Amino Acid Properties Provide Insight to a Protein's Subcellular Location

by

Brian Powell

Submitted in Partial Fulfillment of the Requirements

for the Degree of

Master of Computing and Information Systems

YOUNGSTOWN STATE UNIVERSITY

December, 2016

Declaration of Authorship

I hereby release this thesis to the public. I understand that this thesis will be made available from the OhioLINK ETD Center and the Maag Library Circulation Desk for public access. I also authorize the University or other individuals to make copies of this thesis as needed for scholarly research.

Signature:

Brian Powell, Student	Date

Approvals:

Alina Lazar, Thesis Advisor	Date
Feng Yu, Committee Member	Date
Xiangjia Min, Committee Member	Date
Dr. Salvatore A. Sanders, Dean of Graduate Studies	Date

YOUNGSTOWN STATE UNIVERSITY

Abstract

CSIS
Department of Computer Science & Information Systems

Master of Computing and Information Systems

Amino Acid Properties Provide Insight to a Protein's Subcellular Location

by Brian POWELL

Current approaches of predicting subcellular locations of proteins located in a cell have made some advances but are far from perfect. Accurately predicting these locations result in better annotations of that protein and provide clearer pictures of its functions. We approach this problem by using a chaos game representation of the sequence based on physical and chemical properties of amino acids. We then split the resulting graph into two related discrete series, which is then subjected to wavelet transformation. The wavelet transformation data is then used as input for our classification algorithms. We observe the accuracy of how well each property predicts the correct subcellular location. We aim to achieve above the threshold of 45 percent accuracy, which is the average of existing general sub-cellular predictors. For our study protein sequences were obtained from Uniprot's freely accessible repositories. We parsed data from five different classes, consisting of plant, fungal, mammal, human, and rodent proteins. We accommodate 10 subcellular locations: Nucleus, Membrane, Cytoplasm, Endoplasmic Reticulum, Secreted, Mitochondria, Cell Membrane, Vacuole, Golgi Apparatus, and Chloroplast. Protein sequences comprised of 20 amino acids are sorted into groups of four based on the selected property of amino acids. These groups allow the sequence to be plotted using 2-dimension chaos game theory. The resulting graph retains the sequence order in numerical form. Looking at the graph with a human eye we can't deduce any information. To address this, we split the graph into two related discrete series based on the x-axis and y-axis. We then use a 3-level Haar wavelet transformation. Each level provides us with a detail coefficient vector the length of our sequence. For each detail coefficient vector we calculate the mean, min, max, and standard deviation. This provides us with 24 features to be used as input for classification. We run a variety of classifiers to assess the importance of amino acid properties.

Acknowledgements

My sincere thanks goes to Dr. Min, who provided me an opportunity to join his lab as a research assistant. Without his support and time it would not have been possible to conceive this experiment.
I would also like to thank Dr. Chang for his insight on how to approach this experiment.

Contents

List of Figures

List of Tables

For/Dedicated to/To my...

Chapter 1

Introduction

1.1 The Birth of Bioinformatics

1.1.1 Overview

Bioinformatics is an interdisciplinary field, combining two or more other fields. For bioinformatics researchers should have a grasp on biology, computer science, and statistics. Questions that are conceived in biology are challenged using an informatic approach. To be able to apply this approach, all you need is data readable by a computer and a hypothesis. The first data available was at the top level of biology, the genome, or collection of genes in a species.

DNA sequencing started in the 1970's but the field was slow to catch on. Being able to sequence DNA was one entry point into the field and only a limited number of questions were feasible. As time passed and biology books grew thicker and computers doubled their efficiency so did our yearning for knowledge of everything that the naked eye can't see. Then in 1990, the Human Genome Project became a reality. This was not on a study on the biological details of a lab mouse or insect, but the most intelligent species on Earth. The project unleashed an abundant amount of information in areas of physiology, evolution, and human development [1]. After sorting and organizing through genomic data, scientists soon hit the realization that they had to go deeper to understand the genomic data. Much like reverse engineering of an operating system or a modern day engine, we know what they do, but how do they work? The focus of this study lies a few levels lower than the genome. To give an overview:

- Genome - Study of structure, function, and expression of all genes
- Transcriptome - The study of mRNA within an organism
- Proteome - Study of but not limited to the function and location of proteins in an organism
- Metabolome - Study of metaobolites, which are low molecular weight compounds found in the system, and exhibit significance response to environmental factors [2]

The proteome is able to provide us with the supplemental information that is sought after. The collection of proteins are manufactured in the cell, not ingested through food, and provide insight to which genes are activated simply be telling the cell to make that protein. "Proteomics will add to our understanding of the biochemistry of proteins, processes and pathways for years to come" [3]. Just a little over a decade later from that statement, and the Human genome project, we currently have a draft of the Human proteome [4]. With that nearing completion, tools that operate in the scope of proteomics should be polished to the highest potential.

1.1.2 Areas of Proteomics

There are different branches of computationally trying to predict information about proteins to alleviate the work required by experimental researchers. Protein-protein interactions (PPIs), where either long or short relationships between two or more proteins emerge due to biochemical events. Proteins often work together to accomplish their function, and these interactions highlight the relationships that are formed repeatedly. While the objective is to predict the interaction, inherently the function is cultivated as well. Similar to the wide array of tools available for subcellular location prediction, PPI tools rely on methods ranging from protein structure to text mining [5]. Another being the subcellular localization of proteins. Over the years, several ideologies and tools have been created and published. There have been findings that have turned into facts already in this area. For example, we know that classical secretory proteins almost always contain an N-terminal. Machine learning has alleviated much of this process, revealing trends to the researcher.

1.2 History of Subcellular Prediction

The history of subcellular predicted is still in its infancy. With the rapid growth of sequencing methods, the availability of data is longer an issue. Dating back to 1997, Cedano began exploring that amino acid composition and the cellular location of a protein demonstrates a relationship. Their analysis covers among the following five protein locations: integral membrane proteins, anchored membrane proteins, extracellular proteins, intracellular proteins and nuclear proteins [8]. Similar to the foundation of our experiment, Andrade states in 1998, "Within each subcellular compartment of a given cell type, proteins have co-evolved with the physiochemical environment so that they are stable and functional in that environment. However, the general features of the nuclear, cytoplasmic, and extracellular environments discussed above have been constant factors throughout eukaryotic evolution" [10]. Due to the difficulty of the similarities of protein function and structure

TABLE 1.1: Survey of Prediction Tools

Tool	Notes
PredPlantPTS1 [14]	Focuses on Plant Kingdom
Predotar[15]	Uses the N-terminal to determine Membrane bound Proteins
MitoMiner [16]	Targets prediction of mitchondria proteins
WoLF PSORT [17]	General predictor
MetazSecKB [11]	Mammal KnowledgeBase using multiple tools
FunSecKB [12]	Fungi KnowledgeBase using multiple tools
ProtLock [8]	Amino Acid composition
PredAlgo [18]	Focuses on green algae
ProLoc-GO [19]	Utilizes Gene Ontology (GO) annotations
PredSL [20]	General Predictor

from sequence, bioinformatics hits an obstacle when more proteins diverge in sequence, reducing the effects of sequence homology. Using a large-scale analysis, sequence similarity and identity are fully explored to exemplify this statement [9].

1.2.1 Types of Tools

For subcellular prediction a wide range of approaches can be taken for this computationally difficult problem. The first approach is writing your own algorithm, incorporating any features you think will accurately represent that protein's location without too much overhead. Various features have been used, such as sequence homology, N-terminal existence, sequence manipulation, functional domains, protein families, and amino acid properties to an extent. Different combinations of these features have been used on various types of proteome data, such as prokaryote, eukaryote, bacteria, specific kingdoms, or just specific locations.

