NOVEL DISCOVERY OF BACTERIAL OVARIAN TUMOR DEUBIQUITINASES FROM PRIMARY AMINO ACID SEQUENCE

by

Justine Nguyen

A Thesis

Submitted to Oregon Health & Science University
in partial fulfillment of the requirements for the degree of

Master of Science

School of Medicine
Department of Medical Informatics and Clinical Epidemiology
Portland, Oregon
September 2020

School of Medicine

Oregon Health & Science University

CERTIFICATE OF APPROVAL

This is to certify that the Master's thesis of

Justine Nguyen

has been approved

"Novel Discovery of Bacterial Ovarian Tumor Deubiquitinases from Primary Amino Acid Sequence"

Acknowledgements

I would like to first thank my thesis advisor, Dr. Michael Mooney for all his mentorship and guidance throughout this thesis project. I have learned so much about machine learning throughout this project. I express my sincere gratitude to my thesis advisory committee member Dr. Ted Laderas for his continual encouragement and teaching me how to become a better storyteller. A huge thanks to Dr. Jonathan Pruneda for his leadership within the ubiquitin field and showing me that ubiquitin is in fact the best post-translational modifier to exist.

I am grateful to the Department of Medical Informatics and Clinical Epidemiology, including students, faculty, and staff. I am especially thankful for Diane Doctor for always keeping me on track.

Finally, this all could not have been done without all the love, support, and encouragement I have received from my parents, sister and friends.

Table of Contents

List of Figures and Tables 5

FIGURES 5
TABLES 5

1. Abstract 6

2. Background 7

2.1 UBIQUITIN SYSTEM OVERVIEW 7
2.1 DEUBIQUITINASES 8
2.2 OVARIAN TUMOR DEUBIQUITINASES 9
2.3 PREVIOUS PROTEIN PREDICTION THROUGH MACHINE LEARNING 10
2.4 RESEARCH QUESTION 13
Gaps in Research 13
Research Question 14
Central Hypothesis 14
Data Acquisition 14
Specific Aim 1 15
Specific Aim 2 16

3. Methods 19

3.1 DATA ACQUISITION 19
3.2 FEATURE GENERATION 19
K-mer features with Reduced Amino Acid Encoding 19
Composition Features 20
Composition, Distribution, Transition Features 20
3.3 MODEL BUILDING OVERVIEW 21
3.4 AVERAGE EDIT DISTANCES 23

4. Results 24

4.1 DATA ACQUISITION 24
4.2 SPECIFIC AIM 1 25
Feature Filtering on k-mer Features 25
Feature Selection with RFE 26
Minimal k-mer Models 28
4.3 SPECIFIC AIM 2 29
Composition, Transition, Distribution Features 29
Composition, Dipeptide, and Tripeptide Features 31

5. Discussion 34

CONCLUSIONS 34
LIMITATIONS 34
FUTURE DIRECTIONS 35

6. References 37

List of Figures and Tables

Figures

Figure 1: Various cellular outcomes and linkages of ubiquitin. 7

Figure 2: Structural and functional comparison of known human OTUs............ 9

Figure 3: Sequence comparison of bacterial OTUs............ 10

Figure 4: Overview of Specific Aim 1. 15

Figure 5: Overview of Specific Aim 2. 17

Figure 6: Overview of data partitioning and training. 22

Figure 7: 5-fold cross-validation on rare and variance filters. 26

Figure 8: Validating RFE with the hydrophobicity k-mer. 27

Figure 9: Cross-validation of k-mer features in conjunction with additional CTD feature set. 30

Figure 10: Cross-validation of k-mer features in conjunction with additional composition feature sets. 31

Tables

Table 1: Summary of the different feature types from the literature review. 13

Table 2: Reduced amino acid encodings............ 20

Table 3: Composition, Transition, Distribution feature classification of amino acids based on hydrophobicity. 21

Table 4: k-mer performance on held out test set. 29

Table 5: k-mer and CTD features performance on held out test set. 30

Table 6: k-mer and single amino acid composition features performance on held out test set. 32

Table 7: k-mer and dipeptide composition features performance on held out test set............ 32

Table 8: k-mer and tripeptide composition features performance on held out test set. 32

1. Abstract

The post-translational modifier, ubiquitin, controls many aspects of eukaryotic cell biology, including key aspects of both innate and adaptive immune signaling. Remarkably, despite having no such system of their own, viruses and bacteria have evolved strategies to manipulate ubiquitin signaling of the host cell to support infection. A common strategy to both pathogens is the adaptation of specialized proteases, termed deubiquitinases, that can remove host ubiquitin signals. Though the current body of work suggests that the removal of host ubiquitin signals is a common strategy for virulence, the identification of novel deubiquitinases has been impeded by significant differences in primary sequences that likely indicate an evolutionary convergence in function. Since development of deubiquitinases typically derive from structural mimicry, this allows for diversity in the primary sequences that may make motif searching for deubiquitinases difficult.

To address the problem of identifying novel deubiquitinases despite significant diversity in primary amino acid sequences with a focused approach, we chose to study the Ovarian Tumor (OTU) family of deubiquitinases that contains many essential examples in humans and viruses, and limited validated examples in bacteria. Typically, the discovery of novel protein family members involves generating a sequence motif and manually curating predictions based on a sequence alignment approach. The motif approach is problematic in this case because the sequence homology across eukaryotic, viral, and bacterial OTU domains is quite poor. In fact, beyond unique mechanisms of substrate recognition and catalysis, we observe dramatic sequence permutations in the OTU fold that likely arose through convergent evolution. To address this barrier, our work uses a machine learning approach to identify key features of OTU deubiquitinases through generating features derived from the primary amino acid sequence without the reliance on known motifs. This approach allows for a search of distantly-related examples present in bacteria through the identification of underlying features that define an OTU domain. By identifying novel deubiquitinases in bacteria, we can improve our understanding of how bacteria manipulate the host ubiquitin system in a disease state and contribute methods to the field protein function prediction.

Keywords

ubiquitin, deubiquitinase, Ovarian Tumor, OTU, bacteria, support vector machine, primary sequence prediction

2. Background

2.1 Ubiquitin System Overview

Ubiquitin is a 76 amino acid protein that functions as a post-translational modifier through covalent attachment to proteins that it modifies. Proteins can be modified through a single attachment of ubiquitin known as monoubiquitination or multiple monoubiquitinations. A unique feature of ubiquitin is that it can be itself ubiquitinated through eight different sites: either through the N-terminal methionine (Met1) or any of its seven lysine (K) residues. These multiple attachment points on ubiquitin allow the post-translational modifier to diversely regulate target proteins as it forms unique signals through its various polymeric forms. Each of the different lysine or N-terminal methionine linkage forms are associated with various cellular outcomes. Aside from polyubiquitin, ubiquitin can also be modified with other post-translational modifiers, thus further increasing its diversity through distinct chain types.[1] Below in Figure 1A, the different associated cellular outcomes are depicted with their respective chain types and Figure 1B shows the diversity in regulation of ubiquitin through its diverse chain types.[2]

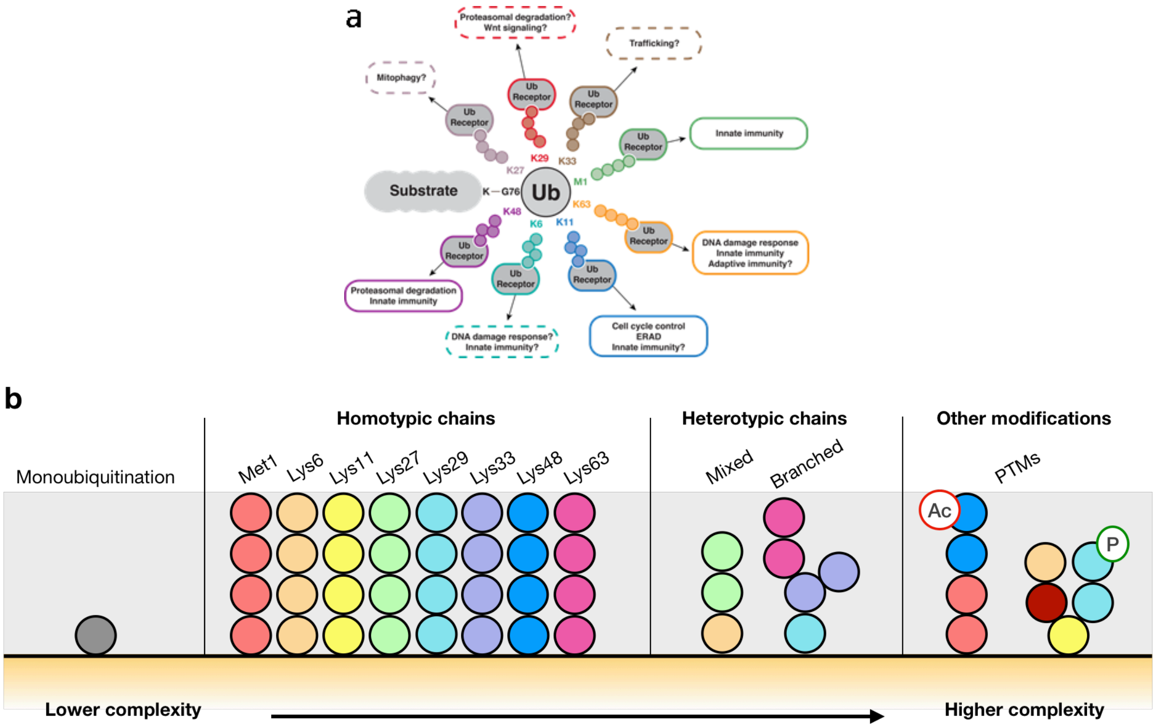


Figure 1: Various cellular outcomes and linkages of ubiquitin. *A) Different cellular outcomes of each ubiquitin chain type. Functions in solid bubbles are known functions of that specific chain types while dotted bubbles are speculative and have some evidence to suggest that functionality. B) Diversity in chain types that can be made with ubiquitin in terms of level of complexity. Figure1B was adapted from Mevissen, et al., 2017.*[2]

The regulation of the ubiquitin signal is cyclic in the eukaryotic host cell, where dedicated families of proteins can 'write', 'read', and 'erase' the signals. The ubiquitin signal is formed through the passing of the ubiquitin molecule onto the following cascade: E1 ubiquitin-activating enzyme, E2 ubiquitin-conjugating enzymes and E3 ubiquitin ligase. The ubiquitin signals can be read by proteins that recognize ubiquitin domains and translate the signal into cellular outcomes. The ubiquitin modifications can be removed through specialized proteases called deubiquitinases (DUBs) that have the ability to disassemble and recycle the ubiquitin molecules. Ubiquitin has evolved uniquely in eukaryotes, but bacteria and viruses have also developed E3 ligases and DUBs that hijack host signaling responses to support infection.

