In silico Analyses of Small Proteins in *Escherichia coli*

Debbie Cerda, McKenzie Burge, Srujana S. Yadavalli

INTRODUCTION

In recent years, there has been a significant increase on the number of small proteins identified in all three kingdoms of life including bacteria. These proteins were missed due to arbitrary size cut-offs for gene lengths while making genome annotations, whereby only open reading frames encoding proteins ≥ 50 amino acids were considered genes. Intuitively, being that proteins smaller than 50 amino acids in length would be unable to form highly complex structures, the assumption that they do not serve any imperative functions in a given organism can be made. Nonetheless, there are several examples of well-characterized small proteins, which serve diverse physiological roles in cellular processes such as cell division, sporulation, regulation of transporters, histidine kinases and other membrane-bound enzymes, and stress responses [1]. Uncovering the identities and functions of small proteins will be beneficial in expanding the amount of information known about important signaling pathways not only in bacteria, but all organisms.

OBJECTIVE

To create a curated database of all novel small proteins in *Escherichia coli* K12 MG1655 as of 2019 and functionally analyze them using bioinformatics.

Starting with *E. coli*, where we can leverage the available small protein datasets identified via bioinformatics as well as high-throughput experimental methods including ribosome-profiling[2], we have created a detailed catalog of small proteins and bioinformatically characterized their functions. The current criteria for entry is a protein that is less than 100 amino acids in length, thus far the database consists of proteins ranging from 8-80 amino acids long (Fig. 1).This will be an ongoing project specific to *Escherichia coli* K12 MG1655 aimed at generating hypotheses, which can be further tested and investigated experimentally.

METHODS

Genes encoding small proteins in *E. coli* discovered starting from the year 2008 to present entered as a comprehensive list.

Bioinformatic analyses performed to provide insights into their functions: NCBI ID, Uniprot ID, amino acid length, molecular weight (kDa), amino acid sequence, co-occurrences with other genes, interactions with other proteins, physical/chemical properties and function (if known), localization predictions, antimicrobial peptide (AMP) activity prediction, relevant information gathered from review of published literature.

General, Basic Information/ Text mining

STRING: (Functional Protein Association Networks, https://string-db.org): All information that pertains to each protein's interactive role with others as well as its proposed function, will be derived from this highly dynamic online database.

EcoCyc (https://ecocyc.org/):, an online database containing information on all known genes and their protein products in *E. coli* K12 MG1655. All information that pertains to each protein's interactive role with others as well as its proposed function, will be derived from a highly dynamic online database

Antimicrobial Peptide Activity Prediction

DBAASP (Database of Antimicrobial Activity and Structure of Peptides, https://dbaasp.org/prediction): General antimicrobial peptide prediction based on the machine learning algorithm, uses Moon and Fleming scale and the physico-chemical characteristics of peptides

APD3 (The antimicrobial peptide database, http://aps.unmc.edu/AP/main.php): A comprehensive database for antimicrobial peptides. Structural information, homology with other known antimicrobial peptides.

Localization Prediction

TMHMM (Prediction of transmembrane helices in proteins, http://www.cbs.dtu.dk/services/TMHMM-2.0/): Will aid in the prediction of the exact sequences that compromise the transmembrane regions of small proteins using Hidden Markov Model.

Phobius (http://phobius.sbc.su.se//): Transmembrane topology, signal peptide and localization predictor based on amino acid sequence

3D Structure Prediction

Robetta (https://robetta.bakerlab.org/): a protein structure prediction service that is continually evaluated through CAMEO. Predicts 3D structure of a query based on amino acid composition, presents results in a 3D model.

Taxonomy Analysis

NCBI's TBLASTn (Translated Basic Local Alignment Search Tool. https://blast.ncbi.nlm.nih.gov/Blast.cgi): Software database compares each query to various matches based on possible nucleotide composition rather than amino acid configuration alone. Used to collect data on taxonomy and multiple sequence alignments.

