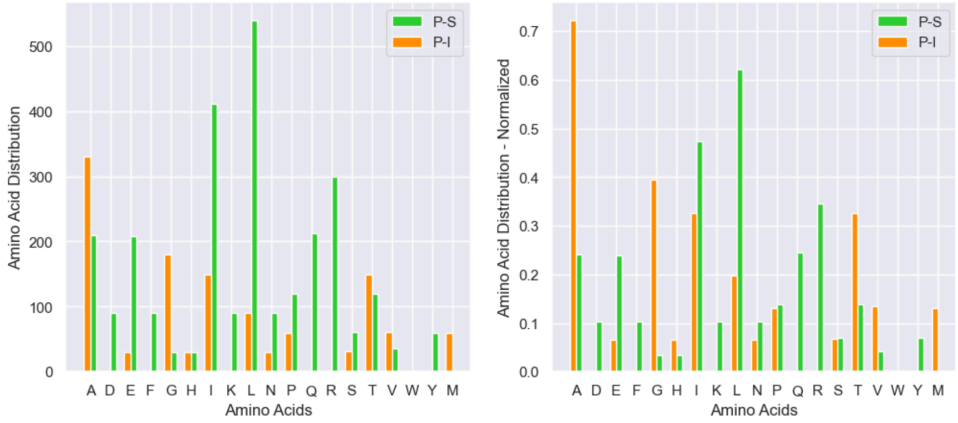
A.6.2. Query 1KPK - plot of frequency of occurrence of amino acid



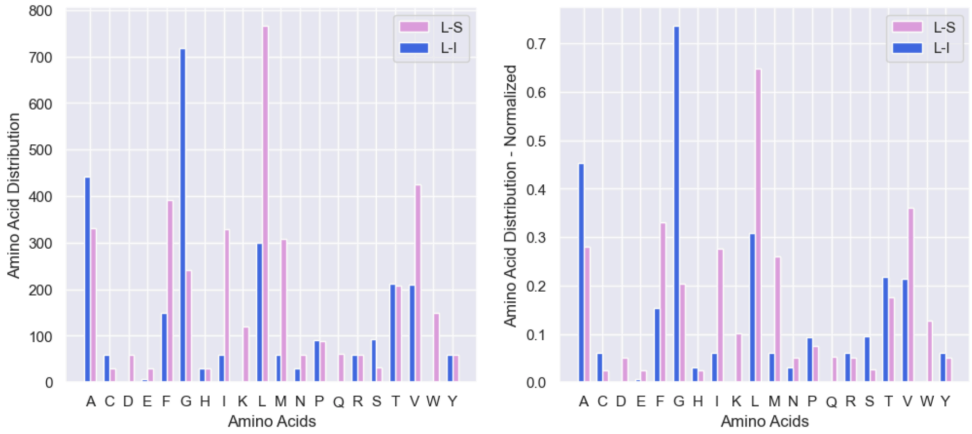
A.6.3. Query 1KPK - plot of frequency of occurrence of annotation classes



A.6.4. Query 1KPK- Amino acid distribution in the P-S and P-I region

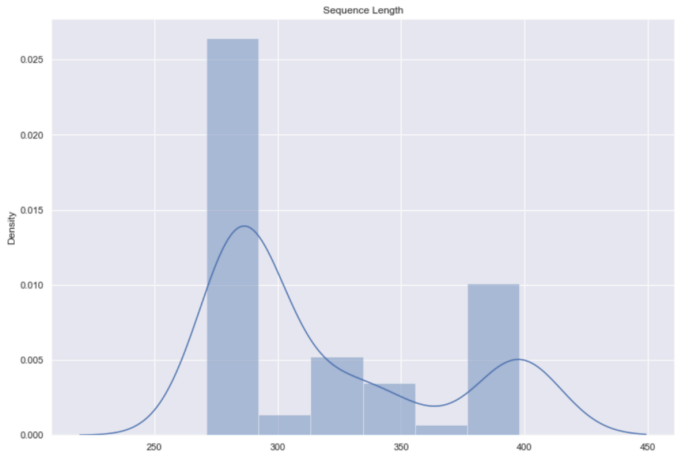


A.6.5. Query 1KPK- Amino acid distribution in the L-S and L-I region

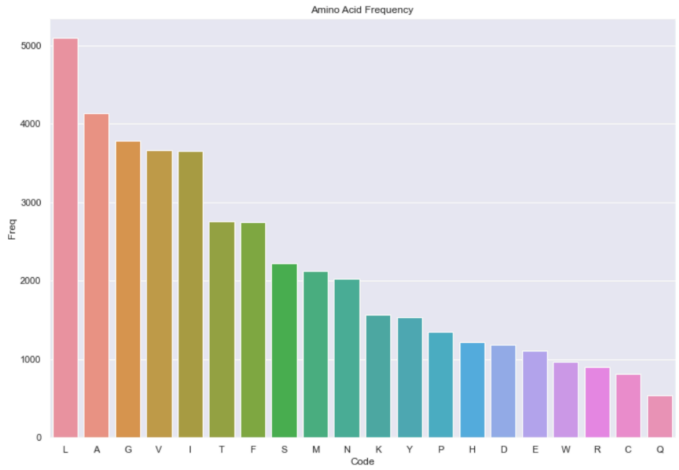


A.7.Query : 3KCU

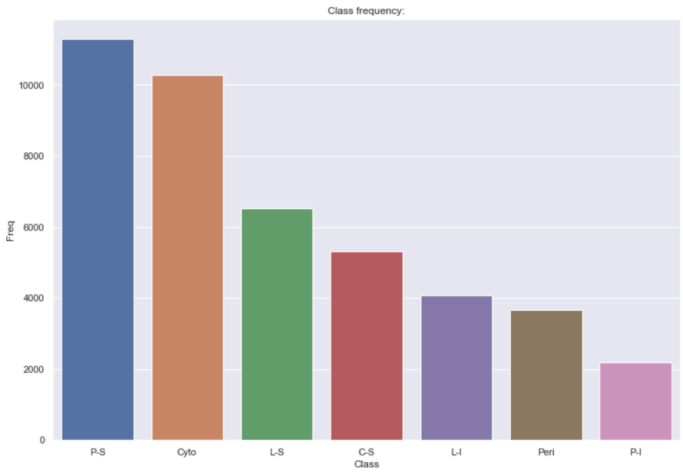
A.7.1.Query 3KCU - plot of sequence length over density



