Fall 12-21-2020

FindFur: A Tool for Predicting Furin Cleavage Sites of Viral Envelope Substrates

Christine Gu
San Jose State University

Follow this and additional works at: https://scholarworks.sjsu.edu/etd_projects

Part of the [Other Computer Sciences Commons](#)

Recommended Citation

Gu, Christine, "FindFur: A Tool for Predicting Furin Cleavage Sites of Viral Envelope Substrates" (2020). *Master's Projects*. 968.

DOI: <https://doi.org/10.31979/etd.4ahv-9jya>

https://scholarworks.sjsu.edu/etd_projects/968

This Master's Project is brought to you for free and open access by the Master's Theses and Graduate Research at SJSU ScholarWorks. It has been accepted for inclusion in Master's Projects by an authorized administrator of SJSU ScholarWorks. For more information, please contact scholarworks@sjsu.edu.

FindFur: A Tool for Predicting Furin Cleavage

Sites of Viral Envelope Substrates

A Project

Presented to

The Faculty of the Department of Computer Science

San José State University

In Partial Fulfilment

Of the Requirements for the

Degree Master of Science in Bioinformatics

By Christine Gu

December 2020

ABSTRACT

FindFur
by Christine Gu

Most biologically active proteins of eukaryotic cells are initially synthesized in the secretory pathway as inactive precursors and require proteolytic processing to become functionally active. This process is performed by a specialized family of endogenous enzymes known as proproteases convertases (PCs). Within this family of proteases, the most notorious and well-researched is furin. Found ubiquitously throughout the human body, typical furin substrates are cleaved at sites composed of paired basic amino acids, specifically at the consensus sequence, R-X-[K/R]-R↓. Furin is often exploited by many pathogens, such as enveloped viruses, for proteolytic processing and maturation of their proteins. Glycoproteins of enveloped viruses often possess the essential basic residues, arginine or lysine, at their recognition site, permitting cleavage and subsequent activation by furin. Recent biochemical research suggests the furin cleavage site encompasses about 20 residues, ranging from P14 to P6', and variations at the site impact viral pathogenicity. Thus, the prediction of furin cleavage sites of viral substrates is an attractive area of research. While prediction methods of furin cleavage sites exist, there is no virus-specific model currently available. This project describes two methods for predicting furin cleavage sites of viral envelope glycoproteins based on profile Hidden Markov Models (HMM) and logistic regression. The logistic regression model was constructed using the hydrophobicity levels of amino acid residues relative to their position at the motif site. The profile HMM predicts furin cleavage sites in independent sequences with a sensitivity of 87% and an accuracy of 89%, and the latter method achieves a sensitivity of 60% and an accuracy of 91%. A Python-based prediction tool called FindFur was designed with the profile HMM and is publicly available at <https://github.com/chwisteeng/FindFur>.

Index terms – Profile Hidden Markov Models (HMM), proteolytic processing, furin cleavage site, consensus sequence, logistic regression, hydrophobicity, prediction tool, viral pathogenesis, enveloped viruses

ACKNOWLEDGMENTS

I want to first thank my graduate advisor, Dr. Wendy Lee, for her mentorship through my project. Our conversations have been a great source of strength and have kept me grounded during this pandemic. I am incredibly appreciative of all the time and care Dr. Lee has provided, not just in helping me navigate my project but also in life. Her passion for the discipline and compassion for others has deeply empowered me.

I want to express my sincerest appreciation to Dr. Philip Heller for all of his support and guidance over the past few years. As the individual who opened the doors to the world of Bioinformatics to me, this work would not have been possible without his help. He has provided constant encouragement to me as a student, researcher, and life mentor; I will take all these experiences with me for the rest of my life and my career. I am thankful for Dr. Heller's belief in me because it has helped build my confidence and inspires me to continue to grow.

I want to recognize Dr. William Murray for serving on my committee and providing invaluable comments and suggestions. I also want to thank him for being my mentor since my undergraduate career. Dr. Murray is an incredibly dedicated and hard-working individual, and his zest for life (plus all his energy!) motivates me to do more with my life.

Finally, I want to thank my family and all my friends for their unwavering love and support. A special thank you to my Mom, my Dad, and Jennifer, for always believing in me. I love you - this degree is for you. □

TABLE OF CONTENTS

1. Introduction	1
2. Methods	6
2.1. Data collection	6
2.2. Multiple sequence alignment	9
2.3. Training and test split	9
2.4. HMMER	9
2.5. Five-fold cross-validation	12
2.6. Testing	13
2.7. Hydrophobicity-based logistic regression model	13
3. Results	15
3.1. Data collection	15
3.2. Sequence logo analysis	16
3.3. Hydrophobicity analysis	20
3.4. Optimization analysis	22
3.5. Prediction performance	28
3.6. Hydropathy-based logistic regression model analysis	31
4. Implementation of FindFur	34
4.1. Requirements	34
4.2. Usage	34
4.2.1. FindFur_Extract.py	36
4.2.2. FindFur_HMMERParse.py	36
5. Discussion	37
References	40

LIST OF FIGURES

Figure 1. Pipeline for building profile HMM from a curated set of viral furin substrates	8
Figure 2. HMMER profile/sequence comparison pipeline	10
Figure 3. Workflow for training and testing profile HMM	12
Figure 4. Sequence logos of aligned peptidase cleavage sites centered at P1: (A): 74 experimentally determined furin cleavage site of viral substrates; (B): 146 general PCs; (C): 414 general proteases	17
Figure 5. Hydrophobicity pattern plot of positive dataset (only viral furin substrates).....	21
Figure 6. Hydrophobicity pattern plot of negative dataset	21
Figure 7. Graphical distribution of hmmsearch hit results for varying independent MSV/Vit/Fwd thresholds	24
Figure 8. ROC diagram based on varying thresholds during parameter optimization shown as TPR (sensitivity) plotted against FPR.....	27
Figure 9. Sequence logo for false positives from five-fold cross-validation.....	30
Figure 10. Boxplot of hydrophobicity of each residue position	31
Figure 11. ROC curve based on varying thresholds during optimization of the final logistic regression model	33
Figure 12. Pipeline for FindFur tool	35
Figure 13. HMMERParser output of test set results	36

LIST OF TABLES

Table I. Viral substrates of furin	4
Table II. Description of adjustable filters in <i>hmmsearch</i>	11
Table III. Physical properties of the 20 amino acid side chains in increasing order of hydrophathy	14
Table IV. Initial distribution of the positive dataset	16
Table V. Initial distribution of the negative dataset.....	16
Table VI. Distribution of training, validation, and testing datasets for profile HMM.....	23
Table VII. Raw scores from <i>hmmsearch</i> hits of threshold combinations for parameter optimization	25
Table VIII. Classification results of threshold combinations for parameter optimization	26
Table IX. Performance evaluation of optimal profile HMM search setting	28
Table X. Distribution of sequences for five-fold cross-validation	28
Table XI. Raw scores from five-fold cross-validation.....	29
Table XII. Predictive performance evaluation from five-fold cross-validation	29
Table XIII. Predictive performance evaluation of final profile HMM from test data	30
Table XIV. Distribution of training and testing datasets for logistic regression model	32
Table XV. Predictive performance evaluation of logistic regression model from test data	32

1. INTRODUCTION

Proteolytic cleavage, or proteolysis, is a ubiquitous reaction involving the degradation of substrates into smaller peptides or amino acids through a family of specialized proteins called proteases [1], [2]. Many biologically active proteins that control fundamental pathways are initially synthesized as inactive secretory precursors termed proproteins and require proteolytic cleavage to be converted into their functional end product [2]. These cleavage events, or proteolytic processing events, result in the irreversible post-translational modification of a protein's structure and activity and lead to the activation or impairment of protein functionality. Consequently, this plays an essential role in regulating various biological processes and potential underlying pathological processes [2], [3]. Proprotein cleavage typically occurs at motifs containing multiple basic arginine or lysine residues by limited endoproteolysis and is cut after the basic residues [4].

The molecular scissors responsible for this type of processing are called proprotein convertases (PCs). They are an essential family of nine mammalian serine proteases responsible for the proteolytic activation of various precursor proteins, including growth factors, hormones, receptors, and adhesion molecules, as well as infectious viral surface glycoproteins. Nine different PCs (furin, PC1, PC2, PC4, PC5, PACE4, PC7, SKI-1, PCSK9) have been identified in humans that cleave proproteins within the general sequence motif $[K/R]-X_n-[K/R] \downarrow$, where $n = 0, 1, 2$, or 3 amino acids, X represents any amino acid except for cysteine or proline, and the arrow ($\downarrow$) denotes the cleavage site. Within this family of enzymes, furin was the first PC to be discovered and is currently the most extensively studied member due to its ubiquity and biological significance [1], [2].