Another approach to predict subcellular locations is to incorporate many existing tools into a master algorithm that uses those results to best fit a protein with it's predicted location. This approach can be considerably better due to the amount of tools created by researchers, and the location or special interest that the tool focuses on. Working examples of this approach are the Secreteome Knowledgebases [11], [12].

1.2.2 Existing Methods and Tools

If the location of a protein is known, or we know where it's traveling to, understanding it's function is natively easier. A majority of the tools are based on sequence analysis. These methods are limited in capacity, and accuracy due to relying on sequence homology [13], which emphasizes the need to understand what other properties determine where, and what a protein's function is. To demonstrate the

variety of existing methods, we surveyed existing tools and the differences among them 1.1.

1.3 Computational Problems

The computational prediction of subcellular localization of proteins is challenging. Not in terms of computational power, but zoning in on what determines a proteins location. At the core can we be given a protein sequence, and identify its subcellular location? To be able to do that a method has to be able to consider multiple facets and incorporate that into it's prediction. For example, a protein is manufactured in the cytoplasm. Therefore even annotated proteins that we use for our model can be in accurate depending on the lifecycle of the protein. Proteins commonly translocate, or move from one location to another to fulfill it's function. Can the model predict both of these locations or is it limited to one, or evolve to the state where it can predict an origin location and where the protein will be going? Achieving this would be paramount in medical research, opening the doorway for the delivery of medicine effectively.

1.3.1 Benefits

There is an abundance of proteins that do not have their function annotated, this is partly caused by orphan genes not being analyzed as part of an organism's genome [22]. Predicting these proteins without having to catalog each and every orphan gene, saves an invaluable amount of time a wet-lab researcher has to spend tagging and then sequencing those proteins. Furthermore, "Shortfalls in the ability of bioinformatics to predict both the existence and function of genes have also illustrated the need for protein analysis" [23]. Annotating the protein and then linking it to it's parent gene, will validate the existence and the function of the genes in question.

1.3.2 Objectives

Our objective is to find relationships between the location a protein is found and the overall composition of a sequence based on the properties of amino acids. Each amino acid has a unique mass, hydrophobicity, defined as the tendency to repel water, and other traits. These properties compromise a protein's overall properties and have to influence the way a protein behaves as there is no controlling brain unit on a protein. At the molecule level, physical and chemical traits, and outside influences should determine where a protein goes.

1.3.3 Approach

We use an approach similar to the experiment conducted by Jia for predicting protein to protein interactions [24]. To incorporate the properties of amino acids, but retain sequence order information we embrace chaos game theory. By placing the amino acids into groups based on the selected property value, we can uniquely plot each sequence. Reading in these coordinates into two separate times series of namely x and y, we can pull out a set of discrete numbers after performing a wavelet transformation. To expand our feature set, we apply basic mathematic operations (mean,average,minimum,maximum,standard deviation) on the number set.

Chapter 2

Data Collection

2.1 Universal Protein Resource

To start our experiment we require a large quantity of protein sequences with subcellular locations annotated to create our models. The Universal Protein Resource (UniProt) is an online repository that curates this protein information from submissions from researchers, publications, or automatic annotations. UniProt is a conjunction of two original knowledgebases, Swiss-Prot and TrEMBL. The differences between the two are that Swiss-Prot is manually annotated information and TrEMBL is of predicted origin [25]. For this experiment we only used proteins with a Swiss-Prot label as attempting to build prediction models on predictions that can not be verified. UniProt gives you the option of downloading the protein sequences categorized by biological kingdoms, and for single species where sufficient data exists. We will be building models for the following protein categories: Fungal Proteins, Plant Proteins, Mammal Proteins. In addition to Mammal Proteins, the two subsets of Human Proteins and Rodent Proteins will be examined.

2.2 Preprocessing

The raw files from UniProt will need to undergo some transformations before being able to be used. A single entry for a protein contains roughly 500 lines of information. To accomplish this stage we will parse out only the information we need, the protein identifier, the sequence, and the subcellular comments which are bundled in the comments section among other things, see 2.1. We used a script called swissknife, which was developed by researchers from either the European Bioinformatics Institute or Swiss Institute of Bioinformatics, and can be downloaded at this link `https://sourceforge.net/projects/swissknife/files/latest/download`. We used this information to create our dataset, and arranged the information using the FASTA format standard. At minimum, the FASTA format is an identifier line beginning with a greater than symbol (>) followed by a new line with the matching sequence. See 2.2, this enables our dataset to be easily parsed into our own experiment and accepted by other existing tools.

```
CC   -!- FUNCTION: Adapter protein implicated in the regulation of a large
CC       spectrum of both general and specialized signaling pathways. Binds
CC       to a large number of partners, usually by recognition of a
CC       phosphoserine or phosphothreonine motif. Binding generally results
CC       in the modulation of the activity of the binding partner.
CC       {ECO:0000269|PubMed:16511572}.
CC   -!- SUBUNIT: Homodimer. Interacts with SAMSN1 (By similarity).
CC       Interacts with RAF1, SSH1 and CRTC2/TORC2. Interacts with ABL1
CC       (phosphorylated form); the interaction retains it in the
CC       cytoplasm. Interacts with GAB2. Interacts with MDM4
CC       (phosphorylated); negatively regulates MDM4 activity toward TP53.
CC       Interacts with PKA-phosphorylated AANAT and SIRT2.Interacts with
CC       the 'Thr-369' phosphorylated form of DAPK2 (PubMed:26047703).
CC       Interacts with PI4KB, TBC1D22A and TBC1D22B (PubMed:23572552).
CC       {ECO:0000250, ECO:0000250|UniProtKB:P61982,
CC       ECO:0000269|PubMed:10433554, ECO:0000269|PubMed:11427721,
CC       ECO:0000269|PubMed:15159416, ECO:0000269|PubMed:15454081,
CC       ECO:0000269|PubMed:15696159, ECO:0000269|PubMed:16511572,
CC       ECO:0000269|PubMed:17085597, ECO:0000269|PubMed:18249187,
CC       ECO:0000269|PubMed:19172738, ECO:0000269|PubMed:23572552,
CC       ECO:0000269|PubMed:26047703}.
CC   -!- INTERACTION:
CC       Q9HC77:CENPJ; NbExp=3; IntAct=EBI-359832, EBI-946194;
CC       O14757:CHEK1; NbExp=7; IntAct=EBI-359832, EBI-974488;
CC       P67828:CSNK1A1 (xeno); NbExp=3; IntAct=EBI-359832, EBI-7540603;
CC       Q9NYF3:FAM53C; NbExp=4; IntAct=EBI-359832, EBI-1644252;
CC       P56524:HDAC4; NbExp=6; IntAct=EBI-359832, EBI-308629;
CC       Q14678-2:KANK1; NbExp=3; IntAct=EBI-359832, EBI-6173812;
CC       Q5S007:LRRK2; NbExp=4; IntAct=EBI-359832, EBI-5323863;
CC       Q7KZI7:MARK2; NbExp=2; IntAct=EBI-359832, EBI-516560;
CC       P27448:MARK3; NbExp=2; IntAct=EBI-359832, EBI-707595;
CC       O15151:MDM4; NbExp=7; IntAct=EBI-359832, EBI-398437;
CC       P61588:Rnd3 (xeno); NbExp=2; IntAct=EBI-359832, EBI-6930266;
CC       Q9BSI4:TINF2; NbExp=2; IntAct=EBI-359832, EBI-717399;
CC       P04637:TP53; NbExp=5; IntAct=EBI-359832, EBI-366083;
CC       P62258:YWHAE; NbExp=4; IntAct=EBI-359832, EBI-356498;
CC   -!- SUBCELLULAR LOCATION: Cytoplasm {ECO:0000250}.
CC   -!- TISSUE SPECIFICITY: Highly expressed in brain, skeletal muscle,
CC       and heart. {ECO:0000269|PubMed:10486217}.
CC   -!- PTM: Phosphorylated by various PKC isozymes.
CC       {ECO:0000269|PubMed:10433554, ECO:0000269|Ref.7}.
CC   -!- SIMILARITY: Belongs to the 14-3-3 family. {ECO:0000305}.
CC   ---------------------------------------------------------------------
CC   Copyrighted by the UniProt Consortium, see http://www.uniprot.org/terms
CC   Distributed under the Creative Commons Attribution-NoDerivs License
CC   ---------------------------------------------------------------------
```