2.1 Deubiquitinases

There are eight known families of DUBs, including ubiquitin-specific proteases (USPs), ubiquitin-like proteases (ULPs), ubiquitin C-terminal hydrolases (UCHs), Josephin, JAB1/MPN/MOV34 metalloenzymes (JAMMs), motif interacting with Ub-containing novel DUB family (MINDY), Zinc finger containing Ub Peptidase (ZUP), and ovarian tumor (OTUs).[2,3] Distinguishing features of the various families include structural differences along with reliance on the differing mechanisms of ubiquitin or ubiquitin-like molecule specificity, recognition, and mechanisms of removal.[4] Dysregulation of DUBs has been implicated in many diseases including cancer, inflammation, neurodegeneration, and viral infection, which makes it a desirable target to study within viruses and bacteria.[3] Although there are not many DUBs in bacteria that are known, a thorough screen through genomes of pathogenic bacteria can allow us to discover DUBs that may play a role in pathogenesis.

Of the eight known families of DUBs, there are well-characterized examples of viral and bacterial proteins in the ULP family, but there is a gap in knowledge in our understanding of viral and bacterial examples across all the other families. Given the lack of data across other DUB families, we chose to focus our study on OTUs, for which there are a few viral and bacterial examples known. Given the available data, OTUs are a logical choice for building and validating the concept that we could learn from eukaryotic and viral examples and apply them to discovering bacterial OTUs. Through discovering and understanding more bacterial examples, we can later see how these bacterial OTUs support invasion and pathogenicity of a diseased state.

2.2 Ovarian Tumor Deubiquitinases

OTUs are the second largest family of DUBs with 16 OTU proteins in humans.[5] It has been shown that OTUs play roles in signaling cascades (including NFκB, interferon, and p97 signaling), the DNA damage response, and inflammation.[3] A common factor dictating classification of OTUs relate to the mechanism of action for deubiquitination. OTUs contain a catalytic triad, consisting of a catalytic cysteine, acidic residue (typically aspartate), and a basic histidine residue that form this active site. The composition of the catalytic triad helps to activate the enzyme in order to catalyze the reaction of hydrolyzing the peptide bond between ubiquitin chains. Although the mechanism is uniform across OTUs, the composition of the protein structure can be very diverse. Figure 2A shows the evolutionary conservation across the human examples of OTUs, but in Figure 2B, the structural makeup of these proteins is very diverse in structure and length, even despite belonging to the same family of proteins within one species. The differences amongst the human OTUs is further highlighted when we look across the preferred ubiquitin chain types these enzymes hydrolyze, as seen in Figure 2C, as they ultimately affect different cellular outcomes through the large range in chain type preferences.[5] Although we have good a understanding of human OTUs, their vast structural and functional differences may be problematic when attempting to predict viral and bacterial OTUs.

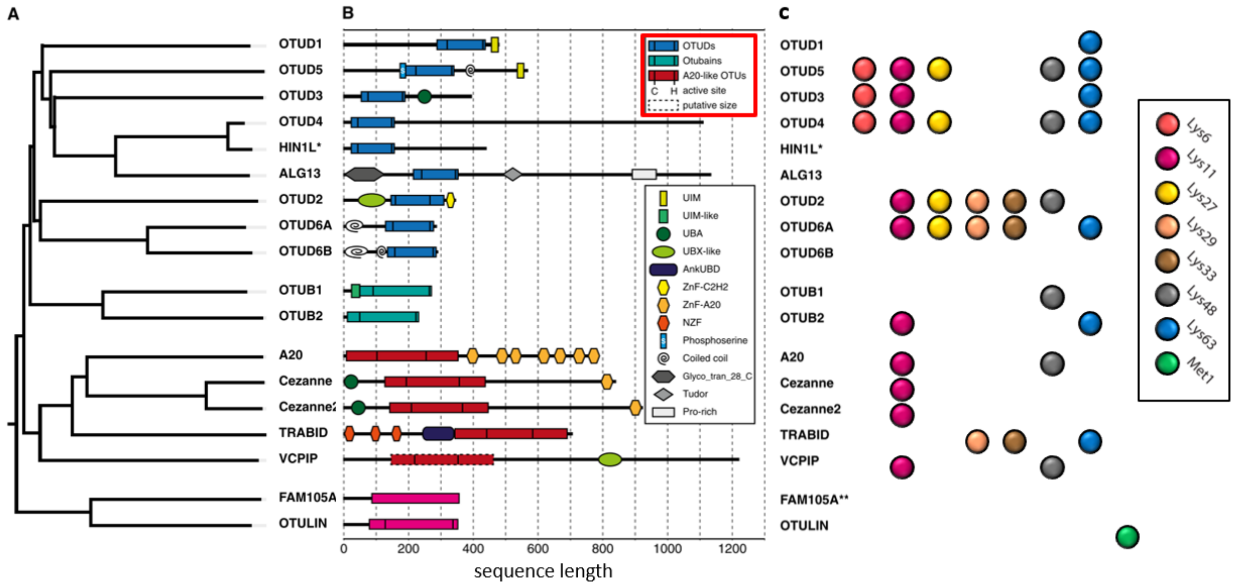
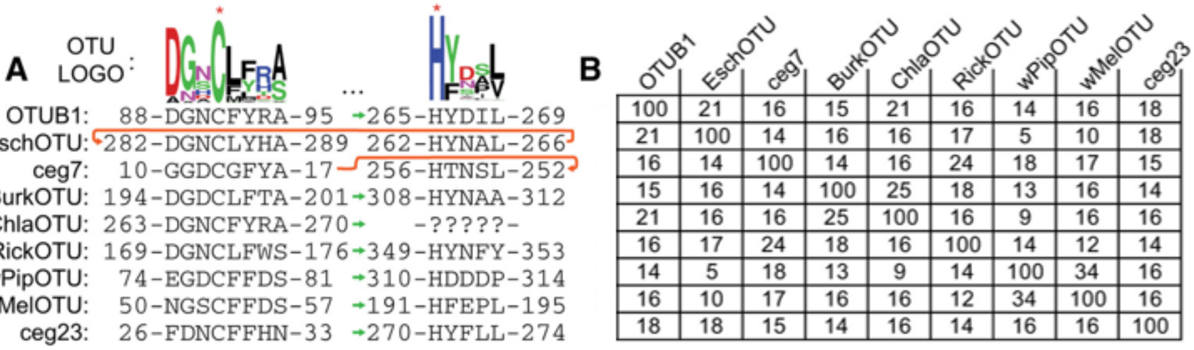


Figure 2: Structural and functional comparison of known human OTUs. *A) Phylogenetic tree of the human OTUs. Figure from Mevissen, et al., 2013[5]. B) Structural comparison of the human OTUs. Figure from Mevissen, et al., 2013[5]. C) Preferred chain types of the human OTUs.*

Previously there were examples of known viral OTUs, but only a few bacterial examples in the literature.[6–8] In our recent study, a set of seven OTUs were predicted and validated in pathogenic bacteria.[9] This further increases the number of examples for known bacterial OTUs. Although all the bacterial examples found by Schubert, *et al.* showed deubiquitinating activity, these bacterial proteins were even more diverse than the human OTUs. Figure 3A shows the sequence logo of the bacterial proteins within the region of the catalytic triad, however, two of the bacterial OTUs, EschOTU and ceg7, contain permutations and rearrangements of the domain. Furthermore, Figure 3B shows the low sequence identity amongst all the proteins in a percent identity matrix.[9] Ultimately, we have a small number of examples of bacterial OTUs, and even within these examples, there is low sequence identity and high diversity amongst the examples.


A
OTU LOGO:
...
OTUB1: 88-DGNCFYRA-95 →265-HYDIL-269
EschOTU: 282-DGNCLYHA-289 262-HYNAL-266
ceg7: 10-GGDCGFYA-17 256-HTNSL-252
BurkOTU: 194-DGDCLFTA-201→308-HYNAA-312
ChlaOTU: 263-DGNCFYRA-270→ -?????-
RickOTU: 169-DGNCLFWS-176→349-HYNFY-353
wPipOTU: 74-EGDCFFDS-81 →310-HDDDP-314
wMelOTU: 50-NGSCFFDS-57 →191-HFEPL-195
ceg23: 26-FDNCFFHN-33 →270-HYFLL-274
B

OTUB1	EschOTU	ceg7	BurkOTU	ChlaOTU	RickOTU	wPipOTU	wMelOTU	ceg23
100	21	16	15	21	16	14	16	18
21	100	14	16	16	17	5	10	18
16	14	100	14	16	24	18	17	15
15	16	14	100	25	18	13	16	14
21	16	16	25	100	16	9	16	16
16	17	24	18	16	100	14	12	14
14	5	18	13	9	14	100	34	16
16	10	17	16	16	12	34	100	16
18	18	15	14	16	14	16	16	100

Figure 3: Sequence comparison of bacterial OTUs. *A) Sequence logo generated surrounding the OTU domain comparing the human OTU, OTUB1, and predicted bacterial OTUs. Asterisks highlight residues that form the catalytic triad. Green and red arrows highlight the arrangement of the sequences to the logo. B) Percent identity matrix of the OTU domains using PSI-Coffee alignment.*[9]

2.3 Previous Protein Prediction through Machine Learning

Protein classification allows for identifying groups of proteins with similar structure, activities, and metabolic roles in the cell.[10] One common approach taken in the protein classification field is to perform a binary classification task of classifying positive examples (family of interest) versus negative examples (proteins not belonging to that family). To this end, an applicable machine learning method used is a support vector machine (SVM) to perform the binary classification task between the two classes. SVMs determine hyperplanes (decision boundaries) by maximizing the margins (distance) between support vectors (points from different classes that are closest to each other). In the case when there is not a linearly separable decision boundary solution, the hyperplane can be mapped into higher dimensions.[11] Within the protein prediction and classification field, there is significant literature on the utilization of SVMs because they allow for interpretation of the relative importance of features through the weights. Table 1 details features generated from primary amino acid sequences used in protein predictions with different applications.