Furin is a 794-residue long transmembrane protein and the product of the *fur* gene located on chromosome 15 in humans. It is expressed in all examined human tissues and cell lines and is localized predominantly in the trans-Golgi network, though it can also be found on the cell surface [5], [6]. Substantial experimental evidence suggests that furin plays a vital role in processing proproteins in the secretory pathway and that its activity has broad biological implications. Day et al. 1993 used gene expression techniques in rat brains to show that furin is ubiquitously expressed throughout the brain with higher abundance in select regions, such as the ventricles and the hippocampus [7]. Schäfer et al. 1993 proposed that this cellular localization of furin in rat brains in specific areas is correlated with its involvement in precursor processing, indicating furin has a more dynamic role in specific enzymatic pathways or tissues [8]. Furthermore, Dubois et al. 1995 and Dubois et al. 2001 implicated furin in possessing cleavage selectivity for certain proproteins based on the presence of its cleavage motifs, such as TGF β -1, which is essential for growth and tissue homeostasis [9], [10]. Altogether, furin appears to regulate a diverse collection of protein functions and critical biological pathways.

The furin cleavage site is typically described by the consensus sequence R-X-[K/R]-R↓ [5], [6]. Furin exhibits a narrower substrate specificity compared to other PCs within the family, preferentially cleaving at the paired basic amino acids, arginine and lysine. It can recognize sites marked by the same canonical PC motif, though often with reduced efficiency [11]. This overlap occurs because their active sites are evolutionarily conserved and highly homologous. Thus, there are similarities in substrate preferences and cell or tissue expression [11], [12]. Due to the variability in furin cleavage site patterns, Tian 2009 biochemically characterized the cleavage site of 130 furin substrates [13]. He determined that the full motif comprises about 20 amino acids ranging from P14 to P6' where the outer regions surrounding the core motif site permit

solvent accessibility [13]. Additionally, variations of the consensus sequence, along with residue variations around the site and additional post-translational modifications, influence the binding interactions between furin and its substrate [13]. This work revealed that furin has an extended substrate-binding site and that differences at motif influence substrate specificity and cleavage efficiency.

In the secretory pathway, furin recognizes the consensus sequence in proproteins as targets and cleaves them to generate bioactive proteins [1], [14]. In the trans-Golgi network, furin typically activates host substrates like growth factors, hormones, serum proteins, extracellular matrix proteins, and many more [15]. When furin is trafficked to the cell surface, it can process both host and pathogenic substrates, such as adhesion proteins and viral fusion peptides. Many viral pathogens have been implicated in exploiting host proteases as a control mechanism for cell entry and infectivity [11], [15]. Glycoproteins of enveloped viruses, such as highly pathogenic Asian avian influenza A (H5N1) and certain human-infecting coronaviruses (CoV), frequently contain cleavage sites with multiple basic arginine or lysine residues that can be processed and activated by furin, leading to viral infection of the host [16], [17]. Furin-mediated cleavage of enveloped glycoproteins has been identified in numerous virus families, including Herpesviridae, Coronaviridae, Togaviridae, Orthomyxoviridae, Paramyxoviridae, and Retroviridae, as seen in Table I [5].

TABLE I. VIRAL SUBSTRATES OF FURIN

Virus family	Virus	Furin substrate
Herpesviridae	Human cytomegalovirus	gB
	Epstein-Barr virus	gB
Coronaviridae	Infectious bronchitis virus	S
	SARS-CoV-2 (?)	S
Togaviridae	Sindbis virus	E2
	Semliki forest virus	E2
Orthomyxoviridae	Influenza A virus (H5/H7/H9)	HA
Paramyxoviridae	Newcastle disease virus	F
	Measles virus	F
Retroviridae	Human immunodeficiency virus 1	env
	Rous sarcoma virus	env

**gB = envelope glycoprotein B, S = spike protein, E2 = regulatory protein E2, HA = hemagglutinin, F = fusion glycoprotein F0 precursor, env = envelope*

Substrate specificity studies have shown that the virulence of these enveloped viruses is correlated with the cleavage sensitivity of their surface glycoproteins [5], [6]. Therefore, the amino acid pattern of cleavage sites can directly impact a viral substrate's cleavability and is an important determinant of viral pathogenicity and tropism. One example is the hemagglutinin (HA) cleavage of avian influenza. Avian influenza A viruses possess HA on their cell surface that can be processed by several host proteases. HA cleavage of mammalian avian influenza virus or low pathogenic strains (LPAIVs) usually occurs at a single basic amino acid by proteases with limited specificity. These infections are generally restricted to cells in the respiratory tract and gastrointestinal tract. LPAIVs, however, can evolve into highly pathogenic forms by acquiring multiple basic amino acids at the cleavage site. Highly pathogenic avian influenza viruses (HPAIVs), like H5N1 and H7N9, are capable of being cleaved by furin and other PCs that recognize multiple basic amino acids at the cleavage site [18], [19]. The capacity

for recognition and activation by the broadly expressed furin enables the virus to widen its cellular tropism and increases the likelihood of rapid dissemination and, potentially, systemic infection. Luczo et al. 2018 found the acquisition of additional basic amino acids by LPAIVs resulted in an extended substrate motif that allowed for furin recognition and was preferentially selected [20]. More recent evidence confirms that certain motifs possess amino acids in particular positions, increasing furin specificity [21]. Thus, the ability to exploit furin can dramatically impact viral pathogenicity.

From a medical and biotechnology perspective, furin cleavage sites are a promising research area due to their importance in modulating protein activity and role in viral infections. Genomic analysis of the novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) suggests its spike glycoprotein contains a furin-like cleavage site not present in other SARS-like CoVs; this site may explain some of its capacity to spread so efficiently through the population [22]. The current SARS-CoV-2 pandemic reveals human vulnerability to viral infections and the lack of global preparedness to deploy medical countermeasures to control them. It emphasizes our need to quickly identify viral factors that contribute to viral pathogenesis to develop effective antiviral therapeutics. Consequently, it is of interest to be able to predict the furin cleavage site accurately.

This project presents work on the characterization and prediction of furin cleavage sites for enveloped viruses using profile Hidden Markov Models (HMM). An HMM for proteins, known as a profile HMM, is a probabilistic model that uses a position-specific scoring system to describe the degree of conservation of the columns in a multiple sequence alignment. For each state, the observed symbols are the weighted distribution of over 20 amino acids. HMMs can be thought of as doubly stochastic models that emit protein sequences by randomly going from state

to state, sometimes emitting amino acids at a state according to a probability distribution unique to that state. For any given sequence, it is possible to compute the most probable way the sequence could be generated by the model and the overall probability of the model generating the sequence [23], [24]. Although previous research suggests there are underlying differences in viral substrates of furin due to their cellular function compared to other furin substrates, there is currently no virus-specific model to detect putative furin cleavage sites in viral precursors [13]. In this project, a Hidden Markov Model (HMM) was designed and built using the HMMER3 software suite, based on the full, 20-residue motif of experimentally verified furin cleavage sites of viral envelope glycoproteins [25]. FindFur is a tool developed specifically for this project. It predicts the presence of putative furin recognition sites in viral protein sequences using a profile HMM. The current study also discusses constructing a logistic regression model for predicting viral furin cleavage sites, based on the hydrophobicity of amino acid residues relative to their position at the motif site. The work is described in four parts: section two discusses the data collection and methods used, section three details the results, section four reviews implementation of FindFur, and section five concludes the project with a discussion and potential future works.

2. METHODS

2.1. *Data collection*

Two primary datasets were collected: (1) known furin cleavage sites of viral envelope glycoproteins and (2) cleavage sites of general PCs and other peptidase substrates cleaved at arginine in position P1 and known mammalian and bacterial furin substrates. The datasets are referred to as the positive and negative datasets, respectively.

Experimentally confirmed furin cleavage data on envelope substrates were compiled for the positive training dataset from FurinDB, MEROPS, and Genbank/Uniprot. FurinDB is a manually curated database containing verified sequence data for substrates of mammalian furin [26]. Developed and maintained by EMBL-EBI, the MEROPS database stores curated data on a vast collection of peptidases, including furin [27]. Additional protein sequences with biochemical evidence of furin cleavage at the essential basic amino acid residues relative to the P1 and P4 positions were collected from literature and obtained through Genbank and Uniprot to increase the signal-to-noise ratio. Any sequence incidentally collected without an R in the P1 and P4 position was excluded from the training dataset. One exception was made for Ebola virus strain Reston, which possesses a lysine at the P4 position because its envelope protein has been demonstrated to be proteolytically processed by furin [28].

The latter (negative) dataset was composed of various protease substrates and a group of known mammalian and bacterial furin substrates. The first group is pulled from MEROPS and represents a collection of PCs within the same family as furin and peptidases cleaved at an arginine residue between the P1 and P1 sites. Only proteins with arginine at the P1 site were selected. No viral substrates were included in the PCs and peptidase group to avoid the possibility of false hits due to overlap in cleavage specificity between the proteases in this set and furin. The second group contained the remaining cleavage data from FurinDB of bacterial and mammalian proteins by furin.