FIGURE 2.1: View of comments section of a protein entry.

```
>Q10428|Cytoplasm|
MKGIKSKMLSRGKSQDTQKSSKKKESKKSNSHDSSKAPKESPSTDPNGSVIGAQNDFLTVPKHSGKKVPIDTTPTPRDEILLENVRTVRK
QRSSLYHISENRNLVRLPSFTDVPVNKWHSLALEKLEQCCVVFDFNDPSTDLYGKEVKREALQDLIDLISVRKEAIDESLYPSIVHMFAV
NVFRPLPPPSNPPGEIMDLEEDEPALEVAWPHLHLVYDFFLRFFESPSLNTSVAKVYINQKFIRKLLVLFDSEDPRERDFLKTTLHRIYG
KFLSLRAFIRRSINNLFLQFVYENEQFNGIAELLEILGSIINGFALPLKEEHKIFLSRVLIPLHKAKSLPLYYPQIAYGIVQFVEKDSSV
TEEVVLGLLRYWPKVNSSKEVLFLNEIEDIIEVMEPSEFLKIQVPLFHKLATSISSQNFQVAERALYFFNNDYFVHLVEENVDIILPIIY
PALFEISKSHWNRVIHSMVCNVLKLFMDINPSLFDEVDAEYSESRRKKEDEEIIREERWTILENIAKENAMKLKSQNPTTVHSTTERLKK
LSLDYTNG
```

FIGURE 2.2: The entry transformed into fasta format.

TABLE 2.1: Subcellular location counts for our datasets.

Location	Plant	Fungal	Mammal	Human	Rodent
Nucleus	4132	6386	2484	3758	4399
Cytoplasm	3005	7697	3358	3505	5286
Endoplasmic	642	1647	650	601	1032
Secreted	1916	1979	2864	1474	2321
Mitochondria	1450	4049	3010	846	1839
Cell membrane	1540	839	1784	2124	2788
Golgi	458	592	340	388	557
Membrane	1994	1521	1416	1975	2575
Vacuole	439	590	0	0	0
Chloroplast	15062	0	0	0	0

2.3 Requirements

Traditionally for protein sequences, they are considered to be complete with the identified start and stop codons. The start codon (AUG), represented by M in a peptide sequence, is a pre-existing requirement for many tools for a protein to be considered a non-fragment. For this experiment, we allow protein sequences that do not have to be part of the dataset. Although not very common, it has been experimentally verified that proteins can have different start codons [26]. We use image analysis for obtaining our features, and therefore need a threshold sequence length of at least 30 amino acids. Anything shorter than this creates too sparse of an image. To consider a subcellular location in the dataset, it must contain at least 300 proteins to provide an ample sample for the classifier.

2.4 Locations

Location definitions can vary in different contexts and be placed under different constraints. We adhere closely to the locations depicted in 2.3, with a few exceptions. The constraints we had to formulate our definitions around were how Uniprot's annotations often listed locations in pairs. For example, plastid is grouped with chloroplast in nearly every occurance. The same applies to cell membrane, and plasma membrane. We have a category tilted simply as membrane, and this includes a range of membrane locations, namely single, and multi-pass membranes.

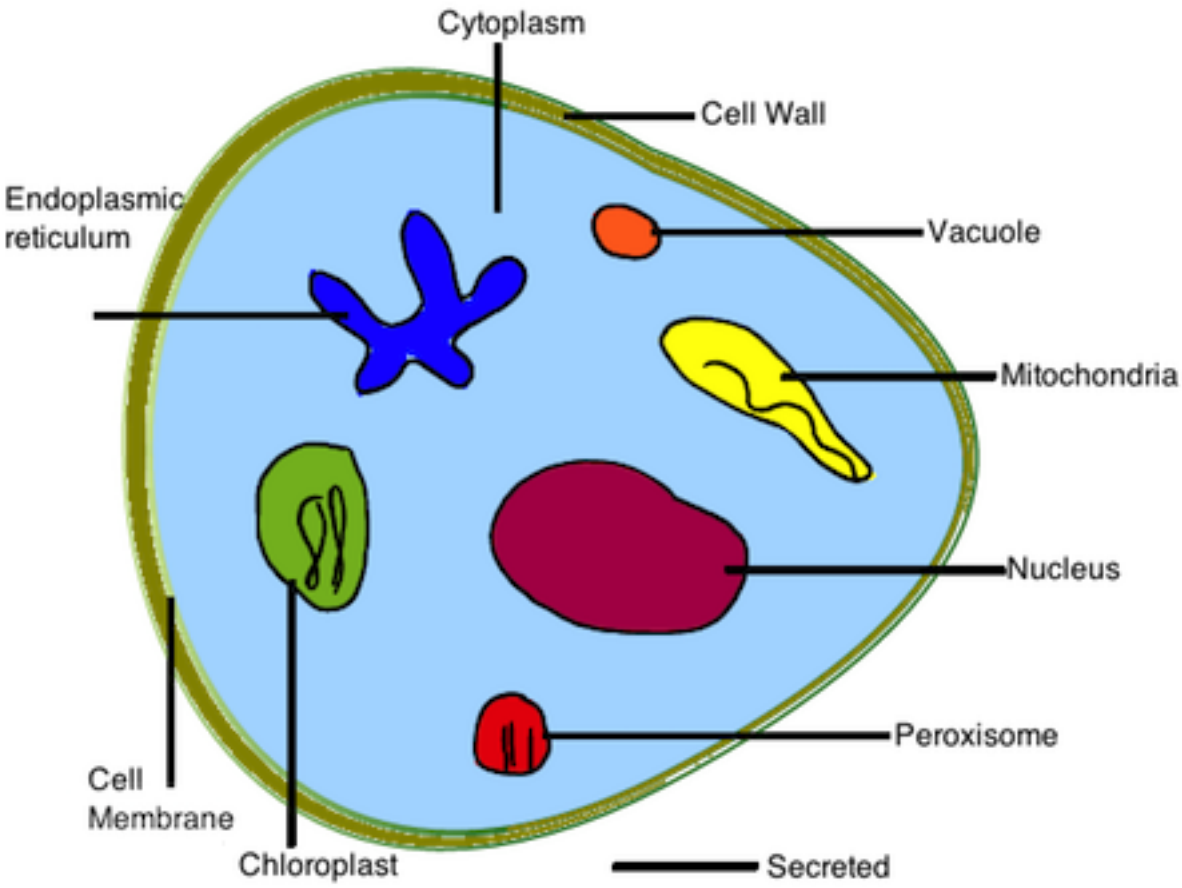

FIGURE 2.3: Locations in a cell visualized.