In terms of features that can be generated strictly from protein sequence, amino acid composition is the simplest feature type. There have been many examples where composition allows for good performance in classification of various protein families. Samudrala *et al.* used amino acid composition (referring to the frequency of occurrence of the natural amino acids in the primary sequence) as a feature set when training their SVM classifier to identify bacterial type III secreted proteins.[12] With amino acid composition as a feature type, they were able to achieve an area under the receiver operating characteristic (ROC) curve (AUC, a representation of accuracy) of 95%, sensitivity (proportion of actual positive cases) of 90%, and specificity (proportion of actual negative cases) of 88%. There have also been other studies where computational modeling used successive dipeptide and tripeptide composition of the amino acids to maintain more integrity of the positions on the amino acids. Although dipeptide and tripeptide composition increases the original feature space from 20 (the number of genetically encoded amino acids) to 20^2 (400) and 20^3 (8000), respectively, there is a gain of information due to the order that exists in protein sequences that dictate their protein family or protein function.[13,14] Bhasin and Raghala were able to achieve a Matthew's correlation coefficient (MCC, metric of accuracy that accounts for unbalanced classes in binary classification tasks) of 0.81 and an AUC of 97.5% in the task of classifying G-coupled protein receptors (GCPRs) with dipeptide composition.[13] Wang, *et al.* were able to improve upon the feature type and achieved an MCC of 0.96, sensitivity of 95.4%, and a specificity of 97.2% while classifying GPCRs with the utilization of tripeptide composition, showing that higher composition could increase the performance of the classification.[14] Amino acid composition features are strictly sequence-based types of features, but have been previously shown to be distinguishable in protein family classification tasks.

Other common types of features extracted from primary amino acid sequence are Composition, Transition, Distribution (CTD) features. These features could be mapped based on different physiochemical properties, especially for the Distribution (D) and Transition (T) features. Once an encoding based on physiochemical property is chosen, (examples include polarity, hydrophobicity, charge, normalized Van der Waals volume, polarity, solvent accessibility), the features can then be built. Composition (C) refers to the proportion of binned amino acids for a specific physiochemical property within the full-length sequence. Distribution features are the proportion of the specific property within the beginning, 25%, 50%, 75%, and full length of the sequence. Transition features are the proportion of changes from one property to another (ie: hydrophobic to hydrophilic switch in a hydrophobicity encoding). Li, *et al.* show that this representation of amino acid sequence can perform well across many protein families with either SVM or K-nearest neighbor (KNN) as machine learning tools.[15] This type of feature encompasses not only sequence-based but also physiochemical-based properties from the full length sequence derived from amino acids.

Previous work by McDermott, *et al.* shows that it is possible to generate features from ubiquitin E3 ligases (enzymes that are writers of the ubiquitin code) from the primary amino acid sequence in a binary classification task of distinguishing bacterial E3 ligase proteins from unrelated bacterial proteins.[16] The goal of the paper was to generate a model that could learn properties of the primary amino acid sequence that distinguish bacterial ligases so predictions could be generated from other bacterial strains lacking a known ligase. Originally the model was trained as a binary classifier with 164 positive cross-kingdom ligase examples curated from UniProt while 235 negative, non-ligase examples were extracted from literature. K-mers were generated for all sequences using reduced amino acid alphabets (RAA), which categorize the amino acids into alphabets based on physiochemical properties, including charge, hydrophobicity, solvent accessibility, and structure to reduce the feature space. An SVM classifier was trained using RAA k-mers ranging in length from three to twenty. In this method, feature generation is a mixture of sequence-based and function-based features because the amino acids were encoded based on physiochemical properties and then the amino acid composition of the features were then extracted after the encoding. To equally sample numbers of positive and negative ligase examples, proteins in the training data were binned into groups based on sequence similarity to account for unbalanced examples of ligases. After cross-validation across all RAAs with a k-mer range of 3 to 20 amino acids, the optimal encoding determined was the binary hydrophobicity RAA with a k-mer length of 14, resulting in an AUC of 90% in the held-out test data. Testing the model showed that the k-mer approach with RAA encoding based on an SVM model was sufficient for detecting known examples of bacterial E3 ligases that were held out in the training phase. This broad application with k-mers and RAA encoding could potentially be applied to learning the OTU family of DUBs without initial bias towards family motifs.[16] Since the goal was to computationally predict more novel bacterial ligases, the final optimized model was applied to the Pathosystems Resource Integration Center (PATRIC) database, which contains over 80,000 genomes of bacteria that are known interactors with the human proteome.[17]

Feature type	Feature	Performance metric	Performance	Application / Reference
Sequence-based & Functional-based	Reduced physiochemical encoding with k-mers	AUC	90%	Ubiquitin ligases (McDermott, et *al.)*
Functional-based	Composition, Distribution, Transition	Precision Sensitivity Specificity	93.3% 95.1% 99.9%	Actin capping (Li, et *al.)*
Functional-based	Amino acid composition	AUC Sensitivity Specificity	95% 90% 88%	Type III secretion (Samudrala, et *al.)*
Sequence-based	Dipeptide composition	AUC MCC	97.5% 0.81	GCPRs (Bhasin & Raghala)
Sequence-based	Tripeptide composition	MCC Sensitivity Specificity	0.96 95.4% 97.2%	GCPRs (Wang, et *al.)*

Table 1: Summary of the different feature types from the literature review. *Amino acid composition, dipeptide composition, and tripeptide composition are representative of sequence-based features, while composition/distribution/transition features are function-based. The reduced encoding with k-mers is a combination of sequence-based and function-based features. The order of features in the table represent increasing information content.*

2.4 Research Question

Gaps in Research

Most protein prediction algorithms strongly rely on sequence homology for predicting proteins. However, OTUs lack homology across known examples, especially in bacteria. This is due to a combination of poor sequence similarity and sequence reorganization that exists within the OTU fold across bacterial examples. Therefore, developing a classifier that uses features derived from the primary amino acid sequence, but is not dependent on sequence homology, is crucial for overcoming this barrier to learning OTUs.

Research Question

Our overarching goal is to increase understanding across all DUB families, especially in the contexts of viruses and bacteria. By narrowing the problem space specifically to the OTU family for this project, we can evaluate the potential of learning from eukaryotic and viral examples for the application of predicting in bacteria. The OTU family has a subset of known bacterial data with experimental validation, but other DUB families lack this information across all kingdoms of life so this could be a proof of principle in training within eukaryotic and viral examples for generalization to bacteria.

Our goal is to refine a computational model that trains on features based off the primary amino acid sequences of known eukaryotic and viral OTU proteins to further predict undiscovered bacterial OTUs. Key considerations in the research question include curating appropriate negative (non-OTU) and positive (OTU) sequences, determining prominent features that distinguish between the two groups in an SVM model, and assessing model performance amongst different models that are generated. The purpose of this project is to build a model that could generalize to bacteria from training with eukaryotic and viral examples. Ideally, the properties of the features will allow for generalizability given that we know that we cannot rely on a model built entirely based on homology.

Central Hypothesis

Our overarching hypothesis is that we can train a classifier with k-mer features encoded based on physiochemical properties to learn OTUs. This initial baseline model using k-mers based on the primary amino acid sequence reduced to physiochemical properties will perform sufficiently to learn features that classify an OTU without domain knowledge on OTUs or protein structure biochemistry. But with the addition of subsequent features that direct the classifier, we can better distinguish OTUs from other proteins without complete reliance on homology. The additional feature sets will be a combination of sequence-based and function-based features.

Data Acquisition

There are many protein databases (including PFAM, MEROPS, and InterPro) that classify proteins into various families or clans mainly through the reliance of computational methods such as pHMMs (profile Hidden Markov Models) or BLAST (Basic Local Alignment Search Tool).[18–20] Although these metrics are sufficient to generate predictions on whether a protein belongs in a specific family, we wanted to train a classifier off the most stringent examples of OTU proteins. UniProt curates protein sequences into two databases: UniProtKB/Swiss-Prot and UniProtKB/TrEMBL, where the former requires manual curation and experimental validation to be determined as a reviewed sequence and the latter can contain any level of evidence.[21]

To curate our negative examples, previous literature in protein family prediction shows that we can sample sequences that are not related to ubiquitin activity from the UniProtKB/Swiss-Prot database or pull sequences from families in the PFAM database that are not related to OTUs nor ubiquitin.[16] Our goal is to obtain only bacterial sequences for the negative set because we ultimately aim to predict novel bacterial OTUs. The full-length sequences will be downloaded from UniProt and are readily available to generate features for subsequent modeling.

Specific Aim 1

Aim 1 – Generate a baseline OTU classifier that utilizes features derived from the primary amino acid sequences of known eukaryotic and viral OTUs and assess its ability to generalize to the prediction of bacterial OTUs.

