University of South Bohemia Faculty of Science

České Budějovice, Czech Republic

and

Johannes Kepler University Linz, Austria

Faculty of Engineering and Natural Sciences

Sequence features indicate a positive net charge is influential for the sub-cellular localization of proteins in organisms with complex plastids

Bioinformatics Thesis

Fuad Sarker Jeem

Supervisor: Dr. Ansgar Gruber

Institute of Parasitology Biology Centre, ASCR · Laboratory of Evolutionary Protistology

České Budějovice

2022

Fuad Sarker Jeem, (2022): Sequence features indicate a positive net charge is influential for the sub-cellular localization of proteins in organisms with complex plastids. Bc. Thesis, in English. – 63 p., Faculty of Science, University of South Bohemia, České Budějovice, Czech Republic and Faculty of Engineering and Natural Sciences, Johannes Kepler University, Linz, Austria.

Annotation

Intracellular targeting of proteins in cells with complex plastids of red algal origin; the targeting signals of plastid proteins in these organisms are different from those in plants and other groups of algae. These signals should be characterized in comparison to proteins that are not targeted to the plastids.

Basic recommended literature:

- Emanuelsson, O., Brunak, S., von Heijne, G., & Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP and related tools. *Nature Protocols*, *2*(4), 953–971. https://doi.org/10.1038/nprot.2007.131
- Felsner, G., Sommer, M. S., & Maier, U. G. (2010). The physical and functional borders of transit peptide-like sequences in secondary endosymbionts. *BMC Plant Biology*, *10*(1), 223. https://doi.org/10.1186/1471-2229-10-223
- Füssy, Z., Faitová, T., & Oborník, M. (2019). Subcellular Compartments Interplay for Carbon and Nitrogen Allocation in Chromera velia and Vitrella brassicaformis. *Genome Biology and Evolution*, *11*(7), 1765–1779. https://doi.org/10.1093/gbe/evz123
- Gruber, A., & Haferkamp, I. (2019). Nucleotide Transport and Metabolism in Diatoms. *Biomolecules*, *9*(12), 761. https://doi.org/10.3390/biom9120761
- Gruber, A., & Kroth, P. G. (2017). Intracellular metabolic pathway distribution in diatoms and tools for genome-enabled experimental diatom research. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1728), 20160402. https://doi.org/10.1098/rstb.2016.0402

I declare that I am the author of this qualification thesis and that in writing it I have used the sources and literature listed in references.

Linz, 07.11.22
Fuad sarker Jeem

Table of Contents

1. Abstract

Proteins are essential elements of each living organism. Research, medicine development, disease diagnostics, and other endeavors require a deeper understanding of protein function. Even though their evolutionary histories are diverse, diatoms have an important role to play in global carbon fixation as photosynthetic heterokonts. Unlike higher plants, their plastids are equipped with four membranes and the complexity of the protein transport system is higher than usual in primary plastids. This study represents a bioinformatic approach by analyzing protein features of three different protein sequence datasets. The marine centric diatom *Thalassiosira pseudonana* and the pennate diatom *Phaeodactylum tricornutum* are chosen for the first dataset primarily because of the superior genetic resources availability of their genome. *Chromera velia* and *Vitrella brassicaformis* are two organisms that share a common ancestor with Apicomplexa and share similar characteristics with diatoms, thus their plastid targeted sequences are equally important for this study. Diatoms acquired their plastids by secondary endosymbiosis. A secondary endosymbiosis occurs when a living cell engulfs a previously endosymbiotic eukaryote. A transit peptide-like domain adjacent to the N-terminus is found along with signal peptide pre-proteins only in the organisms that underwent secondary endosymbiosis. I have extracted various features based on sequence properties, such as net charge, molecular weight, iso-electric point, hydrophobicity etc. in particular sequence motifs. The sequences and their extracted feature sets have been analyzed in search of correlation with the given location using available statistical tools. My results indicate that a positive net charge is required to transport across the plastid envelope. Count of amino acid frequency after cleavage position reveals that in 14th position of the sequence an Arginine is required for efficient transport. Alternatively, instead of Arginine if enough charged residues are present in the transit peptide, then the mature protein can be transported to the stroma. I have discussed why the net charge is essential for the transport mechanism and presented the distribution of amino acids in the target peptide sequences. Additionally, I have provided python and R scripts that were used to conduct these analyses.

Keywords: Plastid, Signal peptide, Transit peptide, Diatoms, N-terminal, Secondary endosymbiosis

2. Introduction

Diatoms, first observed in the 18th century, unicellular photosynthetic organisms, are tiny enough that we cannot see them without a microscope. Still, because of so much of them they can be visible from space. The fact that these organisms can be cultured easily means that they are ideal for research purposes. These aquatic primary producers, yield roughly half of the global primary productivity (Field et al., 1998).

It is believed that chloroplasts are the energy house of plants and are thought to have originated from cyanobacteria taken up by a eukaryotic host. In higher plants, the transferred genes are transcribed in the nucleus, then translated in the cytosol and finally transferred to the plastids. The N-terminal sequence extension is required for correct targeting and is finally removed by a peptidase after the import process. These sequence motifs differ from each organism and have certain features such as a high degree of hydroxylated amino acids (Lang et al., 1998). However, diatoms are different from the higher plants and algae, therefore, the mechanism of protein transportation to their plastids differs significantly.

Diatoms are highly dependent on their internal organelles such as plastids, endoplasmic reticulum, and mitochondria. Plastids are essential component of plant cells. Some plastids store energy in form of currency, some synthesize fatty acids and produce primary and secondary metabolites. Therefore, it is considered as metabolic factory of plant cells. Plastids rely on the host cells, and there is a lot of energy exchange going on between plastid and other organelles. The pathway for intracellular transport has taken a lot of focus in recent years. As I stated earlier, diatoms evolved via a secondary endosymbiosis, meaning a eukaryotic cell engulfed a cell that has already undergone a primary endosymbiosis. As a result, they are surrounded by four membranes.

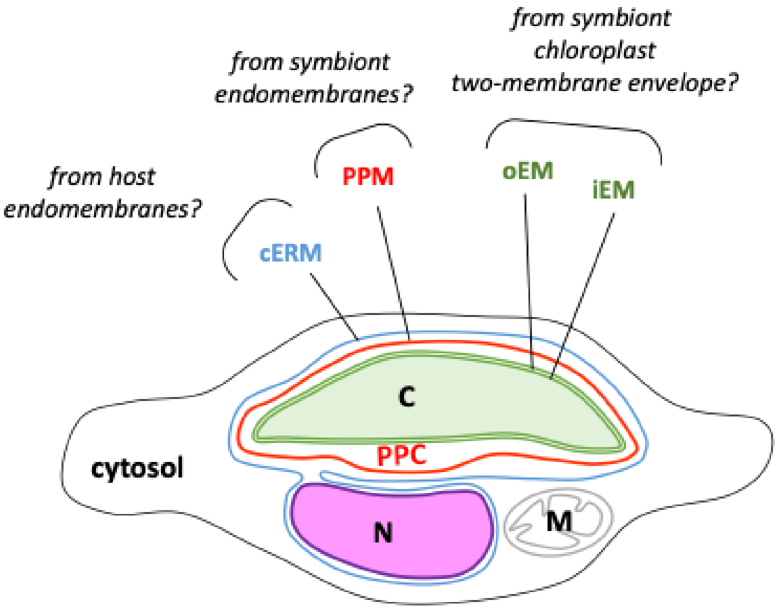

from symbiont chloroplast two-membrane envelope?
from symbiont endomembranes?
from host endomembranes?
PPM
oEM
iEM
cERM
C
PPC
cytosol
N
M

Figure 1. Representation of the secondary plastid in diatoms. In this semantic structure of a diatom, we can see that the plastid is surrounded by 4 membranes. The first membrane is a light blue that is continuous with the outer nuclear envelope membrane also named by chloroplast endoplasmic reticulum membrane(cERM). The periplastidial membrane (PPM) is shown in red. The outer and inner membranes (oEM, iEM) of the plastid stroma are shown in green. Figure from (Flori et al., 2016)

Diatoms are heterokont algae. Because of the secondary endosymbiosis, the first (outermost) membrane of diatoms is continuous with the ER, the innermost two membranes are derived from formar primary plastid (Hirakawa et al., 2009).

Four membranes add extra layers of complexity. For example, organisms containing mitochondria or plastids, require additional gene expression that requires nuclear DNA-encoded proteins transported to several compartments of cell bodies across all those extra membranes. For trafficking, nearly all nuclear-encoded plastid proteins have similar characteristics and topological bipartite ER-like signal sequences namely signal peptides followed by a transit peptide-like sequence (Apt et al., 2002; Felsner et al., 2010; Gruber & Kroth, 2017).

Photosynthetic eukaryotes store photosynthesized carbohydrates in the cytoplasm. When they are in a critical condition such as the absence of light they use stored carbohydrates for energy via respiration (Gruber & Haferkamp, 2019). One way to store sugar is by glycogen, which is special type of starch that many organisms use including humans. Another way to store sugar is by β-glucans, these use chemical bonds between sugar molecules and form gel like structure. Diatoms do not store energy as glycogen instead they use β-glucans to store energy in vacuole. ATP is the store and provider of energy of the whole cell. The chloroplast also has ATP synthase and uses same mechanism to make the ATP like mitochondria. In green algae this ATP synthesis is done in plastids, and they export these ATPs for the whole cell. However, diatoms neither they synthesize the ATP in plastids nor export the energy to other organelles. Therefore, diatoms need to import these energies which are transported via specific transport proteins called NTTs (Gruber & Haferkamp, 2019). These NTTs can be found in the inner membrane and help with exchanging ATP with ADP (Trentmann et al., 2008; Winkler & Neuhaus, 1999). In contrast, the metabolism of diatoms along with the function of NTTs are quite different from bacteria and other photosynthetic eukaryotes. Some nucleotide transporter proteins have similar features as transit peptides in their N-terminal region. Nucleotides are the building block of DNA/RNA and play a vital role in energy exchange with other cellular compartments (Gruber & Haferkamp, 2019). *T. pseudonana* and *P. tricornutum* have specific nucleotide transport systems and NTTs act as the main transporter (Ast et al., 2009), whereas, a different kind of NNT acts as a carrier in higher plants plastid. Thus, NTT proteins are restricted to only a few organismic groups (Witz et al., 2012).

2.1 Signal Peptides

Signal peptides are usually 15 to 30 aa long. They are involved in protein translocation through different membranes. They play an important role in addressing the protein to the endoplasmic reticulum in eukaryotes and to the cytoplasmic membrane in prokaryotes. They initiate transport of the adjacent protein before translation is ended and they are classified in the family of targeting sequence.

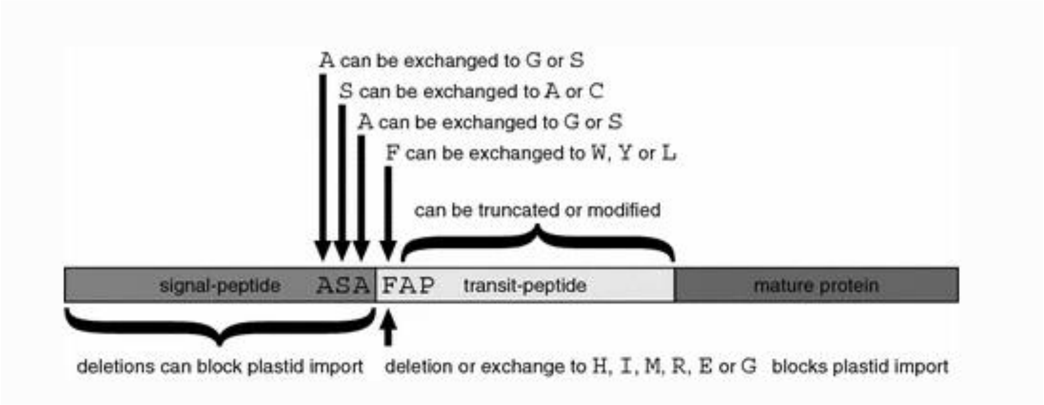

A can be exchanged to G or S
S can be exchanged to A or C
A can be exchanged to G or S
F can be exchanged to W, Y or L
can be truncated or modified
signal-peptide
ASA
FAP
transit-peptide
mature protein
deletions can block plastid import
deletion or exchange to H, I, M, R, E or G blocks plastid import

Figure 2. Diatom plastid proteins and their conserved region. In this case, from the start codon M until the cleavage point between ASA and FAP is a signal peptide and an enzyme cuts the peptide off at that position, downstream of this position is the transit peptide followed by a mature protein. Finally, when the signal peptide cleaves off, some other enzyme reads the transit peptide and is transported in cell compartments. This ASAF motif is conserved in most diatoms, especially the +1 position where phenylalanine can only be replaced by tyrosine, tryptophan or a lucine (Gruber et al., 2007).

In other words, short peptides located in the N-terminus of protein that are carrying information for protein secretion are called signal peptides (Owji et al., 2018). In this study, +1 position to the 20th position means that the sequences from the first position of the signal peptide cleavage site to the 20th position of the signal peptide (window1). This signal peptide is derived from the ASAFind software (Gruber et al., 2015a) which predicts and outputs 25 amino acid sequences. In every predicted signal peptide sequence has a cleavage site after 5th position from the initial amino acid residue, and from 6th position, the transit peptide starts (Figure 3). These sequence windows will be carefully selected for each dataset.

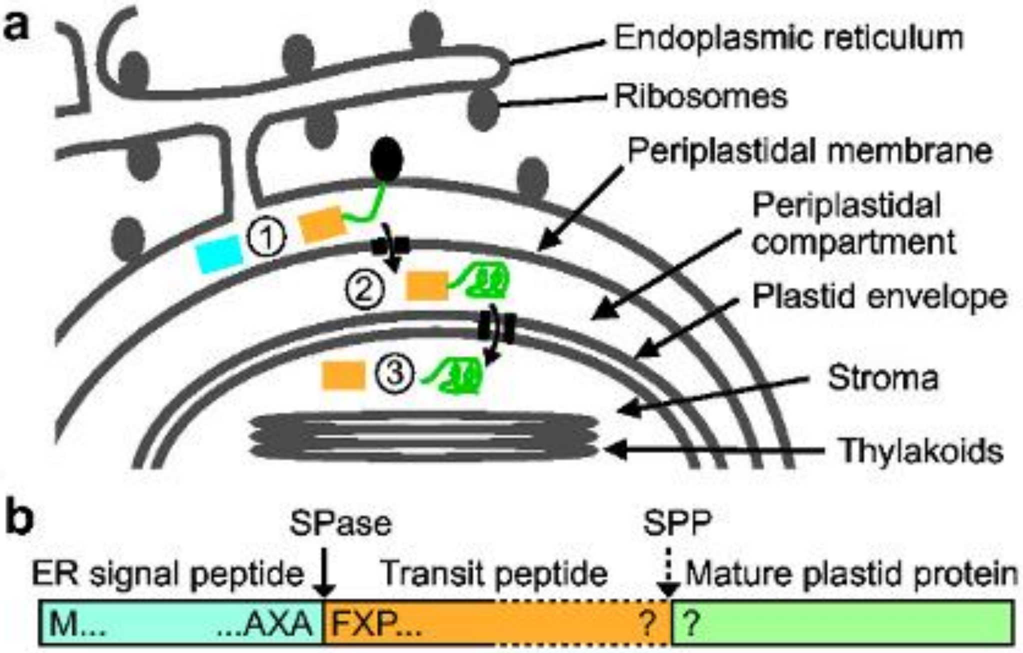


Figure 3.

a. *Here, the transport system through four membranes of the plastid can be observed. Proteins synthesized in the ribosomes are transported to the 1st periplastidal compartment then according to the signal peptide it's cleaved off(blue) with signal peptidase, furthermore, it is transported to the 2nd periplastidal compartment through the membrane. The transit peptide is cleaved off by stromal processing peptidase when it enters the last compartment leaving the mature plastid protein.*
b. *Schematic structure of plastid protein where an ER signal peptide followed by a transit peptide and then a mature plastid protein resides.*
Figure from (Huesgen et al., 2013)

2.2 Transit peptides

Almost 95% of the proteins destined for the plastid are encoded by nuclear DNA (Wojcik & Kriechbaumer, 2021). Proteins are mostly synthesized in the cytosol. The transportation of those proteins to their destined location is a sequential process. In diatoms, targeting pre-sequences of nucleus-encoded proteins are composed of a signal peptide and a transit peptide. Transit peptides are usually 10 to 60 amino acids long. This polypeptide chain works as a recognition system of the remaining mature protein. Essentially, the system knows where to put the mature protein by reading the transit peptide sequence. The signal peptide is cleaved off by a signal peptide peptidase (SiPP), and some enzymes recognize the transit peptide by its properties. Finally, it mediates translocation across the remaining three membranes. The transit peptide targets the proteins across the periplastidial membrane into the periplastidial compartment, and from there to the stroma, crossing two more membranes. In the stroma, the transit peptide is cleaved off, leaving only the mature protein (Huesgen et al., 2013) (Figure 3). In contrast, Transit peptides are transported to organelle membranes after successfully detaching from the signal peptide that initiates the whole transportation system. Despite the values of these algae in nature, we are still in the age of understanding how the cellular protein transport systems in these organisms work. In a research (Emanuelsson et al., 2007) mentioned that these "Zipcode" like signals predictions have become a major task in bioinformatics.

2.3 Organisms mentioned in this study

The importance of diatoms was revealed when scientists successfully sequenced the whole genome of the centric diatom *Thalassiosira pseudonana* (Armbrust et al., 2004; Oudot-Le Secq et al., 2007) and the genome of pennate diatom *Phaeodactylum tricornutum*. The *T. pseudonana* is a widely used marine diatom with a small number of genomic sequence data, showing us insight into ecology and evolution of diatoms. The diverse metabolic pathways of *T. pseudonana* and a small (32MB) genome size represent a large and predominant genus (Armbrust et al., 2004; Oudot-Le Secq et al., 2007; Schober et al., 2019). On the other hand, pennate diatom *P. tricornutum* is the second diatom genome that has been fully sequenced and chosen primarily because it is enriched with genetic resources, and typically found in coastal areas (Armbrust et al., 2004; Janouškovec et al., 2010).

Based on sequence data these two model diatoms along with the closest known photosynthetic relatives of apicomplexan parasites (Kořený et al., 2011) *Chromera velia* holds a unique position in phylogeny as intermediate between photosynthetic dinoflagellate and parasitic apicomplexans (Weatherby & Carter, 2013). The diatoms have complex plastids acquired by secondary endosymbiosis, but they are not the only organisms. Simlar complex plastids have also been identified in variety of other organisms. These organisms include non-photosynthetic, also photosynthetic organisms such as apicomplexan parasites. C. velia and V. brassicaformis are the closest relative that are photosynthetic and has a plastid in common (Figure 4) (Janouškovec et al., 2010; Weatherby & Carter, 2013).