Chapter 3

Experiment

3.1 Representing Sequences

Protein sequences are simplistic when they are represented by their amino acid letters, that correspond to 3 letter nucleotide codons, all of which is decipherable by people and computers. Even though it is understood by us, no additional knowledge about the protein is acquired by reading its sequence. For example, if we gave the prediction model the protein sequence and it is observed subcellular location, the model would look for very similar sequences in a string matching approach to predict locations. While homologous sequences can share the same attributes, this does not accurately represent all proteins. Just quickly browsing a repository we observe that sequences vary by length and composition, yet can be found in the same location as another protein.

3.1.1 Amino Acid Composition

Amino Acid Composition (AAC) is the first method devised to circumvent this. A sequence is transformed into an array of 20 numbers, each representing an amino acid. Each number is calculated by taking the total number of that amino acid in the sequence, and dividing it by the sequence length [28].

3.1.2 Pseudo Amino Acid Composition

One of the pitfalls of AAC is that you lose sequence order information. A sequence that has the same AAC as another sequence could be completely different in structure and order. Chou's Pseudo Amino Acid Composition (PseAA) addresses this by expanding the algorithm. Instead of tallying the counts of single amino acid residues, the algorithm looks at two residues at one time. Amino acids in positions 1, and 2 would be counted as a group, then 2, and 3. This occurs for three iterations, each time incrementing the distance in between the two residues [29].

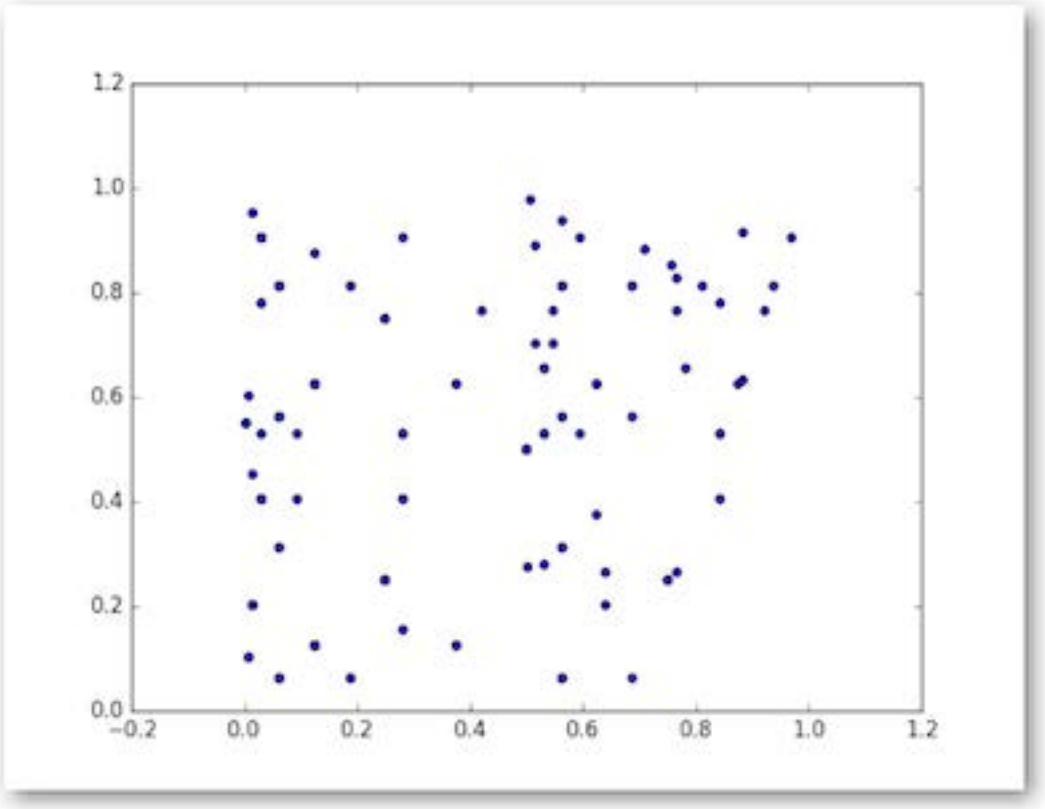

FIGURE 3.1: A protein sequence plotted using Chaos Game theory based on hydrophobic values.

3.1.3 Hidden Markov Models

Hidden Markov Models (HMM) have been around for a long time, and used extensively in other fields. This statistical model behaves like a bayesian network, but looks for hidden states. In bioinformatics, HMM's have a prominent presence in predicting a protein's transmembrane topology, resulting in the tool TMHMM [30].

3.1.4 Chaos Game Theory

To represent the sequences we have chosen a chaos game representation. This is restricted to a two-dimensional representation, meaning the vertex can only move along the x or y axis. For predicting protein-protein interactions this method is utilized, but they transformed the sequence into it's corresponding nucleotide representation [24].

TABLE 3.1: Categorizing Amino Acid Properties

Hydrophobic	Single Letter Amino Acid
Group 1	FIWLVM
Group 2	YCNA
Group 3	THGSQBZ
Group 4	RKDEPD

Acceptor	Single Letter Amino Acid
Group 1	ACGILMFPV
Group 2	DEBZ
Group 3	RKW
Group 4	NQHSTY

Mass	Single Letter Amino Acid
Group 1	GA
Group 2	SPVTC
Group 3	ILNDBQKEZMH
Group 4	FRYW

Pka_1	Single Letter Amino Acid
Group 1	DBCNFPH
Group 2	EZTQYSKR
Group 3	MVGLAI
Group 4	W

pI	Single Letter Amino Acid
Group 1	DEBZ
Group 2	CNFTQYSM
Group 3	WVGLAIP
Group 4	HKR

3.1.5 Amino Acid Properties

For our experiment, we have chosen the following five amino acid properties:

Hydrophobicity - How soluble a given amino acid is

Acceptor - How an amino acid behaves with hydrogen bonds

Mass - The weight of an amino acid

Pka_1 - The carboxyl group pH level of an amino acid

pI - Isoelectic point; pH level necessary for amino acid to be neutral

Since we are restricted to a two-dimensional plane, each amino acid is sorted into a group based on its attribute.

3.2 Wavelets

With a plot for each protein, we can begin to extract features to use as input for our classifiers. To achieve this we will use a type of wavelet for their prowess in analyzing images, or in our case an image of a graph. The Haar wavelet, albeit the most

TABLE 3.2: Classification Algorithms

Classifier	Parameters
K-Nearest Neighbors	Neighbors = Number of Locations
Linear Support Vector Classification	Error Term = 0.025
RBF SVM	Gamma = 2, Error Term = 1
Decision Tree	Max Depth = 5
Random Forest	Max Depth = 5, Estimators = 10
Neural Net	
AdaBoost	
Naive Bayes	
Quadratic Discriminant Analysis	

simple of wavelets, provides us with sufficient features generated from its rescaled square shaped functions [31]. We implemented this wavelet transformation using a python library, http://pywavelets.readthedocs.io/en/latest/ref/dwt-discrete-wavelet-transform.html. Additionally, we did three levels of decomposition which results in one approximation coefficient and three coefficients for each the x and y axis resulting in six coefficients. To expand our feature list, we observed the minimum, maximum, mean, and standard deviation for each coefficient presenting us with twenty-four features.

3.3 Classifiers

For the classification models, we chose to use sklearn's vast library of classifiers [32]. Our data was split using their stratified shuffle split function, which preserves the ratio of classes of each fold. With the larger classes having as many as ten times more samples than the smaller classes, this allows the classifier to adequately train for each class. We selected 9 classifiers to run our predictions. It has been observed in this field, that different classifiers can perform notably better than others [33], [34]. Listed in table 3.2 are the classifiers selected and their parameters, if any.