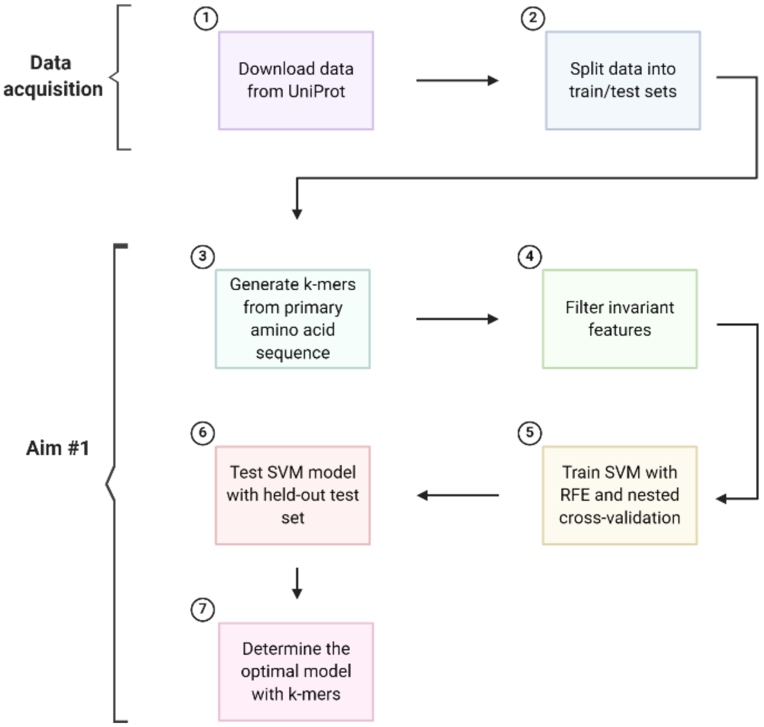

Data acquisition
1
Download data from UniProt
2
Split data into train/test sets
Aim #1
3
Generate k-mers from primary amino acid sequence
4
Filter invariant features
5
Train SVM with RFE and nested cross-validation
6
Test SVM model with held-out test set
7
Determine the optimal model with k-mers

Figure 4: Overview of Specific Aim 1. *Workflow diagram of Aim 1, generating a model based on k-mer features with reduced amino acid encoding.*

Viral and bacterial OTUs can be discovered through sequence alignment to the consensus motif of the OTU family. However, there are instances in bacteria where there are rearrangements of the catalytic triad of the OTU domain, and in these instances a classical sequence-based approach is not suitable. We aim to generate features from the entire primary amino acid sequence though encoding the sequences into k-mers (substrings of the sequence) and train the classifier with an SVM. To reduce the feature space, we will map the sequences into alphabets representing physiochemical properties, such as hydrophobicity, charge, solvent accessibility, and structure because from the native primary amino acid alphabet, with each increase in length, the features increases by a factor of 20. The performance of this initial model will be used as a baseline to compare to subsequent optimized models. The initial k-mer approach should give us flexibility to capture permutations derived through convergent evolution. Our initial approach utilizes the combination of both sequence-based and function-based features to try to classify known OTUs from negative examples.

To build the initial baseline model for binary classification for OTU positive versus negative proteins, counts of k-mers encoded in the various four reduced amino acid alphabets (including hydrophobicity, solvent accessibility, physiochemical properties, and charge), as described in McDermott, et al., 2019 will be generated.[16] The model will be trained using an SVM with recursive feature elimination (RFE) and downsampling of negative examples to balance the two classes in the training data. The performance of this initial model will be used as a baseline to compare to subsequent optimized models. This methodology is depicted in Figure 4.

Specific Aim 2

Aim 2 – Optimize the feature generation from the primary amino acid sequence through adding additional biologically relevant features to the baseline model.

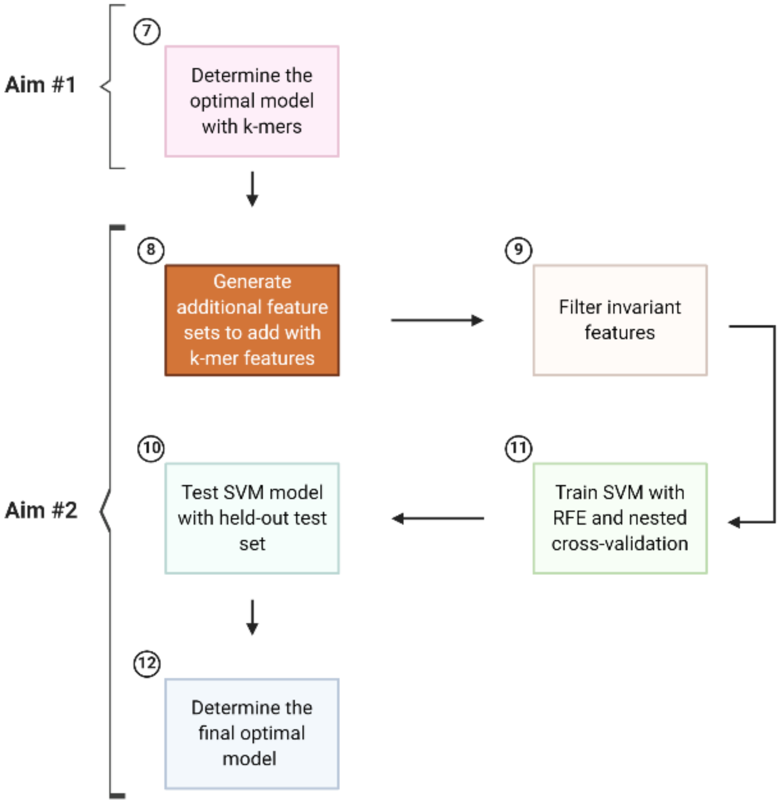

Aim #1
7
Determine the optimal model with k-mers
Aim #2
8
Generate additional feature sets to add with k-mer features
9
Filter invariant features
10
Test SVM model with held-out test set
11
Train SVM with RFE and nested cross-validation
12
Determine the final optimal model

Figure 5: Overview of Specific Aim 2. *Workflow diagram of Aim 2, imputing additional feature sets with the k-mer model generated from Aim 1.*

We aim to improve the initial model built from k-mers by adding additional feature sets. There are many examples using sequence-based and function-based features that are predictive of bacterial effectors, and this may help to prioritize the bacterial OTUs in predictions. We hypothesize that generating features through k-mers from the primary amino acid sequence can be predictive of protein families and by adding feature sets we can increase our prediction accuracies. We aim to generate features that do not rely completely on homology to optimize the model. Our additional sets of features are either more heavily sequence-based or function-based (exclusively) and we are going to test to see which feature type is more successful in predicting bacterial OTUs.

The baseline model will be improved through the addition of other features that are generated from the primary amino acid sequence that are likely to inform the classifier on OTUs. These features will be in addition to the k-mers that are used in the baseline model. We aim to direct the k-mer model with either sequence-based or functional-based features. The sequence-based features include amino acid composition, dipeptide composition, and tripeptide composition. The function-based features contain composition, transition, distribution features. The optimal model will be built by determining the most informative features in the classification task. This is built off the model generated from Aim 1, as shown in Figure 5.

3. Methods

3.1 Data Acquisition

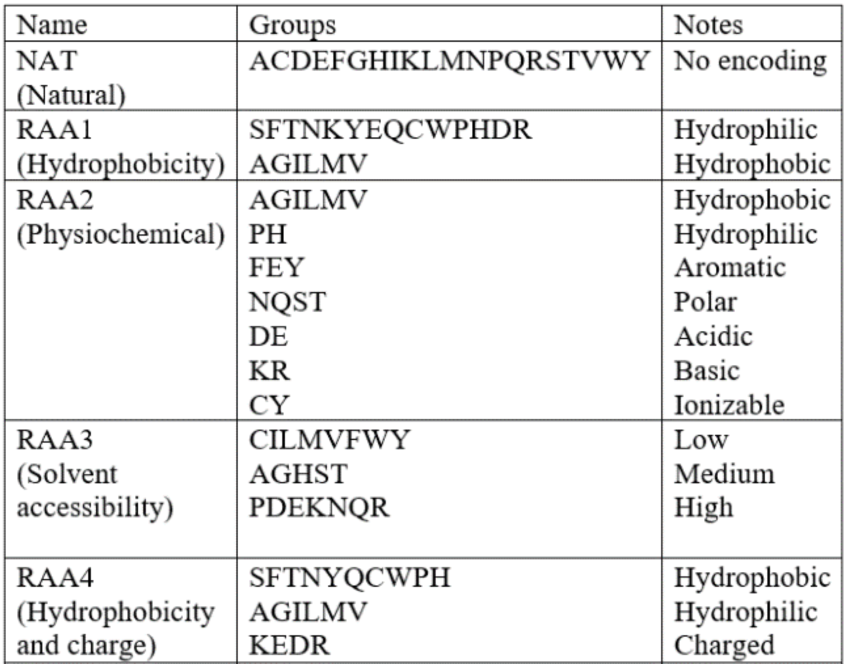
Full-length sequences were downloaded from UniProt. The initial screen for OTU positive sequences was a query for "OTU" with the annotation of "positional domain". Only reviewed sequences were included in the full data set. Other viral and bacterial OTU positive sequences were curated from viral and bacterial screens performed by Dzimianski, *et al.*, and Schubert, *et al.*[8,9]

To curate the OTU negative set, full-length non-redundant bacterial sequences were downloaded from UniProt. The sequences were cleaned to remove any sequences that appeared in the positive OTU set and filtered to remove any sequences with the functional terms relating to "ubiquitin".

3.2 Feature Generation

K-mer features with Reduced Amino Acid Encoding

K-mer features were built based on similar encodings detailed by McDermott, *et al.*[16] The full length sequence was iterated through based on a k-mer length and translated to represent each bin within the encoding type. The k-mer lengths generated were between k=6 and k=16, inclusive, as previously in literature these were tested lengths that were used. Table 2 details the different encodings that were tested in this study, which includes hydrophobicity, physiochemical properties of individual amino acids, solvent accessibility, and hydrophobicity & charge in conjunction. The features were built based on scripts on the public repository: https://github.com/biodataganache/SIEVE-Ub. All the sequences in the positive and negative data set were encoded in all the different RAAs, which encompasses counts data within the full-length sequence for each k-mer and encoded group.



Name	Groups	Notes
NAT (Natural)	ACDEFGHIKLMNPQRSTVWY	No encoding
RAA1 (Hydrophobicity)	SFTNKYEQCWPHDR AGILMV	Hydrophilic Hydrophobic
RAA2 (Physiochemical)	AGILMV PH FEY NQST DE KR CY	Hydrophobic Hydrophilic Aromatic Polar Acidic Basic Ionizable
RAA3 (Solvent accessibility)	CILMVFWY AGHST PDEKNQR	Low Medium High
RAA4 (Hydrophobicity and charge)	SFTNYQCWPH AGILMV KEDR	Hydrophobic Hydrophilic Charged

Table 2: Reduced amino acid encodings. *Encodings used for initial training included hydrophobicity, physiochemical properties (broadly), solvent accessibility, and hydrophobicity & charge.*[16]

Composition Features

Three different composition sets were generated from the primary amino acid sequences: single composition, dipeptide composition, and tripeptide composition. Single amino acid composition encompassed counts of occurrence of each amino acid within the full-length sequence. Dipeptide composition features were generated by iterating through each sequence with a k-mer length of 2 and capturing counts based on occurrence for each of the dipeptide combinations. Tripeptide composition features are similar to dipeptide composition features, except that the k-mer length was 3. Each of these feature types were generated in their own data tables for the entirety of the positive and negative OTU sequences.