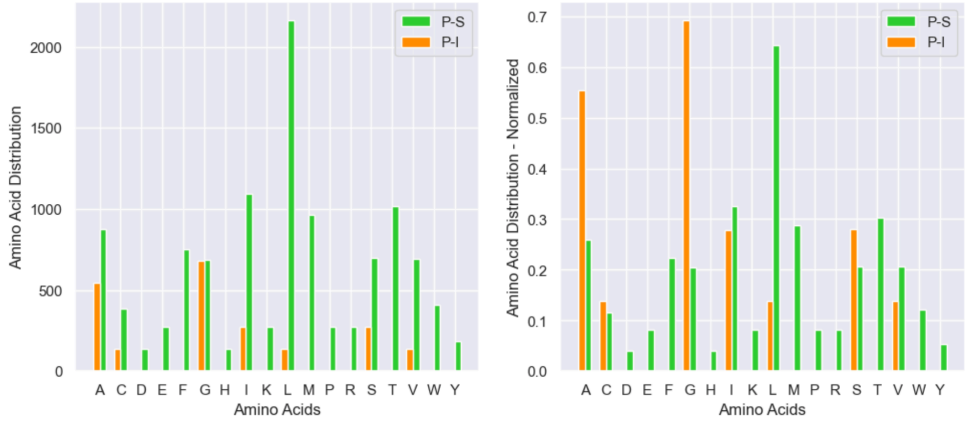
A.7.2. Query 3KCU - plot of frequency of occurrence of amino acid



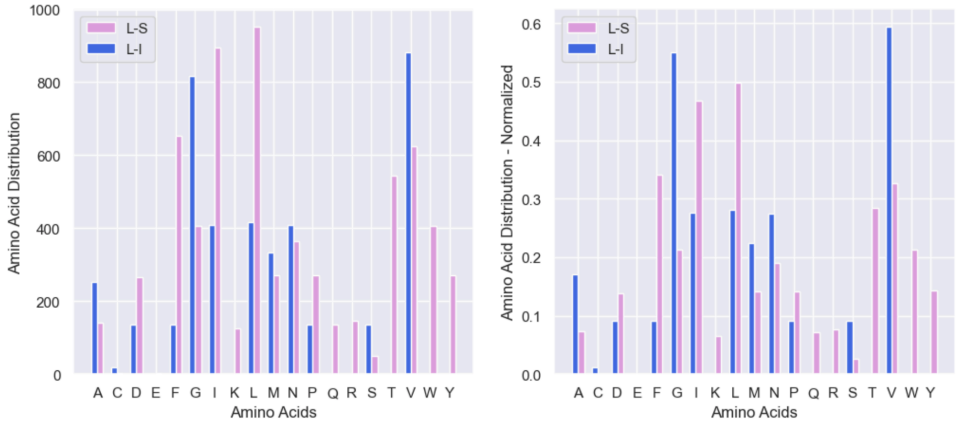
A.7.3. Query 3KCU - plot of frequency of occurrence of annotation classes