Chapter 4

Discussion

4.1 Results

Observing the accuracies, it appears that the plant predictions performed better than the other datasets by 15%. Upon reviewing the confusion matrix, the classifiers overfitted for chloroplast. Excluding the plant kingdom, the other classifiers performed respectfully. The property PKA_1 achieved the highest accuracies regardless of the kingdom. When using machine learning, other metrics are required to adequately represent how well the model performed. A high accuracy alone can depict excellent or mediocre performance. The metrics we include are: Accuracy (Correct Predictions / Total Predictions), Precision (Correct Predictions / Correct Predictions + False Positives), F1-score (2 * (Precision * Recall) / (Precision + Recall)), and Support (Total true samples for specific classes). For prediction models with the highest accuracy we will list these metrics to further examine the results.

4.1.1 Location Evaluation

Viewing the location metrics from the three kingdoms is sufficient enough to see that accuracy alone is misleading, without having to read a confusion matrix. You can see how the plant accuracy reported a high number, with the large support number of 4519 and getting the majority of those correct. The precision metric (53%) provides a better reading than the accuracy (94%). Observing the fungal and mammal metrics, it appears there is a correlation between the number of samples

TABLE 4.1: Average Accuracy by Group.

Group	Accuracy
Plant	46%
Fungi	31%
Mammal	33%
Human	28%
Rodent	28%

TABLE 4.2: Best Performing Classifiers.

Group	Property	Classifier	Accuracy	Precision	F1-Score	Support
Plant	PKA_1	Neural Net	51%	40%	40%	9192
Fungi	PKA_1	Adaboost	35%	30%	29%	7590
Mammal	PKA_1	QDA	41%	42%	40%	4772
Human	PKA_1	QDA	34%	33%	32%	4402
Rodent	PKA_1	QDA	33%	32%	31%	6240