RESULTS

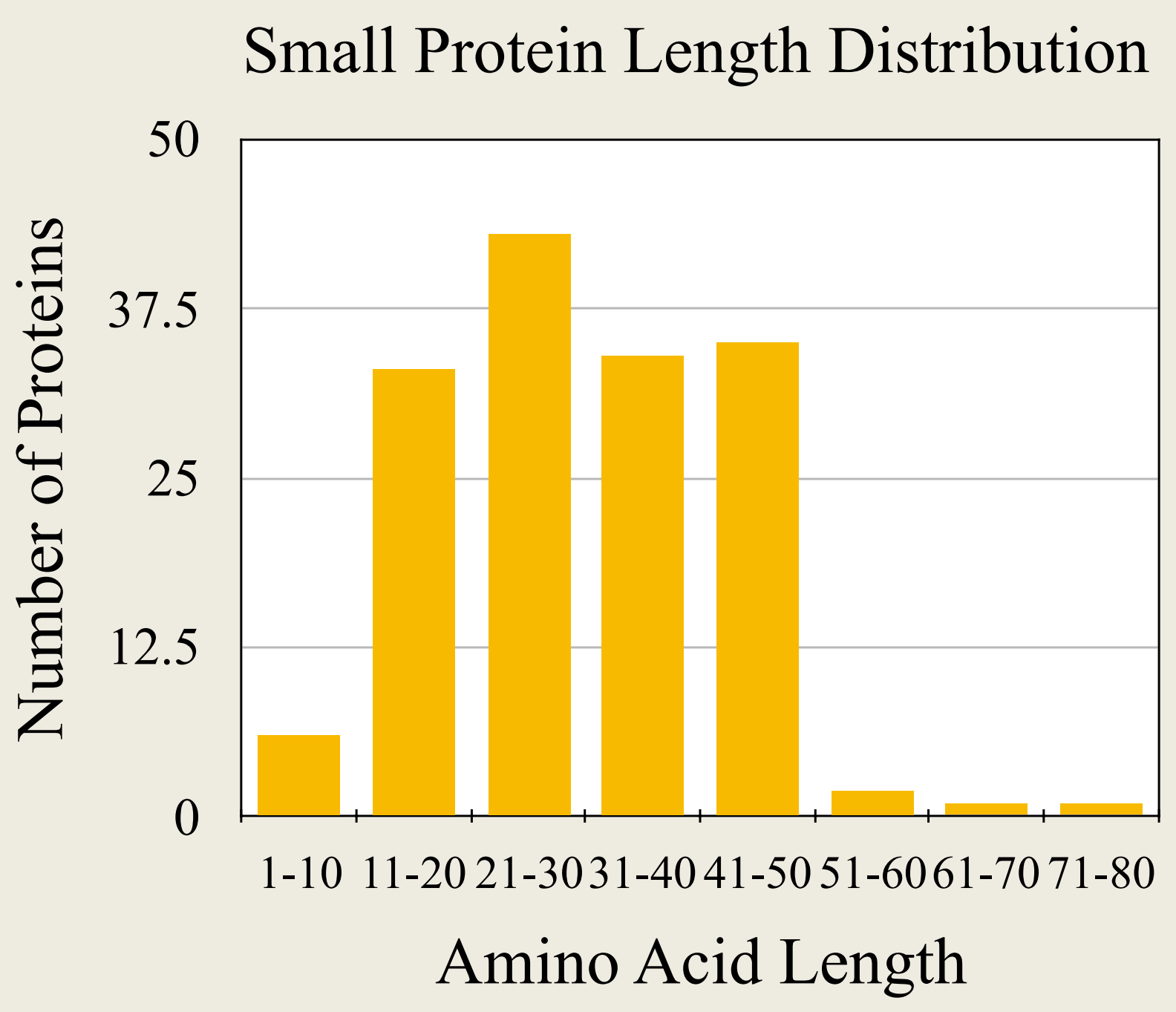
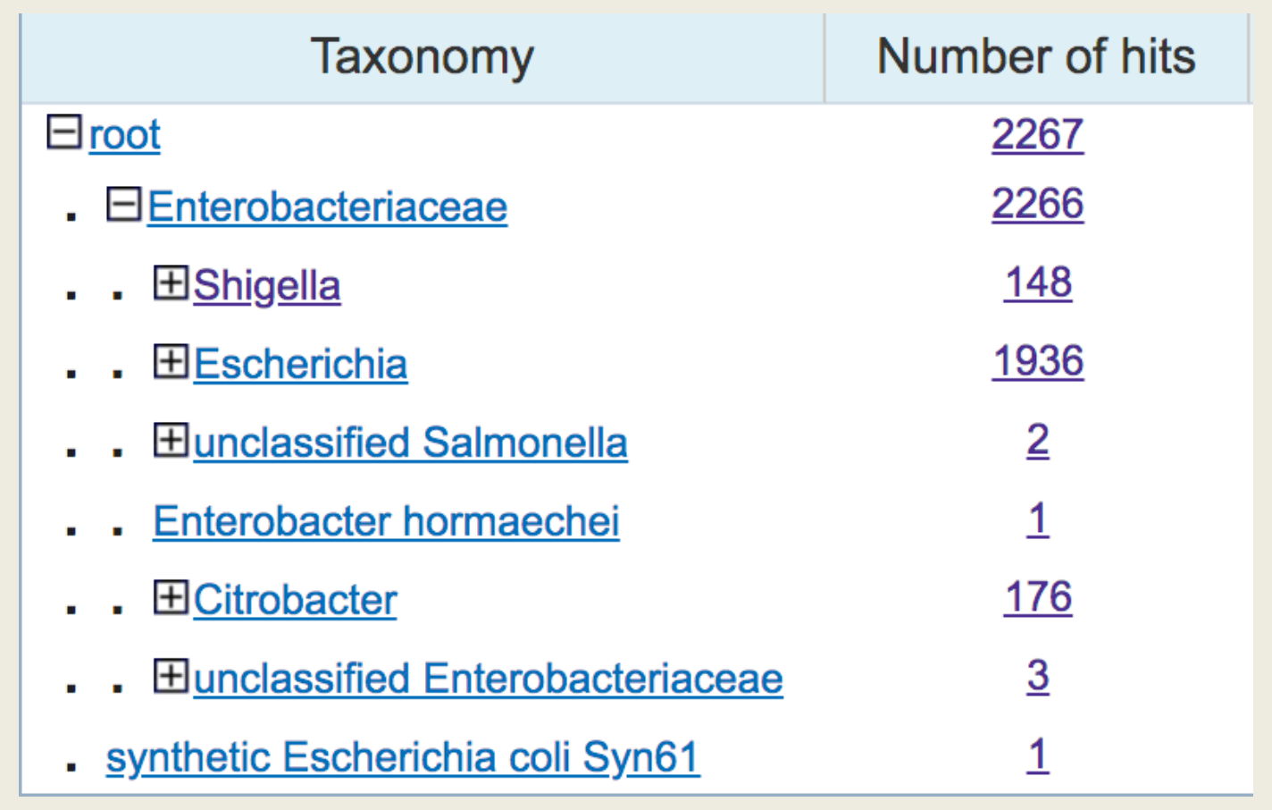