A.7.4. Query 3KCU - Amino acid distribution in the P-S and P-I region

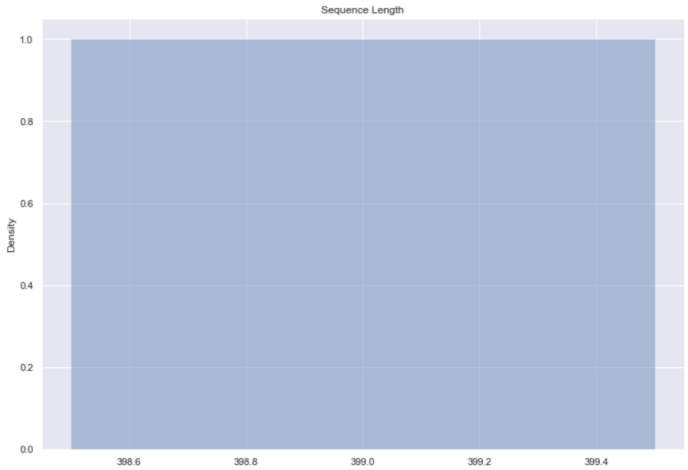


A.7.5. Query 3KCU - Amino acid distribution in the L-S and L-I region

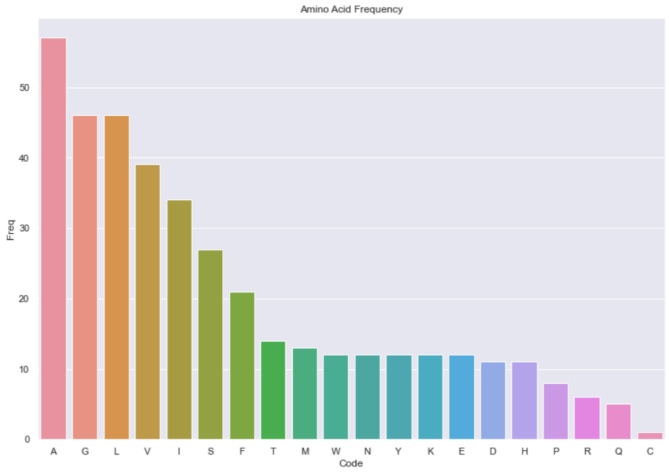


A.8.Query : 2B2F

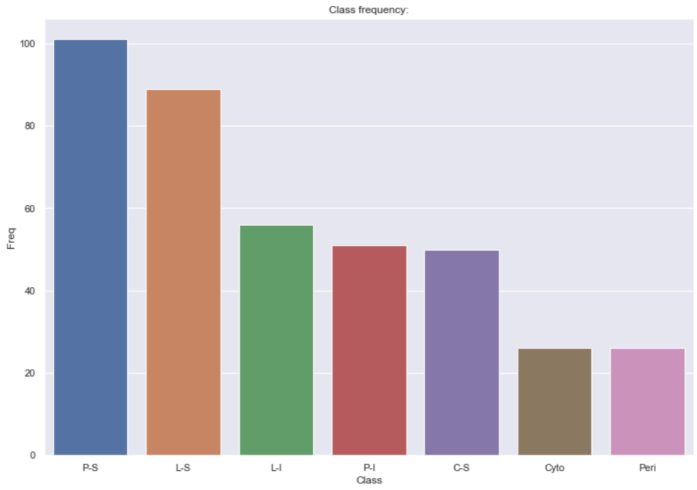
A.8.1.Query 2B2F - plot of sequence length over density



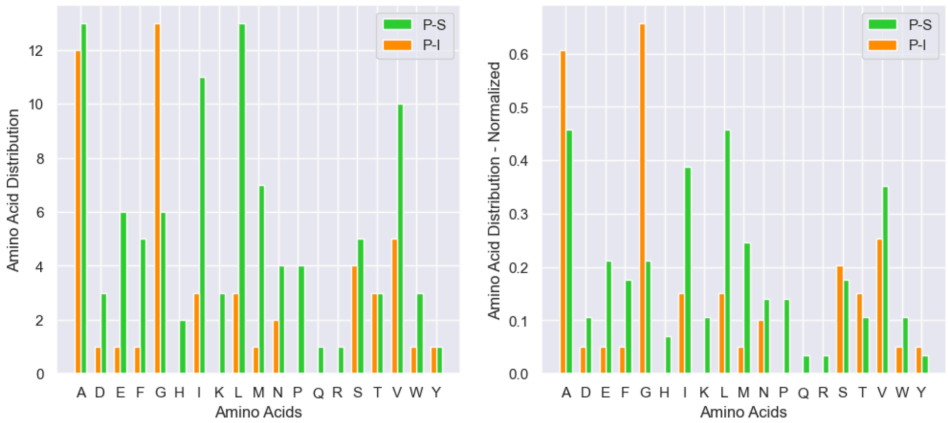
A.8.2. Query 2B2F - plot of frequency of occurrence of amino acid



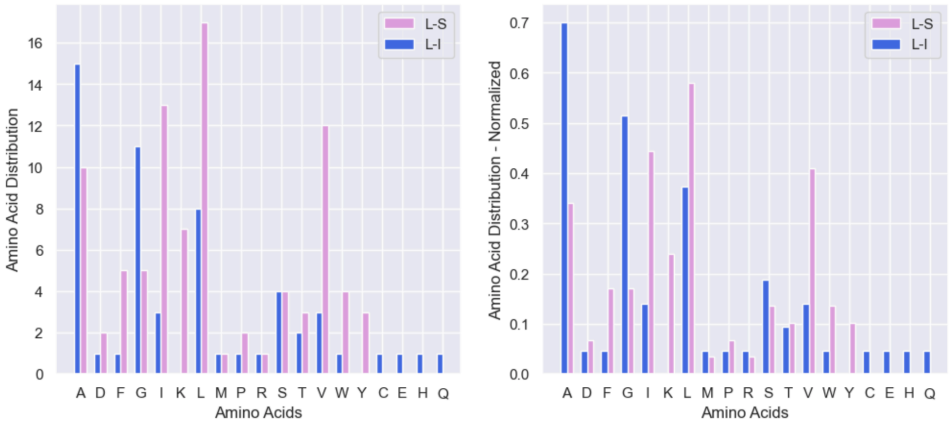
A.8.3. Query 2B2F - plot of frequency of occurrence of annotation classes