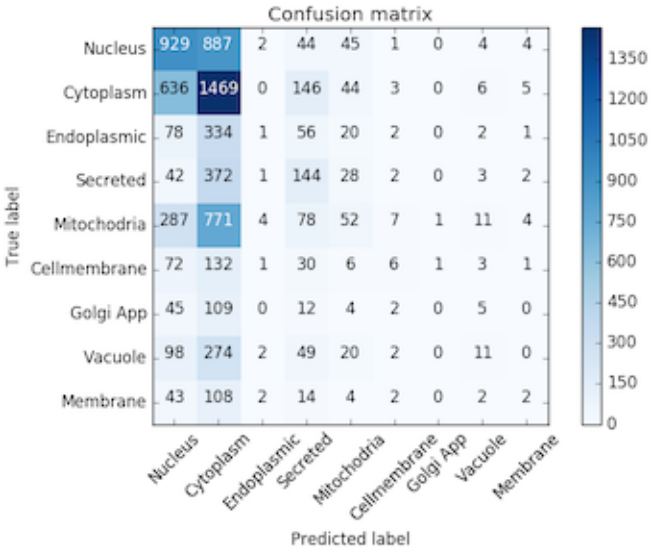

FIGURE 4.1: Adaboost - Fungi

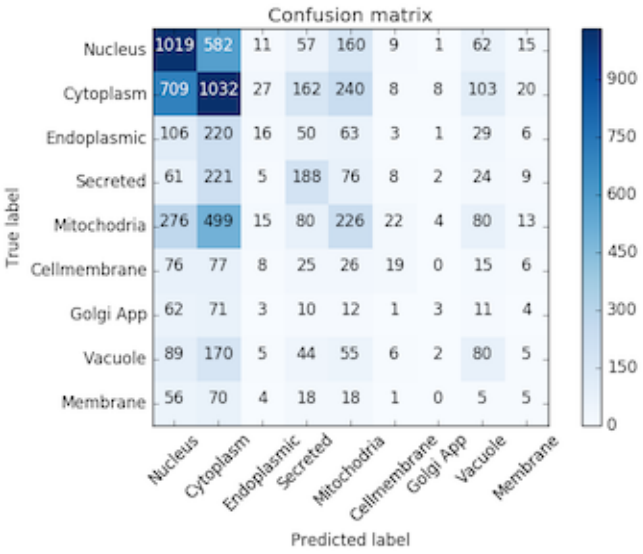

FIGURE 4.2: QDA - Fungi

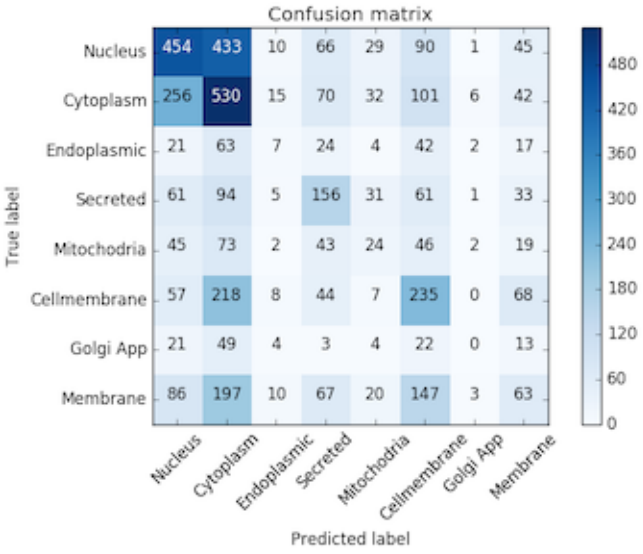

FIGURE 4.3: QDA - Human

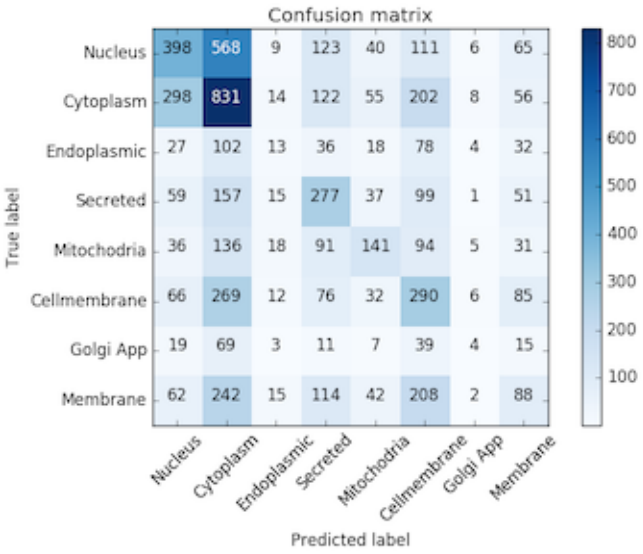

FIGURE 4.4: QDA - Rodent

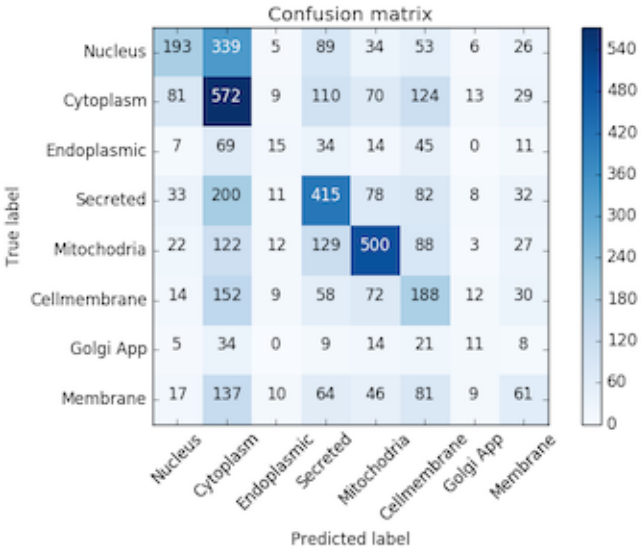

FIGURE 4.5: QDA - Mammal

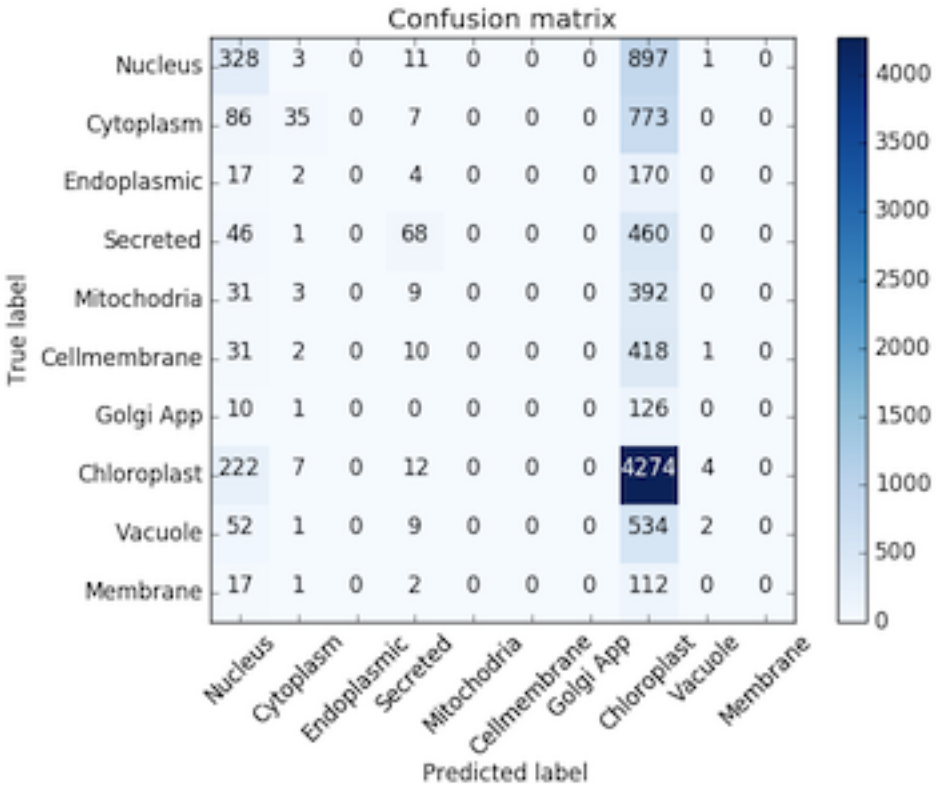

FIGURE 4.6: Neural Net - Plant

TABLE 4.3: Plant Location Classification Metrics.

Location	precision	Accuracy	f1-score	support
Nucleus	0.41	0.22	0.29	1240
Cytoplasm	0.30	0.06	0.10	901
ER	0.00	0.00	0.00	193
Secreted	0.41	0.19	0.26	575
Mitochondria	0.00	0.00	0.00	435
Cell membrane	0.23	0.01	0.01	462
Golgi Apparatus	0.00	0.00	0.00	137
Chloroplast	0.53	0.94	0.68	4519
Membrane	0.22	0.03	0.06	598
Vacuole	0.00	0.00	0.00	132

TABLE 4.4: Fungal Location Classification Metrics.

Location	precision	Accuracy	f1-score	support
Nucleus	0.43	0.49	0.46	1916
Cytoplasm	0.33	0.65	0.44	2309
ER	0.00	0.00	0.00	494
Secreted	0.23	0.22	0.23	594
Mitochondria	0.23	0.04	0.07	1215
Cell membrane	0.14	0.01	0.02	252
Golgi Apparatus	0.50	0.01	0.01	177
Membrane	0.21	0.04	0.07	456
Vacuole	0.17	0.01	0.01	177

and f1-score. From this we conclude that with a larger data set our predictions could improve.

4.2 WoLF PSORT Benchmark

We chose to run our datasets against WoLF PSORT, a general predictor. This evaluation will check the integrity of our dataset, and compare how our classifications performed. Unlike our straightforward approach, WoLF PSORT uses a variety of class features, such as sorting signals, amino acid composition, and functional motifs [17]. This tool covers many of the same areas that we predicted, with the inclusion of peroxisome. Our datasets did not contain enough proteins located in the peroxisome to build a model around. An additional restraint we placed on this benchmark was limiting the WoLF PSORT results to one location to match our experiment. Normally WoLF PSORT predicts several locations for one protein, listing a confidence score to the location.

TABLE 4.5: Mammal Location Classification Metrics.

Location	precision	Accuracy	f1-score	support
Nucleus	0.52	0.26	0.35	745
Cytoplasm	0.35	0.57	0.43	1008
ER	0.21	0.08	0.11	195
Secreted	0.46	0.48	0.47	859
Mitochondria	0.60	0.55	0.58	903
Cell membrane	0.28	0.35	0.31	535
Golgi Apparatus	0.18	0.11	0.13	102
Membrane	0.27	0.14	0.19	425

TABLE 4.6: WoLF PSORT Performance

Location Abbreviations	
Secreted	S
Cytoplasm	C
Cell membrane	Cm
Peroxisome	P
Mitochondria	M
Nucleus	N
Golgi Apparatus	G
Endoplasmic Reticulum	E
Vacuole	V
Chloroplast	Ch

WoLF PSORT Accuracy	
Plant	15%
Fungi	61%
Mammal	43%
Human	56%
Rodent	56%

TABLE 4.7: WoLF PSORT Confusion Matrices.

Plant

Predicted Class \ Actual Class	S	C	Cm	P	M	N	G	E	Ch
S	502	94	53	6	21	116	2	9	696
C	257	1105	224	45	193	727	22	38	4648
Cm	86	174	327	9	66	215	7	6	1457
P	8	16	2	25	3	11	0	0	62
M	56	125	42	5	95	111	18	14	743
N	222	570	227	20	172	1842	27	19	2587
G	2	8	5	0	4	6	1	0	9
E	33	18	16	3	7	20	0	15	221
Ch	637	800	385	33	314	1025	35	44	3798

4.2.1 Plant

For the plant dataset, WoLF PSORT classified the locations poorly, exhibiting the same behavior as our prediction model. Although not to the same degree, as many other locations were predicted to have proteins, but with mixed results. Therefore the dataset can not be blamed because the WoLF PSORT tool has models built into its algorithm, and does not train on our skewed data. The cytoplasm effect is also evident here, responsible for most of the incorrect predictions, including 4,648 from chloroplast alone.

4.2.2 Fungi

Scoring a 61% accuracy, albeit for 8 locations instead of 10 like plant, the tool fared rather well. Again, cytoplasm led the prediction totals for three locations, including its own. For individual locations the tool varied it's accuracy, recording a 91% true positive rate for secreted proteins, but a 26% for peroxisome proteins. This increase for secreted proteins can be attributed to looking for N-terminal which is present on classical secretory proteins.