All sequences were extracted as 20-residue motifs as reported by Tian et al. 2011 [26] in FASTA format using a custom Python script developed for this project, called FindFur_Extract. Motifs of incomplete length ($n < 20$) were filtered from the training set to avoid low-quality data. To reduce sequence redundancy, both the positive and negative datasets were filtered for

duplicates using seqkit, a command-line tool for FASTA file manipulation [29]. The workflow for building and testing the profile HMM is described in Figure 1.

Figure 1. Pipeline for building profile HMM from a curated set of viral furin substrates

2.2. *Multiple sequence alignment*

The two cleavage data sets were formatted and aligned into a multiple sequence alignment around their cleavage site at the P1 position through AliView [30]. Sequence logos were constructed with Skylign to visualize the multiple sequence alignments and observe each motif's position-specific features [31], [32]. NCBI Multiple Sequence Alignment Viewer was used to create a hydrophobicity pattern for the positive and negative datasets [33].

2.3. *Training and test split*

A major challenge for this project was the lack of data available to train, optimize, and test the model effectively. To reduce potential bias during the final model's performance evaluation, the positive and negative datasets were shuffled and split at random into 80% and 20% for training and testing, respectively. This eliminates some risk that the predictive performance is overestimated as a result of overfitting. The negative data's training set was divided again into 80% and 20% for cross-validation and optimization purposes, respectively. This was done to assess the model fit on training data and fine-tune the model's hyperparameters. The positive dataset was not similarly split for the final model due to the limited sample size and underfitting issues.

2.4. *HMMER*

Implementation of HMMER requires two primary programs: *hmmbuild* and *hmmsearch*.

The profile HMM for known furin cleavage site motifs of envelope glycoproteins was built by *hmmbuild* from the aligned positive training [34]. The *hmmsearch* program subsequently identified and scored significantly similar sequence matches of given protein sequences against

the profile HMM. The optimization dataset was used to adjust the *hmmsearch* filter parameters since the default settings failed to allow enough sequences to pass.

HMMER's *hmmsearch* program is an accelerated processing pipeline that compares a profile to a sequence through five heuristic filters, as seen in Figure 2.

Figure 2. HMMER profile/sequence comparison pipeline

Arrows with a p-value specify the adjustable filters. The three thresholds that can be changed are the multiple segment Viterbi (MSV) filter, the Viterbi (Vit) filter, and the forward (Fwd) filter. The objective of these heuristic filters is to accelerate the speed at which results are produced. However, there is a trade-off between speed, specificity, and sensitivity, and it does not guarantee to find all high-scoring hits. The three adjustable filters are briefly described in

Table II as follows (note: the default p-value indicates the fraction of non-homologous sequences that are expected to pass):

TABLE II. DESCRIPTION OF ADJUSTABLE FILTERS IN *HMMSEARCH*

Filter	Option	Default p-value	Description
Multiple Segment Viterbi (MSV)	--F1	0.02	- Looks for 1+ high-scoring ungapped local alignment - Main speed heuristic
Viterbi (Vit)	--F2	0.001	- Looks for optimal gapped alignment
Forward (Fwd)	--F3	$1e^{-5}$	- Given the profile, calculates the likelihood of a target sequence summed over all possible alignments - Most stringent

For each filter, a range of p-values generated results. Only one filter was adjusted at a time. The number of sequences passed through the MSV, Viterbi, and forward filters were recorded, as well as the number of positive and negative hits, the total number of hits, and the lowest and highest e-value. At the end, the positive rate (TPR), false positive rate (FPR), accuracy (ACC), and Matthews correlation coefficient (MCC) were also calculated from the total number of true positives and true negatives. A custom Python script was written for this project to record the outcomes of the different thresholds for each filter parameter. The optimal setting was found to be F1 (MSV) 0.08, F2 (Viterbi) 0.02, and F3 (Forward) 0.00008. The workflow is shown below in Figure 3. A receiver operating characteristic (ROC) curve is presented to illustrate the profile HMM's prediction performance for the varying thresholds and prediction sensitivities.

Figure 3. Workflow for training and testing profile HMM

2.5. Five-fold cross-validation

Five-fold cross-validation was employed to evaluate the profile HMM's training performance with the optimal search settings using scikit-learn's sklearn KFold cross-validation package [35]. The negative cross-validation and positive training datasets were partitioned into five folds: every fold built a profile HMM based on four parts of the positive training data using *hmmbuild*. The remaining positive part was combined with one part of the negative cross-validation dataset and used as the test data. TPR, FPR, ACC, and MCC were determined for each fold and averaged to assess the classifier's overall success.

2.6. *Testing*

To evaluate the ability of the profile HMM to classify putative furin cleavage sites of enveloped virus proteins, the remaining 20% of the positive and negative dataset was used to test the model. A custom Python program called FindFur_HMMERParser was written to parse the output from HMMER in domtblout format. This program's primary purpose is to determine whether or not a hit from HMMER is believed to be a furin recognition site. The cut-off for a putative furin cleavage site is based on the lowest e-value for a positive sequence reported during optimization. Thus, the cut-off is set at 0.0023. The program reports the hit name, description, e-value, bit-score, alignment region, and whether or not the model classifies the hit as a furin cleavage site for each hit.

2.7. *Hydrophobicity-based logistic regression model*

Hydrophobicity is related to the transfer of free energy between polar and nonpolar mediums [40]. It exerts a profound influence on the physicochemical and biological interactions of amino acids, from protein folding to protein function. One apparent influence is the extent to which amino acid side-chains are available for direct interaction with solvents in their environment [41]. Proteins that move through the secretory pathway, such as furin and its substrates, prefer to have hydrophobic residues present at their protein core and hydrophilic residues interfacing the protein surface. Many published hydropathy scales try to capture the nature of these chemical reactions for each amino acid. The most widely used is the Kyte-Doolittle (KD) scale for detecting hydrophobic regions in proteins; where the larger (more positive) the value is, the more hydrophobic the residue [36]-[38]. This, and four other properties, are described of the 20 common amino acids side chain in Table III.

TABLE III. PHYSICAL PROPERTIES OF THE 20 AMINO ACID SIDE CHAINS IN INCREASING ORDER OF HYDROPATHY

Amino Acid			Hydropathy (KD)	Volume	Physiochemical	Charge	Polarity
Arginine	R	Hydrophilic	-4.0	Large	Basic	Positive	Polar
Lysine	K		-3.9	Large	Basic	Positive	Polar
Glutamic acid	E		-3.5	Medium	Acidic	Negative	Polar
Glutamine	Q		-3.5	Small	Amide	Uncharged	Polar
Aspartic acid	D		-3.5	Small	Acidic	Negative	Polar
Asparagine	N		-3.5	Small	Amide	Uncharged	Polar
Histidine	H	Neutral	-3.2	Medium	Basic	Positive	Polar
Proline	P		-1.6	Small	Aliphatic	Uncharged	Polar
Tyrosine	Y		-1.3	Small	Aromatic	Uncharged	Polar
Serine	S		-0.8	Very small	Hydroxyl	Uncharged	Polar
Threonine	T		-0.7	Small	Aliphatic	Uncharged	Nonpolar
Glycine	G		-0.4	Very small	Aliphatic	Uncharged	Nonpolar
Tryptophan	W	Hydrophobic	-0.9	Very large	Aromatic	Uncharged	Nonpolar
Alanine	A		1.8	Very small	Aliphatic	Uncharged	Nonpolar
Methionine	M		1.9	Large	Sulfur	Uncharged	Nonpolar
Cysteine	C		2.5	Small	Sulfur	Uncharged	Nonpolar
Phenylalanine	F		2.8	Very large	Aromatic	Uncharged	Nonpolar
Leucine	L		3.8	Large	Aliphatic	Uncharged	Nonpolar
Valine	V		4.2	Medium	Aliphatic	Uncharged	Nonpolar
Isoleucine	I		4.5	Large	Aliphatic	Uncharged	Nonpolar

Binomial logistic regression is a robust form of regression used when the dependent variable, or target, is dichotomous, and the independent variables are either continuous or categorical [39]. Here, the dichotomy is whether a substrate possesses a furin recognition sequence motif belonging to an enveloped virus precursor or not. The independent variables are the 20 columns in the multiple sequence alignment representing the 20-residue present at the cleavage site. Thus, the goal was to develop a model capable of classifying new, putative furin

cleavage sites into these two groups, depending on the hydrophobicity levels throughout the sequence.

The same data used for the profile HMM was extended to this model. Before applying any statistical learning, all data were converted to its respective KD value and labeled either "1" or "0" for positive and negative, respectively. A boxplot was generated for the positive data to represent the KD score distribution at each residue position.