A.8.4.Query 2B2F - Amino acid distribution in the P-S and P-I region

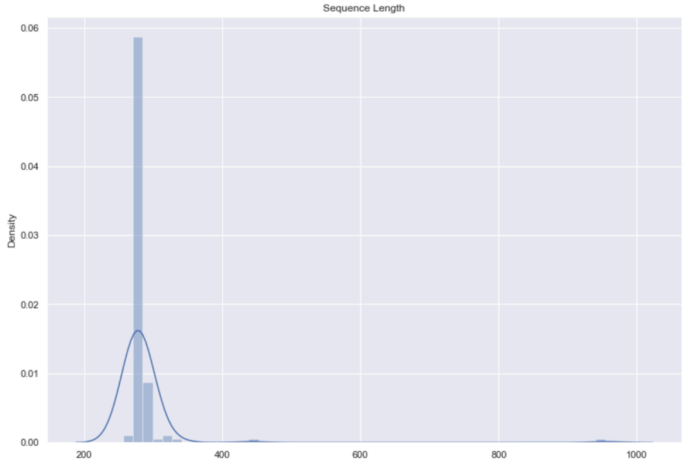


A.8.5.Query 2B2F - Amino acid distribution in the L-S and L-I region

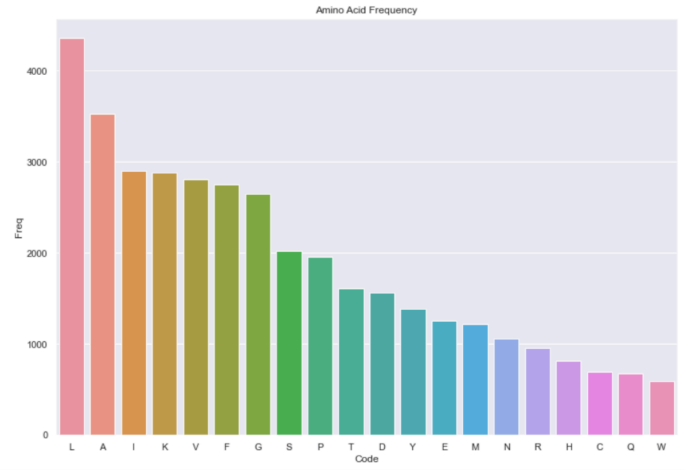


A.9.Query : 5EGI

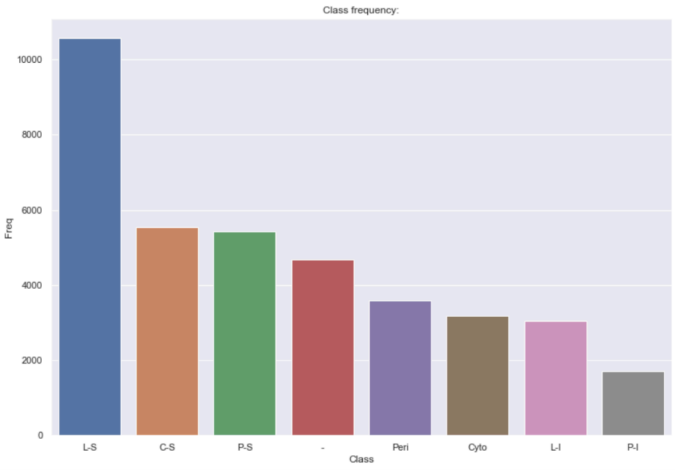
A.9.1.Query 5EGI - plot of sequence length over density



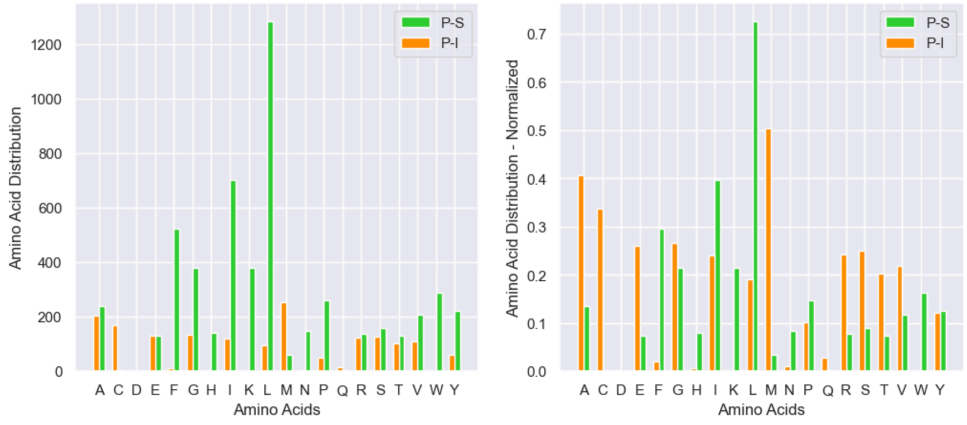
A.9.2. Query 5EGI - plot of frequency of occurrence of amino acid



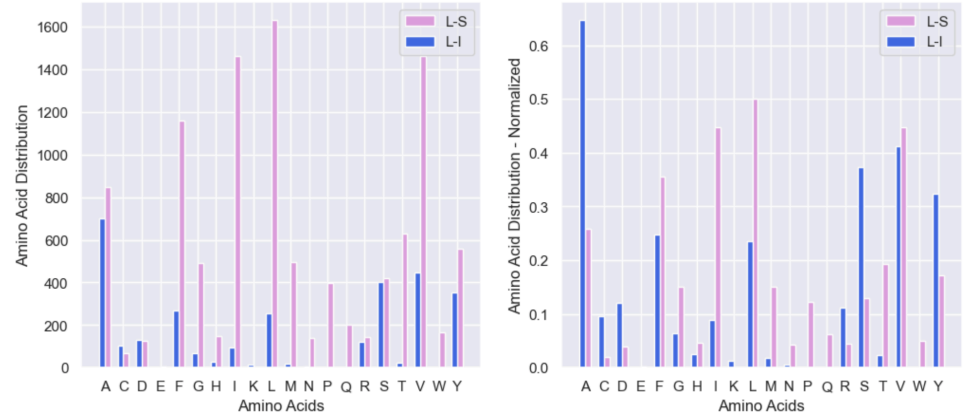
A.9.3. Query 5EGI - plot of frequency of occurrence of annotation classes