Figure 1. Small protein length distribution of all 155 entries in the current database as of July 2020. Small proteins come in varying sizes ranging from 8 to 80 amino acids in length, with a majority of them falling within 15-50 amino acids. Very few small peptides in the list are longer than 50 amino acids, and again emphasizing the difficulties that come with characterizing these small proteins due to their average small size being a limiting factor.

The aim of gathering taxonomic data on the peptides in the list was to observe their distribution among other organisms besides E. coli, and possibly bacteria (Fig. 2). The data ultimately revealed every single entry recognized by the software to be conserved primarily within Enterobacteriaceae. A few outliers worth mentioning include yqhJ and yqgH. Both peptides had a single match within Homo sapiens according to the NCBI's database. Additionally, idlP-3, ysdE, ytiC, ymiB, ymiA and sgrT have single matches in Mammalia, specifically Apteryx mantelli mantelli (North island brown kiwi bird).The most conserved of all the peptides, rpmH (a 50S ribosomal protein) is distributed among 7690 organisms within Bacteria. MgtT, a 34 amino acid long peptide that is potentially involved in Mg^{2+} regulation in E.coli, is an example worth mentioning. This peptide is found in 213 organisms, all within Enterobacteriaceae, and 149 within Escherichia.

Overall, these small peptides seem to be conserved amongst prokaryotes, with few exception being present in Eukaryota.