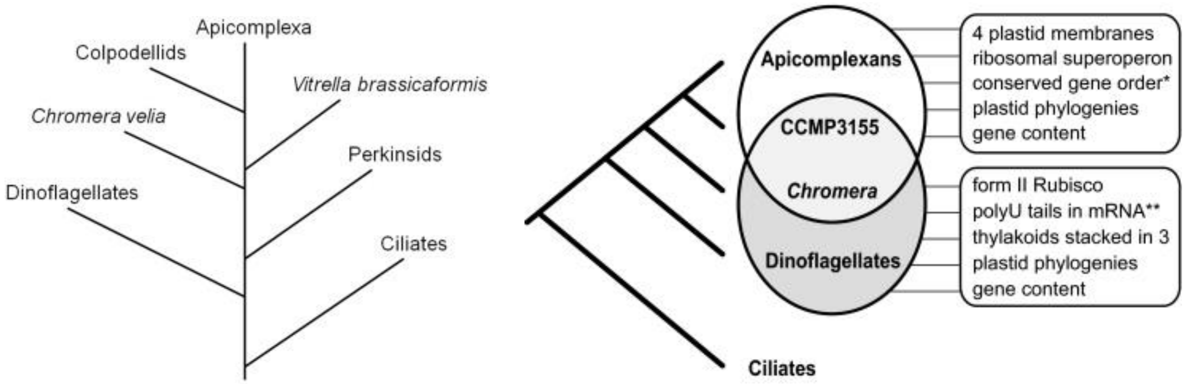


Figure 4. *On the left, is a representation of phylogenetic relationship of C.velia to other apicomplexans. The pictures and data were derived from* (Janouškovec et al., 2010; Weatherby & Carter, 2013). *On the right, a summary of plastid evolution in alveolates shows a direct link between CCMP3155(Vitrella brassicaformis) and C.velia however, the features in the boxes were thought to be different in two lineages but C. velia holds similar features.*

Chromerida is a phylum of unicellular alveolates. *C. velia* and *Vitrella brassicaformis* which are photoautotrophic and evolved from photosynthetic ancestors making them closely related to apicomplexan parasites. They contain secondary plastids like other apicomplexan parasites and both organisms lack chlorophyll-c, yet they differ significantly in the life cycle, morphology, and the plastid genome (Oborník et al., 2012).

Another dataset is made by collecting sequences of typical plastid-targeted proteins as well as proteins with localization to mitochondria, ER, and other compartments, conserved in relatives of diatoms, so that the difference is distinguishable. Organisms such as *Cylindrotheca closterium, Fistulifera solaris, Nitzschia inconspicua, Thalassiosira oceanica* and some other distantly related species have been selected and added to the custom dataset. All the sequences are extracted from NCBI and manually cured for validity. For a valid argument, it is necessary to conduct the analysis on different lineages. This will help us to understand how the feature set of other organisms differs in terms of targeting proteins in various locations.

2.4 Prediction tools

SignalP is a prediction program that predicts the signal peptides for proteins that are targeted to plastids, ER and other membranous locations (Emanuelsson et al., 2007; Huesgen et al., 2013). Furthermore, it predicts the presence and location of the cleavage sites in amino acid sequences. This program extensively uses the hidden Markov model and neural network algorithms trained on separate datasets of prokaryotic and eukaryotic protein sequences (Nielsen et al., 1997). For instance, SignalP 3.0 NN performs better than other versions (Gruber & Kroth, 2017). Although SignalP 3.0 server is currently down, SignalP 4.1 an upgraded version of SignalP 3.0, continues to operate without any issues. SignalP inputs are usually protein sequences in Fasta format of eukaryotes, gram-negative bacteria, or gram-positive bacteria as input and as output, it returns us three different scores, C, S, and Y. The output can be requested as graphical or tabular or both. For each position, it will calculate an S-score indicating that the signal peptide prediction around that position is high. C score is the "cleavage site" score which only is high at the cleavage site.

ASAFind (Gruber et al., 2015a): Based on SignalP scores, this software runs a sliding window to predict the cleavage site specifically for plastid proteins in algae with complex plastids of the red lineage. As a result, this software returns conserved motif "ASAFAP" and transit peptides (Figure 2).

While most proteins of eukaryotic cells are synthesized in the cytosol, many are destined for different compartments by their sequence properties namely signal peptides and transit peptides recognized by translocation machinery. Nonetheless, sequences have features such as length, molecular weight, net charge, isoelectric point, gravy index etc. These features might have a correlation with their localization.

2.5 Sequence features

Sequence length: The more diverse the sequences are the more complex the similarity becomes in protein sequences. One of the reasons for being so diverse is the sequence length. Sequences are conserved in terms of subcellular localization and enzymatic activity (Nair & Rost, 2009). Dozens of algorithms have been invented to measure sequence similarity based on alignment, word frequency, genomic expression etc. however, not all algorithm is perfect to detect similarity. There is always a trade-off between sensitivity and specificity and on top of that computational power. The more complexity, the more computational power is needed to solve the problem. Sequence analysis is not only important for amino acid or DNA functional analysis but also can be crucial for phylogenetics.

Molecular weight: High or low molecular weight could be a key factor in determining intracellular signalling to regulate the transport of protein (Chavrier et al., 1990). Molecular weight is a key parameter to recognizing a primary structure of a protein and can assist in initial assessment of the biomolecule's functionality. Ast et al. in their study of *T. pseudonana* and *P. tricornutum* suggested that many unusual NTT coding sequences are present with sizes ranging from 57 kDa to 92 kDa and 7-11 transmembrane helices (Ast et al., 2009).

Net charge: Charged polypeptides play important role in protein expression. The frequency and distribution of positively and negatively charged residues are calculated to determine the net charge of certain amino acids. In contrast, charged residues of a polypeptide might be a key factor for the recognition of transit peptides. However, most proteins tend to be positively charged in the N-Terminal region (Requião et al., 2017).

Isoelectric point: At a given pH depending on aa sequences, no net migration takes place in an electric field, therefore, the net charge of that protein becomes zero. Proteins are most soluble and least reactive in pH close to their isoelectric point and pH of most cellular compartments is close to 7.5 (Kiraga et al., 2007). As a result, this property could be important to consider for localization.

Aromaticity: Highly variable yet an important feature is the frequency of Phe, Trp and Tyr. In protein expression, there is constant pressure of reducing the aromaticity of proteins in cell body as it is relatively costly for protein biosynthesis. In diatoms, the Phe is conserved at the cleavage position (Figure 2).

Instability index: This is a basic property of a protein that is used to determine whether it will be stable in a test tube. Instability index of less than 40 is stable therefore higher than 40 is unstable. Furthermore, features like instability index can be a key factor to correlate metabolic stability with location as this feature is determined by the order of certain amino acid sequences (Guruprasad et al., 1990).

Gravy Index: This number is a measure of hydrophobicity or hydrophilicity of a given aa sequence. This index resides in between +2 to -2, hydrophobicity scores below zero are more likely to be hydrophilic and above zero are hydrophobic (Magdeldin et al., 2012).

3. Aim

I will extract the sequences from the datasets and perform a sequence analysis with bioinformatics tools. Afterwards perform statistical tests with hypotheses and visual inspection on the extracted feature set.

4. Materials and Methods

For this analysis, I have collected data from two different research papers and another dataset entirely created by myself consisting of publicly available closely related organisms with diatoms from the NCBI protein database. The accession numbers of the proteins can be found in the dataset as well as at (Accession Numbers). The following two papers are already published, and the data used for them are carefully screened.

1. Plastid proteome prediction for diatoms and other algae with secondary plastids of the red lineage (Gruber et al., 2015).
2. Subcellular Compartments Interplay for Carbon and Nitrogen Allocation in *Chromera velia* and *Vitrella brassicaformis* (Füssy et al., 2019).

Finally, these two datasets were merged with my custom dataset for further analysis.

4.1 Data Pre-processing

Both research papers have data in a tabular format including a fasta file each. Fasta files were taken as inputs and run them in SignalP 4.1 server individually. SignalP returns output with or without a graph and returns information on cleavage position. I preferred the short output without plot as I needed to save the output with the extension ".signalp". C- score represents the raw cleavage site score, which is trained to be high at the cleavage site.

S-score is the signal peptide score corresponding to the mature protein within and without signal peptides. Y-score is a better prediction score for the cleavage site by calculating the geometric average of the C-score and the slope of the S-score. This is chosen by the steepness of the slope where multiple C-score peaks occur and take the highest value of it. Source: (https://services.healthtech.dtu.dk/service.php?SignalP-4.1)

Now that I have the short output of the SignalP saved, the next step is to run this in the ASAFind software. ASAFind takes signalp file and the corresponding fasta file and returns a tabular file as the prediction result. The instruction for running the software locally is given on this webpage (https://bitbucket.org/rocaplab/asafind/src/main/). This software is downloaded and the python script was run via the terminal/ command line. As a result, this will identify nuclear-encoded plastid proteins in algae with secondary plastids of the red lineage based on the conserved "ASAFAP" region in the sequences.

4.2 Data cleansing

The first two steps are done individually for all three datasets; therefore, it generated three different tables to work with. For comparison and statistical analysis purposes, I have decided to keep the datasets individual and create a final dataset by merging these three together. The sequences that are SignalP negative, have been eliminated from all the datasets. The next step is to screen through each dataset and find N/A values and clean them as they will raise errors. The NCBI database is a great data source, however, the sequences are not screened and not quality controlled therefore, I have found multiple invalid sequences which are also deleted from the dataset.

The dataset containing *C. velia* and *V. brassicaformis* has only prediction results for plastid destined proteins. Third dataset containing *P. tricornutum* was already marked with localization. Finally, the merged dataset contains the previously mentioned three organisms, as well as the *T. pseudonana* a and some other related organisms (Table 1). A codon/ protein consists of three bases of DNA, and there are 20 proteins expressed alphabetically. Unnecessary, ambiguous letters or characters such as stop codon (*) interrupt the calculation. Any invalid character/ letter which is not a protein has been deleted from the dataset to prepare for the next steps.

4.3 Feature engineering

In this part, I have extracted amino acid features based on sequence data. The tabular files that have been created earlier are converted to a readable csv format and then imported to the Jupyter Notebook where a python environment has been created with the following libraries.

- Pandas (Deals with DataFrame)
- Biopython (Includes essential libraries) (Cock et al., 2009)
- Bio.SeqIO (For sequence manipulation)
- ProtParam/ProteinAnalysis (For protein sequence analysis)

Firstly, the datasets is read using panda‘s read_csv module Appendix 1. By printing out the columns lists will help to determine which columns are necessary. Unnecessary columns have been dropped from the table and made a new copy of the dataset in another data frame so that if we need to change anything inside the dataset, it will not affect the original data Appendix 2.

When the amino acids are retrieved, in some retrieval technologies poorly reads or unknown amino acids are highlighted with "X". Three sequences with the "X" that appeared in the dataset has been deleted Appendix 3.

4.4 Data analysis

Initially, a set of sequence windows has been selected for the features that I am interested in. The sequence windows are described in (Figure 5).

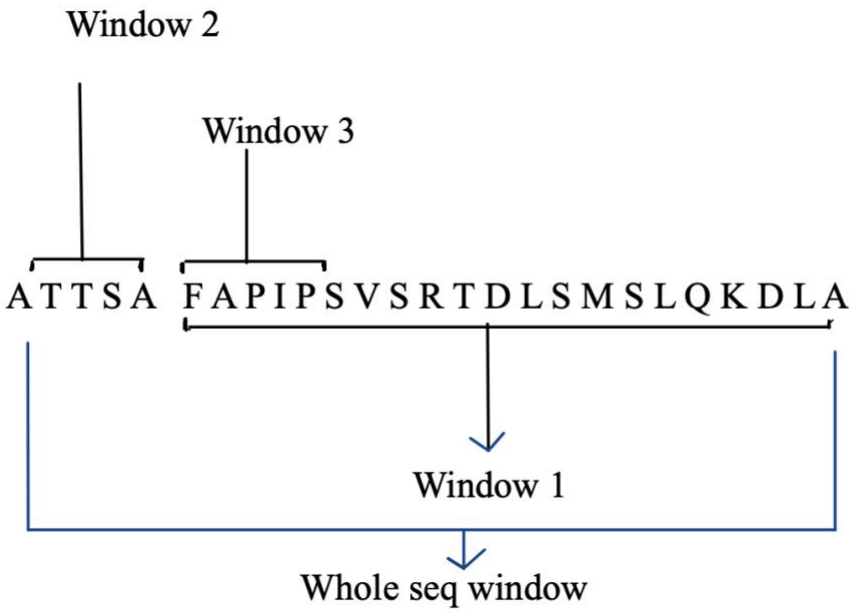


Figure 5: Whole sequence window is the entire 25 amino acid sequence. Window 1 is the sequence after cleavage position, window 2 is the first 5 amino acids of a sequence and finally window 3 is the 5 amino acid sequences of window 1.

These sequence windows have been analysed with ProteinAnalysis library and saved into different lists under different names Appendix 4. Some features such as stability and isoelectric points of amino acids are calculated. I have set a threshold in stability and isoelectric point for categorical values. Appendix 5. All these calculated values are added to the existing data frame. Appendix 6. In the end, the selection of actual features for the dataset by choosing the names of the columns can be made and exported Appendix. 7. This whole process will be carried out each time for each dataset.

Furthermore, the exported datasets will be tested for statistical analysis. Data visualizations were made using tableau desktop software and Rstudio as well as Python. A broader focus will be given to the merged dataset for visual and technical analysis. In Tableau desktop, bins have been created for charges and counted how many amino acid sequences fit in each bin before plotting in the graph. Both the median and mean values have been measured to make the difference visible. The data contains negative values therefore, differences between the median and mean must be addressed. Using R, location-wise charge property Appendix 8 along with an ANOVA test (Appendix 9) has been conducted. However, the ANOVA test does not tell us about the differences in locations, therefore a post hoc analysis has been carried out to see the location-wise differences (Appendix 10). Finally, amino acid sequence counts of each position were calculated to see where the charged amino acids are most frequent in the sequence index Appendix 11. Additionally, visualizations of these amino acid counts have been created in Appendix 12.

5. Results

From chemical point of view, proteins are the most complex structures that has been developed for billions of years. In this time frame, a lot of changes have been occurred such as insertion or deletion or a gene lost/transfer. Therefore, comparing amino acid sequences is an important strategy in molecular biology. In this analysis, The plastid targeted sequences are in the range of 200 to 400 aa long but it is rare more than 400 aa long (Figure 5).

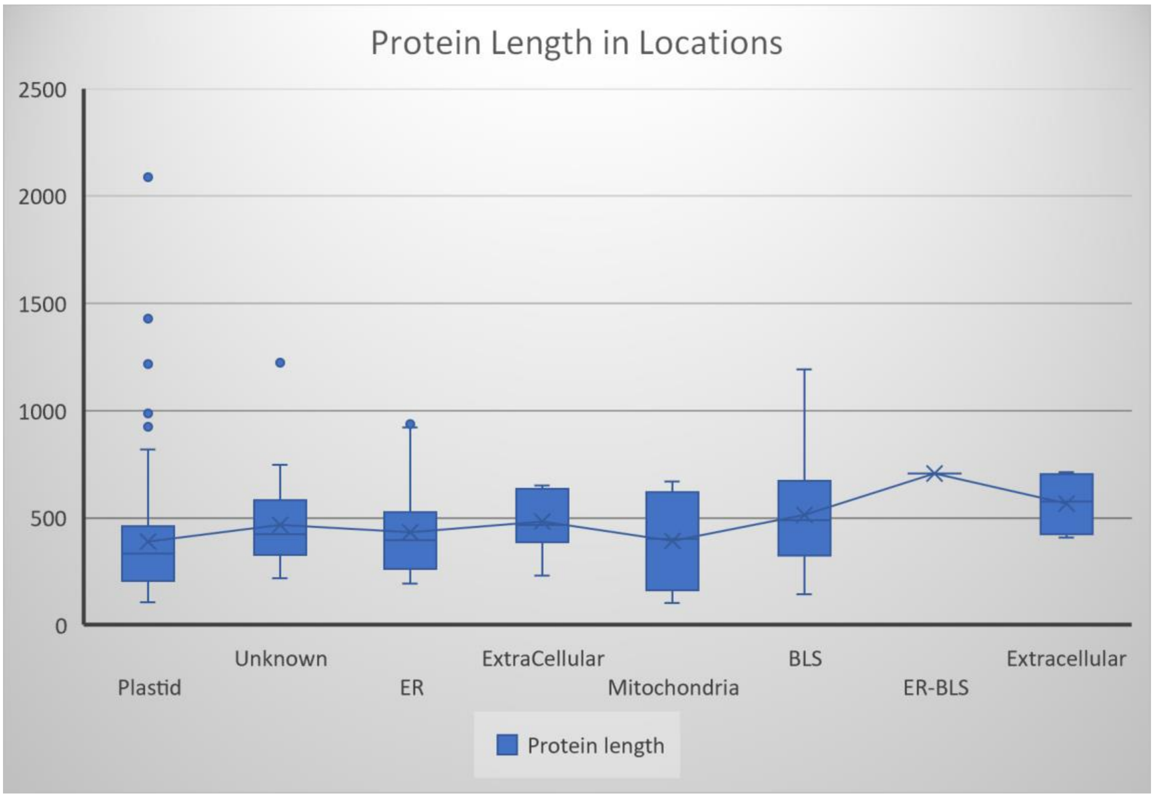


Figure 5: X axis represents the different compartment of the proteins and Y axis represents length of amino acid sequences. Plastid targeted sequences are usually 400 amino acids long, however there, are not much observable differences into other compartments except the ER-BLS. In the dataset has only 1 sequence of ER-BLS sequence, therefore in the graph it just a line and does not have any mean value.

On the other hand, molecular weight (MW) is another feature that completely depends on how long the sequence is. Therefore, the molecular weight tends to be higher along with the sequence length. Additionally, molecular weights in sequence windows are not significant, meaning the plastid-targeted proteins have similar molecular weights to other compartments. Captured protein properties by the proposed feature sets have been conducted to make a decent generalization of the novel sequences. Features, such as protein length and molecular weight are often not discriminative enough. Not many of sequences can be found in databases for different compartments and different organisms. (Table 1) is a count of sequences that are in different locations.
Nevertheless, I found one promising result concerning the net charge of the sequences. I have tested the merged dataset for charge dependency of plastid proteins, and in a result, it shows positive charge dependency in window 1 (Figure 7). Meaning, the net charge of the residues from cleavage position are slightly positive in the merged dataset.