To build the logistic regression model, scikit-learn's sklearn model selection, logistic regression, and metrics packages were used [35]. Data was stratified and split using the same 80% and 20% training and testing ratio. The logistic regression model was created and fit from the training set data. Hyperparameter optimization was performed using Grid Search. The model's robustness was evaluated with the test data and reported with a classification report and ROC curve.

3. RESULTS

3.1. *Data collection*

A bioinformatics pipeline was developed to build a profile HMM from experimentally verified furin cleavage site sequences of viral envelope substrates. This is depicted above in the workflow of Figure 1. To ensure the quality of the profile HMM, the 91 proteins first collected for the positive dataset were filtered to 85 sequences after removing sequences that did not have an essential R in position P1 and P4 or were the incomplete length ($n < 20$). Duplicate sequences were removed with seqkit leaving 74 sequences, which were then split at random for training and testing. For the negative dataset, data were manually filtered to remove any sequence collected

without an arginine residue at P1, and duplicate sequences were also removed with seqkit [29].

Of all 836 sequences initially collected, 414 were selected for the negative dataset.

The number of initial sequences for the positive and negative sets, along with the number of sequences remaining following filtration, are tabulated in Table IV and Table V, respectively.

TABLE IV. INITIAL DISTRIBUTION OF THE POSITIVE DATASET

Positive dataset	Original	Filtered
FurinDB	36	36
MEROPS	18	14
Genbank/Uniprot	37	24
Total	91	74

TABLE V. INITIAL DISTRIBUTION OF THE NEGATIVE DATASET

Negative dataset	Original	Filtered
MEROPS – PCSKs	629	171
MEROPS – Peptidases	207	146
FurinDB – Mammal FCS	91	91
FurinDB – Bacterial FCS	6	6
Total	836	414

3.2. Sequence logo analysis

The residue pattern at the cleavage site for the full positive dataset, full negative dataset, and PCs alone are represented as sequence information logoss in Figure 4. Position 14 or the P1 site designates the central position where cleavage occurs. The sequence logo shows the 6-flanking amino acid residues on the right side of the P1 site (towards the C-terminus), and the 13-amino acid residues on the left side (towards the N-terminus), depicting the frequency and diversity of amino acids frequently present at the cleavage site. Note that sequence logo analysis was performed on the final filtered datasets.

Figure 4. Sequence logos of aligned peptidase cleavage sites centered at P1: (A): 74 experimentally determined furin cleavage site of viral substrates; (B): 146 general PCs; (C): 414 general proteases

The consensus furin site is typically described by the four-amino acid sequence pattern R-X-[K/R]-R ↓, ranging from position P1 to P4. This can be visibly seen from the sequence logo in Figure 4(a). The region between P6 to P2' is often regarded as the core site. The furin cleavage sites of the 74 viral substrates possess the critical arginine residue at position P1. At position P4, the frequency of arginine was 98.6%, with one cleavage motif (1.4%) containing a lysine residue. This belonged to the exception of the training set, Ebola virus strain Reston, and appeared to compensate for the lack of arginine in P4 for one at P5. At position P2, lysine (50%) and arginine (43.2%) appeared most frequently. Glutamine (2.7%), valine (1.4%), serine (1.4%), and alanine (1.4%) were the remaining amino acids found at P2. P3 was considered the most variable region but showed a preference for neutral or slightly hydrophobic residues. While some aliphatic residues, such as alanine (5.4%), isoleucine (2.7%), valine (1.4%), were observed in small amounts here, hydrophobic residues were largely absent. At P1', small, hydrophilic residues tended to be seen with 41.9% of sequences containing serine and 20.3% containing glycine. No bulky, hydrophobic residues (i.e. isoleucine, leucine, and valine) appeared at P1', but did appear in P2'. At P2', hydrophobic with aliphatic residues made up 83.7% of the position.

Recently, the furin cleavage site motif has been characterized as a 20-amino acid motif, extending from position P14 to position P6' [13]. Between P5 and P14, small, hydrophilic, or positively charged residues were favored in this region. Within the ten residues, serine appeared the most frequently in nine out of 10 positions. Glycine and threonine appeared in the top three in seven out of 10 positions. P5 usually was represented by a positively charged residue or small, hydrophilic residue.

On the C-terminal side of the cut site, the region between P3' to P6' demonstrated a preference for hydrophobic residues. Small, hydrophilic residues were also accepted. Positions

closer to the cleavage site tended to be composed of aliphatic residues with P3' containing 94.5%, and some small, hydrophilic residues appeared in greater numbers moving further from the cut site.

The general PC cleavage site recognizes the sequence [K/R]–X–[K/R]–R ↓ and possesses a less stringent substrate specificity than furin. While the pattern is observable in the sequence logo of Figure 4(b), it is also clear that there is greater variability in the distribution of amino acids in the general proprotein convertase group. Arginine representing 100% of the occurrence in 146 proteins at P1 confirms that the dataset was accurately collected. At P2, R and K represented 60.5% and 24.9% of the total frequency, respectively, and at P4, they represented 50.8% and 5.1%. P3 contained primarily positively charged or small, hydrophobic amino acids, including glutamine (13.6%), lysine (11.3%), glycine (10.7%), arginine (10.2%), and serine (7.3%). At position P1', the frequency of serine was 24.9%, which is half of what appeared at the cleavage site of the viral substrates. The remaining P14 to P5 and P3' to P6' was observed to be highly variable. However, similar to the furin cleavage motif, these regions appeared to favor either aliphatic residues like leucine or small, hydrophilic residues like asparagine and glutamate.

The final sequence logo in Figure 4(c) represents the distribution of amino acids in all proteins collected as negative sites. Of 414 proteases, arginine represented 100% of the amino acid distribution at P1, indicating this dataset was also correctly collected. The frequencies of R and K at P2 and P4 are 34.3% and 20.3%, and 39.4% and 4.8%, respectively. Though the amino acid distribution of P3, P14 to P5, and P3' to P6' was also quite variable, it also followed a similar pattern to the general PCs. The furin cleavage motif R–X–[K/R]–R and minimal furin cleavage pattern R-X-X-R↓ appears in almost 30% of the negative dataset. While some sequences may be furin substrates of mammals or bacteria, these observations suggest that the

proteins in this dataset may belong to non-furin PCs and can potentially be classified as furin cleavage sites by the model due to substrate specificity overlap.

3.3. *Hydrophobicity analysis*

To further explore the residue pattern identified during sequence logo analysis, a hydrophobicity pattern plot was generated for the positive and negative dataset with NCBI's Multiple Sequence Alignment Viewer (hydropathy coloring scale) [33]. As seen in Figure 5, the hydrophobicity pattern of the viral substrates follows the general observations made about the residues in the sequence logo. The core site from position P1 to P5 appeared composed primarily of highly hydrophilic residues, with P3 showing the greatest variability, possessing some degree of hydrophobicity. At P6, there was a significant decrease in the hydrophilicity level as more hydrophobic residues began to appear. Following the cut site at P1, P1' demonstrated a preference for neutral or slightly hydrophilic amino acids, and P2' favored highly hydrophobic amino acids. The outer region from P3' to P6 also showed a strong preference for hydrophobic residues, though there are some sequences with more hydrophilic residues. This contrasts with the outer region from P5 to P14, which shows a gradual gradient from hydrophobic residues to hydrophilic residues.

Figure 5. Hydrophobicity pattern plot of positive dataset (only viral furin substrates)

Figure 6. Hydrophobicity pattern plot of negative dataset

While overlap between the hydrophobicity pattern plots of the positive and negative dataset does occur, there are distinct differences that exist. They are seen at position P4 and from P1' to P6' in Figure 6. Unlike the positive dataset in Figure 5, the negative dataset possesses a less noticeable hydrophilic/hydrophobic boundary shift at P4 and has greater variability in the spread of hydrophobic and hydrophilic residues from P5 to P14. The presence of highly hydrophilic residues at P4 is likely due to the other PCs and the furin substrates of mammals and bacteria present. Figure 6 also shows that the outer region from P3' to P6' contains more hydrophilic residues compared to the stretches of hydrophobic residues in Figure 5. These observations support Tian 2009 and highlight the importance of hydrophobicity within the P3' to P6' region for the efficiency of furin cleavage in viral substrates [13]. It also suggests how hydrophobicity profiles can be used to predict and classify potential furin cleavage site motifs.

3.4. Optimization analysis

The positive and negative datasets were randomly shuffled and split into their respective training, validation, and test datasets based on the portions described in the methods above. The distribution of datasets is tabulated in Table VI. The profile HMM model was built from 59 sequences using the *hmmbuild* program in HMMER3 [34]. These sequences and the 67 negative sequences from the optimization dataset were used to assess the optimal filter settings for the *hmmsearch* program.