Taxonomic distribution of representative protein MgtT



Taxonomy	Number of hits
⊟root	2267
. ⊟Enterobacteriaceae	2266
. . ⊞Shigella	148
. . ⊞Escherichia	1936
. . ⊞unclassified Salmonella	2
. . Enterobacter hormaechei	1
. . ⊞Citrobacter	176
. . ⊞unclassified Enterobacteriaceae	3
. synthetic Escherichia coli Syn61	1

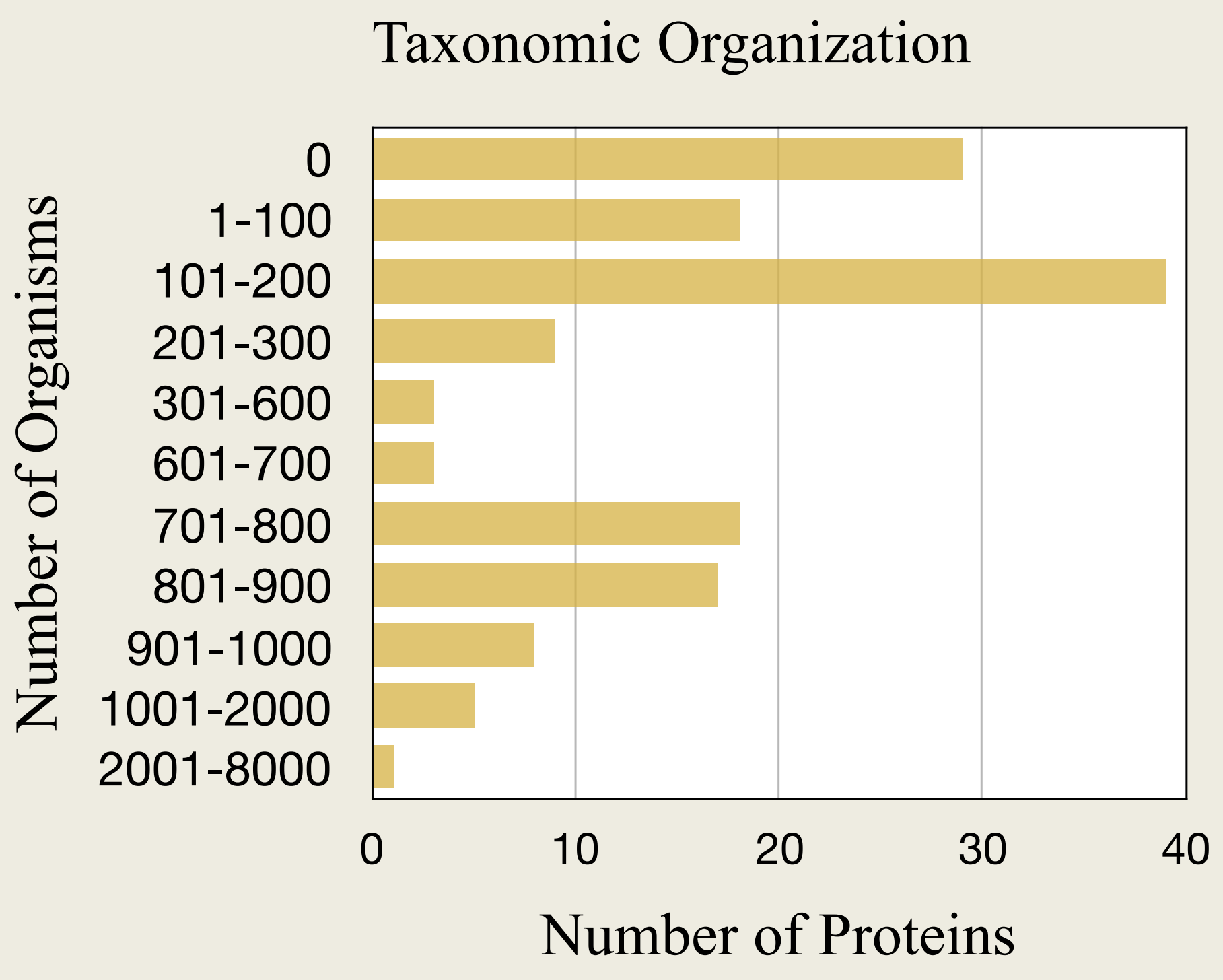


Figure 2. Taxonomic distribution of all small proteins using NCBI's protein translated nucleotide BLAST feature. The algorithm compared each protein query against a translated nucleotide sequence database instead of a protein sequence database. The data display the total number of organisms each entry has been identified in, indicating the extent of conservation of small proteins. A large number of small proteins are present in hundreds of taxonomic groups. As expected, all peptides recognized by the software are widely distributed among the Enterobacteriaceae family of Proteobacteria. The number of organisms a given small protein can be found in ranges from 5 to 7690.