Organism Name	BLS	ER	ExtraCellular	Mitochondria	Plastid	Total
Actinocyclus subtilis					1	1
Biddulphia biddulphiana					1	1
Biddulphia tridens					1	1
Chaetoceros tenuissimus		1				1
Chromera Velia					20	20
Conticribra weissflogii					1	1
Cylindrotheca closterium				3		3
Fistulifera solaris	2	5				7
Guinardia striata					1	1
Leptocylindrus danicus			2			2
Nitzschia inconspicua	1	3				4
Phaeodactylum tricornutum	23	13	4	7	82	129
Rhizosolenia imbricata					1	1
Seminavis robusta	1	1				2
Thalassiosira oceanica		3				3
Thalassiosira pseudonana		1	12		15	28
Vitrella Brassicaformis					12	12
Total	**27**	**27**	**18**	**10**	**135**	**217**

Table 1: Dataset 1 contains 97 Phaeodactylum tricornutum (Gruber et al., 2015)*, dataset 2 contains 32 C. velia and V. brassicaformis* (Füssy et al., 2019)*, and the remaining 86 sequences are collected from NCBI. Out of 217 sequences, 135 sequences are plastid targeted. 10 sequences are targeted to mitochondria. 18 sequences are targeted to Extra-cellular compartment, 27 sequences are targeted to BLS and another 27 sequences are targeted to endoplasmic reticulum.*

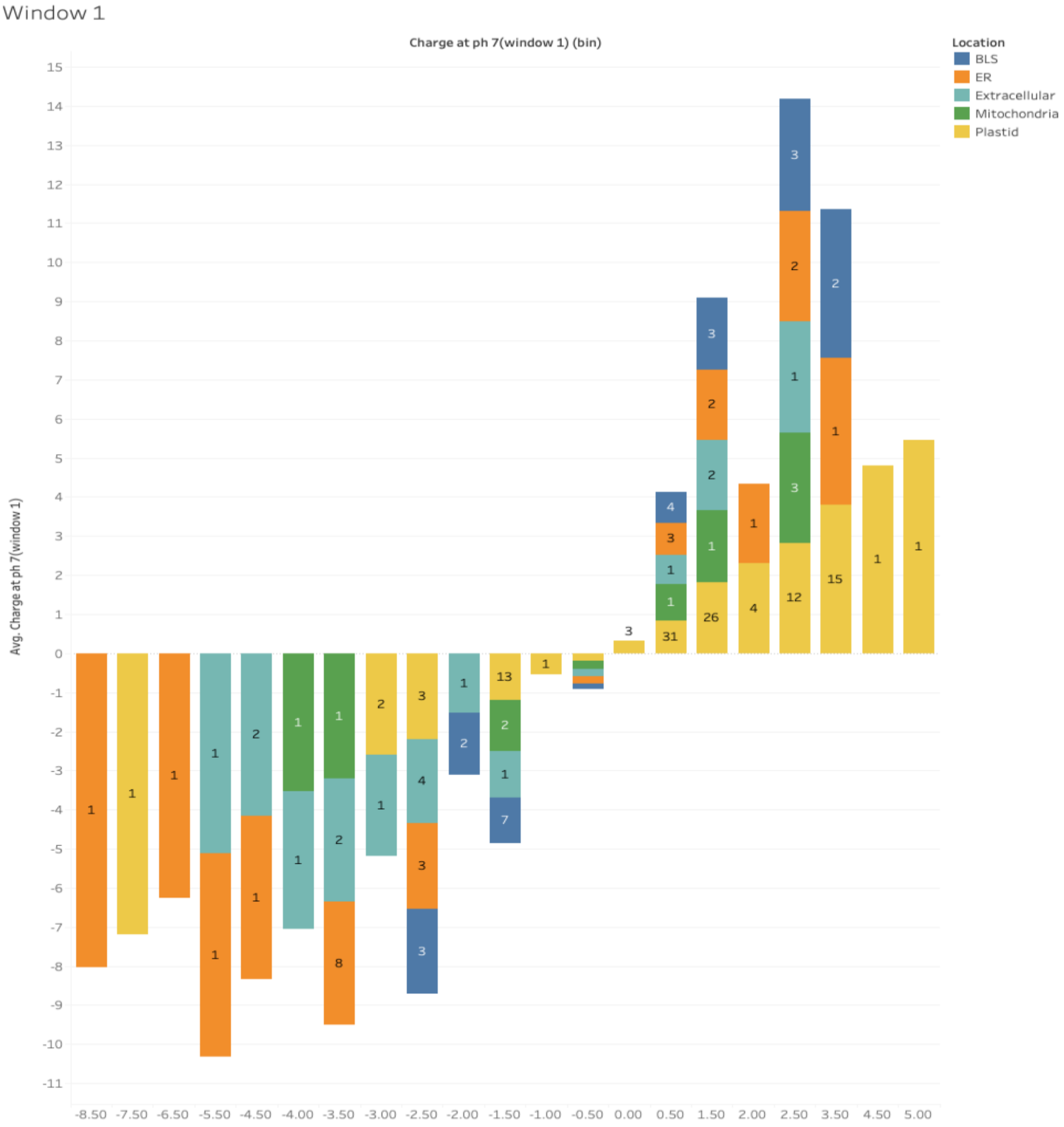


Figure 7. *On the x-axis charge bins and y-axis average charge of each location (Sorted by colour) is marked with count in each bin. Each bar is stacked on top of the other depending on the locations and is the same size because of a similar average value. Most of the plastid-targeted sequences (yellow box) are positive in charge. A few of them are negative but close to zero.*

In contrast, most of the charged amino acids are on the N-terminal of the sequences. Both window 2 and 3 respectively show a net charge slightly negative (less than zero) because there were no or less positive charged residues (Figure 8,9).

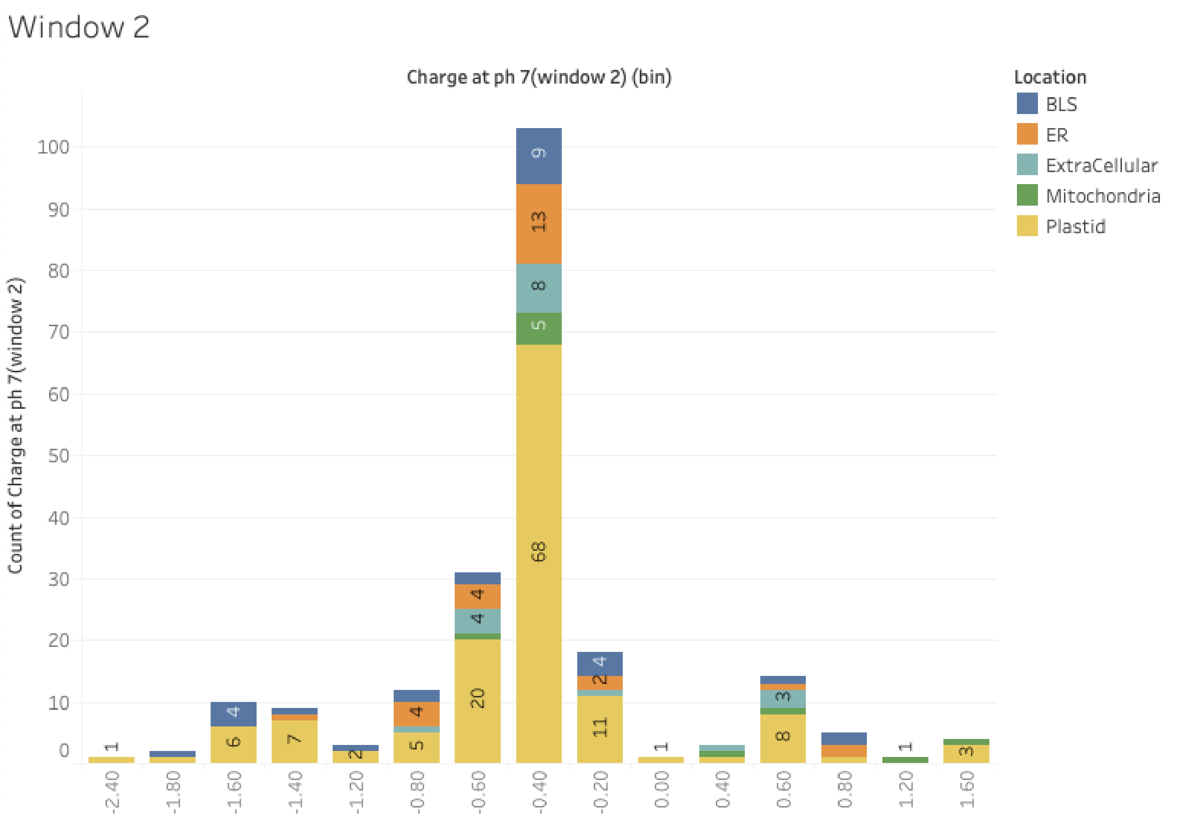


Figure 8: *Window 2 represents the first 5 amino acid of a sequence until cleavage position. These amino acids show that there are none or less amount of charged amino acids present, therefore, the net charge is mostly negative. Plastid-targeted amino acids are mostly in the bin -0.40(charge). The marked texts on the x-axis are the number of counts of location in corresponding charge bins. Only 14 out of 135 of the plastid-targeted proteins are positive and the rest are negative.*

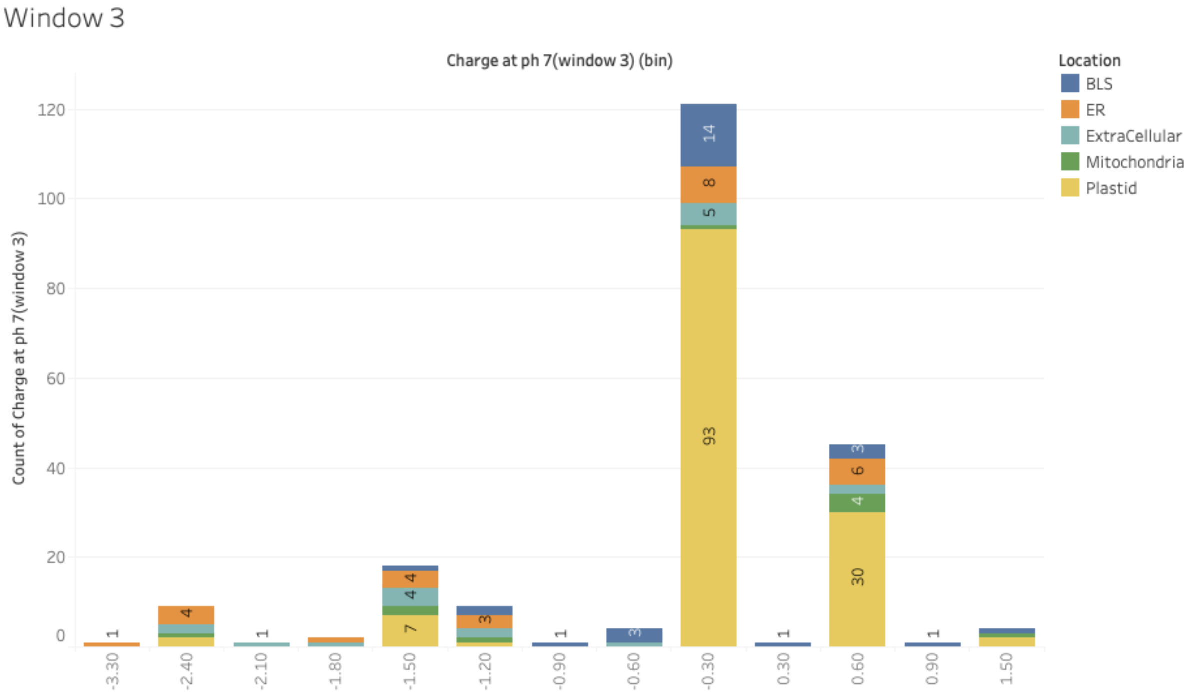


Figure 9: Window 3 which is basically the start of the transit peptide and the 5 amino acids starting from the cleavage position. Most plastid-targeted amino acids are in the bin -0.30(charge). The marked texts on the x-axis are the number of counts in corresponding charge bins. Surprisingly the transit peptide starts showing up positive charge for plastid (yellow), mitochondria (green), Extracellular and ER (orange) because if the protein does not properly recognize they end up in different compartments. In this case, 32 of them are targeted to the plastid and positively charged. Surprisingly, 15 of 32 positively charged plastid-targeted sequences belong to P. tricornutum and 10 sequences from C. velia (Figure 16). The dashboard filter is made with tableau desktop and can be found in supplementary materials.

To know whether the charges are statistically significant or not I have conducted an Anova test on three windows along with location, with the following hypothesis.

H_0 = *There is no relationship between variables in the population*

H_a = *One or more variables are dependent on each other*

The Anova test result shows that the F value of rows is larger than the F critical value as well as on the column side. Therefore, we can confirm that there is a significant difference between the columns and the dependent variables. Furthermore, the P-value is also in both cases less than 5% meaning, I reject the null hypothesis and accept alternative (Table 2). This significance is due to the high charge dependency of the pre-protein import system and a positive net charge is critical for transport through the inner two plastid membranes (Felsner et al., 2010).

ANOVA

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	1658,04	217	7,641	4,107	3,689E-36	1,2094
Columns	58,40	2	29,201	15,696	2,6212E-07	3,0165
Error	807,40	434	1,860			
Total	2523,84	653				

Table 2: P- value in both cases for rows and columns is less than 0.05, therefore; we reject the Null-Hypothesis and accept alternative.

A similar approach has been taken with the aromaticity feature of the amino acid sequences. It is expected that the Phe (F) is crucial for the peptide to be cleaved off (Gruber et al., 2007) (Figure 10). In that study, they experimented by removing the Phe residue which resulted in not proper targeting of the mature protein. Alternative approach by replacing the aromatic amino acid with another nonaromatic residue which also blocks the transport of the protein. Finally, replacing with two aromatic amino acids or by leucine (W, Y, L) can result in proper transport into the stroma.

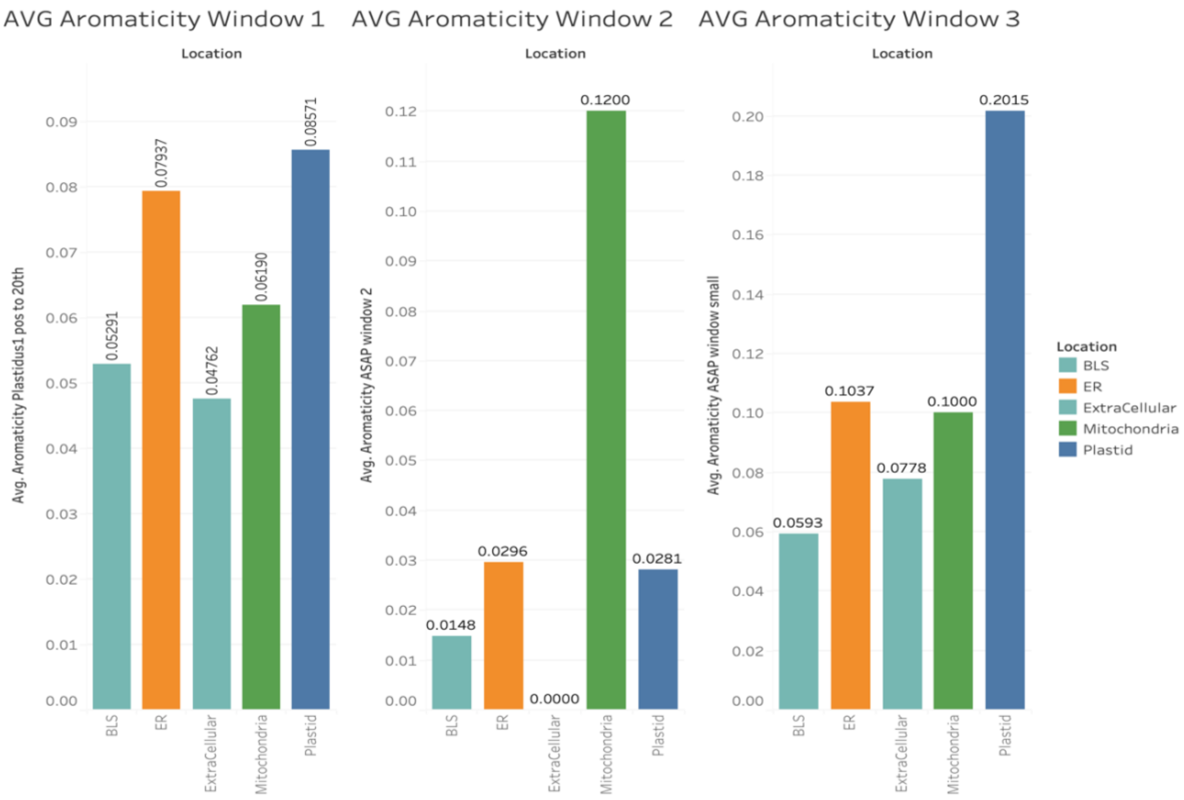


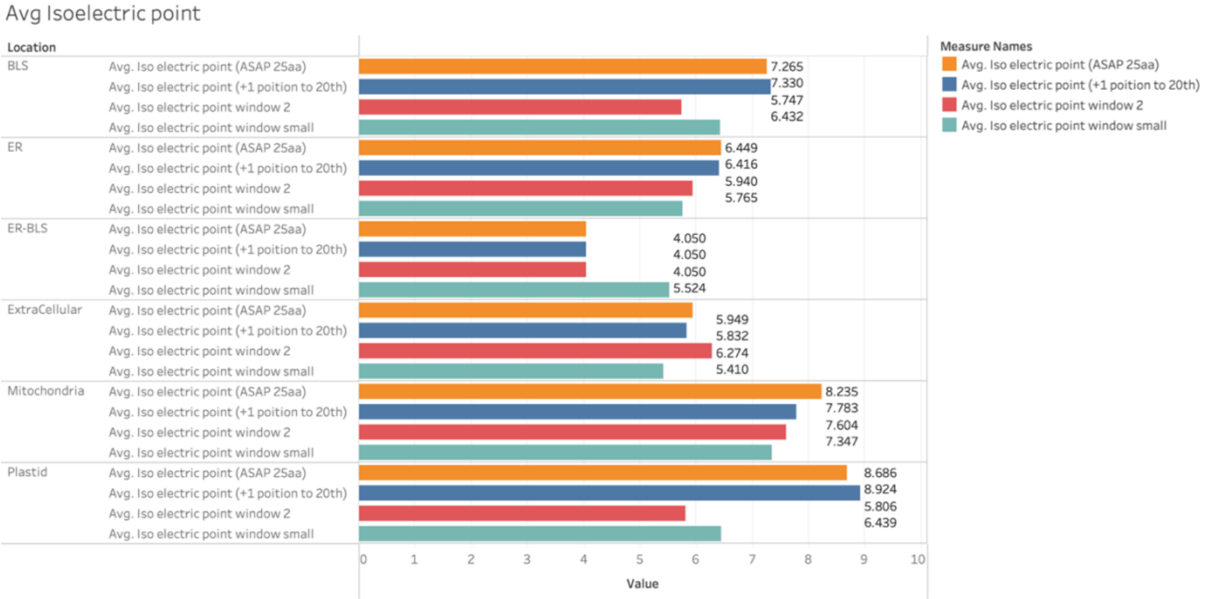
Figure 10: The aromaticity in different windows appears to be highly dependent on the Phe (F) residue which is expected. If there is no Phe residue in the window (window 2), the aromaticity drops down and if there is one or two aromatic residues in the window (window 3) it appears to be higher.

Aromaticity is only higher if there is Phe present. To support the claim, an ANOVA test has been conducted, where the F value and F critical value differ only in columns (window) and the p-value is higher than 5% meaning no difference between the location groups (Table 3).

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	1,594	217	0,0073	1,112	0,1782	1,209
Columns	1,862	2	0,9311	140,966	0,0000	3,017
Error	2,867	434	0,0066			
Total	6,323	653				

Table 3: Anova test for Aromaticity in different windows shows a significant difference in aromaticity window columns however the row value is higher than 5% meaning that there is no location-wise significance. However, the window-wise significance is observed, meaning some window contain aromatic residue and other does not.

On the other hand, there is a significant difference in isoelectric points of different locations and sequence windows. Isoelectric points in plastids are comparatively higher than in other cellular compartments (Figure 11) and the P value is less than the significance level 5% meaning the null hypothesis has been rejected (Table 4).



Avg. Iso electric point (ASAP 25aa), Avg. Iso electric point (+1 poition to 20th), Avg. Iso electric point window 2 and Avg. Iso electric point window small for each Location. Color shows details about Avg. Iso electric point (ASAP 25aa), Avg. Iso electric point (+1 poition to 20th), Avg. Iso electric point window 2 and Avg. Iso electric point window small. The marks are labeled by Avg. Iso electric point (ASAP 25aa), Avg. Iso electric point (+1 poition to 20th), Avg. Iso electric point window 2 and Avg. Iso electric point window small.

Figure 11: Isoelectric point comparison in different sequence windows along with location groups. Significant differences can be observed in different localizations. In the plastid, window 1 (+1 position and whole window) holds the highest isoelectric point but window 2 and window small (window 3) shows similar characteristics with other locations.