4.2.3 Mammal and Mammal subsets

In the mammal kingdom the accuracies dipped compared to fungi, but are still insightful for multiclass prediction. The behavior was actually the opposite of our model, with accuracy rising when handling single specie datasets, which should

Fungi

Predicted Class \ Actual Class	S	C	Cm	P	M	N	G	E
S	1806	92	170	4	25	22	63	60
C	112	2934	76	48	151	763	21	22
Cm	13	55	348	1	28	79	56	28
P	1	12	0	6	2	1	1	0
M	26	649	59	18	1490	403	16	9
N	13	3746	176	20	312	5107	80	23
G	2	8	5	0	4	6	1	0
E	6	4	10	0	1	9	29	15

Actual Class

Mammal

Predicted Class \ Actual Class	S	C	Cm	P	M	N	G	E
S	1806	92	170	4	25	22	63	60
C	112	2934	76	48	151	763	21	22
Cm	13	55	348	1	28	79	56	28
P	1	12	0	6	2	1	1	0
M	26	649	59	18	1490	403	16	9
N	13	3746	176	20	312	5107	80	23
G	2	8	5	0	4	6	1	0
E	6	4	10	0	1	9	29	15

Actual Class

encourage a higher precision level. This encapsulates the differences between the two models, where ours bases predictions off of a single feature source, and where theirs is fine tuned, and have a different set of requirements for each location.

Human

Predicted Class	Actual Class: S	C	Cm	P	M	N	G	E
S	1806	92	170	4	25	22	63	60
C	112	2934	76	48	151	763	21	22
Cm	13	55	348	1	28	79	56	28
P	1	12	0	6	2	1	1	0
M	26	649	59	18	1490	403	16	9
N	13	3746	176	20	312	5107	80	23
G	2	8	5	0	4	6	1	0
E	6	4	10	0	1	9	29	15

Rodent

Predicted Class	Actual Class: S	C	Cm	P	M	N	G	E
S	1806	92	170	4	25	22	63	60
C	112	2934	76	48	151	763	21	22
Cm	13	55	348	1	28	79	56	28
P	1	12	0	6	2	1	1	0
M	26	649	59	18	1490	403	16	9
N	13	3746	176	20	312	5107	80	23
G	2	8	5	0	4	6	1	0
E	6	4	10	0	1	9	29	15

Chapter 5

Conclusion

Not all amino acid properties performed as well as others, but they do provide insight for predicting subcellular locations. It's became clear that the property Pka_1 contains more information than the others, and that quantitative discriminant analysis was the best fit for the features extracted from the wavelet transformations. It's evident to see why prediction tools use a variety of approaches to designing an algorithm to boost accuracy. Although simply using amino acid properties holds some information about where the protein resides, the biological nature of this problem is too complex to approach in that manner.

5.0.1 The Cytoplasm effect

Most of the incorrect predictions were classified as cytoplasm. Labeling those as incorrect may not be the most accurate observation. Proteins originate in the cytoplasm, and then make their way to their desired location, and even then they could cross the cytoplasm again if they translocate. Our experiment has flaws in two areas. The first, upon annotating proteins in a cell, they might not be in their desired location, and this is impossible to account for until we are able to annotate proteins by viewing a live cell in action. The second, proteins are bound to share the same attributes as other proteins that remain in the cytoplasm since they are incepted there.

5.1 Future Directions

Expanding on this experiment in the future, we would allow the chaos game plot to be of varying dimensions. This allows for better groupings of properties, rather than forcing four classes. We would have the freedom to gracefully select the desired amount. Image analysis of the chaos game plot could be improved, whether it be a larger step size or smaller plot, the whitespace was prominent in all plots regardless of sequence length. Even decomposing the amino acids into their respective nucleotide composition and using those properties to expand sequences is under consideration.

Bibliography

[1] e. a. Eric S. Lander, "Initial sequencing and analysis of the human genome", *Nature*, vol. 409, no. 6822, pp. 860–921, Feb. 2001. [Online]. Available: http://dx.doi.org/10.1038/35057062.

[2] R. P. Horgan and L. C. Kenny, "'omic' technologies: Genomics, transcriptomics, proteomics and metabolomics", *The Obstetrician and Gynaecologist*, vol. 13, no. 3, pp. 189–195, 2011, ISSN: 1744-4667. DOI: 10.1576/toag.13.3.189.27672. [Online]. Available: http://dx.doi.org/10.1576/toag.13.3.189.27672.

[3] A Pandey and M Mann, "Proteomics to study genes and genomes.", eng, *Nature*, vol. 405, no. 6788, pp. 837–846, 2000, ISSN: 0028-0836 (Print); 0028-0836 (Linking). DOI: 10.1038/35015709.

[4] M.-S. Kim, S. M. Pinto, D. Getnet, R. S. Nirujogi, S. S. Manda, R. Chaerkady, A. K. Madugundu, D. S. Kelkar, R. Isserlin, S. Jain, J. K. Thomas, B. Muthusamy, P. Leal-Rojas, P. Kumar, N. A. Sahasrabuddhe, L. Balakrishnan, J. Advani, B. George, S. Renuse, L. D. N. Selvan, A. H. Patil, V. Nanjappa, A. Radhakrishnan, S. Prasad, T. Subbannayya, R. Raju, M. Kumar, S. K. Sreenivasamurthy, A. Marimuthu, G. J. Sathe, S. Chavan, K. K. Datta, Y. Subbannayya, A. Sahu, S. D. Yelamanchi, S. Jayaram, P. Rajagopalan, J. Sharma, K. R. Murthy, N. Syed, R. Goel, A. A. Khan, S. Ahmad, G. Dey, K. Mudgal, A. Chatterjee, T.-C. Huang, J. Zhong, X. Wu, P. G. Shaw, D. Freed, M. S. Zahari, K. K. Mukherjee, S. Shankar, A. Mahadevan, H. Lam, C. J. Mitchell, S. K. Shankar, P. Satishchandra, J. T. Schroeder, R. Sirdeshmukh, A. Maitra, S. D. Leach, C. G. Drake, M. K. Halushka, T. S. K. Prasad, R. H. Hruban, C. L. Kerr, G. D. Bader, C. A. Iacobuzio-Donahue, H. Gowda, and A. Pandey, "A draft map of the human proteome", *Nature*, vol. 509, no. 7502, pp. 575–581, May 2014. [Online]. Available: http://dx.doi.org/10.1038/nature13302.

[5] J. Zahiri, J. H. Bozorgmehr, and A. Masoudi-Nejad, "Computational prediction of protein–protein interaction networks: Algo-rithms and resources", *Current Genomics*, vol. 14, no. 6, pp. 397–414, Sep. 2013. DOI: 10.2174/1389202911314060004. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3861891/.

[6] S. Bencharit and M. B. Border, "Where are we in the world of proteomics and bioinformatics?", eng, *Expert Rev Proteomics*, vol. 9, no. 5, pp. 489–491, 2012,

ISSN: 1744-8387 (Electronic); 1478-9450 (Linking). DOI: 10.1586/epr.12.46.

[7] K. Chou and H. Shen, "Recent progress in protein subcellular location prediction", *Anal Biochem*, vol. 370, 2007. DOI: 10.1016/j.ab.2007.07.006. [Online]. Available: http://dx.doi.org/10.1016/j.ab.2007.07.006.

[8] J. Cedano, P. Aloy, J. A. Pérez-Pons, and E. Querol, "Relation between amino acid composition and cellular location of proteins", *J Mol Biol*, vol. 266, 1997. DOI: 10.1006/jmbi.1996.0804. [Online]. Available: http://dx.doi.org/10.1006/jmbi.1996.0804.

[9] R. Nair and B. Rost, "Sequence conserved for subcellular localization", *Protein Sci*, vol. 11, 2002. DOI: 10.1110/ps.0207402. [Online]. Available: http://dx.doi.org/10.1110/ps.0207402.

[10] M. A. Andrade, S. I. O'Donoghue, and B. Rost, "Adaptation of protein surfaces to subcellular location", *J Mol Biol*, vol. 276, 1998. DOI: 10.1006/jmbi.1997.1498. [Online]. Available: http://dx.doi.org/10.1006/jmbi.1997.1498.