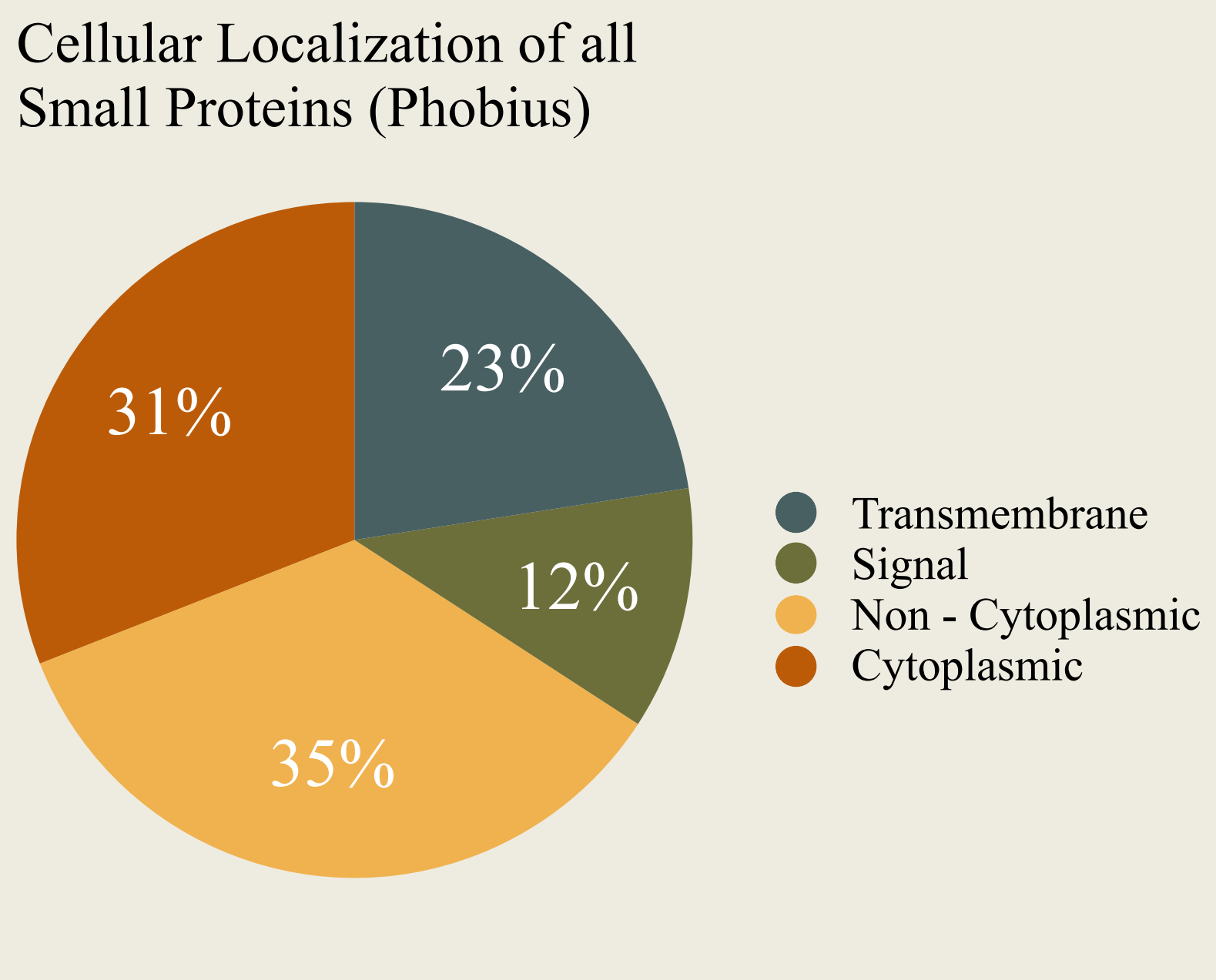