A.9.4. Query 5EGI - Amino acid distribution in the P-S and P-I region

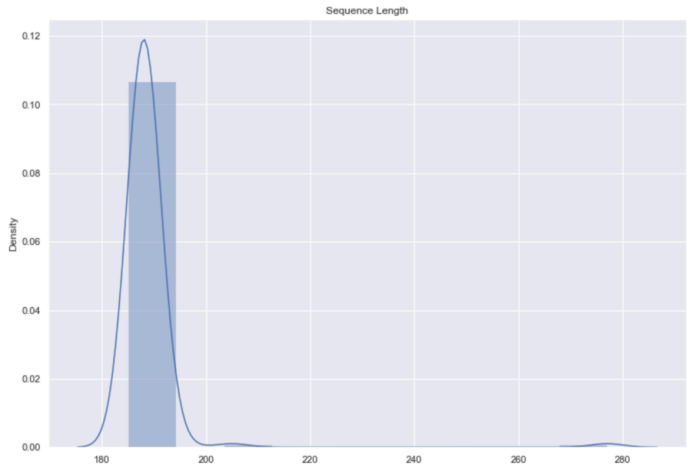


A.9.5. Query 5EGI - Amino acid distribution in the L-S and L-I region

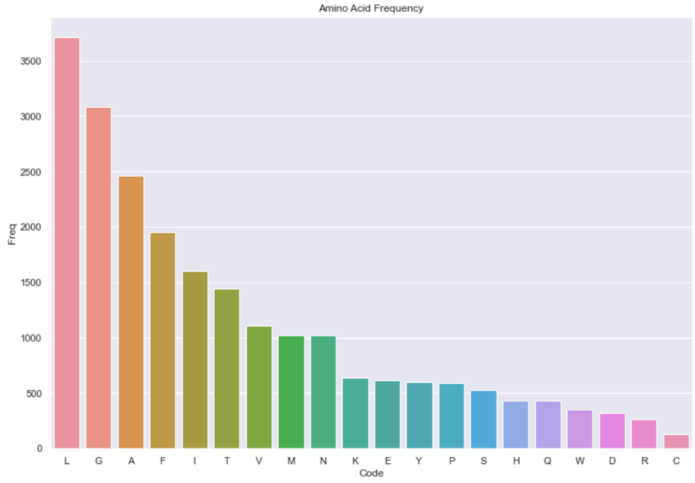


A.10.Query : 5YS3

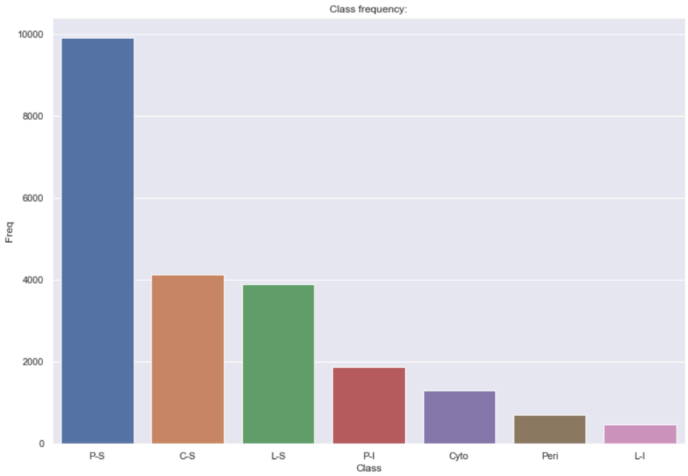
A.10.1.Query 5YS3 - plot of sequence length over density



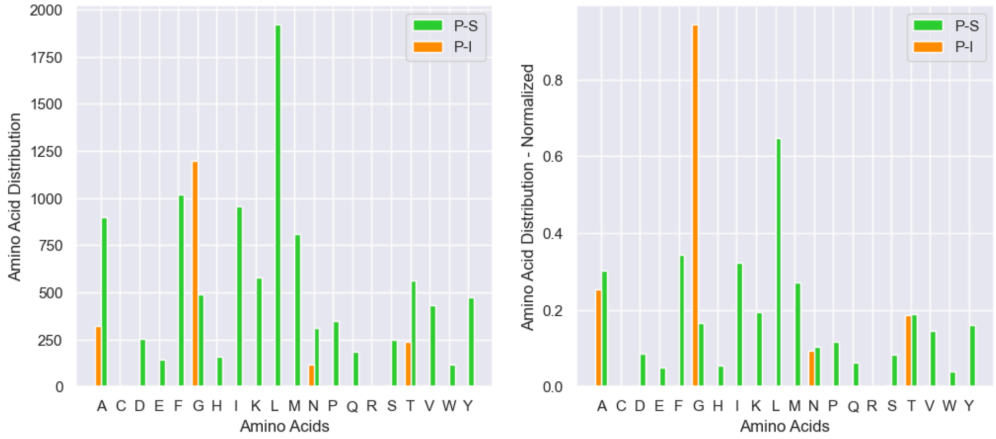
A.10.2. Query 5YS3 - plot of frequency of occurrence of amino acid



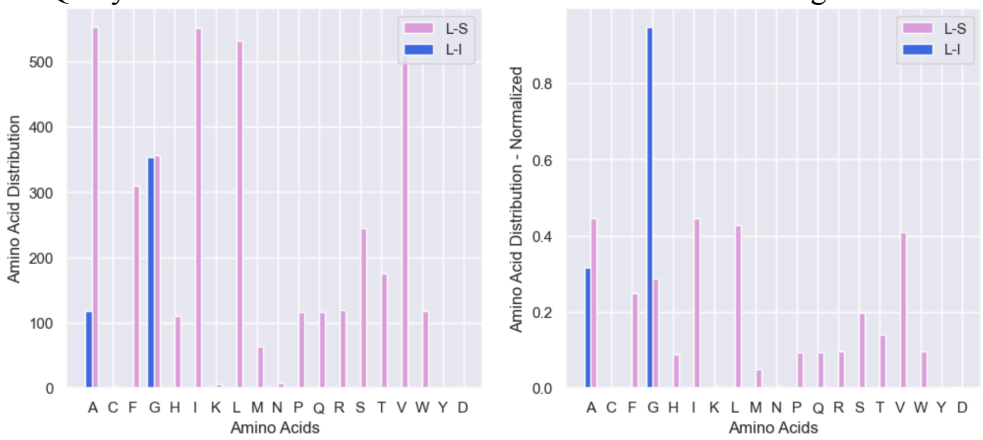
A.10.3. Query 5YS3 - plot of frequency of occurrence of annotation classes