[11] J. Meinken, G. Walker, C. R. Cooper, and X. J. Min, "Metazseckb: The human and animal secretome and subcellular proteome knowledgebase", *Database*, vol. 2015, bav077, 2015.

[12] G. Lum and X. J. Min, "Funseckb: The fungal secretome knowledgebase", *Database*, vol. 2011, bar001, 2011.

[13] Z. Lu, D. Szafron, R. Greiner, P. Lu, D. S. Wishart, B. Poulin, J. Anvik, C. Macdonell, and R. Eisner, "Predicting subcellular localizations of proteins using machine-learned classifiers", *Bioinformatics*, vol. 20, 2004. DOI: 10.1093/bioinformatics/btg447. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/btg447.

[14] S. Reumann, D. Buchwald, and T. Lingner, "Predplantpts1: A web server for the prediction of plant peroxisomal proteins.", eng, *Front Plant Sci*, vol. 3, p. 194, 2012, ISSN: 1664-462X (Electronic); 1664-462X (Linking). DOI: 10.3389/fpls.2012.00194.

[15] I. Small, N. Peeters, F. Legeai, and C. Lurin, "Predotar: A tool for rapidly screening proteomes for n-terminal targeting sequences", *Proteomics*, vol. 4, 2004. DOI: 10.1002/pmic.200300776. [Online]. Available: http://dx.doi.org/10.1002/pmic.200300776.

[16] A. C. Smith and A. J. Robinson, "Mitominer v3.1, an update on the mitochondrial proteomics database.", eng, *Nucleic Acids Res*, vol. 44, no. D1, pp. D1258–61, 2016, ISSN: 1362-4962 (Electronic); 0305-1048 (Linking). DOI: 10.1093/nar/gkv1001.

[17] P. Horton, K.-J. Park, T. Obayashi, N. Fujita, H. Harada, C. Adams-Collier, and K. Nakai, "Wolf psort: Protein localization predictor", *Nucleic Acids Research*, vol. 35, no. Web Server issue, W585–W587, Jul. 2007. DOI: 10.1093/nar/gkm259. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1933216/.

[18] M. Tardif, A. Atteia, M. Specht, G. Cogne, N. Rolland, S. Brugiere, M. Hippler, M. Ferro, C. Bruley, G. Peltier, O. Vallon, and L. Cournac, "Predalgo: A new subcellular localization prediction tool dedicated to green algae.", eng, *Mol Biol Evol*, vol. 29, no. 12, pp. 3625–3639, 2012, ISSN: 1537-1719 (Electronic); 0737-4038 (Linking). DOI: 10.1093/molbev/mss178.

[19] W. L. Huanq, C. W. Tunq, S. W. Ho, S. F. Hwang, and S. Y. Ho, "Proloc-go: Utilizing informative gene ontology terms for sequence-based prediction of protein subcellular localization", *BMC Bioinformatics*, vol. 9, 2008. DOI: 10.1186/1471-2105-9-80. [Online]. Available: http://dx.doi.org/10.1186/1471-2105-9-80.

[20] E. I. Petsalaki, P. G. Bagos, Z. I. Litou, and S. J. Hamodrakas, "Predsl: A tool for the n-terminal sequence-based prediction of protein subcellular localization", *Genomics Proteomics Bioinformatics*, vol. 4, 2006. DOI: 10.1016/S1672-0229(06)60016-8. [Online]. Available: http://dx.doi.org/10.1016/S1672-0229(06)60016-8.

[21] M. Boden and J. Hawkins, "Prediction of subcellular localization using sequence-biased recurrent networks", *Bioinformatics*, vol. 21, 2005. DOI: 10.1093/bioinformatics/bti372. [Online]. Available: http://dx.doi.org/10.1093/bioinformatics/bti372.

[22] E Eisenstein, G. L. Gilliland, O Herzberg, J Moult, J Orban, R. J. Poljak, L Banerjei, D Richardson, and A. J. Howard, "Biological function made crystal clear - annotation of hypothetical proteins via structural genomics.", eng, *Curr Opin Biotechnol*, vol. 11, no. 1, pp. 25–30, 2000, ISSN: 0958-1669 (Print); 0958-1669 (Linking).

[23] P. R. Graves and T. A. J. Haystead, "Molecular biologist's guide to proteomics", *Microbiology and Molecular Biology Reviews*, vol. 66, no. 1, pp. 39–63, Mar. 2002. DOI: 10.1128/MMBR.66.1.39-63.2002. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC120780/.

[24] J. H. Jia, Z Liu, X Chen, X Xiao, and B. X. Liu, "Prediction of protein-protein interactions using chaos game representation and wavelet transform via the random forest algorithm.", eng, *Genet Mol Res*, vol. 14, no. 4, pp. 11 791–11 805, 2015, ISSN: 1676-5680 (Electronic); 1676-5680 (Linking). DOI: 10.4238/2015.October.2.13.

[25] C. H. Wu, R. Apweiler, A. Bairoch, D. A. Natale, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, R. Mazumder, C. O'Donovan, N. Redaschi, and B. Suzek, "The universal protein resource (uniprot): An expanding universe of protein information", *Nucleic Acids Research*, vol. 34, no. Database issue, pp. D187–D191, Jan. 2006. DOI: 10.1093/nar/gkj161. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1347523/.

[26] I. P. Ivanov, A. E. Firth, A. M. Michel, J. F. Atkins, and P. V. Baranov, "Identification of evolutionarily conserved non-aug-initiated n-terminal extensions in human coding sequences", *Nucleic Acids Research*, vol. 39, no. 10, pp. 4220–4234, May 2011. DOI: 10.1093/nar/gkr007. [Online]. Available: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3105428/.

[27] K.-C. Chou and H.-B. Shen, "Plant-mploc: A top-down strategy to augment the power for predicting plant protein subcellular localization", *PloS one*, vol. 5, no. 6, e11335, 2010.

[28] M. H. Smith, "The amino acid composition of proteins", *Journal of Theoretical Biology*, vol. 13, pp. 261 –282, 1966, ISSN: 0022-5193. DOI: http://dx.doi.org/10.1016/0022-5193(66)90021-X. [Online]. Available: http://www.sciencedirect.com/science/article/pii/002251936690021X.

[29] K. Chou and Y. Cai, "Prediction and classification of protein subcellular location - sequence-order effect and pseudo amino acid composition", *J Cell Biochem*, vol. 90, 2003. DOI: 10.1002/jcb.10719. [Online]. Available: http://dx.doi.org/10.1002/jcb.10719.

[30] A. Krogh, B. Larsson, G. Von Heijne, and E. L. Sonnhammer, "Predicting transmembrane protein topology with a hidden markov model: Application to complete genomes", *Journal of molecular biology*, vol. 305, no. 3, pp. 567–580, 2001.

[31] A. Haar, "Zur theorie der orthogonalen funktionensysteme", *Mathematische Annalen*, vol. 69, no. 3, pp. 331–371, 1910. DOI: 10.1007/BF01456326. [Online]. Available: http://dx.doi.org/10.1007/BF01456326.

[32] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python", *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[33] K. S. C. G.-H. Munyon JD Min X, "Prediction of subcellular locations for fungal proteins", in *Proceeding of the Joint Statistics Meeting 2015 (JSM2015)*, American Statistical Association, 2015, pp. 2497–2508.

[34] J. M. X. M. G.-H. C. Kofi Neizer-Ashun Feng Yu, "Prediction of plant protein subcellular locations", in *7th International Conference On Bioinformatics And Computational Biology*, International Society for Computers and Their Applications, 2015, pp. 91–96.