Source of Variation	*SS*	*df*	*MS*	*F*	*P-value*	*F crit*
Rows	1542,58	217	7,109	1,911	6,9253E-09	1,209
Columns	579,02	2	289,509	77,839	1,2937E-29	3,017
Error	1614,19	434	3,719			
Total	3735,78377	653				

Table 4: Anova test for iso electric point in different windows shows that both F (Row and columns) are higher than the F critical and p-value is less than 5% meaning there are significant differences in both location groups and different sequence windows.

To support my finding another approach was made in addition with whole sequence window of ASAFind 25 amino acids. I filtered out only plastid targeted sequences and tested for differences in sequence windows again. In a result, it shows that the differences between:

- Whole ASAFind sequences vs Window 1
- Window 2 vs Window 3

are insignificant. Meanwhile, the differences between:

- Window 1 vs window 2
- Window 1 vs window 3
- Whole ASAFind sequences vs window 2
- Whole ASAFind sequences vs window 3

are statistically significant with p values < 0.05 (Table 5)

```
## # A tibble: 6 x 9
##   term    group1      group2      null.value estimate conf.low conf.high    p.adj
## * <chr>   <chr>       <chr>            <dbl>    <dbl>    <dbl>     <dbl>    <dbl>
## 1 Charges Charge at ~ Charge at~           0  -0.0806   -0.512     0.351 9.63e- 1
## 2 Charges Charge at ~ Charge at~           0  -1.38     -1.81     -0.948 3.37e-10
## 3 Charges Charge at ~ Charge at~           0  -1.14     -1.57     -0.704 5.28e-10
## 4 Charges Charge at ~ Charge at~           0  -1.30     -1.73     -0.867 3.37e-10
## 5 Charges Charge at ~ Charge at~           0  -1.05     -1.49     -0.623 4.1 e- 9
## 6 Charges Charge at ~ Charge at~           0   0.244    -0.188     0.675 4.65e- 1
```

Table 5: Comparison table for plastid targeted sequence windows only. Window 1 and ASAFind 25 aa sequences are not significant, meaning they are similar (Highlighted). More in detail in Supplementary material.

AA	*10th*	*11th*	*12th*	*13th*	*14th*	*15th*	*16th*	*17th*	*18th*	*19th*	*20th*	*21st*	*%*
R	11	7	13	17	27	16	13	16	13	11	12	8	10.12
A	5	13	16	14	25	8	12	16	19	10	19	19	10.86
S	28	20	23	25	19	21	21	31	19	15	19	24	16.35
T	8	15	12	13	15	14	25	19	9	22	6	12	10.49
F	3	3	5	5	8	3		3	5	3	5	9	3.21
V	4	4	10	10	7	22	8	11	15	11	7	7	7.16
Q	19	12	6	6	6	2	8	2	3	5	3	7	4.87
L	8	5	4	5	4	9	9	4	17	17	12	11	6.48
P	13	14	10	14	4	13	8	5	5	12	12	4	7.03
G	5	10	8	4	4	10	4	8	5	6	10	11	5.24
E	2		2	2	3	3	3	1	3	1	4	7	1.91
I	2	2	4	1	3	2	2	4	1	2	3	1	1.67
H	7	6	5	4	3	1	2	2	1	1	2	3	2.28
K	10	6	3	4	3	7	5	7	8	3		3	3.64
N	3	9	10	7	2		7	1	5	8	13	2	4.13
D	4	1	2		2	2	1	3	1		1	1	1.11
Y		2		1		1		2	1		4	2	0.80
W	1		1	2			1		2	1	2	1	0.68
M	2	5	1	1		1	4		2	7	1	6	1.85
C							1		1				0.12

Table 6: Amino acid frequency count in each position of the sequences. Extreme enrichment of Alanine, Arginine, and serine while a decrease of aspartic acid, and glutamine has been observed.

Data is only from the 10th position until the 21st position out of 25 amino acids in each plastid-targeted sequence. In contrast, charge amino acid is required in transit peptide to efficiently transport to the plastid stroma.

Visual representation of the charge differences also suggests that the mean and median of net charge are positive in window 1 and window 2 (Figure 12).

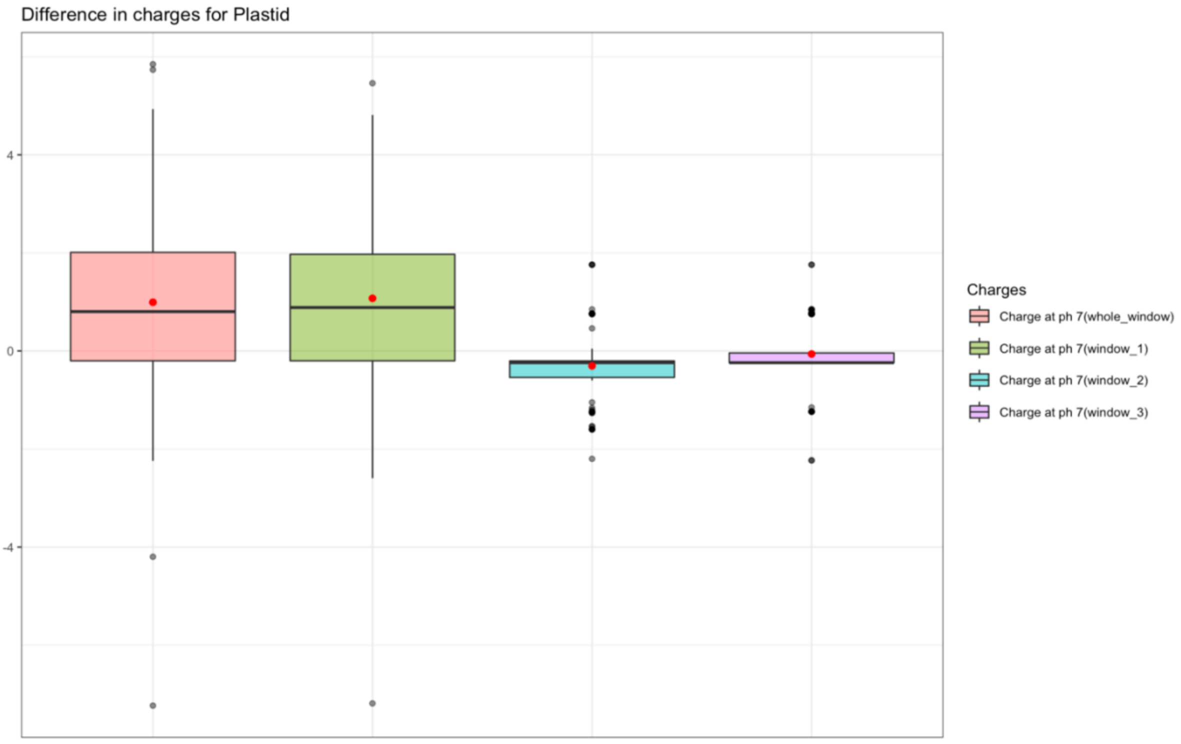


Figure 12: The difference in charges only in plastids targeted sequence windows. The red dot represents the mean as an addition to the median (line). From left Charge window 1, Whole seq window, window 2, window 3. The charge in both window 1 and whole sequences are positive, however, the rest do not contain any positive charge residue therefore negative. In contrast, the charged residues are present only after the transit peptide motif.

Additionally, an Anova test carried out for each location and charges individually (Supplementary Material R file) to confirm that net positive charge is positive and higher than in plastid and mitochondria (Figure 13, 14).

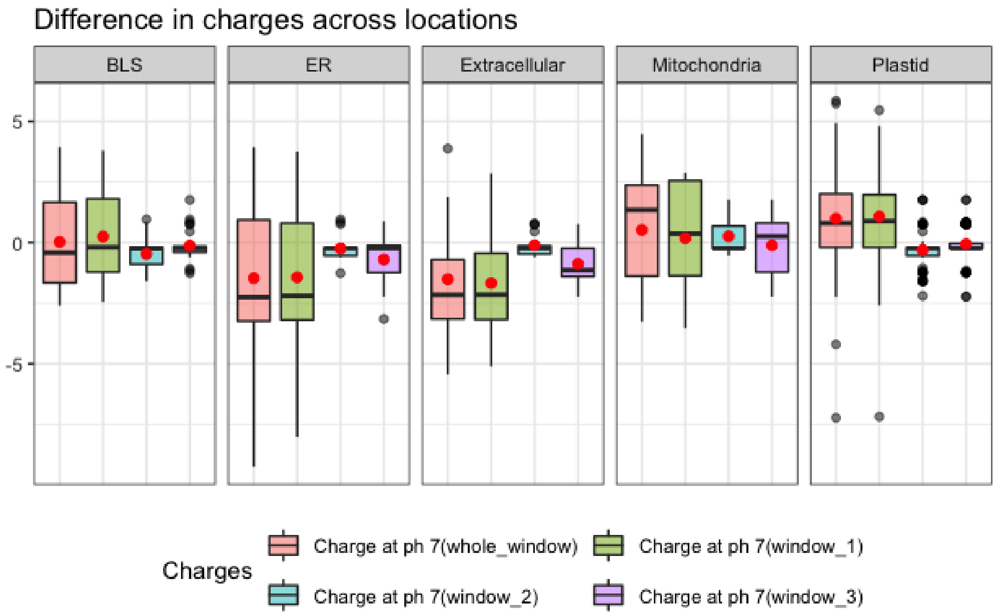


Figure 13: Visual representation of all the protein charges across the cell compartments such as BLS (Blob like structure), Endoplasmic reticulum, Extracellular, Mitochondria, and plastid. The whole window and window 1 are mostly positively charged in plastid and mitochondria on the right two boxes. BLS has some positive charges because when the modified protein (without Phe residue) is not targeted to the plastid it makes a Blob like structure in the plastid membrane.

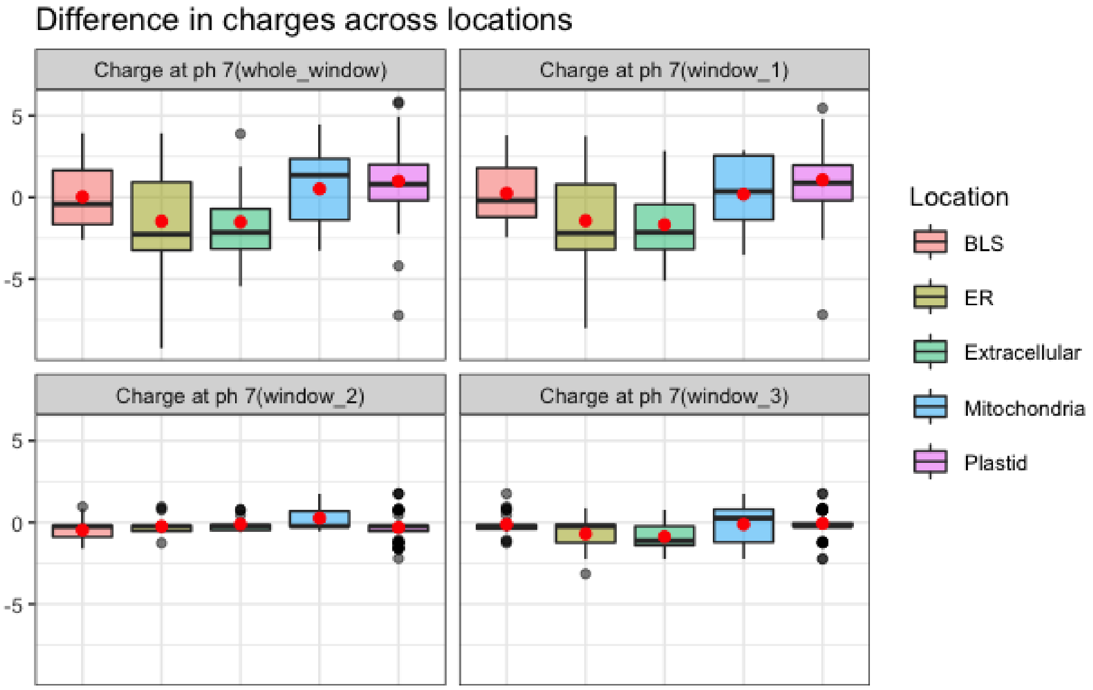


Figure 14: For a different visualization, I switched columns with rows. Same data of (Figure 13) but on x-axis charge windows and y-axis charge values to differentiate the windows.

6. Discussion

Within this study, I used publicly available sequence data from NCBI database. I have found several invalid sequences at the time of performing tests. When looking at the results it is important to keep in mind that the data might be biased at some extent, due to different data sources.

The subcellular localizations of proteins within cell compartments are essential not only for survival but also most of the metabolic actions that take place in them. Genomics analysis and high throughput genomic data suggest that plastid-targeted proteins are either semi-conserved or non-conserved across different lineages (Christian et al., 2020). This means plastid-targeted proteins are differently targeted in different organisms and it is possible only if plastid-targeted proteins are either gained or lost in evolution. The sequences that are located as plastid-targeted proteins are surprisingly short compared to other sequences like BLS or ER-targeted ones. Generally, the protein length represents the functional constraint of proteins (Lipman et al., 2002). Nonetheless, it could be a certain feature that plastid-targeted protein holds for sake of evolution for millions of years. The exact reason for the shorter protein length of these sequences is unknown. In land plants the transit peptides possess a positive charge for transport mature protein into chloroplasts (Soll & Schleiff, 2004), to find out the correlation between land plants and diatoms several studies have been done in recent years. In one research by (Felsner et al., 2010) they tested diatoms sequences transit peptide (NCBI accession Z24768) where they observed two positively charged residues and replaced them with negatively charged residues along with GFP (Green fluorescent protein) reporter, resulting in a substantial amount of green fluorescence in ER. In contrast, it prevented GFP from being translocated to the stroma. In addition, an exchange of another positive charge was introduced thereby, the modified proteins were transported to the stroma. If the net charge is close to zero, the protein locates itself in the PPM. Finally, if modified proteins have a net charge of less than zero the localization is to the ER or PPM. However, that experiment was conducted only on one organism and therefore, leaves questions on other closely related organisms. I suppose that this charge dependency of plastid protein transportation is ancestral and can be found not only in pennate diatoms but also in closely related organisms. Now, this proposal raises another curious question, what role plays net positive charge to send the protein to the stroma?

The evolution of positive net charge in the transit peptide sequence allowed the distinction between the stroma and other compartments. The inner chloroplast membrane has membrane potential which means there is a charge difference between inside of the stroma and outside of the stroma. In contrast, there are more positive charges outside than inside of the innermost membrane. If the peptide has positive charge and inside of the membrane negative, then electrostatically the positively charged peptide is an advantage for the peptide to go to innermost membrane. Because it follows the electrostatic direction in which the peptide is pulled. Therefore, it can be assumed that the membrane potential allows the sequence import across the inner membrane which provides specific signalling for targeting the stroma.

It has been suggested that the difference in the charge on the signal peptide side could favor targeting proteins to plastid or mitochondria (Christian et al., 2020). In this analysis, I have selected the region 10th to 21st of each plastid-targeted sequence to see the average frequency of amino acid composition. Results indicate that the frequency of alanine (10.86%), Arginine (10.12%), serine (16.35%), threonine (10.49%) and proline (7.03%) are quite normal for proteins. Equally important, a decrease in negatively charged aspartic acid (1.11%) and glutamic acid (1.91%) has been observed. However, the aromatic amino acid phenylalanine was relatively low at this part of the peptides because most of it was at the cleavage position, therefore remained neutral. Nevertheless, this R group serves a special purpose, and these amino acids interact with TOC GTPases during translocation (Christian et al., 2020; Vetter & Wittinghofer, 2001). Interestingly, I found substantial amount of arginine at the 14th position of the sequences (Table 6) (Figure 15, 17).

To conclude, I have found that the pre-sequences are positively charged in the beginning of the target peptide. Furthermore, positive charges are statistically overrepresented feature of the plastid targeting sequences that distinguishes itself from the other targeting pre-sequences of other sub-cellular compartments. This holds true for diatoms and closely related organisms such as *Chromera velia and Vitrella brassicaformis* (Figure 18). This analysis has been done on 25 residues long amino acid sequences however, an extension of this study can be done by increasing the transit peptide sequence length up to 60 amino acids, and/or probably another dataset that includes land plants plastid-targeted sequences to get more insight on this feature.

7. References

Apt, K. E., Zaslavkaia, L., Lippmeier, J. C., Lang, M., Kilian, O., Wetherbee, R., Grossman, A. R., & Kroth, P. G. (2002). In vivo characterization of diatom multipartite plastid targeting signals. *Journal of Cell Science*, *115*(21), 4061–4069. https://doi.org/10.1242/jcs.00092

Armbrust, E. V., Berges, J. A., Bowler, C., Green, B. R., Martinez, D., Putnam, N. H., Zhou, S., Allen, A. E., Apt, K. E., Bechner, M., Brzezinski, M. A., Chaal, B. K., Chiovitti, A., Davis, A. K., Demarest, M. S., Detter, J. C., Glavina, T., Goodstein, D., Hadi, M. Z., … Rokhsar, D. S. (2004). The Genome of the Diatom *Thalassiosira Pseudonana* : Ecology, Evolution, and Metabolism. *Science*, *306*(5693), 79–86. https://doi.org/10.1126/science.1101156

Ast, M., Gruber, A., Schmitz-Esser, S., Neuhaus, H. E., Kroth, P. G., Horn, M., & Haferkamp, I. (2009). Diatom plastids depend on nucleotide import from the cytosol. *Proceedings of the National Academy of Sciences*, *106*(9), 3621–3626. https://doi.org/10.1073/pnas.0808862106

Chavrier, P., Parton, R. G., Hauri, H. P., Simons, K., & Zerial, M. (1990). Localization of low molecular weight GTP binding proteins to exocytic and endocytic compartments. *Cell*, *62*(2), 317–329. https://doi.org/10.1016/0092-8674(90)90369-P

Christian, R. W., Hewitt, S. L., Nelson, G., Roalson, E. H., & Dhingra, A. (2020). Plastid transit peptides—where do they come from and where do they all belong? Multi-genome and pan-genomic assessment of chloroplast transit peptide evolution. *PeerJ*, *8*, e9772. https://doi.org/10.7717/peerj.9772

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. L. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Emanuelsson, O., Brunak, S., von Heijne, G., & Nielsen, H. (2007). Locating proteins in the cell using TargetP, SignalP and related tools. *Nature Protocols*, *2*(4), 953–971. https://doi.org/10.1038/nprot.2007.131

Felsner, G., Sommer, M. S., & Maier, U. G. (2010). The physical and functional borders of transit peptide-like sequences in secondary endosymbionts. *BMC Plant Biology*, *10*(1), 223. https://doi.org/10.1186/1471-2229-10-223

Field, C. B., Behrenfeld, M. J., Randerson, J. T., & Falkowski, P. (1998). Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science*, *281*(5374), 237–240. https://doi.org/10.1126/science.281.5374.237

Flori, S., Jouneau, P.-H., Finazzi, G., Maréchal, E., & Falconet, D. (2016). Ultrastructure of the Periplastidial Compartment of the Diatom Phaeodactylum tricornutum. *Protist*, *167*(3), 254–267. https://doi.org/10.1016/j.protis.2016.04.001

Füssy, Z., Faitová, T., & Oborník, M. (2019). Subcellular Compartments Interplay for Carbon and Nitrogen Allocation in Chromera velia and Vitrella brassicaformis. *Genome Biology and Evolution*, *11*(7), 1765–1779. https://doi.org/10.1093/gbe/evz123

Gruber, A., & Haferkamp, I. (2019). Nucleotide Transport and Metabolism in Diatoms. *Biomolecules*, *9*(12), 761. https://doi.org/10.3390/biom9120761