TABLE VI. DISTRIBUTION OF TRAINING, VALIDATION, AND TESTING DATASETS FOR PROFILE HMM

Dataset type	Positive	Negative	Total
Training (80%)	59	331	390
Optimizing threshold (20%)		67	
Cross-validation (80%)		264	
Testing (20%)	15	83	98
Total	74	414	488

Changes to the MSV filter p-value from 0.01 to 0.9 allowed more sequence to pass through the initial filter, but it did not influence the number of sequences passing at the end as either positive or negative hits. Changes to the Viterbi filter p-value from 0.01 to 0.9 showed similar results. Changes to the forward p-value from $2e^{-5}$ to 0.09 demonstrated to have the greatest impact on the number of sequences passed as hits. As seen in Figure 7, incrementing the p-value from $2e^{-5}$ and on showed a gradual increase in the number of positive and negative hits. All positive sequences passed at p-value 0.00008 and three negative sequences passed at p-value 0.0008 as false positives. At p-value 0.0003, six negative sequences passed as false positives, and for all p-values after 0.0004, seven negative sequences would pass as false positives.

It should be mentioned that HMMER also reports results they consider passes as “inclusion threshold.” These are hits they consider reliable enough to be included in an output alignment, or in future searches. For the scope of this project, these results were not considered as hits.

Figure 7. Graphical distribution of *hmmsearch* hit results for varying independent MSV/Vit/Fwd thresholds

To determine the optimal search settings for the profile HMM, an iteration through varying combinations of all three parameters was performed using the results from Table VII as a guide. The numbers of true positives (TP), false positives (FP), true negatives (TN), and false negatives (FN), and true positive rate (TPR) and false positive rates (FPR) were recorded in addition to numbers that were recorded in Table V. Table VII tabulates the raw numbers, and Table VIII tabulates the latter results.

TABLE VII. RAW SCORES FROM *HMMSEARCH* HITS OF THRESHOLD COMBINATIONS FOR PARAMETER OPTIMIZATION

Filter	Total	MSV	Bias	Vit	Fwd	Hits	Pos	Neg
0.05 0.001 0.00002	126	93	87	66	55	55	55	0
0.06 0.001 0.00002	126	95	89	66	55	55	55	0
0.07 0.001 0.00002	126	95	89	66	55	55	55	0
0.08 0.001 0.00002	126	97	90	66	55	55	55	0
0.08 0.002 0.00002	126	97	90	67	55	55	55	0
0.08 0.003 0.00003	126	97	90	68	58	58	57	1
0.08 0.004 0.00003	126	97	90	69	58	58	57	1
0.08 0.005 0.00003	126	97	90	69	58	58	57	1
0.08 0.006 0.00003	126	97	90	70	58	58	57	1
0.08 0.006 0.00003	126	97	90	70	58	58	57	1
0.08 0.007 0.00003	126	97	90	73	58	58	57	1
0.08 0.008 0.00003	126	97	90	73	58	58	57	1
0.08 0.009 0.00003	126	97	90	75	58	58	57	1
0.08 0.01 0.00003	126	97	90	76	58	58	57	1
0.08 0.02 0.00003	126	97	90	80	58	58	57	1
0.08 0.02 0.00004	126	97	90	80	60	60	57	3
0.08 0.02 0.00005	126	97	90	80	60	60	57	3
0.08 0.02 0.00006	126	97	90	80	60	60	57	3
0.08 0.02 0.00007	126	97	90	80	60	60	57	3
0.08 0.02 0.00008	126	97	90	80	62	62	59	3
0.08 0.02 0.00009	126	97	90	80	62	62	59	3
0.08 0.02 0.0001	126	97	90	80	62	62	59	3
0.08 0.02 0.0002	126	97	90	80	64	64	59	5
0.08 0.02 0.0003	126	97	90	80	67	66	59	7
0.08 0.02 0.0004	126	97	90	80	68	67	59	8
0.08 0.02 0.0005	126	97	90	80	68	67	59	8

TABLE VIII. CLASSIFICATION RESULTS OF THRESHOLD COMBINATIONS FOR PARAMETER OPTIMIZATION

Filter	TP	FP	TN	FN	TPR	FPR	ACC	MCC
0.05 0.001 0.00002	55	0	67	4	0.0359	0.0000	0.9683	0.9379
0.06 0.001 0.00002	55	0	67	4	0.0717	0.0000	0.9683	0.9379
0.07 0.001 0.00002	55	0	67	4	0.1076	0.0000	0.9683	0.9379
0.08 0.001 0.00002	55	0	67	4	0.1434	0.0000	0.9683	0.9379
0.08 0.002 0.00002	55	0	67	4	0.1793	0.0000	0.9683	0.9379
0.08 0.003 0.00003	57	1	66	2	0.2164	0.0006	0.9696	0.9402
0.08 0.004 0.00003	57	1	66	2	0.2536	0.0011	0.9705	0.9418
0.08 0.005 0.00003	57	1	66	2	0.2907	0.0017	0.9712	0.9431
0.08 0.006 0.00003	57	1	66	2	0.3279	0.0023	0.9718	0.9440
0.08 0.006 0.00003	57	1	66	2	0.3651	0.0029	0.9722	0.9448
0.08 0.007 0.00003	57	1	66	2	0.4022	0.0034	0.9726	0.9455
0.08 0.008 0.00003	57	1	66	2	0.4394	0.0040	0.9729	0.9460
0.08 0.009 0.00003	57	1	66	2	0.4765	0.0046	0.9731	0.9465
0.08 0.01 0.00003	57	1	66	2	0.5137	0.0051	0.9734	0.9469
0.08 0.02 0.00003	57	1	66	2	0.5508	0.0057	0.9735	0.9473
0.08 0.02 0.00004	57	3	64	2	0.5880	0.0074	0.9727	0.9455
0.08 0.02 0.00005	57	3	64	2	0.6252	0.0091	0.9720	0.9440
0.08 0.02 0.00006	57	3	64	2	0.6623	0.0109	0.9713	0.9426
0.08 0.02 0.00007	57	3	64	2	0.6995	0.0126	0.9708	0.9414
0.08 0.02 0.00008	59	3	64	0	0.7379	0.0143	0.9710	0.9419
0.08 0.02 0.00009	59	3	64	0	0.7764	0.0160	0.9713	0.9424
0.08 0.02 0.0001	59	3	64	0	0.8149	0.0177	0.9715	0.9428
0.08 0.02 0.0002	59	5	62	0	0.8533	0.0206	0.9710	0.9418
0.08 0.02 0.0003	59	7	60	0	0.8918	0.0246	0.9699	0.9396
0.08 0.02 0.0004	59	8	59	0	0.9302	0.0291	0.9686	0.9369
0.08 0.02 0.0005	59	8	67	0	0.9687	0.0337	0.9674	0.9346

When the forward filter remained at p-value $2e^{-5}$, changes to MSV and Viterbi filter for the first five settings did not affect the number of TP, FP, TN, and FN. For all five, 55 out of 59 true positives passed without any false positives, and four sequences were reported as false negatives. When the forward filter initially changed to p-value $3e^{-5}$, three additional hits passed

through. Out of 58 hits, 57 out of 59 were true positives, one was a false positive, and two were false negatives. Subsequent changes to the Viterbi filter did not impact the number of hits passed and maxed out at p-value 0.02 since the MSV filter was not changing. Following this, only the forward filter was adjusted. Two additional sequences passed through as false negative when the p-value was set at $4e^{-5}$ until $7e^{-5}$. At F1 0.08, F2 0.002, and F3 $8e^{-5}$, all 59 sequences passed as true positives, and three negative sites passed as false positives. The number of false positives continued to increase after p-value 0.001 for the forward filter.

A ROC curve was built to illustrate the prediction performance for the varying thresholds, plotting the false positive rate on the x-axis and sensitivity (or TPR) on the y-axis. The ROC curve shown in Figure 9 indicates that F1 0.008, F2 0.002, F3 $8e^{-5}$ is a good threshold choice since the false positive rate is 1.43%, and the true positive is 73.8%. After this threshold, the false positive rate continued to increase, though this was expected since the number of true negative sites surpassed the number of true positive sites.

Figure 8. ROC diagram based on varying thresholds during parameter optimization shown as TPR (sensitivity) plotted against FPR

A summary of the overall performance until F1 0.008, F2 0.002, F3 $8e^{-5}$ is tabulated in Table IX. At this threshold, the model's sensitivity is 95.9%, specificity is 98.1%, and the Matthews correlation coefficient (MCC) is 0.94. It should be noted that the optimization dataset was not completely independent as its constituent proteases were previously used in training the profile HMM search network.

TABLE IX. PERFORMANCE EVALUATION OF OPTIMAL PROFILE HMM SEARCH SETTING

TP	1132		TPR	0.95932
FN	48		FNR	0.04068
TN	1315		TNR	0.98134
FP	1132		FPR	0.01865
P	1180		ACC	0.97103
N	1340		MCC	0.94192
PP	1157			
PN	1363			
Total	5040			

3.5. Prediction performance

Five-fold cross-validation was employed to evaluate the performance of the profile HMM's ability to correctly classify the furin cleavage sites of interest. Sequence distribution is tabulated in Table X for each fold. Each profile searched their respective test set using the optimal *hmmsearch* settings F1 0.008, F2 0.002, F3 $8e^{-5}$.