A.10.4. Query 5YS3 - Amino acid distribution in the P-S and P-I region

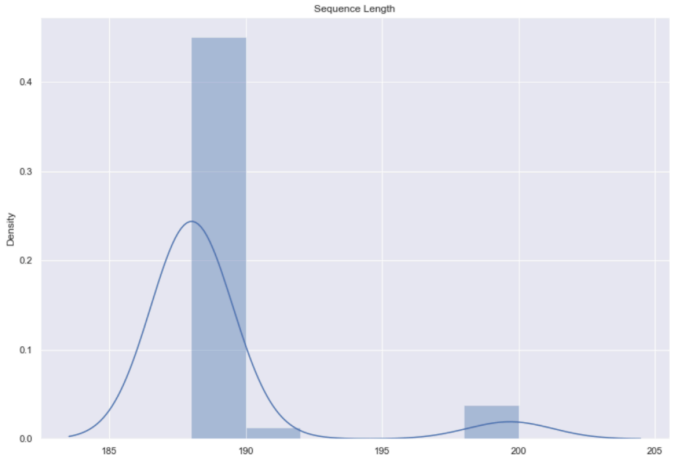


A.10.5. Query 5YS3 - Amino acid distribution in the L-S and L-I region

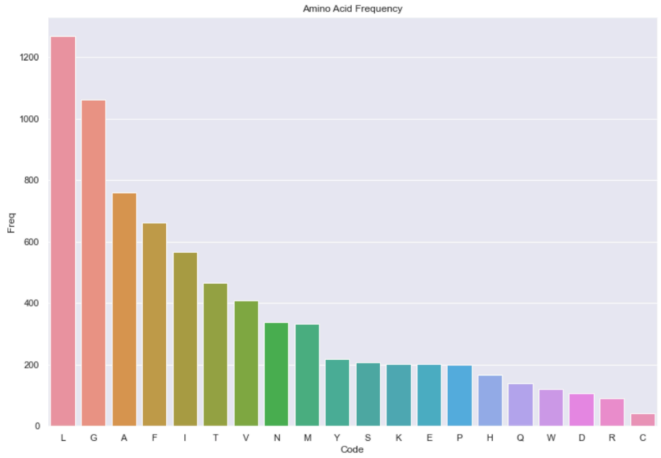


A.11.Query : 5ZUG

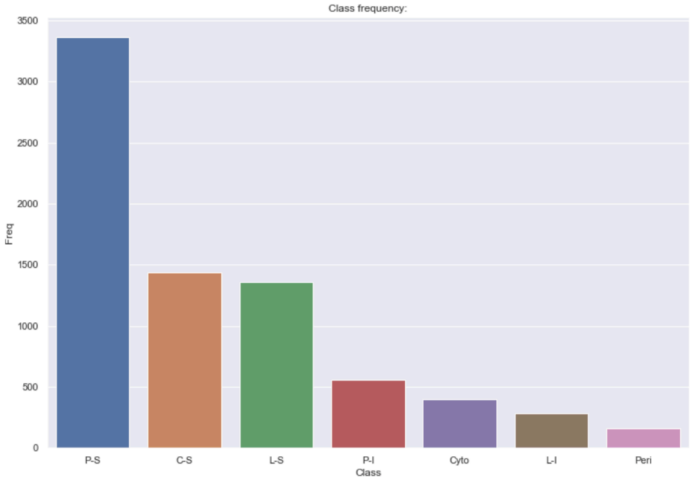
A.11.1.Query 5ZUG - plot of sequence length over density



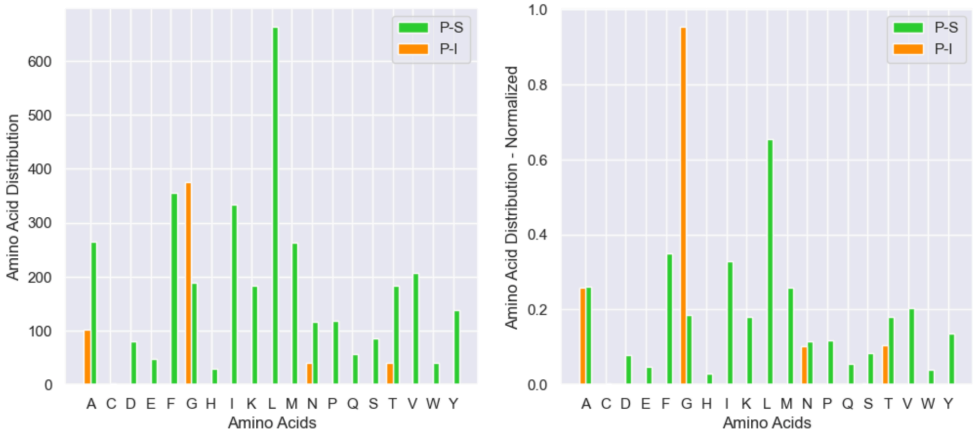
A.11.2. Query 5ZUG - plot of frequency of occurrence of amino acid



A.11.3. Query 5ZUG - plot of frequency of occurrence of annotation classes



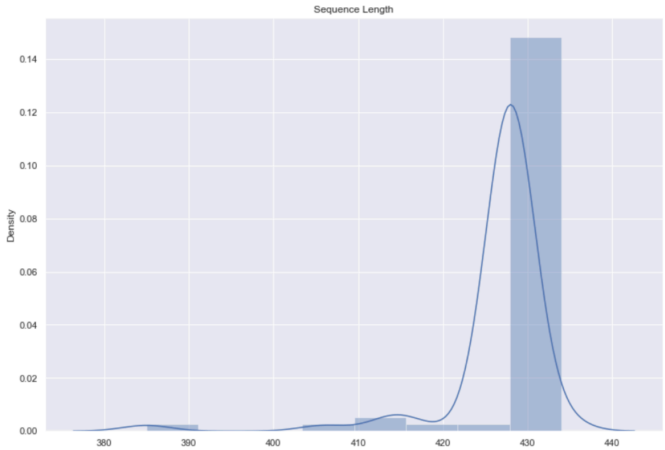
A.11.4. Query 5ZUG- Amino acid distribution in the P-S and P-I region