Gruber, A., & Kroth, P. G. (2017). Intracellular metabolic pathway distribution in diatoms and tools for genome-enabled experimental diatom research. *Philosophical Transactions* *of the Royal Society B: Biological Sciences*, *372*(1728), 20160402. https://doi.org/10.1098/rstb.2016.0402

Gruber, A., Rocap, G., Kroth, P. G., Armbrust, E. V., & Mock, T. (2015). Plastid proteome prediction for diatoms and other algae with secondary plastids of the red lineage. *The Plant Journal*, *81*(3), 519–528. https://doi.org/10.1111/tpj.12734

Gruber, A., Vugrinec, S., Hempel, F., Gould, S. B., Maier, U.-G., & Kroth, P. G. (2007). Protein targeting into complex diatom plastids: functional characterisation of a specific targeting motif. *Plant Molecular Biology*, *64*(5), 519–530. https://doi.org/10.1007/s11103-007-9171-x

Guruprasad, K., Reddy, B. V. B., & Pandit, M. W. (1990). Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting *in vivo* stability of a protein from its primary sequence. *"Protein Engineering, Design and Selection," 4*(2), 155–161. https://doi.org/10.1093/protein/4.2.155

Hirakawa, Y., Nagamune, K., & Ishida, K. (2009). Protein targeting into secondary plastids of chlorarachniophytes. *Proceedings of the National Academy of Sciences*, *106*(31), 12820–12825. https://doi.org/10.1073/pnas.0902578106

Huesgen, P. F., Alami, M., Lange, P. F., Foster, L. J., Schröder, W. P., Overall, C. M., & Green, B. R. (2013). Proteomic Amino-Termini Profiling Reveals Targeting Information for Protein Import into Complex Plastids. *PLoS ONE*, *8*(9), e74483. https://doi.org/10.1371/journal.pone.0074483

Janouškovec, J., Horák, A., Oborník, M., Lukeš, J., & Keeling, P. J. (2010). A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. *Proceedings of the National Academy of Sciences*, *107*(24), 10949–10954. https://doi.org/10.1073/pnas.1003335107

Kiraga, J., Mackiewicz, P., Mackiewicz, D., Kowalczuk, M., Biecek, P., Polak, N., Smolarczyk, K., Dudek, M. R., & Cebrat, S. (2007). The relationships between the isoelectric point and: length of proteins, taxonomy and ecology of organisms. *BMC Genomics*, *8*(1), 163. https://doi.org/10.1186/1471-2164-8-163

Kořený, L., Sobotka, R., Janouškovec, J., Keeling, P. J., & Oborník, M. (2011). Tetrapyrrole Synthesis of Photosynthetic Chromerids Is Likely Homologous to the Unusual Pathway of Apicomplexan Parasites . *The Plant Cell*, *23*(9), 3454–3462. https://doi.org/10.1105/tpc.111.089102

Lang, M., Apt, K. E., & Kroth, P. G. (1998). Protein Transport into "Complex" Diatom Plastids Utilizes Two Different Targeting Signals. *Journal of Biological Chemistry*, *273*(47), 30973–30978. https://doi.org/10.1074/jbc.273.47.30973

Lipman, D. J., Souvorov, A., Koonin, E. v, Panchenko, A. R., & Tatusova, T. A. (2002). The relationship of protein conservation and sequence length. *BMC Evolutionary Biology*, *2*(1), 20. https://doi.org/10.1186/1471-2148-2-20

Magdeldin, S., Yoshida, Y., Li, H., Maeda, Y., Yokoyama, M., Enany, S., Zhang, Y., Xu, B., Fujinaka, H., Yaoita, E., Sasaki, S., & Yamamoto, T. (2012). Murine colon proteome and characterization of the protein pathways. *BioData Mining*, *5*(1), 11. https://doi.org/10.1186/1756-0381-5-11

Nair, R., & Rost, B. (2009). Sequence conserved for subcellular localization. *Protein Science*, *11*(12), 2836–2847. https://doi.org/10.1110/ps.0207402

Oborník, M., Modrý, D., Lukeš, M., Černotíková-Stříbrná, E., Cihlář, J., Tesařová, M., Kotabová, E., Vancová, M., Prášil, O., & Lukeš, J. (2012). Morphology, Ultrastructure and Life Cycle of Vitrella brassicaformis n. sp., n. gen., a Novel Chromerid from the Great Barrier Reef. *Protist*, *163*(2), 306–323. https://doi.org/10.1016/j.protis.2011.09.001

Oudot-Le Secq, M.-P., Grimwood, J., Shapiro, H., Armbrust, E. V., Bowler, C., & Green, B. R. (2007). Chloroplast genomes of the diatoms Phaeodactylum tricornutum and Thalassiosira pseudonana: comparison with other plastid genomes of the red lineage. *Molecular Genetics and Genomics*, *277*(4), 427–439. https://doi.org/10.1007/s00438-006-0199-4

Owji, H., Nezafat, N., Negahdaripour, M., Hajiebrahimi, A., & Ghasemi, Y. (2018). A comprehensive review of signal peptides: Structure, roles, and applications. *European Journal of Cell Biology*, *97*(6), 422–441. https://doi.org/10.1016/J.EJCB.2018.06.003

Requião, R. D., Fernandes, L., de Souza, H. J. A., Rossetto, S., Domitrovic, T., & Palhano, F. L. (2017). Protein charge distribution in proteomes and its impact on translation. *PLOS Computational Biology*, *13*(5), e1005549. https://doi.org/10.1371/journal.pcbi.1005549

Schober, A. F., Rï¿½o Bï¿½rtulos, C., Bischoff, A., Lepetit, B., Gruber, A., & Kroth, P. G. (2019). Organelle Studies and Proteome Analyses of Mitochondria and Plastids Fractions from the Diatom Thalassiosira pseudonana. *Plant and Cell Physiology*, *60*(8), 1811–1828. https://doi.org/10.1093/pcp/pcz097

Soll, J., & Schleiff, E. (2004). Protein import into chloroplasts. *Nature Reviews Molecular Cell Biology*, *5*(3), 198–208. https://doi.org/10.1038/nrm1333

Tardif, M., Atteia, A., Specht, M., Cogne, G., Rolland, N., Brugière, S., Hippler, M., Ferro, M., Bruley, C., Peltier, G., Vallon, O., & Cournac, L. (2012). PredAlgo: A New Subcellular Localization Prediction Tool Dedicated to Green Algae. *Molecular Biology and Evolution*, *29*(12), 3625–3639. https://doi.org/10.1093/molbev/mss178

Trentmann, O., Jung, B., Neuhaus, H. E., & Haferkamp, I. (2008). Nonmitochondrial ATP/ADP Transporters Accept Phosphate as Third Substrate. *Journal of Biological Chemistry*, *283*(52), 36486–36493. https://doi.org/10.1074/jbc.M806903200

Vetter, I. R., & Wittinghofer, A. (2001). The Guanine Nucleotide-Binding Switch in Three Dimensions. *Science*, *294*(5545), 1299–1304. https://doi.org/10.1126/science.1062023

Weatherby, K., & Carter, D. (2013). *Chromera velia* (pp. 119–144). https://doi.org/10.1016/B978-0-12-407672-3.00004-6

Winkler, H. H., & Neuhaus, H. E. (1999). Non-mitochondrial ATP transport. *Trends in Biochemical Sciences*, *24*(2), 64–68. https://doi.org/10.1016/S0968-0004(98)01334-6

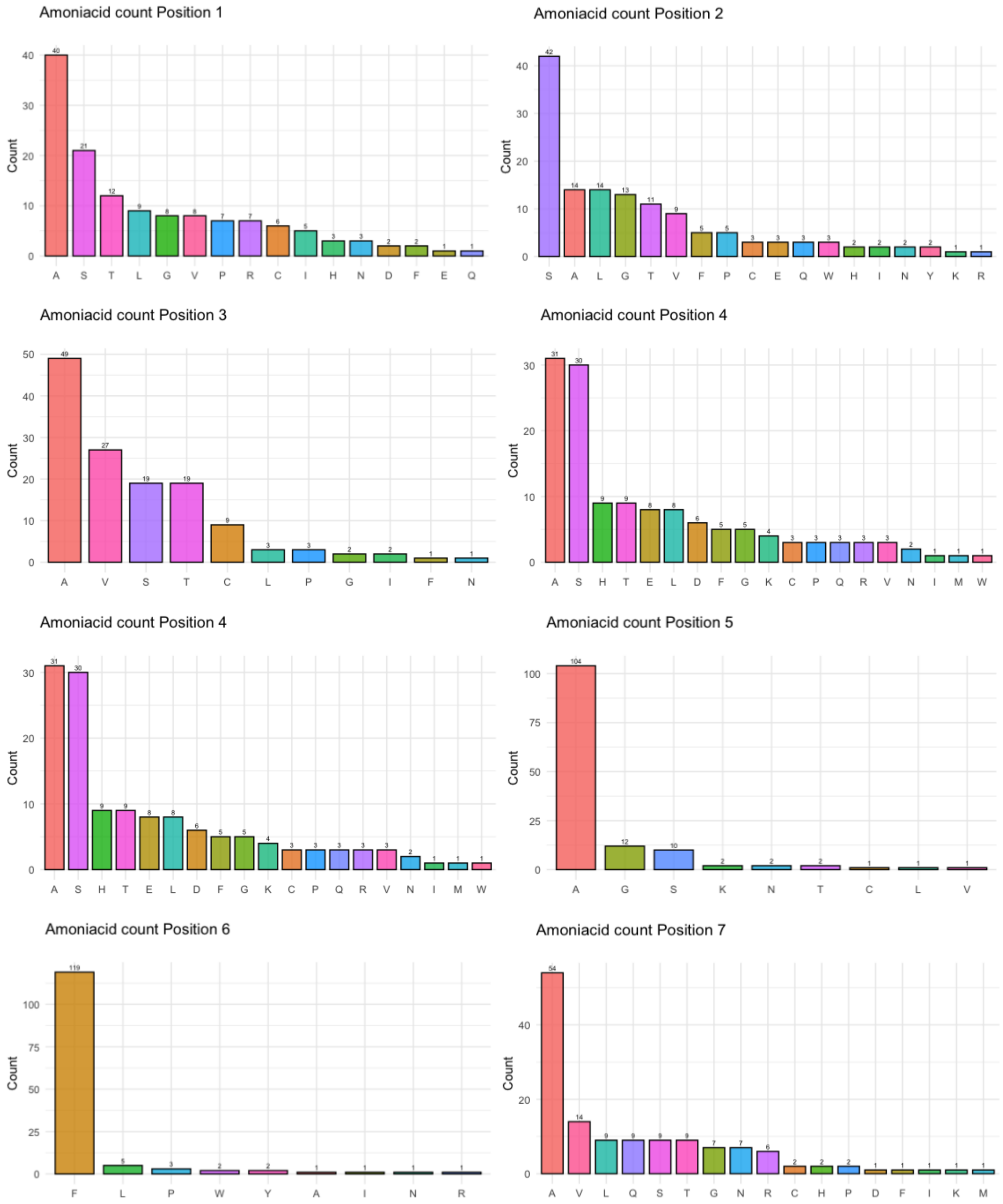
Witz, S., Jung, B., Fürst, S., & Möhlmann, T. (2012). De Novo Pyrimidine Nucleotide Synthesis Mainly Occurs outside of Plastids, but a Previously Undiscovered Nucleobase Importer Provides Substrates for the Essential Salvage Pathway in *Arabidopsis*. *The Plant Cell*, *24*(4), 1549–1559. https://doi.org/10.1105/tpc.112.096743

Wojcik, S., & Kriechbaumer, V. (2021). Go your own way: membrane-targeting sequences. *Plant Physiology*, *185*(3), 608–618. https://doi.org/10.1093/plphys/kiaa058

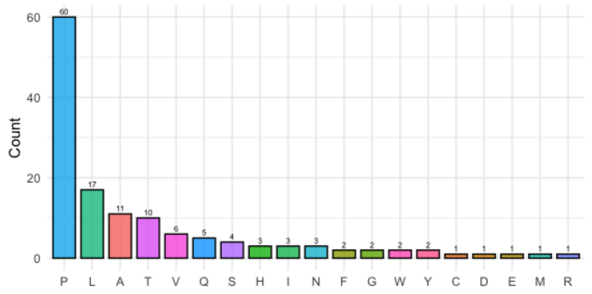
Additional Figures

Figure 15

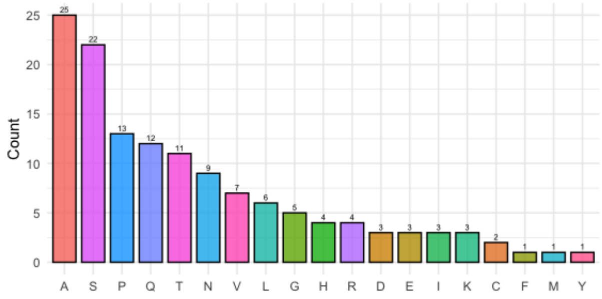
This graph represents amino acid frequency in each position of the amino acid sequences predicted by ASAFind. As we can see in 1st to 6th position the ASAFAP motif and 14th position a increase of Arginine residue.



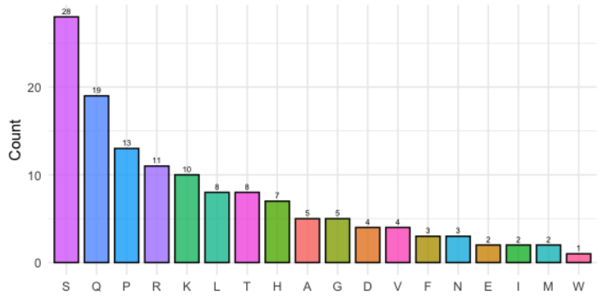
Amoniacid count Position 8



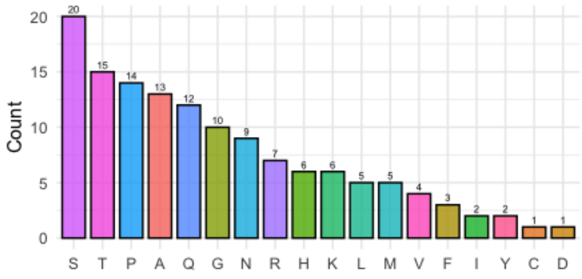
Amoniacid count Position 9



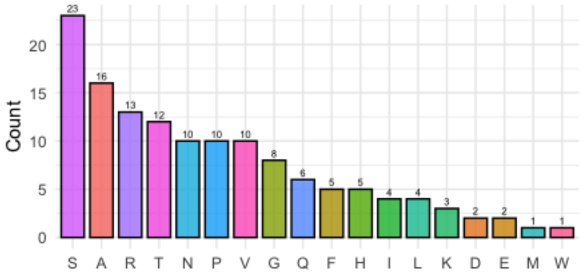
Amoniacid count Position 10



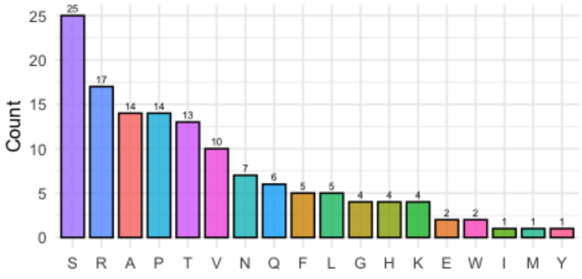
Amoniacid count Position 11



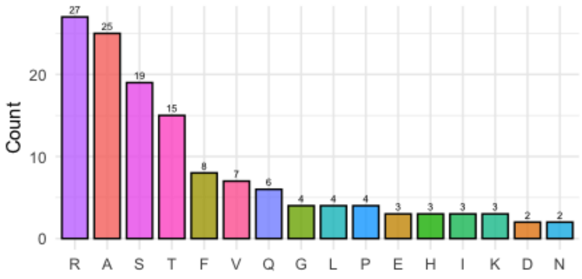
Amoniacid count Position 12



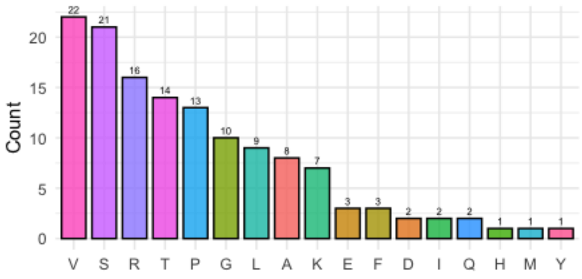
Amoniacid count Position 13



Amoniacid count Position 14



Amoniacid count Position 15



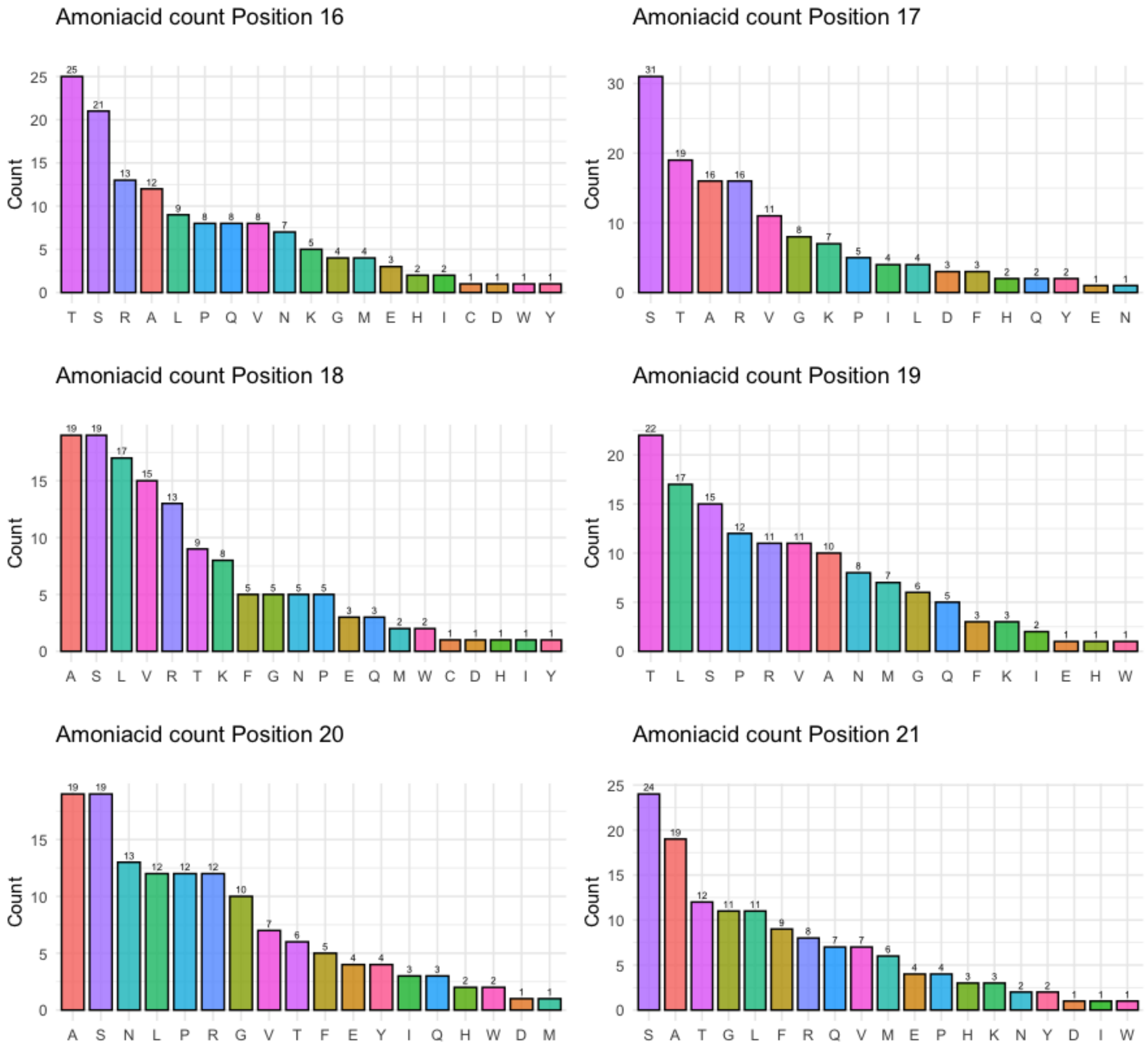
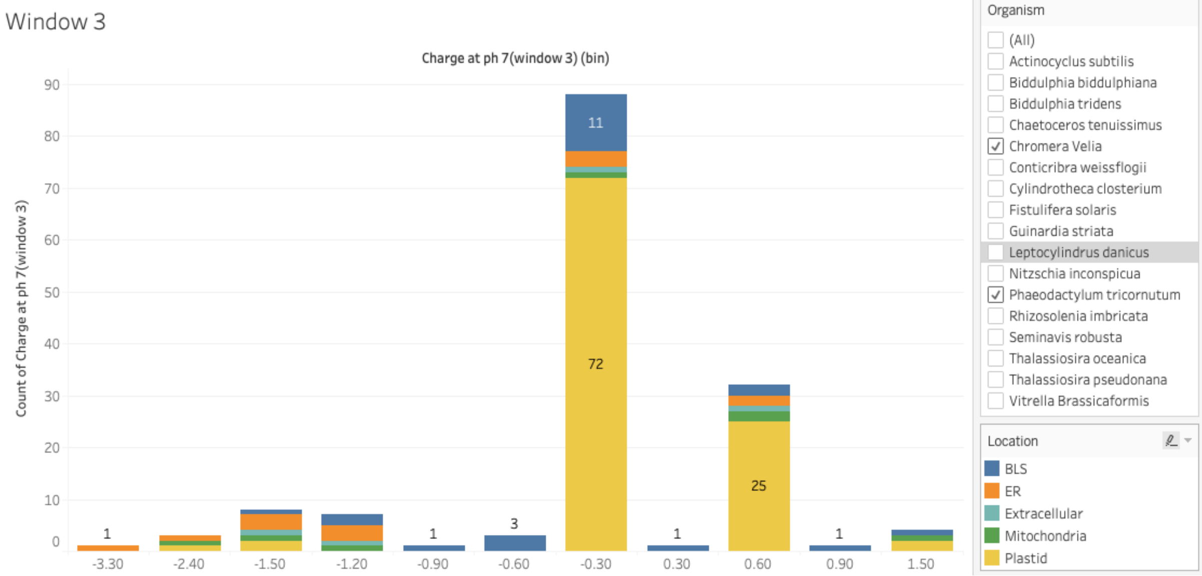