Composition, Distribution, Transition Features

Composition features were based on single amino acid composition, similar to the previous method stated above. The Distribution and Transition features were based on hydrophobicity, classified as shown in Table 3. Distribution features were generated for the beginning, 25%, 50%, 75%, and full length of the sequence based on the occurrence of each of the groups. Transition features were the proportion of changes from one group to another group.

Polar	Neutral	Hydrophobic
R, K, E, D, Q, N	G, A, S, T, P, H, Y	C, L, V, I, M, F, Q

Table 3: Composition, Transition, Distribution feature classification of amino acids based on hydrophobicity. *Bins of each group based on hydrophobicity, which includes polar, neutral, and hydrophobic amino acids.*

3.3 Model Building Overview

Initially, the negative data set was randomly down-sampled to equal the number of OTU positive sequences. The data was split into training and testing sets, where the training set included all the eukaryotic and half the viral OTU positive sequences and an equal proportion of negative OTU sequences. The rest of the down-sampled data was held out in the test set. This is diagramed in Figure 6A.

The initial model built based on k-mers with reduced amino acid encoding included a rare feature filter, which removed features that appeared less than twice in the training data set. Features were then ranked based on variance and only the top 25% most variant features were used in building the model. A 5-fold cross-validation approach was used within the training data to estimate the performance with the rare and variance filters. Performance metrics captured were accuracy, sensitivity, and specificity. This approach is shown in Figure 6B.

To test the validity of RFE as a backwards method of feature selection, we implemented nested cross-validation to estimate the model’s performance compared to the initial 5-fold cross-validation. The nested cross-validation approach split the training into 5-folds, training with RFE in the first 4-folds and testing on the held-out 1-fold to estimate performance. The RFE with cross-validation iterated through each feature and dropped the lowest performing feature until accuracy dropped 10% compared to the initial round of RFE. This is shown in Figure 6C. Accuracy, sensitivity, and specificity were calculated on the held-out data.

The final model utilized all the training data with cross-validation to optimize the number of features. After the optimal number of features was determined based on performance in accuracy, the features underwent a backwards selection to prune to the minimal model.

All models were built in Python with the scikit-learn package, using SVM with default parameters and a linear kernel.

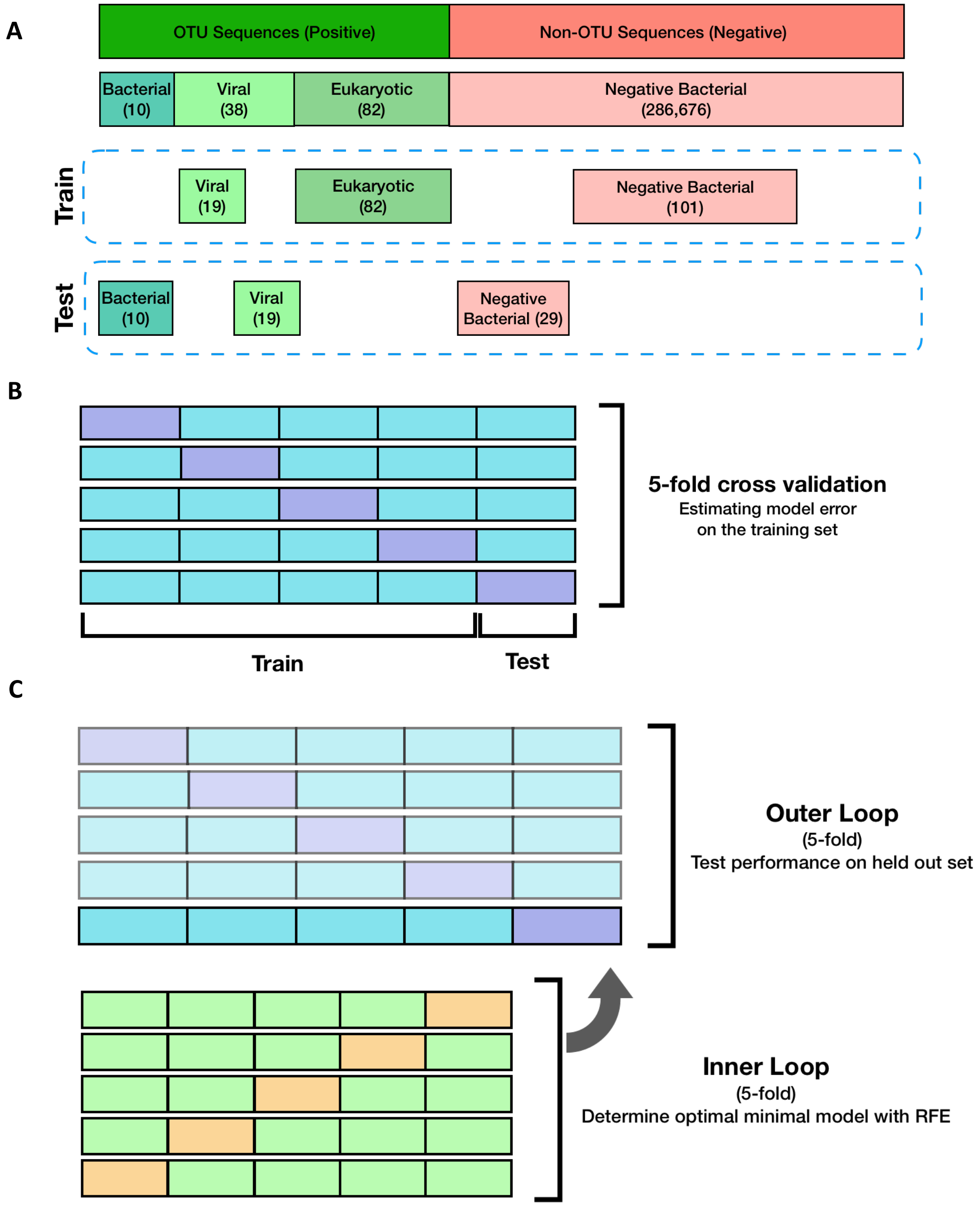


Figure 6: Overview of data partitioning and training. *A) Overview of positive (OTU) and negative (non-OTU) sequences and their division into train and test sets. B) 5-fold cross-validation to estimate model error in the training set. C) Nested cross-validation method used to test performance of RFE.*

3.4 Average Edit Distances

To examine the impact of features retrained by RFE and examine whether the feature selection method reduced redundancy within the set of predictors maintained within the classifiers, we captured all the features used in the model kept at each iteration of RFE. The edit distances were calculated for each pairwise combination of k-mer features within that iteration of RFE. The average was then taken for the group of features within that iteration of RFE.

4. Results

4.1 Data Acquisition

Through a search in UniProt for reviewed sequences of proteins with an annotated OTU domain, there are 94 available sequences, which include 82 eukaryotic, 11 viral, and 1 bacterial example.[22] With an additional literature search of published works with experimental validation showing proteins with an OTU domain and DUB activity, there are an additional 38 viral and 9 bacterial sequences that could be used as positive OTU examples.[6–8,23] In total, that yields 140 positive sequences with OTU domains that could be used to train the classifier.

To curate our negative examples, previous literature in protein family prediction shows that we can sample sequences that are not related to ubiquitin activity from the UniProtKB/Swiss-Prot database. Ultimately, we are interested in predicting novel OTU deubiquitinases in bacteria and by utilizing bacterial sequences, we hope to narrow down the feature space, even despite having a majority of viral and eukaryotic OTU sequences in the positive set. Initially, there were 286,786 bacterial sequences that were reviewed in the non-redundant proteome of UniProt. One sequence was removed because it was a bacterial OTU sequence. Then the sequences were filtered and removed if it contained “ubiquitin” in the functional annotations of the protein descriptors. After the ubiquitin filter, there were 286,676 sequences that were left in the negative training set.

To narrow the training data, the negative OTU sequences were down-sampled to create a balanced data set with the positive OTU examples. The sequences were partitioned such that the training set contained all 82 eukaryotic OTU sequences and half the viral OTU sequences (24 viral sequences) and matched numbers in negative OTU sequences. The testing set contained half of the viral OTU sequences and all 9 bacterial sequences along with matched numbers in negative OTU sequences. This was partitioned so that we can test whether the knowledge on training on eukaryotic and viral examples could be transferred to bacterial sequences. This is due to the lack of validated data on bacterial OTU proteins.

4.2 Specific Aim 1

Feature Filtering on k-mer Features

After the k-mer features were generated and the training and testing sets were sampled, the features were filtered. Within the training data, rare features were filtered out. We defined a feature as being rare if it did not show up at least twice in the training set. Then the variance of the features was calculated across the training data. The top 25% most variant features were then selected to train in an SVM for binary classification. The classifier was trained with 5-fold cross-validation to estimate the accuracy, sensitivity, and specificity across the four encodings, hydrophobicity, physiochemical properties of amino acids, solvent accessibility, and hydrophobicity & charge. This was applied to each k-mer length and encoding combination.

With the filters placed on the feature space, the hydrophobicity encoding performed with above 80% mean cross-validation accuracy across all the k-mer lengths. The sensitivities and specificities across the k-mer lengths and the encodings were consistent in terms of their means, but the hydrophobicity encoding had a smaller standard deviation, as seen in Figure 7A-D. At this point, the performance of the encoded k-mers was similar across the encodings without any RFE applied; none of the combinations of k-mers and encodings outperformed all the others. However, it was promising that with the combinations of k-mers and encodings, the models were able to discern OTU positive and negative sequences better than chance.

At this point, we began to optimize the hydrophobicity encoding because it was the simplest model across all the encodings and understanding the features was also an important factor in the model building. Of all the encodings, the range in number of features is from 64 features to 147,022 features, which is quite a large feature space to try to extrapolate biological understanding.