TABLE X. DISTRIBUTION OF SEQUENCES FOR FIVE-FOLD CROSS-VALIDATION

	Fold 1	Fold 2	Fold 3	Fold 4	Fold 5
<i>HMM</i>	47	47	47	47	48
Positive	12	12	12	12	11
Negative	53	53	53	53	52

On average, 74.7% of the furin cleavage sites of viral envelope viral proteins and 93.5% of the non-furin cleavage sites were accurately predicted across the five models, leaving 6.0% being misclassified. The model achieved 90.4% accuracy and an MCC of 0.68. This suggests the model under consideration was adequate for predicting viral furin substrates; however, it was even better at identifying sequences that did not fit the motif site. These results were tabulated in Table XI and XII.

TABLE XI. RAW SCORES FROM FIVE-FOLD CROSS-VALIDATION

Fold	TP	FP	TN	FN	Total
F1	6	4	49	6	65
F2	10	3	50	2	65
F3	10	3	50	2	65
F4	9	4	49	3	65
F5	9	2	50	2	63
Total	44	16	248	15	323

TABLE XII. PREDICTIVE PERFORMANCE EVALUATION FROM FIVE-FOLD CROSS-VALIDATION

Fold	TPR	FPR	TNR	FNR	ACC	MCC
F1	0.5000	0.0755	0.9245	0.5000	0.8462	0.4565
F2	0.8333	0.0566	0.9434	0.1667	0.9231	0.7534
F3	0.8333	0.0566	0.9434	0.1667	0.9231	0.7534
F4	0.7500	0.0755	0.9245	0.2500	0.8923	0.6543
F5	0.8182	0.0385	0.9615	0.1818	0.9365	0.7797
Average	0.7470	0.0605	0.9395	0.2530	0.9042	0.6807

Of the 16 false positives, six were furin substrates of humans and mice, and 14 were PCs of humans and mice. In Figure 10, a sequence logo of the false positives reveals that the distribution of amino acids within the core region and the outer region mirror closely to the substrate pattern of enveloped viruses. At position P2, the frequency of arginine and lysine are

66.7% and 26.7%, respectively. At P4, 93.3% of sequences possessed an arginine. The distribution of amino acids at the variable P3 is also similar. P1' also did not possess either leucine, isoleucine, or valine in this position. It did appear prominently at P2'.

Figure 9. Sequence logo for false positives from five-fold cross-validation

The surprisingly good performance of the model on independent sequences in correctly identifying cleavage sites of envelope glycoproteins was accomplished at 86.7%, though at the cost of a greater number of false positives (10.8%). From sequence logo analysis of the false positives in Figure 10, it would be reasonable to presume such false positives might appear as sequences that are highly similar to furin cleavage sites of viral envelope proteins. Overall, the model receives a test accuracy of 88.8% and an MCC of 0.65. The classification results are described in Table XII.

TABLE XIII. PREDICTIVE PERFORMANCE EVALUATION OF FINAL PROFILE HMM FROM TEST DATA

TP	FP	TN	FN	TPR	FPR	TNR	FNR	ACC	MCC
13	9	74	2	0.8667	0.1084	0.8916	0.1334	0.88775	0.6543

3.6. *Hydropathy-based logistic regression model analysis*

The sequence logo and hydrophobicity pattern plot analysis revealed that viral substrates of furin possess highly hydrophilic cores with a more hydrophobic C-terminal region from P3' to P6'. This contrasts with the outer N-terminal region from P5 to P14, which shows a more gradual shift from hydrophilic amino acids to hydrophobic residues. The distribution of KD-values at each position are shown as boxplots in Figure 11. For each boxplot, the inside line, the lower hinge, and the upper hinge represents the median, the 25th percentile, and the 75th percentile, respectively. The triangle represents the mean, and the circles beyond the whiskers represent outliers. Overall, the mean hydrophobicity values of the viral furin substrates were the highest from P1' to P6', with P2' being the most hydrophobic. The core region from P1 to P4 was observed to be the most hydrophilic. A gradual increase in hydrophobicity after position P4 can be seen in Figure 11.

Figure 10. Boxplot of hydrophobicity of each residue position

From these observations, it appeared reasonable to build a predictive model based on hydrophobicity. An initial comparison of model accuracy between logistic regression, support vector machine (SVM), Random Forest (RF), and neural networks (NN) suggested the best performing algorithm was logistic regression; therefore, logistic regression was selected to construct the model. It was also the simplest model of all four algorithms. The distribution of the training and testing datasets are tabulated in Table XIII. The logistic regression model was trained on 59 positive sequences and 331 negative sequences.

TABLE XIV. DISTRIBUTION OF TRAINING AND TESTING DATASETS FOR LOGISTIC REGRESSION MODEL

Dataset type	Positive	Negative	Total
Training (80%)	59	331	390
Testing (20%)	16	82	98
Total	75	413	488

Ten-fold cross-validation was employed to evaluate the performance of the classifier. It obtained both an accuracy and mean AUC score of 0.895. The logistic regression hyperparameters were then optimized using a grid search, achieving its highest performance with the regularization parameter $C = 1.0$. Overall, the logistic regression model achieved good validation statistics, with a ten-fold cross-validated accuracy of 90.8% and an MCC value of 0.62. The results are enumerated in Table XIV.

TABLE XV. PREDICTIVE PERFORMANCE EVALUATION OF LOGISTIC REGRESSION MODEL FROM TEST DATA

TP	FP	TN	FN	TPR	FPR	TNR	FNR	ACC	MCC
9	3	80	6	0.60	0.0361	0.9639	0.40	0.9082	0.6193

The predictive power of the model was further evaluated using a ROC curve, which revealed adequate performance on the test set. Figure 12 shows selecting any sensitivity

threshold higher than 60% also increases the FPR. The AUC was 0.78, indicating the model had acceptable predictive power.

Figure 11. ROC curve based on varying thresholds during optimization of the final logistic regression model

In comparison to the profile HMM results, the logistic regression model was more specific than the profile HMM with a difference of 7.19%, but was significantly less sensitive with a difference of 26.7%. Interestingly, the logistic regression model predicted furin cleavage sites in the independent sequence with a slightly higher test accuracy rate at 91%. This provides good evidence that a hydropathy-based furin prediction model would work, though the positive dataset or model's refinement is necessary to improve results.

4. IMPLEMENTATION OF FINDFUR

FindFur is a Python-based prediction tool used to extract, evaluate, and report putative recognition sites of viral furin substrates based on the trained profile HMM. The following subsections detail requirements and usage.

4.1. *Requirements*

- macOS environment
- Python 3+
- Hmmer 3.3.1 installed on system
- Pandas
- Biopython

4.2. *Usage*

Both scripts are executed through the command line. To run the tool is a two-part process and requires the profile HMM called 2POSTITIVE.HMM was built from HMMER's *hmmbuild* [34]. The workflow can be seen in Figure 13. All files can be publically accessed through the Github repository at <https://github.com/chwisteeng/FindFur>.

Figure 12. Pipeline for FindFur tool

4.2.1. FindFur_Extract.py

- **Purpose:** Extract all potential furin cleavage sites in a FASTA file containing the protein sequences a user would like to search with the profile HMM
- **Input:** Protein sequences of interest in FASTA format and path to HMMER's *hmmsearch*
- **Output:** Putative furin cleavage site motifs of length 20-residues, HMMER results in domain table format

4.2.2. FindFur_HMMERParse.py

- **Purpose:** Reports whether motifs of interest are furin cleavage site hits for viral substrates
- **Input:** Previous HMMER results and motif FASTA file
- **Optional output:** Text file of HMMERParse results