Figure 16



Organism

□ (All)
□ Actinocyclus subtilis
□ Biddulphia biddulphiana
□ Biddulphia tridens
□ Chaetoceros tenuissimus
☑ Chromera Velia
□ Conticribra weissflogii
□ Cylindrotheca closterium
□ Fistulifera solaris
□ Guinardia striata
□ Leptocylindrus danicus
□ Nitzschia inconspicua
☑ Phaeodactylum tricornutum
□ Rhizosolenia imbricata
□ Seminavis robusta
□ Thalassiosira oceanica
□ Thalassiosira pseudonana
□ Vitrella Brassicaformis

Location

- 🟦 BLS
- 🟧 ER
- 🟩 Extracellular
- 🟩 Mitochondria
- 🟨 Plastid

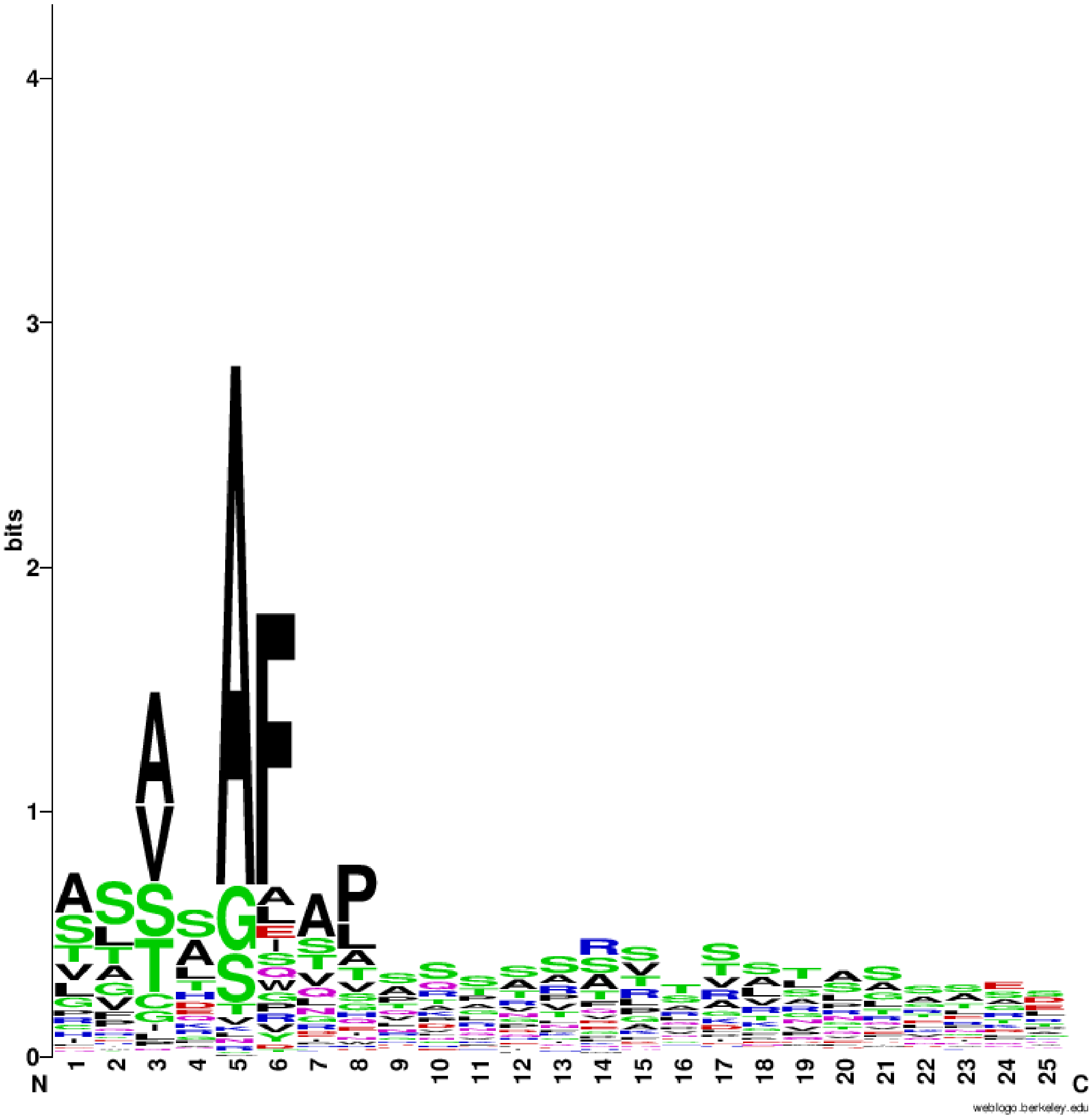


Figure 17
Web logo created with all the sequences from the merged dataset. ASAFAP motif is observed and at 14th position an increase of charged residue is present.

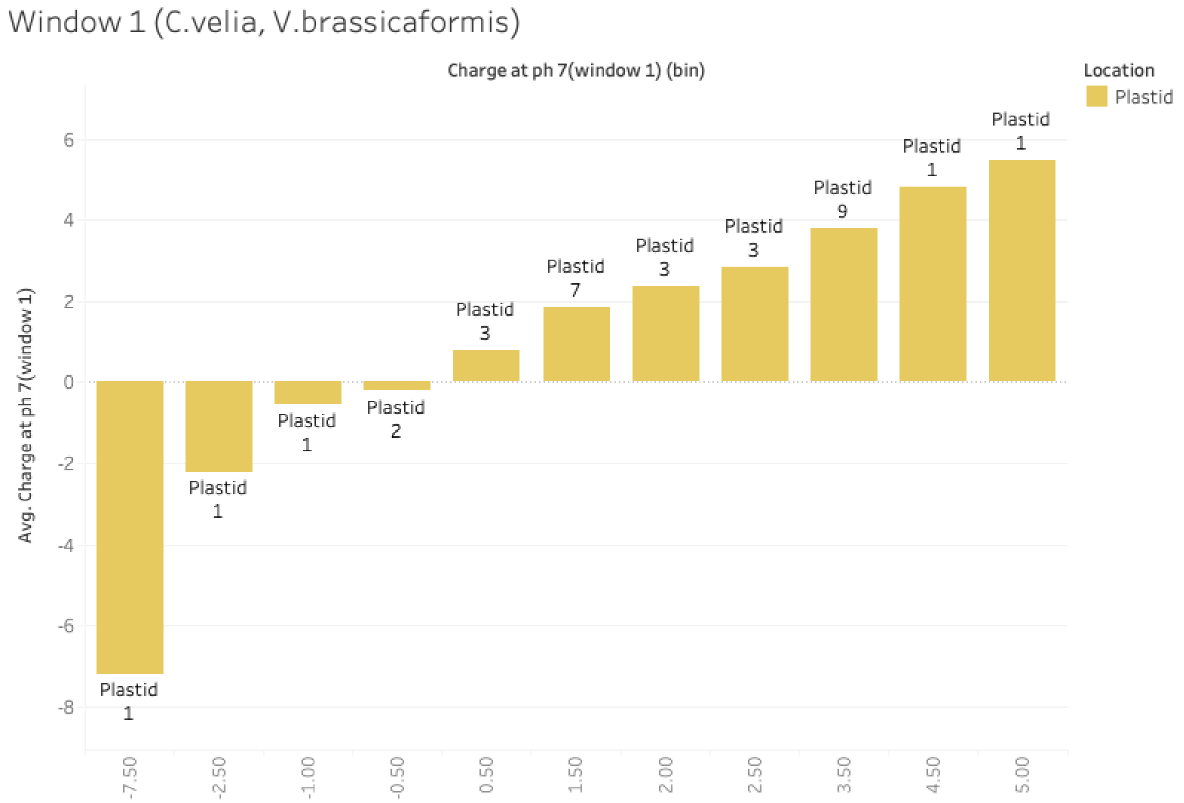


Figure 18: Net charge on window 1 filtered by *C. velia* and *V. brassicaformis*. Here, out of 32 plastid targeted sequences only 5 sequences net charge are negative remaining 27 sequences are positive.

8. Appendices

Following, scripts are written in Jupyter notebook and Rstudio. Some packages need to be installed in the shell environment as well as in R studio before running the scripts. Appendix 1 to 7 is to analyze the datasets. These scripts are used for all three datasets. Finally, these datasets are merged to make one dataset. Remaining scripts are for statistical analysis in Rstudio.

Appendix 1

```python
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
import numpy as np

from scipy.stats import chi2_contingency
from scipy.stats import chi2


from Bio.Seq import Seq
from Bio import SeqIO
from Bio.SeqUtils.ProtParam import ProteinAnalysis

# Reading dataset, Output will return the whole dataset with a table
ref_data = pd.read_csv("Prediction result for native reference sequence org.csv", sep = ";")
ref_data.sample(5)

```

Appendix 2

```python
# The list of columns
for col in ref_data.columns:
    print(col)
# Selecting the columns i need.
df = ref_data[["Name","Location","Organism","SignalP 25aa sequence","ASAFind 25aa sequence","ASAFind cleavage has [FWYL] at +1 position","Protein sequence"]]
#df.dtypes
#print(df["Protein sequence"][:])

# Add additional column with the length of the protein sequences

df["Protein length"] = df.loc[:,("Protein sequence")].str.len()
new_df =df.copy()
```

Appendix 3

```python
# The * represents a stop codons of a protein sequence. therefore, there should be a stop codons end of each
# mRNA sequence. we must delete that in order to protein analysis.
# fowllowing two lines of code is basically replacing the * with white space.
# we can use anything or any character to be replaced by one another.

new_df['SignalP 25aa sequence'] = new_df['SignalP 25aa sequence'].str.replace("*", '')
new_df['ASAFind 25aa sequence'] = new_df['ASAFind 25aa sequence'].str.replace("*", '')
```

```
# After removing all the unambiguous letters, a final check is necessary, this snippet will check if there is any.

def unambigous_checker(protein):
    Unambiguous_Letters=['*', 'Z', 'X', 'J']  # unvalid unambiguous letter for protein
    Letters=list(protein)                     # get the letters from the protein
    dic={}
    checker=1
    for acid in Letters:
        if acid in Unambiguous_Letters:
            dic[acid]=0

        else:
            dic[acid]=1


        checker= checker*dic[acid]
    return checker


N=len(df["Protein sequence"])
check=np.zeros(N)
for i in range(N):
    check[i]=unambigous_checker(df["Protein sequence"][i])


print('There is exactly ', N-sum(check), 'protein(s) with at least one unambigous letter')

```

Appendix 4

```
# Here these lists will contain all the sequence windows

analysed_seq = [] #list to save the all ProteinAnalysis for specific windows

all_prot = (df["Protein sequence"][:])
for i in all_prot:
    analysed_seq.append(ProteinAnalysis(i[:-2]))
#analysed_seq


# In following codes snippets we define multiple sequence window
#List for whole sequence
all_prot_signalP = []
signalP_prot = (df["ASAFind 25aa sequence"][:])
for k in signalP_prot:
    all_prot_signalP.append(ProteinAnalysis(k))
#List for Window 1
signalP_prot_fifteen_seq = []
for m in signalP_prot:
    signalP_prot_fifteen_seq.append(ProteinAnalysis(m[4:]))
# List for window 2
signalP_prot_window_2 = []
for m in signalP_prot:
    signalP_prot_window_2.append(ProteinAnalysis(m[0:-20]))
# List for window 3
signalP_prot_window_small = []
for m in signalP_prot:
    signalP_prot_window_small.append(ProteinAnalysis(m[5:-15]))

```

```python
# These code snippets will analyze and calculate all the features for each sequence window and save them in lists

charges = []
molecular_weight = []
isoelectric_point= []
aromaticity = []
instability_index = []
gravy = []

monoisotopic = []

for j in analysed_seq:
    charges.append(j.charge_at_pH(7.0))

    molecular_weight.append(j.molecular_weight())
    isoelectric_point.append(j.isoelectric_point())

    instability_index.append(j.instability_index())
    gravy.append(j.gravy())


    aromaticity.append(j.aromaticity())
    monoisotopic.append(j.monoisotopic)

charges_signalP = []
molecular_weight_signalP = []
isoelectric_point_signalP = []
aromaticity_signalP = []
instability_index_signalP = []
gravy_signalP = []

#monoisotopic_signalP = []

for j in all_prot_signalP:
    charges_signalP.append(j.charge_at_pH(7.0))
    #print(f"Its charge at pH 7 is {j.charge_at_pH(7.0):.2f}")
    molecular_weight_signalP.append(j.molecular_weight())
    isoelectric_point_signalP.append(j.isoelectric_point())
    aromaticity_signalP.append(j.aromaticity())
    instability_index_signalP.append(j.instability_index())
    gravy_signalP.append(j.gravy())
    #monoisotopic_signalP.append(j.monoisotopic)


charges_plus1_pos = []
molecular_plus1_pos = []
isoelectric_plus1_pos = []
aromaticity_plus1_pos = []
instability_index_plus1_pos = []
gravy_plus1_pos = []


#monoisotopic_signalP = []

for l in signalP_prot_fifteen_seq:
    charges_plus1_pos.append(l.charge_at_pH(7.0))
    #print(f"Its charge at pH 7 is {l.charge_at_pH(7.0):.2f}")
    molecular_plus1_pos.append(l.molecular_weight())
    isoelectric_plus1_pos.append(l.isoelectric_point())
    aromaticity_plus1_pos.append(l.aromaticity())
    instability_index_plus1_pos.append(l.instability_index())
    gravy_plus1_pos.append(l.gravy())
    #monoisotopic_signalP.append(l.monoisotopic)

```

```
charges_window_2 = []
molecular_window_2 = []
isoelectric_window_2 = []
aromaticity_window_2 = []
instability_window_2 = []
gravy_window_2 = []
monoisotopic_window_2 = []

for l in signalP_prot_window_2:
    charges_window_2.append(l.charge_at_pH(7.0))
    #print(f"Its charge at pH 7 is {l.charge_at_pH(7.0):.2f}")
    molecular_window_2.append(l.molecular_weight())
    isoelectric_window_2.append(l.isoelectric_point())
    aromaticity_window_2.append(l.aromaticity())
    instability_window_2.append(l.instability_index())
    gravy_window_2.append(l.gravy())
    monoisotopic_window_2.append(l.monoisotopic)

charges_window_small = []
molecular_window_small = []
isoelectric_window_small = []
aromaticity_window_small = []
instability_window_small = []
gravy_window_small = []
monoisotopic_window_small = []


for l in signalP_prot_window_small:
    charges_window_small.append(l.charge_at_pH(7.0))
    #print(f"Its charge at pH 7 is {l.charge_at_pH(7.0):.2f}")
    molecular_window_small.append(l.molecular_weight())
    isoelectric_window_small.append(l.isoelectric_point())
    aromaticity_window_small.append(l.aromaticity())
    instability_window_small.append(l.instability_index())
    gravy_window_small.append(l.gravy())
    monoisotopic_window_small.append(l.monoisotopic)
```

Appendix 5

```
#Stability catagory of whole sequence
stability_whole_seq = []
for s in instability_index:
    if s < 40:
        stability_whole_seq.append("Stable")
    else:
        stability_whole_seq.append("Not Stable")

#Stability category for signalp Sequences
stability_signalP = []
for s in instability_index_signalP:
    if s < 40:
        #s == "Stable"
        stability_signalP.append("Stable")
    else:
        stability_signalP.append("Not Stable")
```

```
#Stability catagory of +1 pos of 25aa sequence
stability_signalP_Plus1_pos = []
for s in instability_index_plus1_pos:
    if s < 40:
        stability_signalP_Plus1_pos.append("Stable")
    else:
        stability_signalP_Plus1_pos.append("Not Stable")


#Stability catagory of +1 pos of 25aa sequence
stability_signalP_window_2 = []
for s in instability_window_2:
    if s < 40:
        stability_signalP_window_2.append("Stable")
    else:
        stability_signalP_window_2.append("Not Stable")


#This code represents categorization of iso electric point of whole sequence
iso_electric_point_category_wholeseq = []
for ip in isoelectric_point:
    if ip <= 4:
        iso_electric_point_category_wholeseq.append("Less than 4")
    if ip > 4 and ip <= 5:
        iso_electric_point_category_wholeseq.append("4 upto 5")
    if ip > 5 and ip <= 8:
        iso_electric_point_category_wholeseq.append("5 upto 8")
    if ip > 8:
        iso_electric_point_category_wholeseq.append("Greater than 8")

#This code represents categorization of iso electric point of 25 signalp sequence
iso_electric_point_cat = []
for ip in isoelectric_plus1_pos:
    if ip <= 4:
        iso_electric_point_cat.append("Less than 4")
    if ip > 4 and ip <= 6:
        iso_electric_point_cat.append("4 to 6")
    if ip > 6 and ip <= 8:
        iso_electric_point_cat.append("6 to 8")
    if ip > 8:
        iso_electric_point_cat.append("Greater than 8")

#This code represents categorization of iso electric point of the last window we selected of sequence
iso_electric_point_cat_window_2 = []
for ip in isoelectric_window_2:
    if ip <= 4:
        iso_electric_point_cat_window_2.append("Less than 4")
    if ip > 4 and ip <= 5:
        iso_electric_point_cat_window_2.append("4 upto 5")
    if ip > 5 and ip <= 8:
        iso_electric_point_cat_window_2.append("5 upto 8")
    if ip > 8:
        iso_electric_point_cat_window_2.append("Greater than 8")
```

```
# Category of aromaticity of signalP sequences
aroma_signalP = []
for arr in aromaticity_signalP:
    if arr < .04:
        aroma_signalP.append("0")
    if arr == .04:
        aroma_signalP.append(".04")
    if arr > .04 and arr <= .08 :
        aroma_signalP.append("0.08")
    if arr > .08 and arr <= .12:
        aroma_signalP.append(".12 ")
    if arr > .12:
        aroma_signalP.append(".12")