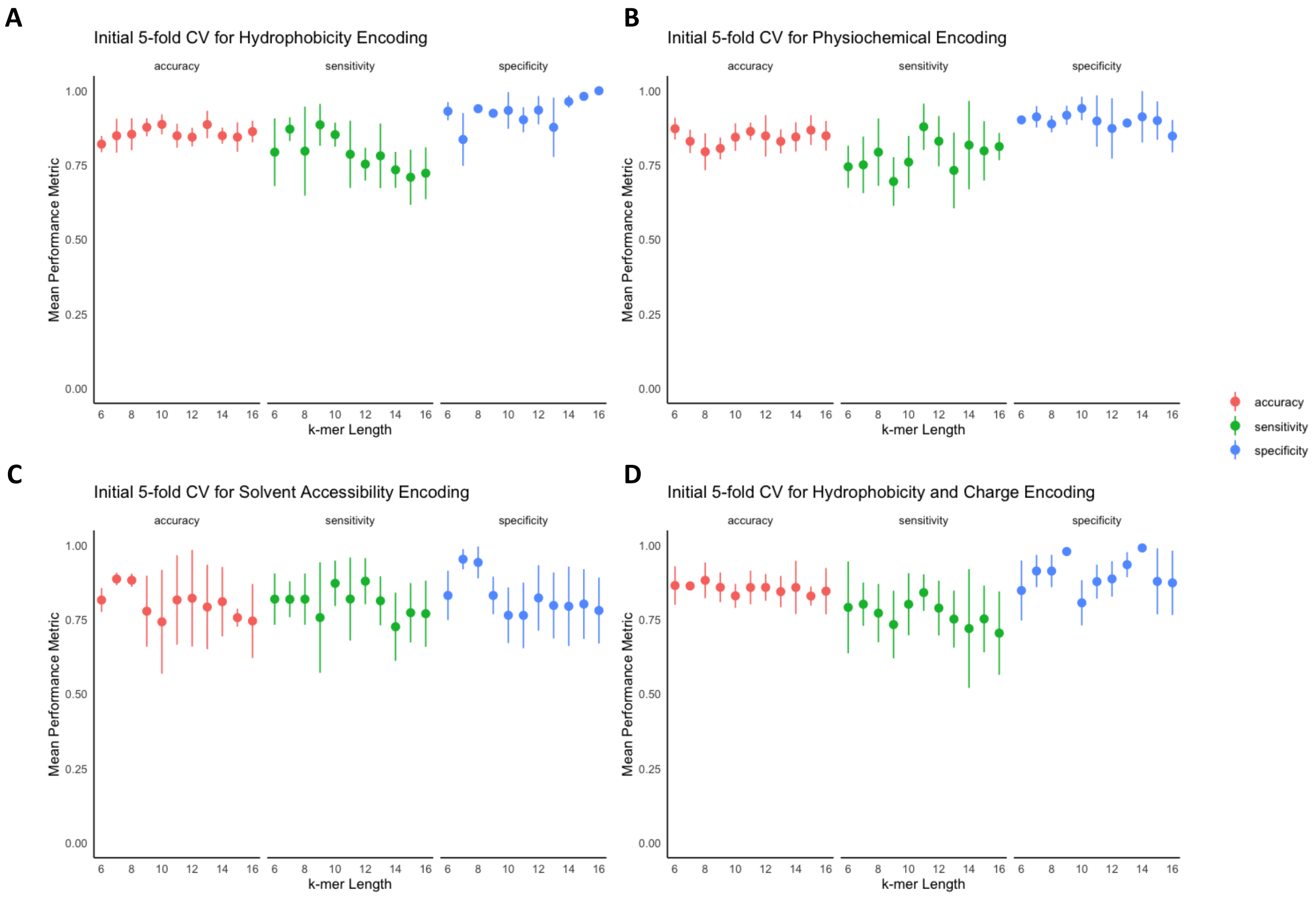
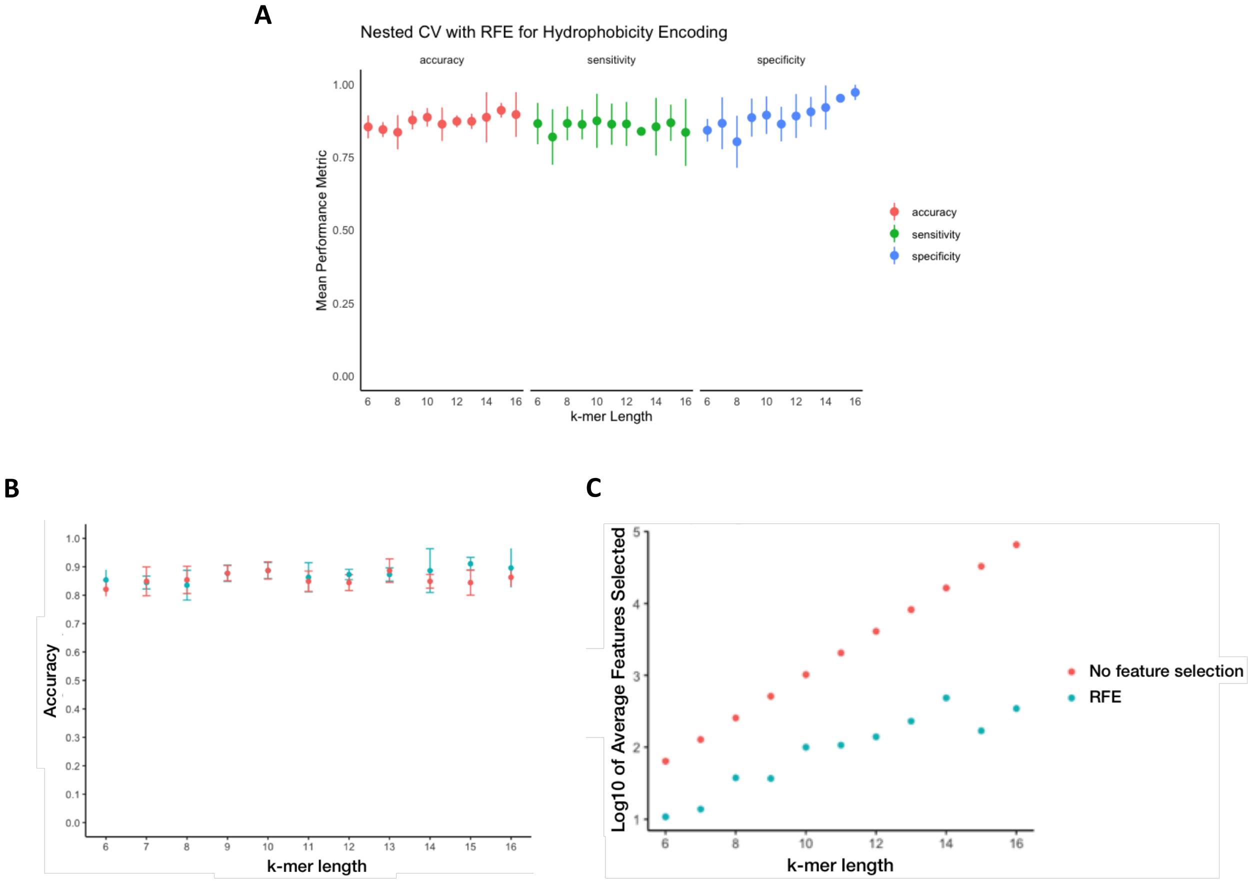


Figure 7: 5-fold cross-validation on rare and variance filters. *Mean accuracy, sensitivity, and specificity across k-mer lengths 6-16 for A) hydrophobicity encoding, B) physiochemical properties encoding, C) solvent accessibility encoding, and D) hydrophobicity and charge encoding.*

Feature Selection with RFE

The next step was to next see if we could use RFE as a feature selection tool. In order to test this, we applied a nested cross-validation approach. This approach encompassed splitting the training data into 5 folds for the outer loop of the nested cross-validation. One of the folds was held out as an unseen test set. The four folds were taken into the inner loop and split into five further folds to run RFE, where the optimal number of features were found within the inner loop. Afterwards, the optimal model from the inner loop training was used to assess performance on the held-out fold in the outer loop.



D

k-mer	6	7	8	9	10	11	12	13	14	15	16
p-value	0.315	0.887	0.669	0.989	0.985	0.31	0.238	0.601	0.256	0.032	0.442

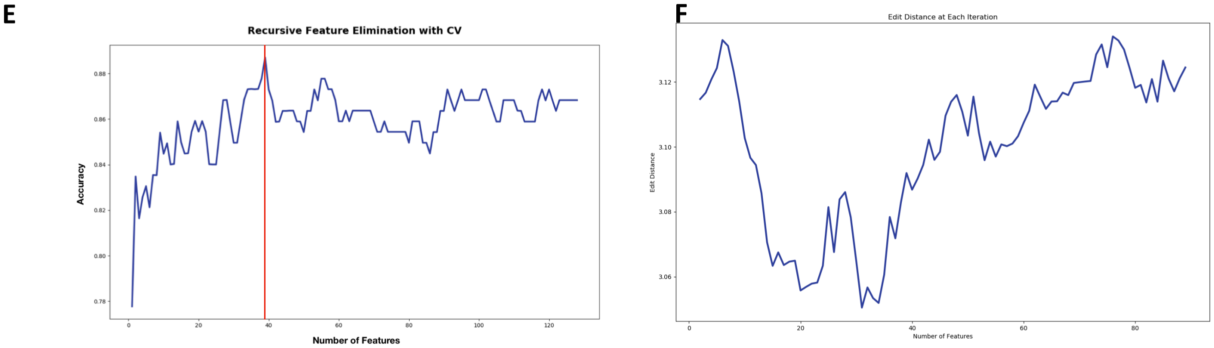


Figure 8: Validating RFE with the hydrophobicity k-mer. *A) Accuracy, sensitivity, and specificity on held-out portion of the training data of all k-mers trained with nested cross-validation B) Mean accuracies of k-mers comparing without feature selection and with RFE as feature selection. C) Number of features used to generate models without feature selection and with RFE as feature selection on a log10 scale. D) p-values of paired t-test between the accuracy performances with and without feature selection across all k-mers E) Example of performance across each iteration of recursive feature elimination, with k-mer 9 in hydrophobicity encoding within one-fold of the inner loop of the nested cross-validation. F) Edit distance across each round of RFE for k-mer 9.*

The accuracy, sensitivity, and specificity were above 80% (similar to the initial training with rare and variance filtering), as seen in Figure 8A. When compared together, as seen in Figure 8B, there were no discernable differences in performances across all the performance metrics between the two model training approaches. There was no statistical difference in the means across the five folds between the initial cross-validation training and the nested cross-validation training, when a paired two-tailed t-test was ran, as seen in Figure *8*D; the p-values were above 0.05 indicating that we reject the null hypothesis that there is a difference in the means between the two groups. At this point, the performances between the two models were comparable, but the number of features selected using RFE was drastically less than without feature selection, as seen in Figure 8D. When we look at one of the training folds, as seen in Figure 8E, for the k-mer length of 9, we see that as we drop features, we see that performance still remains high even when nearly 50% of the features are removed. These data in culmination show that using RFE is able to maintain the performance within the training data while minimizing the number of features in the model.