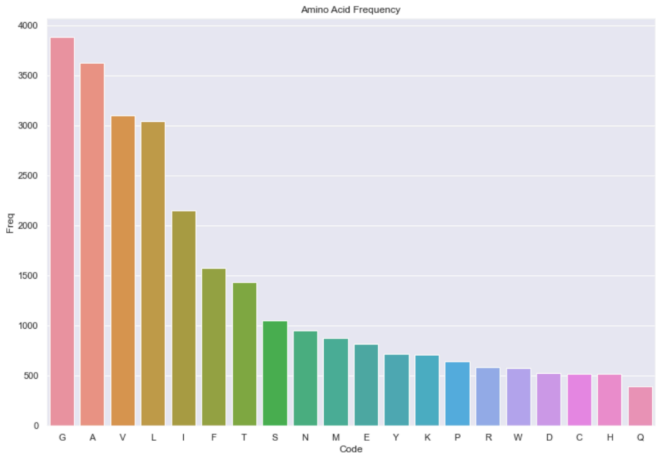
A.11.5. Query 5ZUG- Amino acid distribution in the L-S and L-I region

A.12.Query : 1U7C

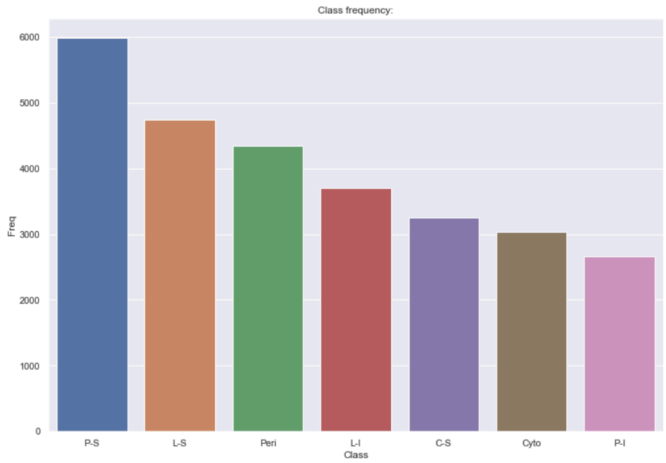
A.12.1.Query 1U7C - plot of sequence length over density



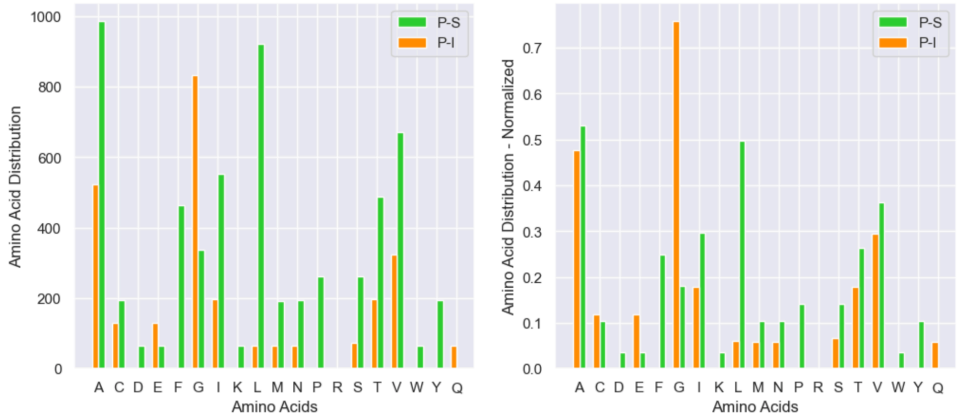
A.12.2. Query 1U7C - plot of frequency of occurrence of amino acid



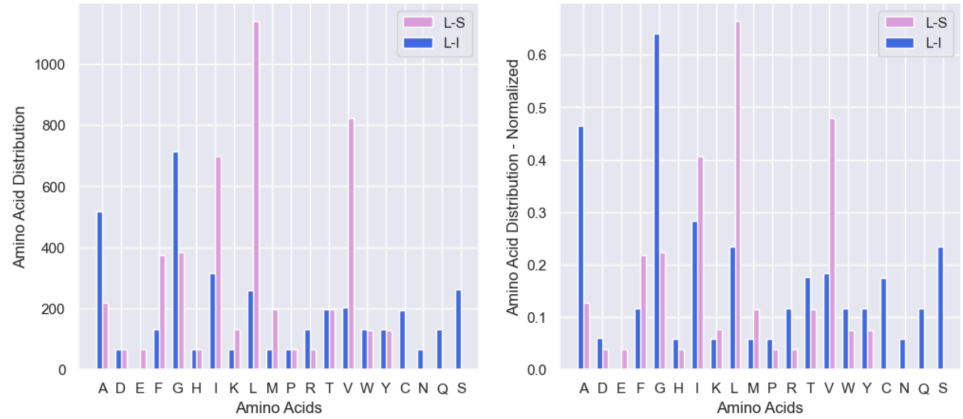
A.12.3. Query 1U7C - plot of frequency of occurrence of annotation classes