An example of the command-line output from FindFur_ParseHMMER.py is shown in Figure 13.

```
(base) Christines-MacBook-Pro:Report_code christinegu$ python3 FindFur_ParseHMMER.py -d final_test.txt -f all_testing.fasta
Hit
0 POSITIVE_74 sp|P29990|POLG_DEN26 2.300000e-07 22.7 0 GTCTTMGEHRRQKRSVAL 18 HIT!
1 POSITIVE_56 P33478.2 Dengue virus 1 Singapore/S275/1990 1.800000e-06 20.0 0 GTCSTQGEHRRQKRSVALA 19 HIT!
2 POSITIVE_46 BAB85129.1 Newcastle disease virus (STRAIN Sato) 2.100000e-06 19.8 2 SVTTSQGRQRFFIGA 18 HIT!
3 POSITIVE_10 sp|O56140|HEMA_I97A1 2.900000e-06 19.4 1 RNTPQRRRRKKRGLFGAI 20 HIT!
4 POSITIVE_66 28849361 Influenza A virus (A/Pheasant/Hong Ko... 2.900000e-06 19.4 1 RNTPQRRRRKKRGLFGAI 20 HIT!
5 POSITIVE_13 sp|P11223|SPIKE_IBVB 3.600000e-06 19.1 3 KITNGTRRRFRRSITEN 19 HIT!
6 POSITIVE_1 sp|O56861|ENV_FFV 9.700000e-06 17.8 1 MGGNTSSSSRRRRDIQY 18 HIT!
7 POSITIVE_19 P12650.1 Avian infectious bronchitis virus (st... 1.300000e-05 17.4 3 KITNGTRRRFRRSITE 18 HIT!
8 POSITIVE_11 sp|P11132|HEMA_I83A3 2.100000e-05 16.8 3 RNTPQRRKKRGLFGAI 20 HIT!
9 POSITIVE_64 130579 Sindbis virus 3.800000e-05 16.0 2 ILRCGSSGRKRSVIDD 19 HIT!
10 POSITIVE_29 NP_040154.2 Human alphaherpesvirus 3 6.700000e-05 15.3 2 QKHPTNTRRSRSPVPE 19 HIT!
11 POSITIVE_5 sp|P04578|ENV_HV1H2 3.400000e-04 13.2 3 AKRRVVQREKRAVGIG 19 HIT!
12 NEGATIVE_275 NEGATIVE_179 sp|P30885|BMP4_XENLA 3.500000e-04 13.1 6 ALTRRSKRSPKQ 18 HIT!
13 NEGATIVE_40 14717127 Xenopus tropicalis 3.500000e-04 13.1 6 ALTRRSKRSPKQ 18 HIT!
14 POSITIVE_6 tr|Q9MMU9|Q9MMU9_9HIV1 3.700000e-04 13.1 1 TRARRRVQREKRAVGL 18 HIT!
15 NEGATIVE_74 33112647 Human 6.900000e-04 12.2 6 PRPSRKRRSLGD 18 HIT!
16 NEGATIVE_292 NEGATIVE_202 sp|P01137|TGFBI_HUMAN 9.800000e-04 11.8 2 RAQHLQSSRRRLD 17 HIT!
17 NEGATIVE_32 62906884 Human 1.000000e-03 11.7 4 GLSARNRQKRFV 16 HIT!
18 NEGATIVE_31 1352125 Mouse 1.400000e-03 11.3 6 NARNRQKRFV 16 HIT!
19 NEGATIVE_291 NEGATIVE_201 sp|Q76LX8|ATS13_HUMAN 1.500000e-03 11.2 6 RQRRQRRRAAGGI 19 HIT!
20 NEGATIVE_337 NEGATIVE_250 sp|Q9Y275|TN13B_HUMAN 2.000000e-03 10.8 2 EGNSSQNSRNKRAVGQP 19 HIT!
21 NEGATIVE_23 12644170 Human 2.000000e-03 10.8 4 TESHNSRKKREIT 17 HIT!

Number of queries: 98
Number of hits: 22
Number of potential FCS: 22
```

Figure 13. HMMERParser output of test set results

5. DISCUSSION

Furin is an essential protease that promotes the proteolytic maturation of many inactive precursors into their biologically active forms. Substrates of furin are typically recognized and cleaved at the consensus site R-X-[K/R]-R↓, marked by a pair of basic amino acids. Although cleavage at the general PC motif R-X-X-R by furin has also been reported, it is usually less efficient. Many families of viruses exploit furin for proteolytic processing of their viral particles to help mediate their cell entry and infectivity. Due to its ubiquity and capacity to identify its substrates, this suggests being able to predict the variation at furin recognition sites is important because the processing of these viral particles by furin can lead to widespread disease throughout the body.

In this project, a dataset of known furin cleavage sites of viral envelope glycoproteins were used to build two different statistical models. The first used protein sequences to construct a profile HMM and the latter was constructed by converting the sequences to their respective KD-values and modeled using logistic regression. The data was unique in that it attempted to cover all known furin consensus sites in viral envelope precursors. However, due to the small sample size, the scope of the dataset may have limited predictive performance in both models. Optimizing the dataset will require a greater number of confirmed positive sequences to better capture the sequence variation at the cleavage site, thereby improving model performance. It may also be interesting to see how more drastic changes to the data selection criteria may affect predictive performance.

Testing the final profile HMM's classifier performance on the test set revealed a low false positive rate (10.8%) with relatively good sensitivity (86.7%), indicating the classifier's ability to identify the recognition site of viral furin substrates adequately. This is reinforced by the test

accuracy score of 88.7% and the MCC score of 0.65. Given the sequence similarity seen in the false positives during five-fold cross-validation, it is reasonable to believe that the false positives of the test set may have also been similar to the target. Here, it was known that motifs were primarily substrates of PCs or mammalian and bacterial substrates of furin because it was within the project's boundaries and was somewhat expected since overlap can occur. In an external setting, depending on the sequences' source, presumed false positive results should be interpreted with caution but could be further evaluated.

The Python-based tool developed for this project, FindFur, has two thresholds, one for each part of the tool. FindFur_Extract filters sequences through the optimized *hmmsearch* parameters (F1 0.08, F2 0.02, F3 0.00008) and FindFur_HMMERParse reports only hits whose e-value are above 0.0023. The test scores suggest that the classifier's cut-off thresholds may be too stringent. However, it is better to have a more stringent cut-off with a narrow target to have more confidence in the hits being reported.

FindFur also reports the e-value, bit-score, aligned sequence, and the unique identifier for each hit. Both the e-value and bit scores are parameters to describe sequence similarity. The e-value represents the likelihood that a given sequence match is purely due to chance when searching a particular size database. In contrast, the bit score acts as a constant statistical indicator of homology because it is independent of query sequence length and database size. The two parameters have an opposite relationship, where a smaller e-value represents a better match, and a higher bit-score represents a more significant match. These reported values can help guide future studies to identify the probability of furin cleavage in viral substrates. The degree of significance may also help to uncover information about cleavage efficiency and cell tropism as research has shown that the residue pattern at the cut site correlates with both those aspects.

The proposed logistic regression performed fairly well with an extremely low test false negative rate at 3.61% and a decent test sensitivity rate at 60.0%. While these scores are acceptable, the false negative rate is high at 40.0% compared to the profile HMM, which indicates a moderate chance that the model will misclassify a viral furin cleavage site. Probable reasons for the erroneous predictions may be attributed to underfitting and the model hyperparameters, but it does not account for all. At this point, it is unclear as to what are the best means to improve this particular model. However, definitive steps include refinement of the dataset and the model hyperparameters.

This project demonstrated the ability to build an enveloped virus-specific prediction model for furin cleavage. Future studies involving furin cleavage sites should involve the full 20 amino acid motifs. From this research, it is clear that there are significant differences at the cut site for viral substrates, which could be exploited as targets for treatment. Expanding this tool to include substrates of naked viruses would be beneficial. In future work, optimizing the profile HMM search settings will be done through five or ten-fold cross-validation. Splitting the monolithic profile HMM into separate modules may better capture the biology at the cleavage site. The advantage of this is that if the model becomes more precise, the more accurate the hits will be. The logistic regression model also shows significant potential and indicates that other statistical models may be applied to use hydrophobicity as a predictive tool. It should be mentioned that future studies involving furin cleavage sites should include the full 20 amino acids motif.