In order to examine whether RFE removed features based on redundancy, we attempted to look at edit distances across each round of RFE for a few of the k-mer encodings. For each round of RFE, the optimal features were captured. The Levenshtein edit distance was calculated for each pairwise combination of k-mer features and the mean was taken for that group of k-mer features. Figure 8F shows a plot of the edit distances for each iteration of RFE. What we would predict is that through each round of RFE, we are removing redundant features so as we continue through with RFE, the average edit distance would increase as we decrease the number of features in our model. However, Figure 8F shows that there is small change in the average edit distance across all iterations of RFE. The average edit distances range from 3.05 to 3.13 across each iteration of RFE for the k-mer length of 9 in the hydrophobicity encoding. It appears that with RFE, the edit distances are slightly minimized towards the optimal model, which is the opposite result of what is expected so further exploration is needed to examine why RFE slightly increases the redundancy with the feature set.

Minimal k-mer Models

After it was determined that RFE is a valid approach for feature selection, we built the minimal optimal models for the hydrophobicity encodings with k-mer lengths 9, 10, and 11. We chose to optimize these k-mer lengths because they were the minimal lengths with the best performance within the encoding, with cross-validation accuracies between 85% to 87% with the RFE. All the training data was used to find the optimal minimal features using the RFE as a feature selection tool. The final models were tested against the held out bacterial OTU set that also contains half of the viral examples.

The results are depicted in Table 4. We see across the three k-mer lengths of 9, 10, and 11, the performances are very similar. They have accuracies of 87.9%, sensitivities ranging between 81.8% to 87.9%, and specificities of 90.9%. The minimal number of features after the RFE is approximately 10-fold less than the starting number of k-mer features that were generated for each k-mer length in the hydrophobicity encoding (Table 4). Our next step is to add different feature sets in conjunction with the filtered k-mer encoded features to try to improve the model performance in predicting bacterial OTUs in Specific Aim 2.

k-mer	Accuracy	Sensitivity	Specificity	Starting features	Minimal features
9	87.9%	81.8%	90.9%	512	39
10	87.9%	87.9%	90.9%	1024	102
11	87.9%	84.8%	90.9%	2048	322

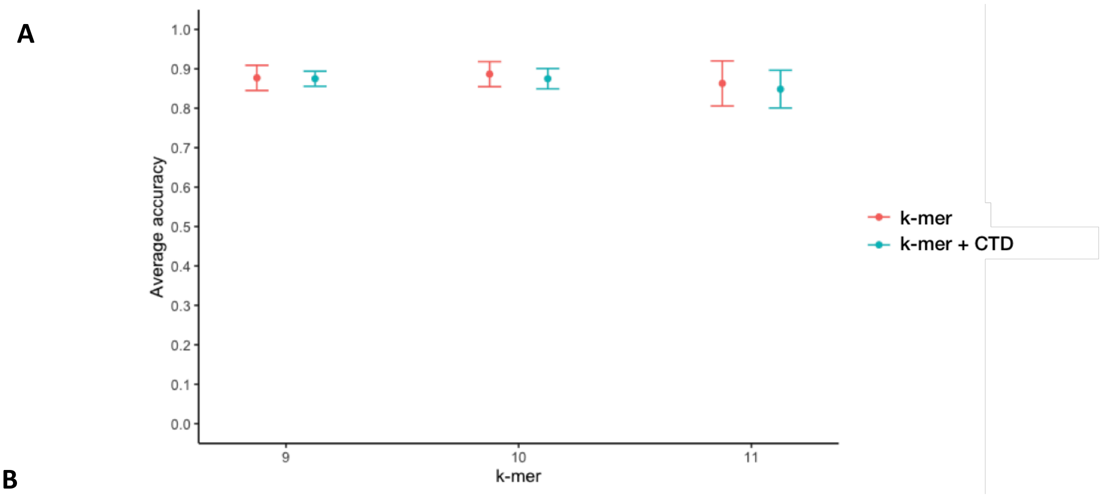
Table 4: k-mer performance on held out test set. *Performance (accuracy, sensitivity, specificity) on held out test set of viral and bacterial sequences, with k-mers 9, 10, 11 encoded based on hydrophobicity.*

4.3 Specific Aim 2

Composition, Transition, Distribution Features

Since the hydrophobicity encoding had performed the best compared to the other encodings within the k-mer feature set, we opted to use the hydrophobicity encoding of the CTD features to see whether it could improve the model's performance on predicting OTUs from non-OTUs. This feature set is a physiochemical-based feature set and also encompassed proportions of where the characteristics were found within the sequence.

After generating the CTD features and running RFE together with the k-mer features, we see in Figure 9A that there is no significant difference in mean cross-validation accuracies for any of the three k-mer lengths. Results of t-tests comparing the two groups of k-mer only features versus k-mer features with the additional hydrophobicity CTD features are shown in Figure 9B.



B

k-mer	9	10	11
p-value	0.908	0.565	0.691

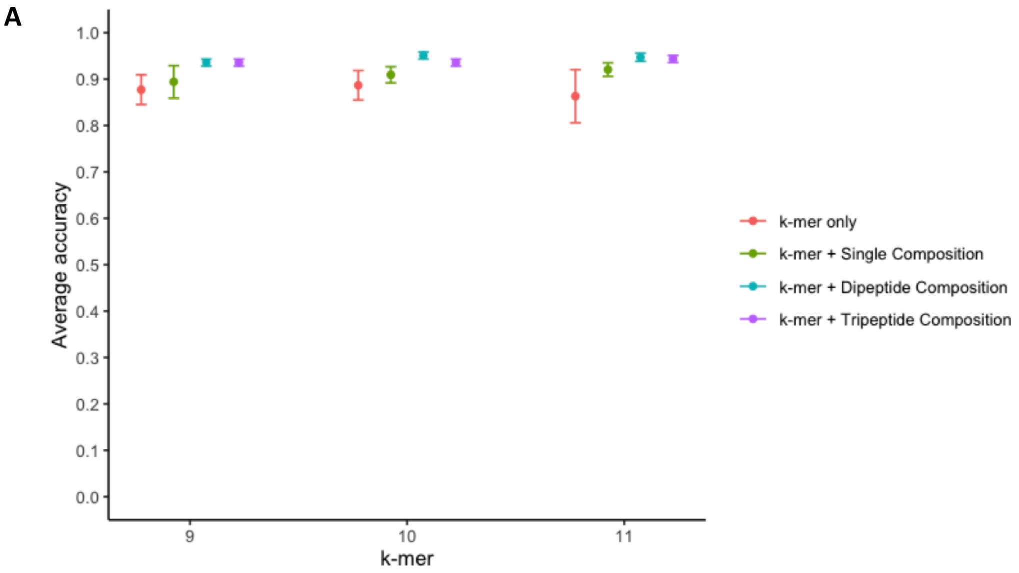
Figure 9: Cross-validation of k-mer features in conjunction with additional CTD feature set. *A). Cross-validation accuracy of k-mer features encoded in hydrophobicity versus k-mer features with additional CTD features. B) P-values of a t-test comparing the means of the accuracies between k-mer only features and k-mer with additional CTD features.*

After generating the cross-validation accuracies for the k-mer features with the CTD features, the classifier was retrained with all the sequences held out in the training set and then the classifier predicted within the testing set. Table 5 depicts the results for the k-mer features with the additional CTD features. We see that the final model produced is within the cross-validation accuracies we observed from Figure 9B, with accuracies ranging from 86.4% to 90.9%.

k-mer	Accuracy	Sensitivity	Specificity	Starting features	Minimal features
9	86.4%	84.8%	87.9%	523	24
10	87.9%	84.8%	90.9%	1033	207
11	90.9%	90.9%	90.9%	2059	233

Table 5: k-mer and CTD features performance on held out test set. *Performance (accuracy, sensitivity, specificity) on held out test set, with k-mers 9, 10, 11 with RFE.*

Composition, Dipeptide, and Tripeptide Features



B

k-mer	9	10	11
Adjusted p-value	0.0075	0.0019	0.0067

Figure 10: Cross-validation of k-mer features in conjunction with additional composition feature sets. *A) Graph of cross-validation accuracies of k-mer features versus k-mer features with additional composition features. B) P-values of an ANOVA comparing the group means of each group of cross-validation accuracy.*

The single amino acid composition was calculated for each of the sequences in the complete data set. This encompassed counts of individual amino acids in the full length of each sequence. After the k-mer features were filtered for rarity and variance, the composition features were added and applied to RFE for feature selection. A similar approach was taken for dipeptide composition and tripeptide composition features.

Above in Figure 10 are the results comparing the k-mer only features to k-mer composition features with the additional composition features. The result of the cross-validation accuracies improves across the addition of all composition features as seen in Figure 10A. In Figure 10B are the results of an ANOVA comparing the group mean accuracies of the cross-validation, where we see a significance in the difference of the means across the k-mer lengths of 9, 10, and 11. These data show that we see that the addition of the composition features, including single amino acid composition, dipeptide composition, and tripeptide composition appear to improve the model's ability to distinguish OTUs from non-OTUs.

k-mer	Accuracy	Sensitivity	Specificity	Starting features	Minimal features
9	87.9%	81.8%	93.9%	532	47
10	87.9%	90.9%	90.9%	1044	81
11	93.9%	90.9%	97.0%	2088	137

Table 6: k-mer and single amino acid composition features performance on held out test set. *Performance (accuracy, sensitivity, specificity) on held out test set, with k-mers 9, 10, 11 and single amino acid composition features with RFE.*

k-mer	Accuracy	Sensitivity	Specificity	Starting features	Minimal features
9	95.5%	97.0%	93.4%	912	133
10	95.5%	97.0%	93.4%	1424	245
11	95.5%	97.0%	93.4%	2448	149

Table 7: k-mer and dipeptide composition features performance on held out test set. *Performance (accuracy, sensitivity, specificity) on held out test set, with k-mers 9, 10, 11 and dipeptide composition features with RFE.*

k-mer	Accuracy	Sensitivity	Specificity	Starting features	Minimal features
9	92.4%	93.9%	90.9%	8512	269
10	93.94%	93.9%	90.9%	9024	474
11	93.94%	96.7%	90.9%	10048	629

Table 8: k-mer and tripeptide composition features performance on held out test set. *Performance (accuracy, sensitivity, specificity) on held out test set, with k-mers 9, 10, 11 wand tripeptide composition features with RFE.*

Above in Table 6, Table 7, and Table 8 are the fitted final models for the k-mer lengths 9, 10, and 11 encoded with hydrophobicity with the addition of single amino acid composition features, dipeptide composition features, and tripeptide composition features, respectively. We see that there is improvement in accuracy and specificity across all the k-mer lengths with the addition of the composition features compared to the k-mer only classifier in Table 4 and the k-mer classifier with CTD features. The optimized classifiers with the additional composition feature sets performed within the range of the cross-validation accuracies. For the single amino acid, dipeptide, and tripeptide composition features, the accuracies range from 87.9% to 95.5% accuracy. The sensitivities range from

81.8% to 97.0% and the specificities range from 90.9% to 97.0%. The minimal features chosen for each of the classifiers are more than ten-fold less than the initial features generated through the use of RFE as a feature selection tool.