A.12.4. Query 1U7C- Amino acid distribution in the P-S and P-I region

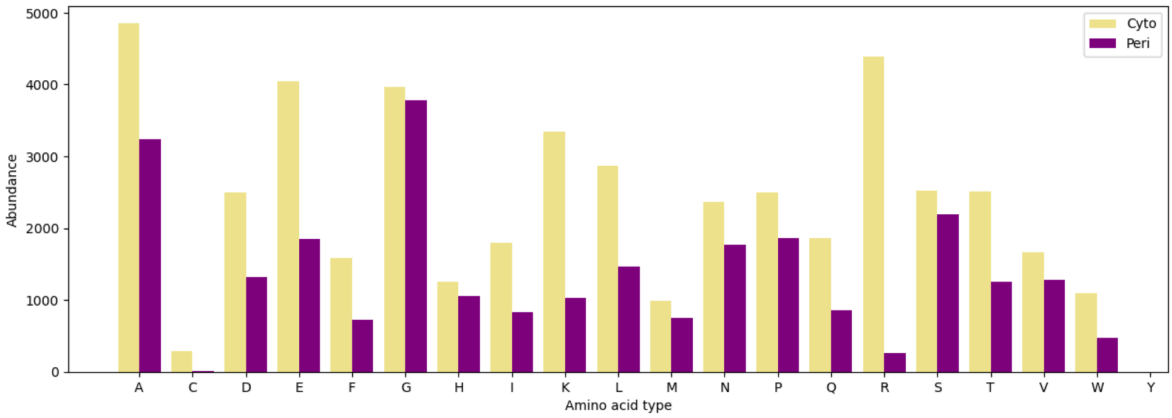


A.12.5. Query 1U7C- Amino acid distribution in the L-S and L-I region

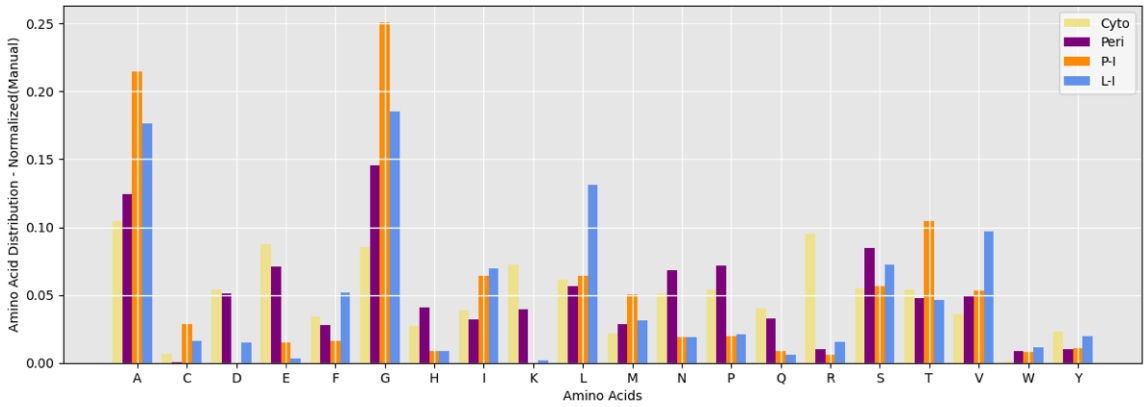


B. Amino acid distribution in specific regions

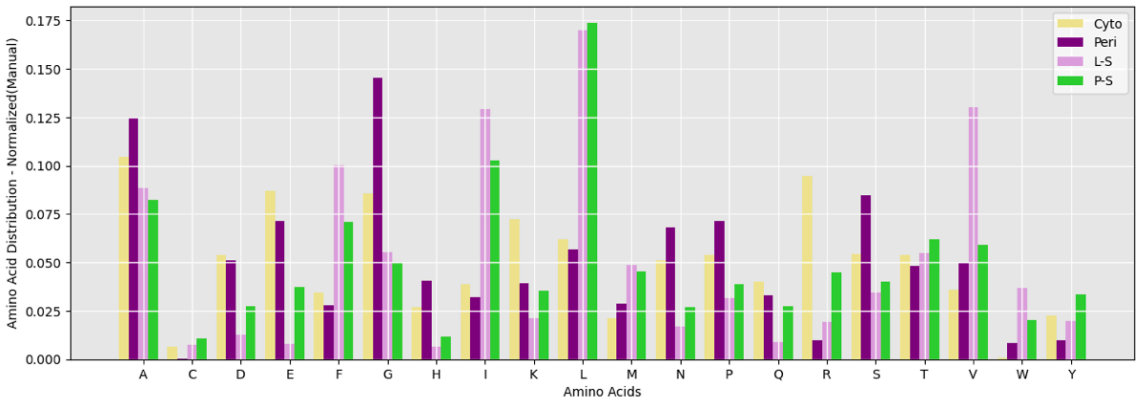
B.1. Comparison of amino acid distributions in Peri and Cyto region



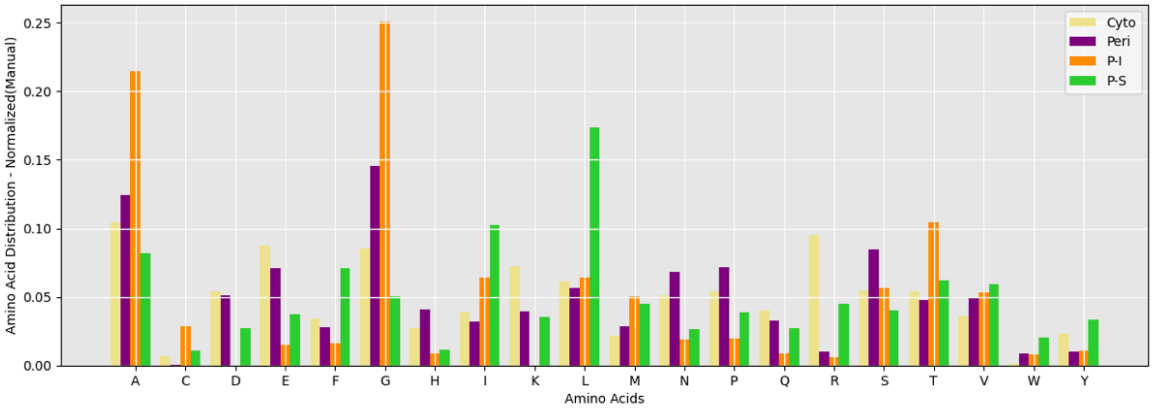
B.2. Comparison of amino acid distributions of Cyto, Peri, P-I and L-I region



B.3. Comparison of amino acid distributions of Cyto, Peri, L-S and P-S region



B.4. Comparison of amino acid distributions of Cyto, Peri, P-I and P-S region



B.5. Comparison of amino acid distributions of Cyto, Peri, L-I and L-S region