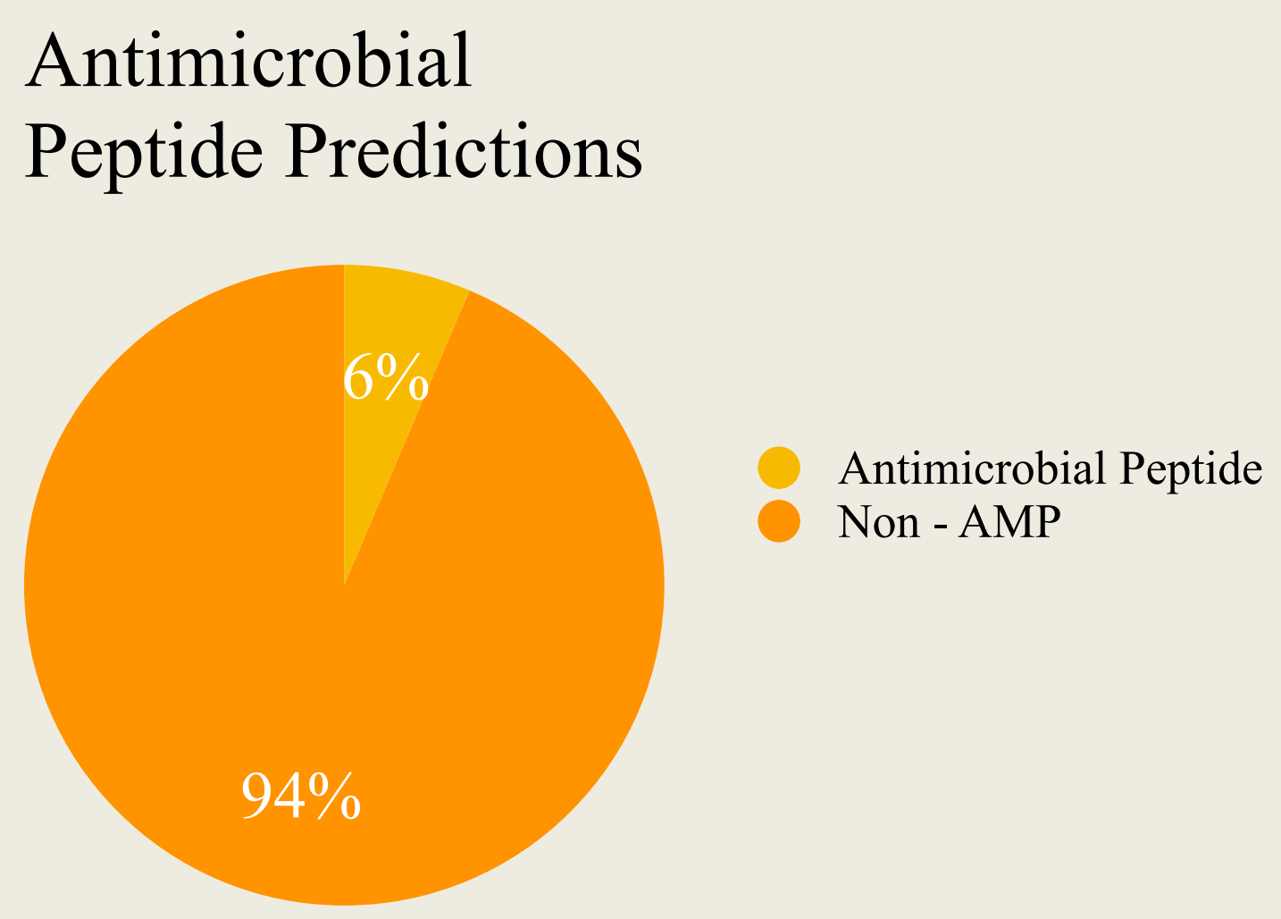