5. Discussion

We initially sought to find a method of predicting novel bacterial deubiquitinases, however, it has been observed that there is low sequence similarity within the OTU domain and even sequence rearrangement shuffling the catalytic triad within the domain. We proposed using machine learning based on features generated from the primary amino acid sequence to overcome the reliance on sequence homology, which is typically used with multiple alignment techniques to sequence motifs. Many of the proposed features previously used in machine learning techniques to classify families of proteins included k-mers encoded in RAAs, composition features, and CTD features generated from the primary amino acid sequence.

Conclusions

Based on Specific Aim 1, we see that using k-mer features with hydrophobicity encoding can be predictive of bacterial OTU sequences from non-OTU sequences with the addition of feature filtering based on variance. Furthermore, we can use RFE to select the most informative features, thereby significantly decreasing the feature space and simplifying our classifier models, while still maintaining classifier performance.

From Specific Aim 2, we see that the addition of CTD features does not appear to improve the classification of OTUs from non-OTUs, through comparing the cross-validation accuracies between the two feature sets. However, with the addition of single amino acid composition, dipeptide composition, or tripeptide composition does appear to improve the model performance in the classification of OTUs.

Limitations

A limitation to this study is that the models lack generalizability to other DUB families. This particular application is only usable in the context of predicting OTUs, considering the lack of examples across a majority of the DUB families. Therefore, we cannot say whether the types of features identified as important for predicting OTUs would also serve well for predicting other bacterial DUBs.

This brings it to the context of the current study itself, where a large limitation is the limited size in the training data, especially for validated, non-redundant proteins for viruses and bacteria. As more is understood in the field of ubiquitin DUBs, we could in the future improve the machine learning application of protein family prediction for OTUs.

Another limitation to this study is that the method of sampling data may not be the best representation. Due to the limitation in unique sequences annotated in UniProt, we down-sampled non-OTU sequences to balance to the OTU sequences. Classifier performance may be dependent on the selection of non-OTU sequences used in the training set. We also trained against eukaryotic, viral, and bacterial sequences to predict ultimately in bacterial sequences. This transfer learning may not be completely comprehensive and more work needs to be done in evaluating larger groups of non-OTU sequences.

A computational limitation to the study is within the generation of the various feature sets and the model building, which lack scalability for larger sizes. A majority of the models were built with the maximum of about 400 sequences at a time, so more work would need to be done in order to address the problem of scalability.

Future Directions

Future work from this project includes examining how the selection of negative non-OTU sequences affects the model performance. The original method of down-sampling non-OTU sequences may affect the model performance and may not be representative of the population. There is also the consideration that the OTU sequences encompass eukaryotic, viral, and bacterial sequences, while the non-OTU sequences are only bacterial. It would be interesting to compare training a non-OTU set across all kingdoms versus just bacteria, although it is an important consideration that the goal is to discover more bacterial OTU proteins.

Next steps from this work also include comparing the prominent features from each of the generated final models to known OTU sequence logos to see if parts of the catalytic triad were determined to be distinguishing for OTUs versus non-OTUs. An extension of this would be to examine the model's performance compared to other computational tools that are used to predict domains and protein families, such as HMMs on InterPro to see if there is some consensus to this method. Importantly, more work needs to be done in order to examine the effect of RFE as a feature selection tool to determine whether redundant features are removed and whether there are biological insights regarding the distinguishing features of OTUs that can be gathered from this method.

Within the scope of analysis done for this project so far, another aspect of the that can be examined more closely is the OTUs that were not classified correctly. There may be features within the sequences of these OTUs that could guide future prediction studies for OTU if the structure of these proteins is drastically different than the other known OTUs.

Although there is more work to be done to refine the models in terms of improving performance and understanding the biological significance of the RFE, this work has shown that k-mer features can be useful for distinguishing OTUs from non-OTUs. This takes a step outside of relying heavily on homology and the construction of sequence logos for predicting more proteins within a protein family of interest.

6. References

1. Komander D, Rape M. The Ubiquitin Code. Annu Rev Biochem. 2012 Jul 7;81(1):203–29. Available from: http://www.annualreviews.org/doi/10.1146/annurev-biochem-060310-170328
2. Mevissen TET, Komander D. Mechanisms of Deubiquitinase Specificity and Regulation. Annu Rev Biochem. 2017;86:159–92. Available from: https://doi.org/10.1146/annurev-biochem-
3. Clague MJ. CJM. US. Cellular Functions of the DUBs. J Cell Sci . 2012;124:277–86. Available from: https://pdfs.semanticscholar.org/6d6d/d454c05b36dab95e13c0ea82a6a8c04af66b.pdf
4. Wilkinson KD. DUBs at a glance. J Cell Sci. 2009 Jul 15;122(14):2325–9.
5. Mevissen TET, Hospenthal MK, Geurink PP, Elliott PR, Akutsu M, Arnaudo N, et al. OTU deubiquitinases reveal mechanisms of linkage specificity and enable ubiquitin chain restriction analysis. Cell. 2013 Jul 3;154(1):169.
6. Kubori T. LotA, a Legionella deubiquitinase, has dual catalytic activity and contributes to intracellular growth. Cell Microbiol . 2018;20(7). Available from: https://doi.org/10.1111/cmi.12840
7. Furtado AR, Essid M, Perrinet S, Balañá ME, Yoder N, Dehoux P, et al. The chlamydial OTU domain-containing protein Chla OTU is an early type III secretion effector targeting ubiquitin and NDP52. Cell Microbiol . 2013 Dec 1; 15(12):2064–79. Available from: http://doi.wiley.com/10.1111/cmi.12171
8. Dzimianski J V., Beldon BS, Daczkowski CM, Goodwin OY, Scholte FEM, Bergeron É, et al. Probing the impact of nairovirus genomic diversity on viral ovarian tumor domain protease (vOTU) structure and deubiquitinase activity. Mirazimi A, editor. PLOS Pathog . 2019 Jan 10; 15(1):e1007515. Available from: http://dx.plos.org/10.1371/journal.ppat.1007515
9. Schubert AF, Nguyen J V, Franklin TG, Geurink PP, Roberts CG, Sanderson DJ, et al. Identification and characterization of diverse OTU deubiquitinases in bacteria. EMBO J . 2020 Jun 22; 39(15):e105127–e105127. Available from: https://europepmc.org/articles/PMC7396840
10. Wu CH, Huang H, Yeh LSL, Barker WC. Protein family classification and functional annotation. Comput Biol Chem. 2003 Feb 1;27(1):37–47.
11. Bishop C. Pattern Recognition and Machine Learning. Jordan M, Kleinberg J, Schölkopf B, editors. Springer; 2006.
12. Samudrala R, Heffron F, Mcdermott JE. Accurate Prediction of Secreted Substrates and Identification of a Conserved Putative Secretion Signal for Type III Secretion Systems. Stebbins CE, editor. PLoS Pathog . 2009 Apr 24; 5(4):1000375. Available from: https://dx.plos.org/10.1371/journal.ppat.1000375
13. Bhasin M, Raghava GPS. GPCRpred: an SVM-based method for prediction of families and subfamilies of G-protein coupled receptors. Nucleic Acids Res . 2004 Jul 1; 32(Web Server):W383–9. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkh416
14. Wang YF, Chen H, Zhou YH. Prediction and classification of human G-protein coupled receptors based on support vector machines. Genomics, Proteomics Bioinforma. 2005 Jan 1;3(4):242–6.

15. Li YH, Xu JY, Tao L, Li XF, Li S, Zeng X, et al. SVM-prot 2016: A web-server for machine learning prediction of protein functional families from sequence irrespective of similarity. PLoS One. 2016 Aug 1;11(8).

16. McDermott JE, Cort JR, Nakayasu ES, Pruneda JN, Overall C, Adkins JN. Prediction of bacterial E3 ubiquitin ligase effectors using reduced amino acid peptide fingerprinting. PeerJ. 2019 Jun 7;7:e7055.

17. Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, et al. Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. Nucleic Acids Res . 2017 Jan 4; 45(D1):D535–42. Available from: https://academic.oup.com/nar/article-lookup/doi/10.1093/nar/gkw1017

18. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, et al. InterPro: The integrative protein signature database. Nucleic Acids Res. 2009;37(SUPPL. 1):D211.

19. Bateman A. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res . 2019 Jan 8; 47(D1):D506–15. Available from: https://academic.oup.com/nar/article/47/D1/D506/5160987

20. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, et al. The Pfam protein families database in 2019. Nucleic Acids Res . 2019 Jan 8; 47(D1):D427–32. Available from: https://academic.oup.com/nar/article/47/D1/D427/5144153

21. UniProt Consortium. Biocuration in UniProt . 2018. Available from: https://www.uniprot.org/help/biocuration

22. UniProt Consortium. UniProtKB results: annotation:(type:"positional domain" otu) AND reviewed:yes. Available from: https://www.uniprot.org/uniprot/?query=annotation%3A(type%3A%22positional+domain%22+otu)&fil=reviewed%3Ayes&offset=25&sort=score#orgViewBy

23. Ma K, Zhen X, Zhou B, Gan N, Cao Y, Fan C, et al. The bacterial deubiquitinase Ceg23 regulates the association of Lys-63-linked polyubiquitin molecules on the Legionella phagosome. J Biol Chem . 2020; Available from: http://www.jbc.org/cgi/doi/10.1074/jbc.RA119.011758