```

Appendix 6

```
# back to our original dataframe
# we need to assign the features to the dataframe

charge_list = pd.Series(charges)
df["Charge at ph 7(whole sequence)"] = charge_list.values

charge_list_signalP = pd.Series(charges_signalP)
df["Charge at ph 7(ASAP 25aa)"] = charge_list_signalP.values

charges_plus1_pos = pd.Series(charges_plus1_pos)
df["Charge at ph 7(+1 poition to 20th)"] = charges_plus1_pos.values

charges_window_2 = pd.Series(charges_window_2)
df["Charge at ph 7(charges_window_2)"] = charges_window_2.values

charges_window_small = pd.Series(charges_window_small)
df["Charge at ph 7(charges_window_small)"] = charges_window_small.values



molecular_weight = pd.Series(molecular_weight)
df["MW Whole seq"] = molecular_weight.values

molecular_weight_signalP = pd.Series(molecular_weight_signalP)
df["MW ASAP 25seq"] = molecular_weight_signalP.values

molecular_plus1_pos = pd.Series(molecular_plus1_pos)
df["MW ASAP +1 pos"] = molecular_plus1_pos.values

molecular_window_2 = pd.Series(molecular_window_2)
df["MW ASAP_window_2"] = molecular_window_2.values

molecular_window_small = pd.Series(molecular_window_small)
df["MW ASAP_window_small"] = molecular_window_small.values

```

```python
isoelectric_point = pd.Series(isoelectric_point)
df["Iso electric point (whole sequence)"] = isoelectric_point.values

isoelectric_point_signalP = pd.Series(isoelectric_point_signalP)
df["Iso electric point (ASAP 25aa)"] = isoelectric_point_signalP.values

isoelectric_plus1_pos = pd.Series(isoelectric_plus1_pos)
df["Iso electric point (+1 poition to 20th)"] = isoelectric_plus1_pos.values

isoelectric_window_2 = pd.Series(isoelectric_window_2)
df["Iso electric point_window_2"] = isoelectric_window_2.values

isoelectric_window_small = pd.Series(isoelectric_window_small)
df["Iso electric point_window_small"] = isoelectric_window_small.values



iso_electric_point_category_wholeseq = pd.Series(iso_electric_point_category_wholeseq)
df["Iso electric point Category"] = iso_electric_point_category_wholeseq.values


iso_electric_point_cat_window_2 = pd.Series(iso_electric_point_cat_window_2)
df["Iso electric point Cat_window_2"] = iso_electric_point_cat_window_2.values



#Aromaticity Calculates the aromaticity value of a protein according to Lobry,
# It is simply the relative frequency of Phe+Trp+Tyr.

aromaticity = pd.Series(aromaticity)
df["Aromaticity (whole seq)"] = aromaticity.values

aromaticity_signalP = pd.Series(aromaticity_signalP)
df["Aromaticity ASAP(25aa)"] = aromaticity_signalP.values

aromaticity_plus1_pos = pd.Series(aromaticity_plus1_pos)
df["Aromaticity plus1_pos to 20th"] = aromaticity_plus1_pos.values

#aromaticity_category = pd.Series(aroma_signalP)
#df["Aromaticity ASAP category"] = aromaticity_category.values

aromaticity_window_2 = pd.Series(aromaticity_window_2)
df["Aromaticity ASAP window_2"] = aromaticity_window_2.values

aromaticity_window_small = pd.Series(aromaticity_window_small)
df["Aromaticity ASAP window_small"] = aromaticity_window_small.values

```

```
# Calculate the instability index according to Guruprasad et al 1990.
# Any value above 40 means the protein is unstable (has a short half life).


instability_index = pd.Series(instability_index)
df["Instability index(whole seq)"] = instability_index.values

instability_index_signalP = pd.Series(instability_index_signalP)
df["Instability index ASAP(25aa)"] = instability_index_signalP.values

instability_index_plus1_pos = pd.Series(instability_index_plus1_pos)
df["Instability index plus1_pos to 20th"] = instability_index_plus1_pos.values

instability_window_2 = pd.Series(instability_window_2)
df["Instability index_window_2"] = instability_window_2.values

instability_window_small = pd.Series(instability_window_small)
df["Instability index_window_small"] = instability_window_small.values



stability_whole_seq = pd.Series(stability_whole_seq)
df["Stability whole seq"] = stability_whole_seq.values

stability_signalP = pd.Series(stability_signalP)
df["Stability ASAP(25aa)"] = stability_signalP.values

stability_signalP_Plus1_pos = pd.Series(stability_signalP_Plus1_pos)
df["Stability +1 pos"] = stability_signalP_Plus1_pos.values

stability_signalP_window_2 = pd.Series(stability_signalP_window_2)
df["Stability signalP_window_2"] = stability_signalP_window_2.values



gravy = pd.Series(gravy)
df["Gravy Whole seq"] = gravy.values

gravy_signalP = pd.Series(gravy_signalP)
df["Gravy ASAP(25aa)"] = gravy_signalP.values

gravy_plus1_pos = pd.Series(gravy_plus1_pos)
df["Gravy plus1_pos"] = gravy_plus1_pos.values

gravy_window_2 = pd.Series(gravy_window_2)
df["Gravy window_2"] = gravy_window_2.values

```

Appendix 7

```
final_df = df[[
         'Name',
         'Organism',
         'Location',

         #'Charge Bin new_window2',
         #'Charge Bin',
         #'Charge Bin new',
         'MW ASAP +1 pos',
         'MW ASAP_window_2',
         'MW ASAP_window_small',
         'Charge at ph 7(whole sequence)',
         'Charge at ph 7(ASAP 25aa)',
         'Charge at ph 7(+1 poition to 20th)',
         'Charge at ph 7(charges_window_2)',
         'Charge at ph 7(charges_window_small)',
         'ASAFind cleavage has [FWYL] at +1 position',
         'ASAFind 25aa sequence',


         'Iso electric point (whole sequence)',
         'Iso electric point (ASAP 25aa)',
         'Iso electric point (+1 poition to 20th)',
         'Iso electric point_window_2',
         'Iso electric point_window_small',
         #'Iso electric point Category',
         #'Iso electric point Cat_window_2',
         'Aromaticity ASAP(25aa)',
         'Aromaticity (whole seq)',
         'Aromaticity plus1_pos to 20th',
         'Aromaticity ASAP window_2',
         'Aromaticity ASAP window_small',

         'Instability index(whole seq)',
         'Instability index ASAP(25aa)',
         'Instability index plus1_pos to 20th',
         'Instability index_window_2',
         'Instability index_window_small',

         'Stability whole seq',
         'Stability ASAP(25aa)',
         'Stability signalP_window_2',
         'Stability ASAP(25aa)',
         'Stability +1 pos',

         'Gravy Whole seq',
         'Gravy ASAP(25aa)',
         'Gravy plus1_pos',
         'Gravy window_2',
         ]]
file_name = "Project_output_for_thesis_final0000001.csv"
final_df.to_csv(file_name)
file_name = "Project_output_for_thesis_final0000001.xlsx"
final_df.to_excel(file_name)

```

Appendix 8

First we select the relevant columns (charges and Location) and filter the data for Plastid and then pivot from wide to long before plotting the boxplot. The red dot represents the mean as addition to the median.

```
data_comp <- data %>%
 dplyr::select(Location, `Charge at ph 7(ASAP 25aa)`,
               `Charge at ph 7(charges_window_2)`,
               `Charge at ph 7(+1 poition to 20th)`,
               `Charge at ph 7(charges_window_small)`)
data_plast <- data_comp %>%
 filter(Location == "Plastid")
# pivot from wide to long
data_plast2 <- data_plast %>%
 pivot_longer(!Location, names_to = "Charges", values_to = "value")

# boxplot

data_plast2 %>%
 ggplot(aes(x = Charges , y = value, fill = Charges)) +
 geom_boxplot(outlier.colour="black", alpha = 0.5) +
 stat_summary(fun=mean, geom="point", shape=20, size=3, color="red", fill="red") +
 labs(title="Difference in charges for Plastid",
      x=" ", y = "") +
 theme_bw() +
  theme(legend.position = "right",
    axis.title.x=element_blank(),
    axis.text.x=element_blank(),
    axis.ticks.x=element_blank()) +
guides(fill=guide_legend(nrow=6,byrow=TRUE, reverse=FALSE))
```

Appendix 9

```
# ANOVA model

plastid_aov <- aov(value ~ Charges,
 data = data_plast2

)

# summary
summary(plastid_aov) # we can see from summary that there are significant differences
```

Appendix 10

```
# display it as a data frame and perform post hoc analysis
plastid_aov <- data_plast2 %>%
 tukey_hsd(value ~ Charges)

plastid_aov
```

Appendix 11

To see the position of each letter we create a new column (x1, x2 , x3 ...) and simply split the column off at the first location (first letter), the second location etc. so we are left with the single Letter from each position in a new column.

```
# glimpse(data)

data1 <- data

# create new columns with each respective letter from position 1, 2, 3 and so on ...
data1 <- data1 %>%
 mutate(x1 = substr(data1$`ASAFind 25aa sequence`, start = 1, stop = 1),
       x2 = substr(data1$`ASAFind 25aa sequence`, start = 2, stop = 2),
       x3 = substr(data1$`ASAFind 25aa sequence`, start = 3, stop = 3),
       x4 = substr(data1$`ASAFind 25aa sequence`, start = 4, stop = 4),
       x5 = substr(data1$`ASAFind 25aa sequence`, start = 5, stop = 5),
       x6 = substr(data1$`ASAFind 25aa sequence`, start = 6, stop = 6),
       x7 = substr(data1$`ASAFind 25aa sequence`, start = 7, stop = 7),
       x8 = substr(data1$`ASAFind 25aa sequence`, start = 8, stop = 8),
       x9 = substr(data1$`ASAFind 25aa sequence`, start = 9, stop = 9),
       x10 = substr(data1$`ASAFind 25aa sequence`, start = 10, stop = 10),
       x11 = substr(data1$`ASAFind 25aa sequence`, start = 11, stop = 11),
       x12 = substr(data1$`ASAFind 25aa sequence`, start = 12, stop = 12),
       x13 = substr(data1$`ASAFind 25aa sequence`, start = 13, stop = 13),
       x14 = substr(data1$`ASAFind 25aa sequence`, start = 14, stop = 14),
       x15 = substr(data1$`ASAFind 25aa sequence`, start = 15, stop = 15),
       x16 = substr(data1$`ASAFind 25aa sequence`, start = 16, stop = 16),
       x17 = substr(data1$`ASAFind 25aa sequence`, start = 17, stop = 17),
       x18 = substr(data1$`ASAFind 25aa sequence`, start = 18, stop = 18),
       x19 = substr(data1$`ASAFind 25aa sequence`, start = 19, stop = 19),
       x20 = substr(data1$`ASAFind 25aa sequence`, start = 20, stop = 20),
       x21 = substr(data1$`ASAFind 25aa sequence`, start = 21, stop = 21),
       x22 = substr(data1$`ASAFind 25aa sequence`, start = 22, stop = 22),
       x23 = substr(data1$`ASAFind 25aa sequence`, start = 23, stop = 23),
       )
```

Appendix 12

Last step we create a count of the letters and visualize it in a descending bar chart for each new column.

Position 1.

```
data1 %>%
  dplyr::count(x1) %>%
  ggplot(aes(x = reorder(x1,desc(n)), y = n, fill = x1)) +
  geom_bar(stat = "identity", position = "dodge", color = "black", alpha = 0.8, width = 0.
  geom_text(aes(label=n), position=position_dodge(width=0.8), vjust= -0.5, size = 2) +
    labs(title="Amoniacid count Position 1",
        subtitle = "",
           x=" ", y = "Count") +
  theme_minimal() +
  theme(legend.position = "none")
```

Position 2.

```
data1 %>%
 dplyr::count(x2) %>%
 ggplot(aes(x = reorder(x2,desc(n)), y = n, fill = x2)) +
 geom_bar(stat = "identity", position = "dodge", color = "black", alpha = 0.8, width = 0.
 geom_text(aes(label=n), position=position_dodge(width=0.8), vjust= -0.5, size = 2) +
  labs(title="Amoniacid count Position 2",
    subtitle = "",
      x=" ", y = "Count") +
 theme_minimal() +
 theme(legend.position = "none")
```

Position 3.

```
data1 %>%
 dplyr::count(x3) %>%
 ggplot(aes(x = reorder(x3,desc(n)), y = n, fill = x3)) +
 geom_bar(stat = "identity", position = "dodge", color = "black", alpha = 0.8, width = 0.
 geom_text(aes(label=n), position=position_dodge(width=0.8), vjust= -0.5, size = 2) +
  labs(title="Amoniacid count Position 3",
    subtitle = "",
      x=" ", y = "Count") +
 theme_minimal() +
 theme(legend.position = "none")
```

Position 4.

```
data1 %>%
 dplyr::count(x4) %>%
 ggplot(aes(x = reorder(x4,desc(n)), y = n, fill = x4)) +
 geom_bar(stat = "identity", position = "dodge", color = "black", alpha = 0.8, width = 0.
 geom_text(aes(label=n), position=position_dodge(width=0.8), vjust= -0.5, size = 2) +
  labs(title="Amoniacid count Position 4",
    subtitle = "",
      x=" ", y = "Count") +
 theme_minimal() +
 theme(legend.position = "none")
```

Position 5.

```
data1 %>%
 dplyr::count(x5) %>%
 ggplot(aes(x = reorder(x5,desc(n)), y = n, fill = x5)) +
 geom_bar(stat = "identity", position = "dodge", color = "black", alpha = 0.8, width = 0.
 geom_text(aes(label=n), position=position_dodge(width=0.8), vjust= -0.5, size = 2) +
  labs(title="Amoniacid count Position 5",
    subtitle = "",
      x=" ", y = "Count") +
 theme_minimal() +
 theme(legend.position = "none")
```

Position 6.

```
data1 %>%
 dplyr::count(x6) %>%
 ggplot(aes(x = reorder(x6,desc(n)), y = n, fill = x6)) +
 geom_bar(stat = "identity", position = "dodge", color = "black", alpha = 0.8, width = 0.
 geom_text(aes(label=n), position=position_dodge(width=0.8), vjust= -0.5, size = 2) +
  labs(title="Amoniacid count Position 6",
     subtitle = "",
       x=" ", y = "Count") +
 theme_minimal() +
 theme(legend.position = "none")
```

Position 7.

```
data1 %>%
 dplyr::count(x7) %>%
 ggplot(aes(x = reorder(x7,desc(n)), y = n, fill = x7)) +
 geom_bar(stat = "identity", position = "dodge", color = "black", alpha = 0.8, width = 0.
 geom_text(aes(label=n), position=position_dodge(width=0.8), vjust= -0.5, size = 2) +
  labs(title="Amoniacid count Position 7",
     subtitle = "",
       x=" ", y = "Count") +
 theme_minimal() +
 theme(legend.position = "none")
```

Position 8.

```
data1 %>%
 dplyr::count(x8) %>%
 ggplot(aes(x = reorder(x8,desc(n)), y = n, fill = x8)) +
 geom_bar(stat = "identity", position = "dodge", color = "black", alpha = 0.8, width = 0.
 geom_text(aes(label=n), position=position_dodge(width=0.8), vjust= -0.5, size = 2) +
  labs(title="Amoniacid count Position 8",
     subtitle = "",
       x=" ", y = "Count") +
 theme_minimal() +
 theme(legend.position = "none")
```

Position 9.

```
data1 %>%
 dplyr::count(x9) %>%
 ggplot(aes(x = reorder(x9,desc(n)), y = n, fill = x9)) +
 geom_bar(stat = "identity", position = "dodge", color = "black", alpha = 0.8, width = 0.
 geom_text(aes(label=n), position=position_dodge(width=0.8), vjust= -0.5, size = 2) +
  labs(title="Amoniacid count Position 9",
     subtitle = "",
       x=" ", y = "Count") +
 theme_minimal() +
 theme(legend.position = "none")
```

Position 10.

```
data1 %>%
 dplyr::count(x10) %>%
 ggplot(aes(x = reorder(x10,desc(n)), y = n, fill = x10)) +
 geom_bar(stat = "identity", position = "dodge", color = "black", alpha = 0.8, width = 0.
 geom_text(aes(label=n), position=position_dodge(width=0.8), vjust= -0.5, size = 2) +
   labs(title="Amoniacid count Position 10",
      subtitle = "",
        x=" ", y = "Count") +
 theme_minimal() +
 theme(legend.position = "none")
```

Position 11.

```
data1 %>%
 dplyr::count(x11) %>%
 ggplot(aes(x = reorder(x11,desc(n)), y = n, fill = x11)) +
 geom_bar(stat = "identity", position = "dodge", color = "black", alpha = 0.8, width = 0.
 geom_text(aes(label=n), position=position_dodge(width=0.8), vjust= -0.5, size = 2) +
   labs(title="Amoniacid count Position 11",
      subtitle = "",
        x=" ", y = "Count") +
 theme_minimal() +
 theme(legend.position = "none")
```

Position 12.

```
data1 %>%
 dplyr::count(x12) %>%
 ggplot(aes(x = reorder(x12,desc(n)), y = n, fill = x12)) +
 geom_bar(stat = "identity", position = "dodge", color = "black", alpha = 0.8, width = 0.
 geom_text(aes(label=n), position=position_dodge(width=0.8), vjust= -0.5, size = 2) +
   labs(title="Amoniacid count Position 12",
      subtitle = "",
        x=" ", y = "Count") +
 theme_minimal() +
 theme(legend.position = "none")
```

Position 13.

```
data1 %>%
 dplyr::count(x13) %>%
 ggplot(aes(x = reorder(x13,desc(n)), y = n, fill = x13)) +
 geom_bar(stat = "identity", position = "dodge", color = "black", alpha = 0.8, width = 0.
 geom_text(aes(label=n), position=position_dodge(width=0.8), vjust= -0.5, size = 2) +
   labs(title="Amoniacid count Position 13",
      subtitle = "",
        x=" ", y = "Count") +
 theme_minimal() +
 theme(legend.position = "none")
```

Position 14.

```
data1 %>%
 dplyr::count(x14) %>%
 ggplot(aes(x = reorder(x14,desc(n)), y = n, fill = x14)) +
 geom_bar(stat = "identity", position = "dodge", color = "black", alpha = 0.8, width = 0.
 geom_text(aes(label=n), position=position_dodge(width=0.8), vjust= -0.5, size = 2) +
  labs(title="Amoniacid count Position 14",
    subtitle = "",
      x=" ", y = "Count") +
 theme_minimal() +
 theme(legend.position = "none")
```