Figure 3. Predicted localization of all small peptides through Phobius program. Phobius predicts the location of a protein based on its amino acid sequence as being either: Transmembrane, Signal, Cytoplasmic and Non cytoplasmic. Out of 155 entries, 18 are signal, 35 are transmembrane, 54 are non cytoplasmic, and 48 are cytoplasmic. Signal peptides predicted by Phobius are said to contain an H - region, which is otherwise identified as an alpha helical structure, meaning these peptides could very well be localized to the membrane as well.

Various softwares were used to predict the localization of all the peptides in the list. Among these are: Phobius, TMHMM, SignalP 5.0, Secretome P, PSL Pred. Phobius and TMHMM. SignalP 5.0 and Secretome P specifically predicted signal peptides, 7 and 10 out of 155 small peptides respectively. PSL Pred predicted non cytoplasmic proteins, of which there were 7. TMHMM was useful in identifying potential alpha helix structures based on a query's amino acid composition, and 47 out of 155 predicted transmembrane proteins resulted. Phobius was the most comprehensive localization prediction software (Fig. 3) and yielded results that aligned closely with the previous programs. Roughly a third of all the small proteins are presumably cytoplasmic proteins, while a fourth of the proteins are predicted to localize to the membrane, and the rest either contain signal peptides, or are considered non - cytoplasmic.

RESULTS (cont.)



All entries were run through the DBAASP and APS (Antimicrobial Peptide Scanner). Each of these predictors gave either an affirmative or negative result for antimicrobial peptide activity with no further information. Out of the 10 high confidence antimicrobial peptides that were later obtained (Fig 4.), 7 potentially formed alpha helix structures. Considering once more these peptides' size and the fact that those ≤20 amino acids in length could not be recognized by certain programs, 6% of the total being possible antimicrobial peptides is a significant finding.

Figure 4. Antimicrobial peptides predicted with high confidence. Using DBAASP and APS AMP prediction softwares, all small proteins were analyzed for potential antimicrobial activity. Those that were given a positive result by both softwares were taken into consideration for further analysis with APD3 and AntiBP2. A total of 10 out of the 155 peptides on the list are high confidence predicted AMPs. APD3 in particular provided additional information on protein structure; If an alpha helix structure was predicted, hydrophobic residues were identified as well as residues that may be on the same surface. It provided information on each query such as homology with other known AMPs, hydrophobicity and total net charge, amino acid composition, among others.

DISCUSSION

The information gathered from these small proteins may aid in revealing what is yet to be discovered about important signaling pathways in a wide variety of organisms. It is also possible to take advantage of those small peptides that are well studied in E. coli and predict relevant functions of their unknown neighbors, especially those involved in well studied regulation pathways that are responsive to certain external stressors. Many of these small proteins are thought to be regulators of gene expression and stress response.

For example, YceO is a peptide presumed to be involved in regulatory mechanisms, acid stress in E. Coli, where little to no information about its exact function or structure has been uncovered. This peptide is 46 amino acids in length, and more than half are hydrophobic residues (27 total). Curiosly, despite this, yceO was not predicted to exhibit any antimicrobial activity, and is not a designated antimicrobial peptide within the list. It has however, been recognized to potentially form an alpha helix and localize to the membrane by both Phobius and TMHMM. Using Robetta, 3D models have been generated for certain peptides of interest.

It is one of the more widely conserved peptides in the list, found in 884 organisms within Enterobacteriaceae, and 1 among Eukaryota.

Ultimately, through in silico analysis of these peptides it is possible to gather enough information about them to create a narrative for their possible structure, role and significance within the organisms they find themselves in. Predictions from bioinformatic analysis are useful in this way, and can be validated in the lab.

Future Directions:

Further analysis (literature searches, function and location predictions) centered on a few select candidates of interest that could be potentially studied in vitro.

Continue to expand the list by investigating scholarly internet sources and scientific papers relevant to the project for previously discovered and novel small peptides.

ACKNOWLEDGEMENTS

My sincere thanks to:

Division of Life Sciences Summer Undergraduate Research Fellowship (DLS SURF 2020),

Mr. Dan MacMillan & Mrs.Nancy MacMillan,

Dr. Yadavalli and members of Yadavalli Lab.

REFERENCES

1. Storz, G., Wolf, Y. I., Ramamurthi, K. S. (2014). Small Proteins Can No Longer be Ignored. *Annual Review of Biochemistry.* doi: 10.1146/annurev-biochem-070611-102400
2. Weaver, J., Mohammad, F., Buskirk, A. R., Storz, G. (2019). Identifying Small Proteins by Ribosome Profiling with Stalled Initiation Complexes. *mBio.* doi: e02819-18