REFERENCES

- [1] N. G. Seidah and A. Prat, "The biology and therapeutic targeting of the proprotein convertases," *Nat. Rev. Drug Discov.*, vol. 11, no. 5, pp. 367–383, May 2012.
- [2] M. Ehrmann and T. Clausen, "Proteolysis as a regulatory mechanism," *Annu. Rev. Genet.*, vol. 38, no. 1, pp. 709–724, 2004, doi: 10.1146/annurev.genet.38.072902.093416.
- [3] L. D. Rogers and C. M. Overall, "Proteolytic post-translational modification of proteins: proteomic tools and methodology," *Mol. Cell. Proteomics*, vol. 12, no. 12, pp. 3532–3542, Dec. 2013.
- [4] N. G. Seidah and M. Chrétien, "Proprotein and prohormone convertases: a family of subtilases generating diverse bioactive polypeptides," *Brain Res.*, vol. 848, no. 1–2, pp. 4562, Nov. 1999.
- [5] E. Braun and D. Sauter, "Furin-mediated protein processing in infectious diseases and cancer," *Clin. Transl. Immunol.*, vol. 8, no. 8, 2019, doi: 10.1002/cti2.1073.
- [6] G. Thomas, "Furin at the cutting edge: from protein traffic to embryogenesis and disease," *Nat. Rev. Mol. Cell Biol.*, vol. 3, no. 10, pp. 753–766, Oct. 2002.
- [7] R. Day, M. K. Schäfer, W. E. Cullinan, S. J. Watson, M. Chrétien, and N. G. Seidah, "Region specific expression of furin mRNA in the rat brain," *Neurosci. Lett.*, vol. 149, no. 1, pp. 27–30, Jan. 1993.
- [8] M. K. Schäfer, R. Day, W. E. Cullinan, M. Chrétien, N. G. Seidah, and S. J. Watson, "Gene expression of prohormone and proprotein convertases in the rat CNS: a comparative in situ hybridization analysis," *J. Neurosci.*, vol. 13, no. 3, pp. 1258–1279, Mar. 1993.
- [9] C. M. Dubois, M.-H. Laprise, F. Blanchette, L. E. Gentry, and R. Leduc, "Processing of transforming growth factor 1 precursor by human furin convertase," *J. Biol. Chem.*, vol. 270, no. 18, pp. 10618–10624, 1995, doi: 10.1074/jbc.270.18.10618.
- [10] C. M. Dubois, F. Blanchette, M.-H. Laprise, R. Leduc, F. Grondin, and N. G. Seidah, "Evidence that furin is an authentic transforming growth factor- β 1-converting enzyme," *Am. J. Pathol.*, vol. 158, no. 1, pp. 305–316, 2001, doi: 10.1016/s0002-9440(10)63970-3.
- [11] N. I. Solovyeva, T. A. Gureeva, O. S. Timoshenko, T. A. Moskvitina, and E. V. Kugaevskaya, "Furin as proprotein convertase and its role in normal and pathological biological processes," *BK1HA8*, vol. 62, no. 6, pp. 609–621, 2016, doi: 10.18097/pbmc20166206609.
- [12] K. Nakayama, "Furin: a mammalian subtilisin/Kex2p-like endoprotease involved in processing of a wide variety of precursor proteins," *Biochem. J.*, vol. 327, no. 3, pp. 625–635, 1997, doi: 10.1042/bj3270625.
- [13] S. Tian, "A 20 Residues motif delineates the furin cleavage site and its physical properties may influence viral fusion," *Biochem. Biophys. Res. Commun.*, vol. 391, no. 2, pp. 2049–2054, 2009, doi: 10.4137/bci.s2049.
- [14] A.-M. Khatib, *The Proprotein Convertases: Discovery, Characteristics, and Link to Tumor Progression and Metastasis*. Morgan & Claypool Publishers, 2013.
- [15] M. Hosaka *et al.*, "Arg-X-Lys/Arg-Arg motif as a signal for precursor cleavage catalyzed by furin within the constitutive secretory pathway," *J. Biol. Chem.*, vol. 266, no. 19, pp. 12127–12130, Jul. 1991.

- [16] X.-L. Guo, L. Li, D.-Q. Wei, Y.-S. Zhu, and K.-C. Chou, "Cleavage mechanism of the H5N1 hemagglutinin by trypsin and furin," *Amino Acids*, vol. 35, no. 2, pp. 375–382, Aug. 2008.
- [17] J. K. Millet and G. R. Whittaker, "Host cell entry of Middle East respiratory syndrome coronavirus after two-step, furin-mediated activation of the spike protein," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 111, no. 42, pp. 15214–15219, Oct. 2014.
- [18] P. Decha *et al.*, "Source of high pathogenicity of an avian influenza virus H5N1: why H5 is better cleaved by furin," *Biophys. J.*, vol. 95, no. 1, pp. 128–134, Jul. 2008.
- [19] V. J. Munster *et al.*, "Insertion of a multibasic cleavage motif into the hemagglutinin of a low-pathogenic avian influenza H6N1 virus induces a highly pathogenic phenotype," *J. Virol.*, vol. 84, no. 16, pp. 7953–7960, Aug. 2010.
- [20] J. M. Luczo *et al.*, "Evolution of high pathogenicity of H5 avian influenza virus: haemagglutinin cleavage site selection of reverse-genetics mutants during passage in chickens," *Sci. Rep.*, vol. 8, no. 1, p. 11518, Aug. 2018.
- [21] Y. Zhang *et al.*, "A single amino acid at the hemagglutinin cleavage site contributes to the pathogenicity and neurovirulence of H5N1 influenza virus in mice," *J. Virol.*, vol. 86, no. 12, pp. 6924–6931, Jun. 2012.
- [22] B. Coutard, C. Valle, X. de Lamballerie, B. Canard, N. G. Seidah, and E. Decroly, "The spike glycoprotein of the new coronavirus 2019-nCoV contains a furin-like cleavage site absent in CoV of the same clade," *Antiviral Res.*, vol. 176, p. 104742, Apr. 2020.
- [23] S. R. Eddy, "What is a hidden Markov model?," *Nat. Biotechnol.*, vol. 22, no. 10, pp. 1315–1316, 2004, doi: 10.1038/nbt1004-1315.
- [24] A. Krogh, B. Larsson, G. von Heijne, and E. L. Sonnhammer, "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes," *J. Mol. Biol.*, vol. 305, no. 3, pp. 567–580, Jan. 2001.
- [25] S. R. Eddy, "Profile hidden Markov models," *Bioinformatics*, vol. 14, no. 9, pp. 755–763, 1998.
- [26] S. Tian, Q. Huang, Y. Fang, and J. Wu, "FurinDB: A database of 20-residue furin cleavage site motifs, substrates and their associated drugs," *Int. J. Mol. Sci.*, vol. 12, no. 2, pp. 1060–1065, Feb. 2011.
- [27] N. D. Rawlings, A. J. Barrett, P. D. Thomas, X. Huang, A. Bateman, and R. D. Finn, "The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database," *Nucleic Acids Res.*, vol. 46, no. D1, pp. D624–D632, Jan. 2018.
- [28] V. E. Volchkov, H. Feldmann, V. A. Volchkova, and H. D. Klenk, "Processing of the Ebola virus glycoprotein by the proprotein convertase furin," *Proc. Natl. Acad. Sci. U. S. A.*, vol. 95, no. 10, pp. 5762–5767, May 1998.
- [29] W. Shen, S. Le, Y. Li, and F. Hu, "SeqKit: A cross-platform and ultrafast toolkit for FASTA/Q file manipulation," *PLOS ONE*, vol. 11, no. 10, p. e0163962, 2016, doi: 10.1371/journal.pone.0163962.
- [30] A. Larsson, "AliView: a fast and lightweight alignment viewer and editor for large datasets," *Bioinformatics*, vol. 30, no. 22, pp. 3276–3278, Nov. 2014.
- [31] T. D. Schneider and R. M. Stephens, "Sequence logos: a new way to display consensus sequences," *Nucleic Acids Res.*, vol. 18, no. 20, pp. 6097–6100, Oct. 1990.

- [32] T. J. Wheeler, J. Clements, and R. D. Finn, "Skyalign: a tool for creating informative, interactive logos representing sequence alignments and profile hidden Markov models," *BMC Bioinformatics*, vol. 15, p. 7, Jan. 2014.
- [33] "NCBI Multiple Sequence Alignment Viewer 1.18.1." <https://www.ncbi.nlm.nih.gov/projects/msaviewer/> (accessed Nov. 30, 2020).
- [34] "HMMER: Fast and sensitive sequence similarity searches." doi: 10.6019/tol.hmmerr-w.2018.00001.1.
- [35] R. Garreta and G. Moncecchi, *Learning scikit-learn: Machine Learning in Python*. Packt Publishing Ltd, 2013.
- [36] J. Kyte and R. F. Doolittle, "A simple method for displaying the hydropathic character of a protein," *J. Mol. Biol.*, vol. 157, no. 1, pp. 105–132, May 1982.
- [37] P. L. Kastiris, K. M. Visscher, A. D. J. van Dijk, and Alexandre M J, "Solvated protein-protein docking using Kyte-Doolittle-based water preferences," *Proteins: Struct., Funct., Bioinf.*, vol. 81, no. 3. pp. 510–518, 2013, doi: 10.1002/prot.24210.
- [38] S. Simm, J. Einloft, O. Mirus, and E. Schleiff, "50 years of amino acid hydrophobicity scales: revisiting the capacity for peptide classification," *Biol. Res.*, vol. 49, no. 1, p. 31, Jul. 2016.
- [39] A. A. Diaz, E. Tomba, R. Lennarson, R. Richard, M. J. Bagajewicz, and R. G. Harrison, "Prediction of protein solubility in *Escherichia coli* using logistic regression," *Biotechnol. Bioeng.*, vol. 105, no. 2, pp. 374–383, Feb. 2010.
- [40] S. Schauerl, M. Podewitz, B. J. Waldner, and K. R. Liedl, "Enthalpic and entropic contributions to hydrophobicity," *J. Chem. Theory. Comput.*, vol. 12, no. 9, pp.4600-4610, Sep. 2016.
- [41] G. J. Lesser and G. D. Rose, "Hydrophobicity of amino acid subgroups in proteins," *Proteins*, vol. 8, no. 1, pp.6-13, Feb. 1990.