Position 15.

```
data1 %>%
 dplyr::count(x15) %>%
 ggplot(aes(x = reorder(x15,desc(n)), y = n, fill = x15)) +
 geom_bar(stat = "identity", position = "dodge", color = "black", alpha = 0.8, width = 0.
 geom_text(aes(label=n), position=position_dodge(width=0.8), vjust= -0.5, size = 2) +
  labs(title="Amoniacid count Position 15",
    subtitle = "",
      x=" ", y = "Count") +
 theme_minimal() +
 theme(legend.position = "none")
```

Position 16.

```
data1 %>%
 dplyr::count(x16) %>%
 ggplot(aes(x = reorder(x16,desc(n)), y = n, fill = x16)) +
 geom_bar(stat = "identity", position = "dodge", color = "black", alpha = 0.8, width = 0.
 geom_text(aes(label=n), position=position_dodge(width=0.8), vjust= -0.5, size = 2) +
  labs(title="Amoniacid count Position 16",
    subtitle = "",
      x=" ", y = "Count") +
 theme_minimal() +
 theme(legend.position = "none")
```

Position 17.

```
data1 %>%
 dplyr::count(x17) %>%
 ggplot(aes(x = reorder(x17,desc(n)), y = n, fill = x17)) +
 geom_bar(stat = "identity", position = "dodge", color = "black", alpha = 0.8, width = 0.
 geom_text(aes(label=n), position=position_dodge(width=0.8), vjust= -0.5, size = 2) +
  labs(title="Amoniacid count Position 17",
    subtitle = "",
      x=" ", y = "Count") +
 theme_minimal() +
 theme(legend.position = "none")
```

Position 18.

```
data1 %>%
 dplyr::count(x18) %>%
 ggplot(aes(x = reorder(x18,desc(n)), y = n, fill = x18)) +
 geom_bar(stat = "identity", position = "dodge", color = "black", alpha = 0.8, width = 0.
 geom_text(aes(label=n), position=position_dodge(width=0.8), vjust= -0.5, size = 2) +
  labs(title="Amoniacid count Position 18",
    subtitle = "",
      x=" ", y = "Count") +
 theme_minimal() +
 theme(legend.position = "none")
```

Position 19.

```
data1 %>%
 dplyr::count(x19) %>%
 ggplot(aes(x = reorder(x19,desc(n)), y = n, fill = x19)) +
 geom_bar(stat = "identity", position = "dodge", color = "black", alpha = 0.8, width = 0.
 geom_text(aes(label=n), position=position_dodge(width=0.8), vjust= -0.5, size = 2) +
  labs(title="Amoniacid count Position 19",
    subtitle = "",
      x=" ", y = "Count") +
 theme_minimal() +
 theme(legend.position = "none")
```

Position 20.

```
data1 %>%
 dplyr::count(x20) %>%
 ggplot(aes(x = reorder(x20,desc(n)), y = n, fill = x20)) +
 geom_bar(stat = "identity", position = "dodge", color = "black", alpha = 0.8, width = 0.
 geom_text(aes(label=n), position=position_dodge(width=0.8), vjust= -0.5, size = 2) +
  labs(title="Amoniacid count Position 20",
    subtitle = "",
      x=" ", y = "Count") +
 theme_minimal() +
 theme(legend.position = "none")
```

Position 21.

```
data1 %>%
 dplyr::count(x21) %>%
 ggplot(aes(x = reorder(x21,desc(n)), y = n, fill = x21)) +
 geom_bar(stat = "identity", position = "dodge", color = "black", alpha = 0.8, width = 0.
 geom_text(aes(label=n), position=position_dodge(width=0.8), vjust= -0.5, size = 2) +
  labs(title="Amoniacid count Position 21",
    subtitle = "",
      x=" ", y = "Count") +
 theme_minimal() +
 theme(legend.position = "none")
```

Position 22.

```
data1 %>%
 dplyr::count(x22) %>%
 ggplot(aes(x = reorder(x22,desc(n)), y = n, fill = x22)) +
 geom_bar(stat = "identity", position = "dodge", color = "black", alpha = 0.8, width = 0.
 geom_text(aes(label=n), position=position_dodge(width=0.8), vjust= -0.5, size = 2) +
  labs(title="Amoniacid count Position 22",
     subtitle = "",
       x=" ", y = "Count") +
 theme_minimal() +
 theme(legend.position = "none")
```

Position 23.

```
data1 %>%
 dplyr::count(x23) %>%
 ggplot(aes(x = reorder(x23,desc(n)), y = n, fill = x23)) +
 geom_bar(stat = "identity", position = "dodge", color = "black", alpha = 0.8, width = 0.
 geom_text(aes(label=n), position=position_dodge(width=0.8), vjust= -0.5, size = 2) +
  labs(title="Amoniacid count Position 23",
     subtitle = "",
       x=" ", y = "Count") +
 theme_minimal() +
 theme(legend.position = "none")
```

9. Accession numbers and genes

Accession Numbers	Organism	Location
XP_002184128.1	Phaeodactylum tricornutum	Plastid
XP_002288939.1	Thalassiosira pseudonana	Plastid
XP_002185582.1	Thalassiosira pseudonana	Plastid
XP_002288023.1	Thalassiosira pseudonana	Plastid
XP_002291225.1	Thalassiosira pseudonana	Plastid
EEC45554.1	Phaeodactylum tricornutum	Plastid
GAX23844.1	Fistulifera solaris	ER
EJK59214.1	Thalassiosira oceanica	ER
EEC44953.1	Phaeodactylum tricornutum	Plastid
XP_002290241.1	Thalassiosira pseudonana	ER
EEC45309.1	Phaeodactylum tricornutum	Plastid
EEC51157.1	Phaeodactylum tricornutum	Plastid
EEC51500.1	Phaeodactylum tricornutum	Plastid
XP_002297355.1	Thalassiosira pseudonana	ExtraCellular
XP_002290634.1	Thalassiosira pseudonana	Plastid
EED94157.1	Thalassiosira pseudonana	Plastid
EEC51049.1	Phaeodactylum tricornutum	ER
XP_002296083.1	Thalassiosira pseudonana	Plastid
EEC48498.1	Phaeodactylum tricornutum	Plastid
EEC47913.1	Phaeodactylum tricornutum	Plastid
EED93668.1	Thalassiosira pseudonana	Plastid
EED90771.1	Thalassiosira pseudonana	Plastid
AWT40073.1	Biddulphia biddulphiana	Plastid
AIR76125.1	Conticribra weissflogii	Plastid
YP_009497492.1	Biddulphia tridens	Plastid
GAX21350.1	Fistulifera solaris	ER
KAG7352660.1	Nitzschia inconspicua	ER

EEC50402.1	Phaeodactylum tricornutum	Plastid
QYB22995.1	Cylindrotheca closterium	Mitochondria
YP_009029310.1	Leptocylindrus danicus	ExtraCellular
GAX10111.1	Fistulifera solaris	ER
EEC42998.1	Phaeodactylum tricornutum	Plastid
EEC43856.1	Phaeodactylum tricornutum	ER
EEC46981.1	Phaeodactylum tricornutum	Plastid
CAB9501036.1	Seminavis robusta	ER
KAG7352045.1	Nitzschia inconspicua	ER
EED87472.1	Thalassiosira pseudonana	Plastid
EJK66825.1	Thalassiosira oceanica	ER
AIR75739.1	Rhizosolenia imbricata	Plastid
AWT39260.1	Actinocyclus subtilis	Plastid
AWT38845.1	Guinardia striata	Plastid
EEC51482.1	Phaeodactylum tricornutum	Plastid
YP_874535.1	Thalassiosira pseudonana	ExtraCellular
XP_002292195.1	Thalassiosira pseudonana	ExtraCellular
EED87828.1	Thalassiosira pseudonana	Plastid
EEC45046.1	Phaeodactylum tricornutum	Plastid
EEC42835.1	Phaeodactylum tricornutum	Plastid
EEC51753.1	Phaeodactylum tricornutum	Plastid
GAX20925.1	Fistulifera solaris	ER
XP_002294577.1	Thalassiosira pseudonana	ExtraCellular
XP_002295563.1	Thalassiosira pseudonana	Plastid
EEC49058.1	Phaeodactylum tricornutum	Plastid
XP_002296338.1	Thalassiosira pseudonana	ExtraCellular
EEC51053.1	Phaeodactylum tricornutum	Plastid
YP_009495409.1	Cylindrotheca closterium	Mitochondria
GFH59446.1	Chaetoceros tenuissimus	ER
EEC50220.1	Phaeodactylum tricornutum	Plastid
XP_002297376.1	Thalassiosira pseudonana	ExtraCellular
XP_002288372.1	Thalassiosira pseudonana	ExtraCellular
XP_002295986.1	Thalassiosira pseudonana	ExtraCellular

EEC47439.1	Phaeodactylum tricornutum	Plastid
EED90928.1	Thalassiosira pseudonana	Plastid
KAG7358328.1	Nitzschia inconspicua	ER
XP_002293552.1	Thalassiosira pseudonana	ExtraCellular
EEC50356.1	Phaeodactylum tricornutum	Plastid
ACI65867.1	Phaeodactylum tricornutum	Plastid
EEC47366.1	Phaeodactylum tricornutum	Plastid
QYB22988.1	Cylindrotheca closterium	Mitochondria
GAX18283.1	Fistulifera solaris	ER
XP_002289865.1	Thalassiosira pseudonana	Plastid
EEC48995.1	Phaeodactylum tricornutum	Plastid
ACI65438.1	Phaeodactylum tricornutum	Plastid
XP_002291468.1	Thalassiosira pseudonana	Plastid
EEC46729.1	Phaeodactylum tricornutum	ER
EJK62482.1	Thalassiosira oceanica	ER
XP_002293611.1	Thalassiosira pseudonana	ExtraCellular
EEC42979.1	Phaeodactylum tricornutum	Plastid
XP_002297377.1	Thalassiosira pseudonana	ExtraCellular
EEC43542.1	Phaeodactylum tricornutum	Plastid
YP_009029440.1	Leptocylindrus danicus	ExtraCellular
XP_002289239.1	Thalassiosira pseudonana	ExtraCellular
XP_002177121.1	Phaeodactylum tricornutum	Plastid
EEC43382.1	Phaeodactylum tricornutum	Plastid
EEC51495	Phaeodactylum tricornutum	Mitochondria
EEC42858.1	Phaeodactylum tricornutum	Plastid
EEC46387.1	Phaeodactylum tricornutum	Plastid
Cvel_639.t1__M_Transcript_8572extd	Chromera Velia	Plastid
Cvel_4018.t1__JO800247.1extd	Chromera Velia	Plastid
Cvel_7872.t1__M_Transcript_43023extd	Chromera Velia	Plastid
Cvel_9313.t1__M_Transcript_33772extd	Chromera Velia	Plastid
Cvel_9474.t1__M_Transcript_11840extd	Chromera Velia	Plastid

Cvel_9830.t1__JO807406.1extd	Chromera Velia	Plastid
Cvel_13572.t1__M_Transcript_21954extd	Chromera Velia	Plastid
Cvel_14048.t1__M_Transcript_50549extd	Chromera Velia	Plastid
Cvel_21680.t1__M_Transcript_38580extd	Chromera Velia	Plastid
Cvel_21762.t1__M_Transcript_49573extd	Chromera Velia	Plastid
Cvel_22466.t1__M_Transcript_38997extd	Chromera Velia	Plastid
Cvel_22643.t1__M_Transcript_36939extd	Chromera Velia	Plastid
Cvel_22994.t1__M_Transcript_14600extd	Chromera Velia	Plastid
Cvel_23943.t1__HQ222934.1extd	Chromera Velia	Plastid
Cvel_25057.t1__JO809437.1extd	Chromera Velia	Plastid
Cvel_26177.t1__M_Transcript_41167extd	Chromera Velia	Plastid
Cvel_27782.t1__M_Transcript_38155extd	Chromera Velia	Plastid
Cvel_28848.t1__M_Transcript_27382extd	Chromera Velia	Plastid
Cvel_31936.t1__HQ222929.1extd	Chromera Velia	Plastid
Cvel_35011.t1__M_Transcript_17069extd	Chromera Velia	Plastid
Vbra_1060.t1__M_Transcript_17770extd	Vitrella Brassicaformis	Plastid
Vbra_1322.t1__M_Transcript_19109extd	Vitrella Brassicaformis	Plastid
Vbra_4131.t1__M_Transcript_49435extd	Vitrella Brassicaformis	Plastid
Vbra_7099.t1__M_Transcript_52951extd	Vitrella Brassicaformis	Plastid
Vbra_8768.t1__M_Transcript_30622extd	Vitrella Brassicaformis	Plastid
Vbra_11289.t1__M_Transcript_51653extd	Vitrella Brassicaformis	Plastid
Vbra_14926.t1__M_Transcript_51215extd	Vitrella Brassicaformis	Plastid
Vbra_15476.t1__M_Transcript_43618extd	Vitrella Brassicaformis	Plastid
Vbra_15949.t1__M_Transcript_3353extd	Vitrella Brassicaformis	Plastid
Vbra_16645.t1__M_Transcript_40282extd	Vitrella Brassicaformis	Plastid
Vbra_17448.t1__M_Transcript_5652extd	Vitrella Brassicaformis	Plastid
Vbra_22408.t1__M_Transcript_27098extd	Vitrella Brassicaformis	Plastid
Lhl1/RedCAP	Phaeodactylum tricornutum	Plastid
unnamed	Phaeodactylum tricornutum	Plastid
SybA	Phaeodactylum tricornutum	ER
Lhcr4	Phaeodactylum tricornutum	Plastid
PtSP1	Phaeodactylum tricornutum	Plastid
Lhcf1-FcpA	Phaeodactylum tricornutum	Plastid
sCdc48-2	Fistulifera solaris	BLS
OEE1	Phaeodactylum tricornutum	Plastid

GLR-Gsr2	Phaeodactylum tricornutum	Mitochondria
AtpC	Phaeodactylum tricornutum	Plastid
MGD1	Phaeodactylum tricornutum	Plastid
Lhcf10	Phaeodactylum tricornutum	Plastid
sbeta7	Phaeodactylum tricornutum	BLS
GapC1	Phaeodactylum tricornutum	Plastid
Lhcf8	Phaeodactylum tricornutum	Plastid
Lhcr2	Phaeodactylum tricornutum	Plastid
FbaC2	Phaeodactylum tricornutum	Plastid
Lhcr11	Phaeodactylum tricornutum	Plastid
sTLP-1	Phaeodactylum tricornutum	BLS
Tpt1	Phaeodactylum tricornutum	Plastid
NDK3	Phaeodactylum tricornutum	Mitochondria
Lhcf4-FcpD	Phaeodactylum tricornutum	Plastid
Lhcf2	Phaeodactylum tricornutum	Plastid
Lhcx1	Phaeodactylum tricornutum	Plastid
Lhcr13	Phaeodactylum tricornutum	Plastid
Lhcf9	Phaeodactylum tricornutum	Plastid
Lhcf5-FcpE	Phaeodactylum tricornutum	Plastid
FBPC3	Phaeodactylum tricornutum	Plastid
Der1-1	Phaeodactylum tricornutum	BLS
TrxO	Phaeodactylum tricornutum	Mitochondria
Trx-y	Phaeodactylum tricornutum	Plastid

Lhcf16	Phaeodactylum tricornutum	Plastid
CPF2	Phaeodactylum tricornutum	ER
sDer1-2	Phaeodactylum tricornutum	BLS
PGL	Phaeodactylum tricornutum	Plastid
PtCA1	Phaeodactylum tricornutum	Plastid
FBPC2	Phaeodactylum tricornutum	Plastid
predicted, PGRL	Phaeodactylum tricornutum	Plastid
CA-VII-så±CA-2	Phaeodactylum tricornutum	ER-BLS
sSMC	Phaeodactylum tricornutum	BLS
FBPC1	Phaeodactylum tricornutum	Plastid
HAP	Phaeodactylum tricornutum	ER
CA-II	Phaeodactylum tricornutum	BLS
Lhcr1	Phaeodactylum tricornutum	Plastid
sSec14	Phaeodactylum tricornutum	BLS
sDTC	Phaeodactylum tricornutum	BLS
PGDH	Phaeodactylum tricornutum	BLS
sbeta2	Nitzschia inconspicua	BLS
PtCa2	Phaeodactylum tricornutum	Plastid
SybD	Phaeodactylum tricornutum	ER
unnamed	Phaeodactylum tricornutum	Extracellular
sORF532a	Phaeodactylum tricornutum	BLS
PPG1	Phaeodactylum tricornutum	Plastid
unnamed	Phaeodactylum tricornutum	Extracellular
sPUB	Fistulifera solaris	BLS

Tom70	Phaeodactylum tricornutum	Mitochondria
FolC	Phaeodactylum tricornutum	Mitochondria
Lhcr14	Phaeodactylum tricornutum	Plastid
ptE3P	Phaeodactylum tricornutum	BLS
Fru	Phaeodactylum tricornutum	ER
PtFAD6	Phaeodactylum tricornutum	Plastid
TrxH	Phaeodactylum tricornutum	BLS
sDPC	Seminavis robusta	BLS
unnamed	Phaeodactylum tricornutum	Plastid
sORF534	Phaeodactylum tricornutum	BLS
unknown protein	Phaeodactylum tricornutum	ER
Ntt1	Phaeodactylum tricornutum	Plastid
ORF387	Phaeodactylum tricornutum	ER
Lhcf12	Phaeodactylum tricornutum	Plastid
Lhcf3-FcpC	Phaeodactylum tricornutum	Plastid
Lhcr3	Phaeodactylum tricornutum	Plastid
Tkl	Phaeodactylum tricornutum	Extracellular
FtrB	Phaeodactylum tricornutum	Plastid
GSII	Phaeodactylum tricornutum	Plastid
Lhcf11	Phaeodactylum tricornutum	Plastid
Trx-m	Phaeodactylum tricornutum	Plastid
Rpe	Phaeodactylum tricornutum	Plastid
RecA	Phaeodactylum tricornutum	Plastid

Lhcr12	Phaeodactylum tricornutum	Plastid
Bip	Phaeodactylum tricornutum	ER
CA-VI-å±CA-2	Phaeodactylum tricornutum	ER
Ubi	Phaeodactylum tricornutum	BLS
TRD1	Phaeodactylum tricornutum	BLS
CA-III	Phaeodactylum tricornutum	ER
Hlip2	Phaeodactylum tricornutum	Plastid
Ank5	Phaeodactylum tricornutum	Mitochondria
Hsp70_2	Phaeodactylum tricornutum	BLS
Lhcf17	Phaeodactylum tricornutum	Plastid
GLRX2	Phaeodactylum tricornutum	BLS
NTRC	Phaeodactylum tricornutum	BLS
unknown protein	Phaeodactylum tricornutum	BLS
sDrp (Drp5b)	Phaeodactylum tricornutum	BLS
Lhcx2	Phaeodactylum tricornutum	Plastid
ptDUP	Phaeodactylum tricornutum	BLS
unnamed	Phaeodactylum tricornutum	Extracellular
sORF261	Phaeodactylum tricornutum	BLS
ptOmp85	Phaeodactylum tricornutum	Plastid
unnamed	Phaeodactylum tricornutum	Plastid
sPEL	Phaeodactylum tricornutum	BLS
FbaC1	Phaeodactylum tricornutum	Plastid